

Rozaida Ghazali
Mustafa Mat Deris
Nazri Mohd Nawi
Jemal H. Abawajy *Editors*

Recent Advances on Soft Computing and Data Mining

Proceedings of the Third International
Conference on Soft Computing and
Data Mining (SCDM 2018), Johor,
Malaysia, February 06–07, 2018

Advances in Intelligent Systems and Computing

Volume 700

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within “Advances in Intelligent Systems and Computing” are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

e-mail: nikhil@isical.ac.in

Members

Rafael Bello Perez, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba

e-mail: rbellop@uclv.edu.cu

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

e-mail: escorchado@usal.es

Hani Hagra, University of Essex, Colchester, UK

e-mail: hani@essex.ac.uk

László T. Kóczy, Széchenyi István University, Győr, Hungary

e-mail: koczy@sze.hu

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA

e-mail: vladik@utep.edu

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan

e-mail: ctlin@mail.nctu.edu.tw

Jie Lu, University of Technology, Sydney, Australia

e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico

e-mail: epmelin@hafsamx.org

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil

e-mail: nadia@eng.uerj.br

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland

e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong

e-mail: jwang@mae.cuhk.edu.hk

More information about this series at <http://www.springer.com/series/11156>

Rozaida Ghazali · Mustafa Mat Deris
Nazri Mohd Nawi · Jemal H. Abawajy
Editors

Recent Advances on Soft Computing and Data Mining

Proceedings of the Third International
Conference on Soft Computing and Data
Mining (SCDM 2018), Johor, Malaysia,
February 06–07, 2018

Editors

Rozaida Ghazali
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Johor
Malaysia

Nazri Mohd Nawi
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Johor
Malaysia

Mustafa Mat Deris
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Johor
Malaysia

Jemal H. Abawajy
School of Information Technology
Deakin University
Geelong, VIC
Australia

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-3-319-72549-9

ISBN 978-3-319-72550-5 (eBook)

<https://doi.org/10.1007/978-3-319-72550-5>

Library of Congress Control Number: 2017960913

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Rapid advancements in data collection and storage technology have enabled the access to vast amount of data. However, extracting useful information has proven extremely challenging. This is due to the fact that more complex systems arising are often remained intractable to conventional mathematical and analytical methods. For this, data mining which supports a wide range of business intelligence applications has opened up opportunities for exploring and analyzing various type of data. With the deployment of data and soft computing techniques to scour large database, novel and useful patterns can be found, otherwise remain unknown. Soft computing which refers to a consortium of computational techniques in computer science can deal with imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness, and low solution cost. Soft computing tools, individually or in integrated manner, are turning out to be strong candidates for performing tasks in the area of data mining, decision support systems, supply chain management, medicine, business, financial systems, automotive systems and manufacturing, image processing and data compression, etc.

The SCDM 2018 is so significant in a sense that it starts off a host of activity which collaborates Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), and Soft Computing and Data Mining research group. After the success of our two previous SCDM conferences in 2014 until 2016, we hope to continuously moving on this journey of success through this third international conference. Indeed, we are honored to host this event and the fact that we are getting more papers, commitments, contributions, and partnerships that indicates a continuous support from researchers throughout the globe.

The SCDM 2018, with the theme “Concise and Informative” was held in Johor Baharu Malaysia on August 6–7, 2018. We received 75 paper submissions from 13 countries around the world. The conference also approved three special sessions that are Intelligent Human-Centred Computing; Web Mining and Content Analytics; and Web Mining, Services, and Security. Each paper in regular submission and special sessions was screened by the Proceeding’s Chair and carefully peer-reviewed by two experts from the Program Committee. Finally, only 49 papers

with the highest quality and merit were accepted for oral presentation and publication in this volume proceeding, giving an acceptance rate of 65%.

The papers in this proceeding are grouped into five tracks:

- Soft Computing
- Data Mining
- Intelligent Human-Centred Computing
- Web Mining and Content Analytics
- Web Mining, Services, and Security

On behalf of SCDM 2018, we would like to express our highest gratitude to the Faculty of Computer Science and Information Technology, UTHM, and also to Soft Computing and Data Mining research group, Steering Committee, Conference Chair, Program Committee Chair, Organizing Chairs, Special Session Chair, all Program and Reviewer Committee members for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

We would also like to express our thanks to the four keynote speakers, Dr. Dzaharudin Mansor from Microsoft Malaysia, Assoc. Prof. Dr. David Taniar from Monash University, Prof. Junzo Watada from Waseda University, Japan, and Prof. Dr. Mustafa Mat Deris from Universiti Tun Hussein Onn Malaysia. Our special thanks are also due to Mr. Suresh Rettagunta and Dr. Thomas Ditzinger for publishing the proceeding in *Advances in Intelligent Systems and Computing*, Springer. We wish to thank the members of the Organizing and Student Committees for their very substantial work, especially those who played essential roles.

Lastly, we cordially thank all the authors for their valuable contributions and other participants of this conference. The conference would not have been possible without them.

Batu Pahat, Malaysia
Batu Pahat, Malaysia
Batu Pahat, Malaysia
Geelong, Australia

Rozaida Ghazali
Mustafa Mat Deris
Nazri Mohd Nawi
Jemal H. Abawajy

Conference Organization

Patron

Wahid Razzaly, Vice-Chancellor, Universiti Tun Hussein Onn Malaysia

Honorary Chair

Witold Pedrycz, University of Alberta, Canada

Junzo Watada, Waseda University, Japan

Ajith Abraham, Machine Intelligence Research Labs, USA

Nikola Kasabov, KEDRI, Auckland University of Technology, New Zealand

Hamido Fujita, Iwate Prefectural University, Japan

Hojjat Adeli, University of Ohio, USA

Steering Committee

Mustafa Mat Deris, Universiti Tun Hussein Onn Malaysia

Jemal H. Abawajy, Deakin University, Australia

Nazri Mohd Nawi, Universiti Tun Hussein Onn Malaysia

Rozaida Ghazali, Universiti Tun Hussein Onn Malaysia

Chair

Hairulnizam Mahdin, Universiti Tun Hussein Onn Malaysia

Organizing Committee

Azizul Azhar Ramli, Universiti Tun Hussein Onn Malaysia

Noorhaniza Wahid, Universiti Tun Hussein Onn Malaysia

Noraini Ibrahim, Universiti Tun Hussein Onn Malaysia

Norhalina Senan, Universiti Tun Hussein Onn Malaysia

Yana Mazwin Mohmad Hassim, Universiti Tun Hussein Onn Malaysia

Mohamad Aizi bin Salamat, Universiti Tun Hussein Onn Malaysia

Muhaini Othman, Universiti Tun Hussein Onn Malaysia

Program Committee Chair

Nureize Arbaiy, Universiti Tun Hussein Onn Malaysia

Proceeding Chair

Rozaida Ghazali, Universiti Tun Hussein Onn Malaysia

Special Session Chair

Mohd Helmy Abd Wahab, Universiti Tun Hussein Onn Malaysia

Sponsorship

Shahreen Kasim, Universiti Tun Hussein Onn Malaysia

Technical Support

Azizul Zamri Muhamed Amin, Universiti Tun Hussein Onn Malaysia

Program Committee

Nureize Arbaiy, Universiti Tun Hussein Onn Malaysia

Chuah Chai Wen, Universiti Tun Hussein Onn Malaysia

Mohd. Najib Mohd. Salleh, Universiti Tun Hussein Onn Malaysia

Mohd. Farhan Md. Fudzee, Universiti Tun Hussein Onn Malaysia

Mohd Helmy Abd Wahab, Universiti Tun Hussein Onn Malaysia

Ruhaidah Samsudin, Universiti Teknologi Malaysia

Aida Mustapha, Universiti Tun Hussein Onn Malaysia

José M. Merigó, University of Chile

Norfaradilla Wahid, Universiti Tun Hussien Onn Malaysia

Alwis Nazir, State Islamic University of Sultan Syarif Kasim Riau

Elena Benderskaya, St. Petersburg State Polytechnical University

Gai-Ge Wang, School of Computer Science and Technology, Jiangsu Normal University

Saima Anwar Lashari, Universiti Tun Hussein Onn Malaysia

Riswan Efendi, Universiti Tun Hussein Onn Malaysia

Mohd Amin Mohd Yunus, Universiti Tun Hussein Onn Malaysia

Mohd Zainuri Saringat, Universiti Tun Hussein Onn Malaysia

Mustafa Mat Deris, Universiti Tun Hussein Onn Malaysia

Nazri Mohd Nawi, Universiti Tun Hussein Onn Malaysia

Hairulnizam Mahdin, Universiti Tun Hussein Onn Malaysia

Rozaida Ghazali, Universiti Tun Hussein Onn Malaysia

Isredza Rahmi Ab Hamid, Universiti Tun Hussein Onn Malaysia

Nurul Hidayah Ab Rahman, Universiti Tun Hussein Onn Malaysia

Azizul Azhar Ramli, Universiti Tun Hussein Onn Malaysia

Mohamad Aizi Salamat, Universiti Tun Hussein Onn Malaysia
Rosziati Ibrahim, Universiti Tun Hussein Onn Malaysia
Noraini Ibrahim, Universiti Tun Hussein Onn Malaysia
Muhaini Othman, Universiti Tun Hussein Onn Malaysia
Noor Azah Samsudin, Universiti Tun Hussein Onn Malaysia
Yana Mazwin Mohmad Hassim, Universiti Tun Hussein Onn Malaysia
Noorhaniza Wahid, Universiti Tun Hussein Onn Malaysia
Norhalina Senan, Universiti Tun Hussein Onn Malaysia
Noryusliza Abdullah, Universiti Tun Hussein Onn Malaysia
Sapi'ee Jamel, Universiti Tun Hussein Onn Malaysia
Nurul Azma Abdullah, Universiti Tun Hussein Onn Malaysia
Kamaruddin Malik Mohamad, Universiti Tun Hussein Onn Malaysia
Pei-Chun Lin, Feng Chia University
Hazlina Hamdan, Universiti Putra Malaysia
Zhiang Wu, Jiangsu Provincial Key Laboratory of E-Business, Nanjing University
of Finance and Economics, Nanjing, P.R. China
Hamidah Ibrahim, Universiti Putra Malaysia
Dayang N. A. Jawawi, Universiti Teknologi Malaysia
Maslina Zolkepli, Universiti Putra Malaysia
Haza Nuzly Abdull Hamed, Universiti Teknologi Malaysia
El-Sayed M. El-Alfy, King Fahd University of Petroleum and Minerals
Katsuhiro Honda, Osaka Prefecture University
Jose Santos Reyes, University of A Coruña
Siti Yuhaniz, Universiti Teknologi Malaysia
Mohd Saberi Mohamad, Universiti Teknologi Malaysia
Zaidah Ibrahim, Universiti Teknologi MARA
Paulus Insap Santosa, Gadjah Mada University
Elpida Tzafestas, University of Athens
Nadjet Kamel, University Ferhat Abbas Setif1
Rahmat Hidayat, Politeknik Negeri Padang, Indonesia

Special Session Committee

Intelligent Human-Centred Computing

Azizi Ab Aziz, Universiti Utara Malaysia
Natalie van der Wal, Vrije Universiteit Amsterdam
Somnuk Phon-Amnuaisuk, Universiti Teknologi Brunei
Waqar ul Qounain, University of the Punjab
Afizan Azman, Multimedia University
Husniza Husni, Universiti Utara Malaysia

Web Mining and Content Analytics

Nurfadhlina Mohd Sharef, Universiti Putra Malaysia
Kazutaka Shimada, Kyushu Institute of Technology, Japan
Patricia Anthony, Lincoln University, New Zealand

Web Mining, Services and Security

Dr. Mohd Farhan Md Fudzee, Universiti Tun Hussein Onn Malaysia
Dr. Hairulnizam Mahdin, Universiti Tun Hussein Onn Malaysia
Isredza Rahmi Ab Hamid, Universiti Tun Hussein Onn Malaysia
Kamaruddin Malik Mohamad, Universiti Tun Hussein Onn Malaysia
Shahreen Kasim, Universiti Tun Hussein Onn Malaysia
Nurul Hidayah Ab Rahman, Universiti Tun Hussein Onn Malaysia
Izuan Hafez Ninggal, Universiti Putra Malaysia
Masitah Ahmad, Universiti Teknologi MARA (UiTM), Malaysia
Mohd Zalisham Jali, Universiti Sains Islam Malaysia
Izzatdin Abdul Aziz, Universiti Teknologi Petronas
Harinda Fernando, APIIT, Sri Lanka

Conference Logo

SCDM-2018

Conference Banner



SCDM-2018
6th-7th Feb 2018
Johor Baharu, Malaysia

Contents

Part I Soft Computing

A Percentile Transition Ranking Algorithm Applied to Binarization of Continuous Swarm Intelligence Metaheuristics	3
José García, Broderick Crawford, Ricardo Soto, and Gino Astorga	
An Improved Hybrid Firefly Algorithm for Solving Optimization Problems	14
Fazli Wahid, Rozaida Ghazali, and Habib Shah	
Exploration and Exploitation Measurement in Swarm-Based Metaheuristic Algorithms: An Empirical Analysis	24
Mohd Najib Mohd Salleh, Kashif Hussain, Shi Cheng, Yuhui Shi, Arshad Muhammad, Ghufuran Ullah, and Rashid Naseem	
Classification of JPEG Files by Using Extreme Learning Machine	33
Rabei Raad Ali, Kamaruddin Malik Mohamad, Sapiee Jamel, and Shamsul Kamal Ahmad Khalid	
Evaluating the Performance of Three Classification Methods in Diagnosis of Parkinson's Disease	43
Salama A. Mostafa, Aida Mustapha, Shihab Hamad Khaleefah, Mohd Sharifuddin Ahmad, and Mazin Abed Mohammed	
Some New Results on the Stability of Fractional Integro-Differential Equations Under Uncertainty	53
A. Ahmadian, S. Salahshour, N. Senu, and F. Ismail	
Semantic Approach for Web-Based Presentation Mining Based Ontology Support	64
Vinothini Kasinathan, Aida Mustapha, and Imran Medi	

A Relative Tolerance Relation of Rough Set for Incomplete Information Systems	72
Rd. Rohmat Saedudin, Hairulnizam Mahdin, Shahreen Kasim, Edi Sutoyo, Iwan Tri Riyadi Yanto, and Rohayanti Hassan	
Fuzzy Evaluation Scheme for KDF Based on Stream Ciphers	82
Hamijah Mohd. Rahman, Nureize Arbaiy, and Chuah Chai Wen	
A Similarity Precision for Selecting Ontology Component in an Incomplete Sentence	95
Fatin Nabila Rafei Heng, Mustafa Mat Deris, and Nurlida Basir	
Mitigating Manual Final Year Project (FYP) Management to Be Centralized Electronically	105
Noryusliza Abdullah, Shahril Nazim Mohamed Salleh, Hairulnizam Mahdin, Rozanawati Darman, Basil David Daniel, and Ely Salwana Mat Surin	
An Algorithm Design of Kansei Recommender System	115
Pei-Chun Lin and Nureize Arbaiy	
Warehouse Picking Model for Single Picker Routing Problem in Multi Dimensional Warehouse Layout Using Genetic Algorithm Approach to Minimize Delay	124
Dida Diah Damayanti, Erlangga Bayu Setyawan, Luciana Andrawina, and Budi Santosa	
Relationship Between Angiotensin Converting Enzyme Gene and Cardiac Autonomic Neuropathy Among Australian Population	135
Ahmad Shaker Abdalrada, Jemal H. Abawajy, Morshed U. Chowdhury, Sutharshan Rajasegarar, Tahsien Al-Quraishi, and Herbert F. Jelinek	
Modeling of Consumer Interest on E-commerce Products Using Eye Tracking Methods	147
Juni Nurma Sari, Lukito Edi Nugroho, P. Insap Santosa, and Ridi Ferdiana	
Part II Data Mining	
A New Concept of Fuzzy TOPSIS and Fuzzy Logic in a Multi-criteria Decision	161
Ratih Fitria Jumarni and Nurnadiah Zamri	
Comparative Studies of Information Retrieval Approaches in User-Centered Health Information System	171
Ibrahim Umar Kontagora and Isredza Rahmi A. Hamid	

A Framework to Cluster Temporal Data Using Personalised Modelling Approach	181
Muhaini Othman, Siti Aisyah Mohamed, Mohd Hafizul Afifi Abdullah, Munirah Mohd Yusof, and Rozlini Mohamed	
Measurement of the Pitch Exploration Amongst Elite Professional Soccer Players: Official Match Analysis	191
Filipe Manuel Clemente, Adam Owen, Aida Mustapha, Cornelis M. I. (Niels) van der Linden, João Ribeiro, Bruno Mendes, and Jelle Reichert	
RMIL/AG: A New Class of Nonlinear Conjugate Gradient for Training Back Propagation Algorithm	200
Sri Mazura Muhammad Basri, Nazri Mohd Nawi, Mustafa Mamat, and Norhamreeza Abdul Hamid	
A Regulative Norms Mining Algorithm for Complex Adaptive System	213
Moamin A. Mahmoud, Mohd Sharifuddin Ahmad, Mohd Zaliman M. Yusoff, and Salama A. Mostafa	
Violence Video Classification Performance Using Deep Neural Networks	225
Ashikin Ali and Norhalina Senan	
Fibonacci Polynomials Based Functional Link Neural Network for Classification Tasks	234
Umer Iqbal, Rozaida Ghazali, and Habib Shah	
Decision Support Model in Determining Factors and Its Dominant Criteria Affecting Cholesterol Level Based on Rough-Regression	243
Riswan Efendi and Mustafa Mat Deris	
A Numerical Classification Technique Based on Fuzzy Soft Set Using Hamming Distance	252
Iwan Tri Riyadi Yanto, Rd Rohmat Saedudin, Saima Anwar Lashari, and Haviluddin	
Is SVM+FS Better to Satisfy Decision by Majority?	261
Yao Lin, Kohei Yamaguchi, Tsunenori Mine, and Sachio Hirokawa	
Exploiting LabVIEW FPGA in Implementation of Real-Time Sensor Data Acquisition for Rowing Monitoring System	272
Zarina Tukiran and Afandi Ahmad	

A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses. 282
Abdullahi O. Adeleke, Noor Azah Samsudin, Aida Mustapha, and Nazri Mohd Nawi

A Review: Image Analysis Techniques to Improve Labeling Accuracy of Medical Image Classification 298
Mazniha Berahim, Noor Azah Samsudin, and Shelena Soosay Nathan

A New Adaptive Energy-Aware Job Scheduling in Cloud Computing 308
Ali Aghababaeipour and Shamsollah Ghanbari

Breast Cancer Recurrence Prediction Using Random Forest Model . . . 318
Tahsien Al-Quraishi, Jemal H. Abawajy, Morshed U. Chowdhury, Sutharshan Rajasegarar, and Ahmad Shaker Abdalrada

A New Theoretical Framework for Testing Consciousness in a Machine 330
Azree Nazri, Abdul Azim Abd Ghani, Izuan Hafez, and Keng-Yap Ng

Temporal Based Factorization Approach for Solving Drift and Decay in Sparse Scoring Matrix 340
Al-Hadi Ismail Ahmed Al-Qasem, Nurfadhlina Mohd Sharef, Sulaiman Md Nasir, and Mustapha Norwati

Part III Intelligent Human-Centred Computing (IHCC)

Preliminary Design of a Dual-Sensor Based Sign Language Translator Device. 353
Radzi Ambar, Chan Kar Fai, Chew Chang Choon, Mohd Helmy Abd Wahab, Muhammad Mahadi Abdul Jamil, and Ahmad Alabqari Ma'Radzi

M-DCocoa: M-Agriculture Expert System for Diagnosing Cocoa Plant Diseases 363
Munirah Mohd Yusof, Nur Fazliyana Rosli, Muhaini Othman, Rozlini Mohamed, and Mohd Hafizul Afifi Abdullah

Dyslexia Adaptive Learning Model: Student Engagement Prediction Using Machine Learning Approach. 372
Siti Suhaila Abdul Hamid, Novia Admodisastro, Noridayu Manshor, Azrina Kamaruddin, and Abdul Azim Abd Ghani

Towards Designing Tangible Interaction for Children with Dyslexia in Learning the Malay Language	385
Siti Nurliana Jamali, Novia Admodisastro, Siti Suhaila Abdul Hamid, Azrina Kamaruddin, Abdul Azim Abd Ghani, and Sa'adah Hassan	
Comparison of Approaches Made to Enhance Pupils' Numeracy Skill	396
Nur Faizura Ahmad Fuadi, Muhammad Fakri Othman, and Norhalina Senan	
 Part IV Web Mining and Content Analytics (WMCA)	
Stock Market Prediction Using Keywords from Expert Articles	409
Ko Ichinose and Kazutaka Shimada	
Sarcasm Detection Using Features Based on Indicator and Roles	418
Satoshi Hiai and Kazutaka Shimada	
Reducing Computational Effort for Plagiarism Detection with Approximate String Matching	429
Tetsuya Nakatoh and Toshiro Minami	
Weighting of Noun Phrases Based on Local Frequency of Nouns	436
Yasuhiro Yamada, Yuusuke Himeno, and Tetsuya Nakatoh	
Multi-layers Convolutional Neural Network for Twitter Sentiment Ordinal Scale Classification	446
Muath ALALI, Nurfadhlina Mohd Sharef, Hazlina Hamdan, Masrah Azrifah Azmi Murad, and Nor Azura Husin	
Instance-Based Ontology Matching: A Literature Review	455
Mansir Abubakar, Hazlina Hamdan, Norwati Mustapha, and Teh Noranis Mohd Aris	
 Part V Web Mining, Services and Security (WMSS)	
Maximum Attribute Relative Approach of Soft Set Theory in Selecting Cluster Attribute of Electronic Government Data Set	473
Deden Witarsyah Jacob, Iwan Tri Riyadi Yanto, Mohd Farhan Md Fudzee, and Mohamad Aizi Salamat	
Android Malware Detection Based on Network Traffic Using Decision Tree Algorithm	485
Aqil Zulkifli, Isredza Rahmi A. Hamid, Wahidah Md Shah, and Zubaile Abdullah	

An Improved Low Contrast Image in Normalization Process for Iris Recognition System	495
Abdulrahman Aminu Ghali, Sapiee Jamel, Kamaruddin Malik Mohamad, Shamsul Kamal Ahmad Khalid, Zahraddeen Abubakar Pindar, and Mustafa Mat Deris	
Presenting a New Method of Authentication for the Internet of Things Based on RFID	506
Farshad Asadpour and Shamsollah Ghanbari	
Author Index	517

Contributors

Jemal H. Abawajy Deakin University, Burwood, VIC, Australia

Ahmad Shaker Abdalrada Deakin University, Burwood, VIC, Australia

Siti Suhaila Abdul Hamid University Putra Malaysia, Seri Kembangan, Malaysia

Mohd Hafizul Afifi Abdullah Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Noryusliza Abdullah Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Zubaile Abdullah Information Security Interest Group (ISIG), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Johor, Malaysia

Mansir Abubakar Faculty of Computer Science and Information Technology, University Putra Malaysia, Seri Kembangan, Malaysia

Abdullahi O. Adeleke Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Novia Admodisastro University Putra Malaysia, Seri Kembangan, Malaysia

Afandi Ahmad Department of Computer Engineering, Faculty of Electrical & Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia; Reconfigurable Computing for Analytics Acceleration (ReCAA) Research Laboratory, Microelectronics and Nanotechnology—Shamsuddin Research Centre (MiNT-SRC), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

Mohd Sharifuddin Ahmad College of Computer Science and Information Technology, Universiti Tenaga Nasional, Kajang, Selangor, Malaysia; Business Development Unit, TNB Integrated Learning Solution Sdn. Bhd. (ILSAS), Kajang, Malaysia

Nur Faizura Ahmad Fuadi Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

A. Ahmadian Laboratory of Computational Sciences and Mathematical Physics, Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Al-Hadi Ismail Ahmed Al-Qasem Amran University, Amran, Yemen

Tahsien Al-Quraishi Deakin University, Burwood, VIC, Australia

Muath ALALI Faculty of Computer Science and Information Technology, Intelligent Computing Research Group, Unversiti Putra Malaysia UPM Serdang, Selangor, Malaysia

Ashikin Ali Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Batu Pahat, Johor, Malaysia

Rabei Raad Ali Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn Malaysia, Parit Raj, Johor, Malaysia

Radzi Ambar Department of Computer Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Luciana Andrawina Faculty of Industrial and System Engineering, Telkom University, Bandung, Indonesia

Nureize Arbaiy Soft Computing and Data Mining (SMC), Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Teh Noranis Mohd Aris Faculty of Computer Science and Information Technology, University Putra Malaysia, Seri Kembangan, Malaysia

Farshad Asadpour Department of Computer Science, Islamic Azad University, Ashtian Branch, Iran; Iranian Non-profit Association of Distributed Computing and Sytems, Qom, Iran

Gino Astorga Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile; Universidad de Valparaíso, Valparaíso, Chile

Masrah Azrifah Azmi Murad Faculty of Computer Science and Information Technology, Intelligent Computing Research Group, Unversiti Putra Malaysia UPM Serdang, Selangor, Malaysia

Nurlida Basir Universiti Sains Islam Malaysia (USIM), Nilai, Negeri Sembilan, Malaysia

Sri Mazura Muhammad Basri Faculty of Science, Technology and Human Development, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Mazniha Berahim Department of Information Technology, Center for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Pt Raja, Batu Pahat, Johor, Malaysia

Shi Cheng School of Computer Science, Shaanxi Normal University, Xian, China

Chew Chang Choon Department of Computer Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Morshed U. Chowdhury Deakin University, Burwood, VIC, Australia

Filipe Manuel Clemente School of Sport and Leisure, Viana do Castelo Polytechnic Institute, Melgaço, Portugal; Instituto de Telecomunicações, Delegação da Covilhã, Covilhã, Portugal

Broderick Crawford Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Dida Diah Damayanti Faculty of Industrial and System Engineering, Telkom University, Bandung, Indonesia

Basil David Daniel Faculty of Civil and Environmental Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Rozanawati Darman Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Mustafa Mat Deris Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

Riswan Efendi Faculty of Computer Science, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia; Mathematics Department, Faculty of Science and Technology, State Islamic University of Sultan Syarif Kasim Riau, Panam, Pekanbaru, Indonesia

Chan Kar Fai Department of Computer Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Ridi Ferdiana Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Jogjakarta, Indonesia

José García Telefónica Investigación y Desarrollo, Santiago, Chile; Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Abdulrahman Aminu Ghali Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

Shamsollah Ghanbari Department of Computer Science, Islamic Azad University, Ashtian Branch, Iran; Iranian Non-profit Association of Distributed Computing and Systems, Qom, Iran

Abdul Azim Abd Ghani Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Rozaida Ghazali Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Izuan Hafez Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Hazlina Hamdan Faculty of Computer Science and Information Technology, Intelligent Computing Research Group, Universiti Putra Malaysia UPM Serdang, Selangor, Malaysia

Isredza Rahmi A. Hamid Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn, Parit Raja, Malaysia

Norhamreeza Abdul Hamid Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Siti Suhaila Abdul Hamid Universiti Putra Malaysia, Serdang, Malaysia

Rohayanti Hassan Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

Sa'adah Hassan Universiti Putra Malaysia, Serdang, Malaysia

Haviluddin Faculty of Computer Science and Information Technology, Mulawarman University, Samarinda, Indonesia

Fatin Nabila Rafei Heng Universiti Sains Islam Malaysia (USIM), Nilai, Negeri Sembilan, Malaysia

Satoshi Hiai Department of Artificial Intelligence, Kyushu Institute of Technology, Fukuoka, Japan

Yuusuke Himeno Interdisciplinary Faculty of Science and Engineering, Shimane University, Matsue-shi, Shimane, Japan

Sachio Hirokawa Department of Advanced Information Technology, Kyushu University, Nishi-ku, Fukuoka, Japan

Nor Azura Husin Faculty of Computer Science and Information Technology, Intelligent Computing Research Group, Universiti Putra Malaysia UPM Serdang, Selangor, Malaysia

Kashif Hussain Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Ko Ichinose Department of Artificial Intelligence, Kyushu Institute of Technology, Fukuoka, Japan

P. Insap Santosa Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Jogjakarta, Indonesia

F. Ismail Laboratory of Computational Sciences and Mathematical Physics, Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Umer Iqbal Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Deden Witarsyah Jacob Department of Industrial Engineering, Telkom University, Bandung, West Java, Indonesia

Siti Nurliana Jamali Universiti Putra Malaysia, Serdang, Malaysia

Sapiee Jamel Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Parit Raj, Johor, Malaysia

Muhammad Mahadi Abdul Jamil Department of Electronic Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Herbert F. Jelinek Charles Sturt University, Albury, NSW, Australia

Ratih Fitria Jumarni Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut, Terengganu, Malaysia

Azrina Kamaruddin University Putra Malaysia, Seri Kembangan, Malaysia

Shahreen Kasim Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

Vinothini Kasinathan Faculty of Computing Engineering and Technology, Asia Pacific University of Innovation and Technology, Bukit Jalil, Kuala Lumpur, Malaysia

Shihab Hamad Khaleefah Faculty of Computer Science, Almaaref University College, Anbar, Iraq

Shamsul Kamal Ahmad Khalid Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Parit Raj, Johor, Malaysia

Ibrahim Umar Kontagora Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn, Parit Raja, Johor, Malaysia; Department of Computer Science, Niger State Polytechnic, Zungeru, Niger State, Nigeria

Saima Anwar Lashari Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Johor, Malaysia

Pei-Chun Lin Department of Information Engineering and Computer Science, Feng Chia University, Seatwen, Taichung, Taiwan

Yao Lin Department of Advanced Information Technology, Kyushu University, Nishi-ku, Fukuoka, Japan

Cornelis M. I. (Niels) van der Linden Department of Sports Sciences, JOHAN Sports, Noordwijk, The Netherlands

Hairulnizam Mahdin Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Moamin A. Mahmoud College of Computer Science and Information Technology, Universiti Tenaga Nasional, Kajang, Malaysia; Business Development Unit, TNB Integrated Learning Solution Sdn. Bhd. (ILSAS), Kajang, Malaysia

Mustafa Mamat Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Besut, Terengganu Darul Iman, Malaysia

Noridayu Manshor University Putra Malaysia, Seri Kembangan, Malaysia

Ahmad Alabqari Ma'Radzi Department of Electronic Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Mohd Farhan Md Fudzee Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia

Imran Medi Faculty of Computing Engineering and Technology, Asia Pacific University of Innovation and Technology, Bukit Jalil, Kuala Lumpur, Malaysia

Bruno Mendes BenficaLab, Sport Lisboa e Benfica, Lisbon, Portugal

Toshiro Minami Kyushu Institute of Information Sciences, Fukuoka, Japan

Tsunenori Mine Department of Advanced Information Technology, Kyushu University, Nishi-ku, Fukuoka, Japan

Kamaruddin Malik Mohamad Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Parit Raj, Johor, Malaysia

Rozlini Mohamed Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Siti Aisyah Mohamed Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Mazin Abed Mohammed Planning and Follow-up Department, University of Anbar, Anbar, Iraq

Hamijah Mohd. Rahman Soft Computing and Data Mining (SMC), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Batu Pahat, Johor, Malaysia; Information Security Interest Group (ISIG), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Batu Pahat, Johor, Malaysia

Nurfadhlina Mohd Sharef Faculty of Computer Science and Information Technology, Intelligent Computing Research Group, University Putra Malaysia, UPM, Serdang, Selangor, Malaysia

Salama A. Mostafa Business Development Unit, TNB Integrated Learning Solution Sdn. Bhd. (ILSAS), Kajang, Malaysia; Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Arshad Muhammad Faculty of Computing and Information Technology, Sohar University, Sohar, Oman

Aida Mustapha Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Norwati Mustapha Faculty of Computer Science and Information Technology, University Putra Malaysia, Seri Kembangan, Malaysia

Tetsuya Nakatoh Research Institute for Information Technology, Kyushu University, Nishi-ku, Fukuoka, Japan

Rashid Naseem Department of Computer Science, City University of Science and Information Technology, Peshawar, Pakistan

Sulaiman Md Nasir Faculty of Computer Science and Information Technology, University Putra Malaysia, UPM, Serdang, Selangor, Malaysia

Shelena Soosay Nathan Department of Information Technology, Center for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Pt Raja, Batu Pahat, Johor, Malaysia

Nazri Mohd Nawi Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Azree Nazri Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Keng-Yap Ng Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Mustapha Norwati Faculty of Computer Science and Information Technology, University Putra Malaysia, UPM, Serdang, Selangor, Malaysia

Lukito Edi Nugroho Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Jogjakarta, Indonesia

Muhaini Othman Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Muhammad Fakri Othman Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Adam Owen Centre de Recherche et d'Innovation sur le Sport, Université Claude Bernard Lyon. 1, Lyon, France

Zahraddeen Abubakar Pindar Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

Sutharshan Rajasegarar Deakin University, Burwood, VIC, Australia

Jelle Reichert Department of Sports Sciences, JOHAN Sports, Noordwijk, The Netherlands

João Ribeiro Gabinete de Otimização Desportiva, Sporting Clube de Braga, Braga, Portugal

Nur Fazliyana Rosli Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Rd. Rohmat Saedudin School of Industrial Engineering, Telkom University, Bandung, West Java, Indonesia

S. Salahshour Young Researchers and Elite Club, Mobarakeh Branch, Islamic Azad University, Mobarakeh, Iran

Mohamad Aizi Salamat Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia

Mohd Najib Mohd Salleh Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Shahril Nazim Mohamed Salleh Information Technology Centre, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Noor Azah Samsudin Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Budi Santosa Faculty of Industrial and System Engineering, Telkom University, Bandung, Indonesia

Juni Nurma Sari Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Jogjakarta, Indonesia; Department of Informatics Engineering, Politeknik Caltex Riau, Pekanbaru, Indonesia

Norhalina Senan Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Batu Pahat, Johor, Malaysia

N. Senu Laboratory of Computational Sciences and Mathematical Physics, Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Erlangga Bayu Setyawan Faculty of Industrial and System Engineering, Telkom University, Bandung, Indonesia

Habib Shah Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia

Wahidah Md Shah Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Malaysia

Yuhui Shi Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

Kazutaka Shimada Department of Artificial Intelligence, Kyushu Institute of Technology, Fukuoka, Japan

Ricardo Soto Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Ely Salwana Mat Surin Institute of Visual Informatics, University Kebangsaan Malaysia, Bangi, Malaysia

Edi Sutoyo School of Industrial Engineering, Telkom University, Bandung, West Java, Indonesia

Zarina Tukiran Department of Computer Engineering, Faculty of Electrical & Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia; Reconfigurable Computing for Analytics Acceleration (ReCAA) Research Laboratory, Microelectronics and Nanotechnology—Shamsuddin Research Centre (MiNT-SRC), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

Ghufran Ullah Department of Computer Science, City University of Science and Information Technology, Peshawar, Pakistan

Mohd Helmy Abd Wahab Department of Computer Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Fazli Wahid Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia

Chuah Chai Wen Information Security Interest Group (ISIG), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Batu Pahat, Johor, Malaysia

Yasuhiro Yamada Interdisciplinary Graduate School of Science and Engineering, Shimane University, Matsue-shi, Shimane, Japan

Kohei Yamaguchi Department of Advanced Information Technology, Kyushu University, Nishi-ku, Fukuoka, Japan

Iwan Tri Riyadi Yanto Department of Information Systems, University of Ahmad Dahlan, Kampus III UAD, Yogyakarta, Indonesia

Munirah Mohd Yusof Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Mohd Zaliman M. Yusoff College of Computer Science and Information Technology, Universiti Tenaga Nasional, Kajang, Malaysia; Business Development Unit, TNB Integrated Learning Solution Sdn. Bhd. (ILSAS), Kajang, Malaysia

Nurnadiah Zamri Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut, Terengganu, Malaysia

Aqil Zulkifli Information Security Interest Group (ISIG), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Johor, Malaysia

Part I

Soft Computing

A Percentile Transition Ranking Algorithm Applied to Binarization of Continuous Swarm Intelligence Metaheuristics

José García^{1,2(✉)}, Broderick Crawford², Ricardo Soto², and Gino Astorga^{2,3}

¹ Telefónica Investigación y Desarrollo, Santiago, Chile
joseantonio.garcia@telefonica.com

² Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
broderick.crawford@pucv.cl, ricardo.soto@pucv.cl

³ Universidad de Valparaíso, 2361864 Valparaíso, Chile
gino.astorga@uv.cl

Abstract. The binarization of continuous swarm-intelligence metaheuristics is an area of great interest in operational research. This interest is mainly due to the application of binarized metaheuristics to combinatorial problems. In this article we propose a general binarization algorithm called Percentil Transition Ranking Algorithm (PTRA). PTRA uses the percentile concept as a binarization mechanism. In particular we apply this mechanism to the Cuckoo Search metaheuristic to solve the Set Covering Problem (SCP). We provide necessary experiments to investigate the role of key ingredients of the algorithm. Finally to demonstrate the efficiency of our proposal, Set Covering benchmark instances of the literature show that PTRA competes with the state-of-the-art algorithms.

Keywords: Combinatorial optimization · Set covering problem
Binary metaheuristics · Percentile ranking

1 Introduction

In recent years, the areas of physics and particle intelligence have generated a large number of metaheuristic algorithms [4]. These metaheuristics are suitable for solving a broad class of complex optimization problems. Examples of these algorithms are Firefly Algorithm (FA) [24], magnetic optimization [22], gravitational search algorithm (GSA) [19], Cuckoo Search Algorithm (CSA) [25], Particle Swarm Optimization (PSO) [15]. Part of the success of modern metaheuristics is because they are simple to understand and to implement given their nature inspired behaviour. However many of these algorithms are specifically designed to solve continuous problems.

On the other hand combinatorial problems arise in many areas of computer science and application domains. For example in protein structure prediction, grouping routing, planning, scheduling and timetabling problems to mention only some examples. It is natural to try to apply these algorithms inspired by physics and particle intelligence in combinatorial problems. In the process of adaptation, a series of difficulties arise when moving from continuous spaces to discrete spaces. Examples of these difficulties are spacial disconnect, hamming cliffs, loss of precision and curse of dimension [17]. This has the consequence that binarizations are not always effective and efficient [7, 16].

In this paper, a Percentile Transition Ranking Algorithm (PTRA) is proposed as a general mechanism to binarize continuous metaheuristics. This algorithm is composed of two operators. The main operator corresponds to the percentile ranking transition operator. This operator transfers the exploration/exploitation capabilities of continuous metaheuristics to their discrete or binary adapted version. The second operator corresponds to a perturbation operator. The main goal of this work corresponds to evaluate our binarization algorithm when dealing with an well-known NP-hard combinatorial optimization problem such as the SCP.

To develop the evaluation, we used the Cuckoo Search metaheuristic. Experiments were developed that shed light on the contribution of the different operators to the effectiveness of the algorithm. Moreover our algorithm was compared with algorithms that use a specific Teaching-learning binarization [18], a transfer function Binary Shuffled Frog Leaping Algorithm (BSFLA) [6], and a Lagrangian-based heuristic algorithm [3]. For this purpose we use tests problems from the OR-Library¹ and the instances from Balas and Carrera [1]. The numerical results show that PTRA achieves highly competitive results.

The remainder of this paper is organized as follows. Section 2 briefly introduces the Set Covering problem. In Sect. 3 we explain the transition ranking binarization algorithm. The results of numerical experiment are presented in Sect. 4. Finally we provide the conclusions of our work.

2 Set Covering Problem

The SCP is one of the oldest and most studied optimization problems. SCP is well-known to be NP-hard [10]. Nevertheless, different algorithms for solving it have been developed. There exist exact algorithms who generally rely on the branch-and-bound and branch-and-cut methods to obtain optimal solutions. These methods, however, need an effort for solving an SCP instance that grows exponential with the problem size. Then, even medium-sized problem instances often become intractable and cannot be solved any more using exact algorithms. To overcome this issue, the use of different heuristics have been proposed.

For example, [13] presented a number of greedy algorithms based on a Lagrangian relaxation (called the Lagrangian heuristics), Caprara et al. [3] intro-

¹ OR-Library: <http://www.brunel.ac.uk/mastjjb/jeb/orlib/mknapinfo.html>.

duced relaxation-based Lagrangian heuristics applied to the set covering problem. Metaheuristics also have been applied to solve SCP, some examples are genetic algorithm [26], simulated annealing [2] and ant colony optimization [23]. More recently swarm based metaheuristics as cuckoo search [20], black hole [8], and a Meta-optimization approach [5] were also proposed.

SCP has many practical applications in engineer, e.g., vehicle routing, facility location, railway, and airline crew scheduling [5, 12, 14, 21] problems.

The SCP can be formally defined as follows. Let $A = (a_{ij})$, be a $n \times m$ zero-one matrix, where a column j cover a row i if $a_{ij} = 1$, besides a column j is associated with a non-negative real cost c_j . Let $I = \{1, \dots, n\}$ and $J = \{1, \dots, m\}$, be the row and column set of A , respectively. The SCP consists in searching a minimum cost subset $S \subset J$ for which every row $i \in I$ is covered by at least one column $j \in J$, i.e.,:

$$\text{Minimize } f(x) = \sum_{j=1}^m c_j x_j \quad (1)$$

$$\text{Subject to } \sum_{j=1}^m a_{ij} x_j \geq 1, \forall i \in I, \text{ and } x_j \in \{0, 1\}, \forall j \in J \quad (2)$$

where $x_j = 1$ if $j \in S$, $x_j = 0$ otherwise.

3 Percentile Transition Ranking Algorithm

The application of statistics and data mining, can be found in many areas such as transports, smart cities, agriculture and computational intelligence [7, 9, 11]. In this section, we explore the application of the percentile concept to the binarization of continuous metaheuristics. The Proposed PTR algorithm has three modules. The first module corresponds to the initialization of the feasible solutions. Once the initialization of the particles is performed, it is consulted if the detention criterion is satisfied. This criterion includes a maximum of iterations. Subsequently if the criterion is not satisfied, the percentile transition ranking operator is executed (Sect. 3.1). This module is responsible for performing the iteration of solutions. Once the transitions of the different solutions are made, we compare the resulting solutions with the best solution previously obtained. In the event that a superior solution is found, this replaces the previous one. Finally having met a number of iterations where there has not been a replacement for the best solution, a perturbation operator is used. The general algorithm scheme is detailed in Fig. 1. In the following subsection we will explain in detail the percentile transition ranking operator. A detail explanation of the other operators will be left for an extended version.

3.1 Percentile Transition Ranking Operator

Considering that our metaheuristic is a continuous and swarm intelligence. Due to its iterative nature, it needs to update the position of particles at each iteration. When the metaheuristic is continuous, this update is performed in \mathbb{R}^n

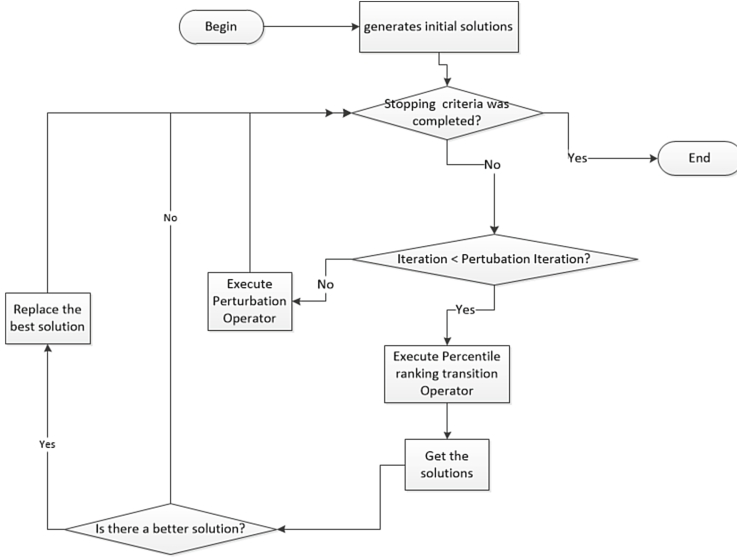


Fig. 1. Flowchart of the percentile transition ranking algorithm

space. In Eq. 3, the update position is presented in a general manner. The $x(t+1)$ variable represents the x position of the particle at time $t+1$. This position is obtained from the position x at time t plus a Δ function calculated at time $t+1$. The function Δ is proper to each metaheuristic and produces values in \mathbb{R}^n . For example in Cuckoo Search $\Delta(x) = \alpha \oplus Levy(\lambda)(x)$, in Black Hole $\Delta(x) = rand \times (x_{bh}(t) - x(t))$ and in the Firefly, Bat and PSO algorithms Δ can be written in simplified form as $\Delta(x) = v(x)$.

$$x(t+1) = x(t) + \Delta_{t+1}(x(t)) \quad (3)$$

In the percentile transition ranking operator, we considering the movements generated by the metaheuristic in each dimension for all particles. $\Delta^i(x)$ corresponds to the magnitude of the displacement $\Delta(x)$ in the i -th position for the particle x . Subsequently these displacement are grouped using $\Delta^i(x)$, the magnitude of the displacement. This grouping is done using the percentile list. In our case the percentile list used the values $\{20, 40, 60, 80, 100\}$.

The percentile operator has as entry the parameters percentile list (percentileList) and the list of values (valuesList). Given an iteration, the list of values corresponds to the magnitude $\Delta^i(x)$ of the displacements of the particles in each dimension. As a first step the operator uses the valueList and obtains the values of the percentiles given in the percentileList. Later, each value in the valueList is assigned the group of the smallest percentile to which the value belongs. Finally, the list of the percentile to which each value belongs is returned (percentileGroupValue). The algorithm is shown in Algorithm 1.

A transition probability through the function P_{tr} is assigned to each element of the valueList. This assignment is done using the percentile group assigned to

each value. For the case of this study, we particularly use the Step function given in Eq. 4.

$$P_{tr}(x^i) = \begin{cases} 0.1, & \text{if } x^i \in \text{group } \{0, 1\} \\ 0.5, & \text{if } x^i \in \text{group } \{2, 3, 4\} \end{cases} \quad (4)$$

Afterwards the transition of each particle is performed. In the case of Cuckoo search the Eq. 5 is used to perform the transition, where \hat{x}^i is the complement of x^i . Finally, each solution is repaired using the repair operator.

$$x^i(t+1) := \begin{cases} \hat{x}^i(t), & \text{if } rand < P_{tg}(x^i) \\ x^i(t), & \text{otherwise} \end{cases} \quad (5)$$

Algorithm 1 percentile ranking operator

```

1: Function percentileRankingTransition(valueList, percentileList)
2: Input valueList, percentileList
3: Output percentileGroupValue
4: percentileValue = getPercentileValue(valueList, percentileList)
5: for each value in valueList do
6:   percetileGroupValue = getPercentileGroupValue(percentileValue, valueList)
7: end for
8: return percetileGroupValue

```

4 Results

In this section, we present computational experiments with the proposed PTRA. For the execution of the instances we use a PC with windows 10, Intel Core i7-4770 processor with 16GB in RAM, and programmed in Python 2.7. Experiments were developed to analyze the key ingredients of our algorithm. To perform the statistical analysis in this study, the non-parametric test of wilcoxon signed-rank test and violin charts are used. The analysis is performed by comparing the dispersion, median and the interquartile range of the distributions. Finally, comparisons were made with different algorithms, a Binary Teaching-Learning Based Optimization [18], a Binary Shuffled Frog Leaping Algorithm (BSFLA) [6], and a Lagrangian-based heuristic algorithm [3] were selected.

4.1 Insight of PTRA Algorithm

In this section we investigate some important ingredients of PTRA to get insight into the behavior of the proposed algorithm. To carry out this comparison the instances from Balas and Carrera were chosen. The contribution of the percentile transition ranking operator on the final performance of the algorithm was studied. The contribution of the perturbation operator will be developed in an extended version. To compare the distributions of the results of the different experiments we use violin Chart. The horizontal axis corresponds to the problems. While Y axis uses the measure % - Gap defined in Eq. 6

$$\% - Gap = 100 \frac{SolutionValue - BestKnown}{BestKnown} \quad (6)$$

Furthermore, a non-parametric test, Wilcoxon signed-rank test is carried out to determine if the results of PTRA with respect to other algorithms have significant difference or not. The parameter settings and browser ranges are shown in Table 1.

Table 1. Setting of parameters for Cuckoo Search Algorithm

Parameters	Description	Value	Range
ν	Coefficient for the perturbation operator	20%	[15, 20, 25]
N	Number of nest	20	[15, 20, 25]
G	Number of percentiles	5	[4, 5, 6]
γ	Step length	0.01	[0.009, 0.01, 0.011]
κ	Levy distribution parameter	1.5	[1.4, 1.5, 1.6]
Iteration number	Maximum iterations	500	[500]

Evaluation of Percentile Transition Ranking Operator: To evaluate the contribution of the percentile transition ranking operator to the final result, we designed a random operator. This random operator executes the transition with a fixed probability (0.5) without considering the ranking of the particle in each dimension. In the evaluation the perturbation operator is included. PTRA corresponds to our standard algorithm. *05.pe* is the random variant that includes the perturbation operator.

When we compared the best values and averages between PTRA and *05.pe* algorithms in Table 2. PTRA outperforms to *05.pe* in all problems. The comparison of distributions is shown in Fig. 2. We see the dispersion of the *05.pe* distributions are bigger than the dispersions of PTRA. In particular this can be appreciated in the problems AA06, AA11, AA13, AA15, AA16 and AA19. When the interquartile ranges are compared, it is observed that the PTRA ranges are closer to 0 than those of *05.pe*. All this suggests that the percentile transition ranking operator together with perturbation operator, contribute to improve the precision and the quality of the results. When we evaluate the behaviour of the algorithms through the Wilcoxon test, this indicates that there is a significant difference between the two algorithms.

4.2 PTRA Comparisons

In this section, we describe the comparisons that were perform to evaluate our algorithm. Two groups of problems were chosen. The first group was obtained from the OR-library. The TBLO [18] and BSFLA [6] algorithms were chosen

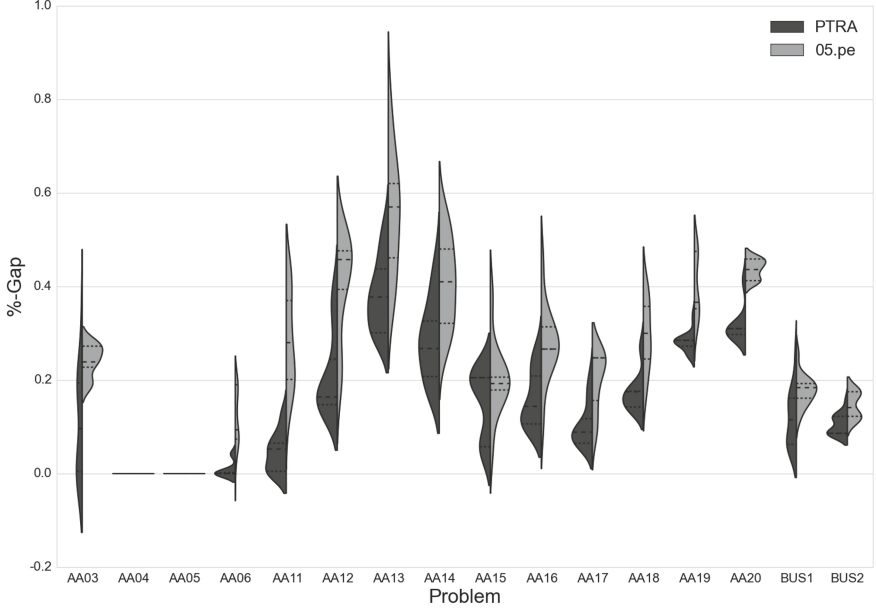


Fig. 2. Evaluation of percentile transition operator with perturbation operator

to perform the comparison. The second group corresponds to the instances of Balas and Carrera [1], and a Lagrangian-based heuristics algorithm [3] was used to perform the comparison.

Comparison with TLBO and BSFLA: The TLBO algorithm adapts in a specific way the principles of the Teaching Learning metaheuristic which works naturally in a continuous space, to solve combinatorial problems such as set covering. In the case of BSFLA, it uses transfer functions that correspond to a general binarization mechanism. The results are shown in the Table 3. PTR found 13 of the 20 Best Known values. When we compare the best values of PTR with respect to TBLO and BSFLA, PTR is superior in all problems. When we compare the averages TLBO has better results in problems E.4, F.3, F.5, G.3, G.5, H.1 and H.4. Regarding convergence times despite not being directly comparable due to differences in implementation and hardware, TBLO turns out to be more efficient than PTR.

Comparison with Lagrangian-based heuristics: In the case of Lagrangian-based heuristics algorithm, the results are shown in the Table 2. The Lagrangian-based heuristics completely outperforms PTR regarding the quality of the solutions. the difference in % - Gap was 0.57 for the Best Value and 0.78 for the average. However the Lagrangian-based heuristic approach is not nearly as straightforward for practitioners to implement and use. On the other

Table 3. OR-Library benchmarks E, F, G, H

Instance	Row	Col	Density (%)	Best known	PTRA (best)	PTRA (avg)	PTRA (time)	TBLO	BSFLA (Best)
E.1	500	5000	10	29	29	29	12	29	29
E.2	500	5000	10	30	30	30	16	31	31
E.3	500	5000	10	27	27	27.8	31	28	28
E.4	500	5000	10	28	28	28.1	7	28	29
E.5	500	5000	10	28	28	28	21	28	28
F.1	500	5000	20	14	14	14	21.5	14	15
F.2	500	5000	20	15	15	15	11.2	15	15
F.3	500	5000	20	14	14	14.5	31.5	14	16
F.4	500	5000	20	14	14	14	21.3	14	15
F.5	500	5000	20	13	13	13.8	16.4	13	15
G.1	1000	10000	2	176	176	177.9	192	179	182
G.2	1000	10000	2	154	155	155.9	143	156	161
G.3	1000	10000	2	166	167	168.7	176	168	173
G.4	1000	10000	2	168	170	171.1	198	172	173
G.5	1000	10000	2	168	168	169.7	143	168	174
H.1	1000	10000	5	63	64	64.9	185	64	68
H.2	1000	10000	5	63	64	64	165	64	66
H.3	1000	10000	5	59	60	60.8	185	61	62
H.4	1000	10000	5	58	59	60.4	154	59	63
H.5	1000	10000	5	55	55	55.4	120	56	59
Average				67.1	67.5	68.2	92.5	68.1	70.1

hand PTRa uses mechanisms quite simple to implement, besides easily adapting to different combinatorial problems and any continuous swarm-intelligence metaheuristics.

5 Conclusion and Future Work

In this article, we proposed a algorithm whose main function is to binarize continuous swarm-intelligence metaheuristics. To evaluate the performance of our algorithm, the set covering problem was used together with the Cuckoo Search metaheuristic. The contribution of the main operator of the algorithm was evaluated, finding that the percentile transition ranking operator contributes significantly to improve the precision and quality of the solutions. Finally, in comparison with state of the art algorithms our algorithm showed a good performance.

As a future works we want to investigate the behaviour of other metaheuristics. Furthermore, the algorithm must be verified with other NP-hard problems. Moreover to simplify the choice of the appropriate configuration, it is important to explore adaptive techniques. From an understanding point of view of how the framework performs binarization, it is interesting to understand how the algorithm alters the properties of exploration and exploitation. Also is interesting to study how the velocities and positions generated by continuous metaheuristics are mapped to positions in the discrete space.

References

1. Balas, E., Carrera, M.C.: A dynamic subgradient-based branch-and-bound procedure for set covering. *Oper. Res.* **44**(6), 875–890 (1996)
2. Brusco, M.J., Jacobs, L.W., Thompson, G.M.: A morphing procedure to supplement a simulated annealing heuristic for cost-and-coverage-correlated set-covering problems. *Ann. Oper. Res.* **86**, 611–627 (1999)
3. Caprara, A., Fischetti, M., Toth, P.: A heuristic method for the set covering problem. *Oper. Res.* **47**(5), 730–743 (1999)
4. Crawford, B., Soto, R., Astorga, G., García, J., Castro, C., Paredes, F.: Putting continuous metaheuristics to work in binary search spaces. *Complexity* **2017** (2017)
5. Crawford, B., Soto, R., Monfroy, E., Astorga, G., García, J., Cortes, E.: A meta-optimization approach for covering problems in facility location. In: *Workshop on Engineering Applications*, pp. 565–578. Springer, Berlin (2017)
6. Crawford, B., Soto, R., Peña, C., Palma, W., Johnson, F., Paredes, F.: Solving the set covering problem with a shuffled frog leaping algorithm. *Intelligent Information and Database Systems*, pp. 41–50. Springer, Berlin (2015)
7. García, J., Crawford, B., Soto, R., Carlos, C., Paredes, F.: A k-means binarization framework applied to multidimensional knapsack problem. *Appl. Intell.* 1–24 (2017)
8. García, J., Crawford, B., Soto, R., García, P.: A multi dynamic binary black hole algorithm applied to set covering problem. In: *International Conference on Harmony Search Algorithm*, pp. 42–51. Springer, Berlin (2017)
9. García, J., Pope, C., Altimiras, F.: A distributed k-means segmentation algorithm applied to lobesia botrana recognition. *Complexity* **2017** (2017)

10. Gary, M.R., Johnson, D.S.: *Computers and Intractability: A Guide to the Theory of NP-Completeness* (1979)
11. Graells-Garrido, E., García, J.: Visual exploration of urban dynamics using mobile data. In: *International Conference on Ubiquitous Computing and Ambient Intelligence*, pp. 480–491. Springer, Berlin (2015)
12. Horváth, M., Kis, T.: Computing strong lower and upper bounds for the integrated multiple-depot vehicle and crew scheduling problem with branch-and-price. *Cent. Eur. J. Oper. Res.* 1–29 (2017)
13. John, B.: A lagrangian heuristic for set-covering problems. *Nav. Res. Logist. (NRL)* **37**(1), 151–164 (1990)
14. Kasirzadeh, A., Saddoune, M., Soumis, F.: Airline crew scheduling: models, algorithms, and data sets. *EURO J. Transp. Logist.* **6**(2), 111–137 (2017)
15. Kennedy, J.: Particle swarm optimization. *Encyclopedia of Machine Learning*, pp. 760–766. Springer, Berlin (2011)
16. Lanza-Gutierrez, J.M., Crawford, B., Soto, R., Berrios, N., Gomez-Pulido, J.A., Paredes, F.: Analyzing the effects of binarization techniques when solving the set covering problem through swarm optimization. *Expert Syst. Appl.* **70**, 67–82 (2017)
17. Leonard, B.J., Engelbrecht, A.P., Cleghorn, C.W.: Critical considerations on angle modulated particle swarm optimisers. *Swarm Intell.* **9**(4), 291–314 (2015)
18. Lu, Y., Vasko, F.J.: An or practitioner’s solution approach for the set covering problem. *Int. J. Appl. Metaheuristic Comput. (IJAMC)* **6**(4), 1–13 (2015)
19. Rashedi, E., Nezamabadi-Pour, H., Saryazdi, S.: GSA: a gravitational search algorithm. *Inf. Sci.* **179**(13), 2232–2248 (2009)
20. Soto, R., Crawford, B., Olivares, R., Barraza, J., Johnson, F., Paredes, F.: A binary cuckoo search algorithm for solving the set covering problem. *Bioinspired Computation in Artificial Systems*, pp. 88–97. Springer, Berlin (2015)
21. Stojković, M.: The operational flight and multi-crew scheduling problem. *Yugoslav J. Oper. Res.* **15**(1) (2016)
22. Totonchi, A., Reza, M.: Magnetic optimization algorithms, a new synthesis. In: *IEEE International Conference on Evolutionary Computations* (2008)
23. Valenzuela, C., Crawford, B., Soto, R., Monfroy, E., Paredes, F.: A 2-level meta-heuristic for the set covering problem. *Int. J. Comput. Commun. Control* **7**(2), 377–387 (2014)
24. Yang, X.-S.: Firefly algorithm, stochastic test functions and design optimisation. *Int. J. Bio-Inspired Comput.* **2**(2), 78–84 (2010)
25. Yang, X.-S., Deb, S.: Cuckoo search via lévy flights. In: *World Congress on Nature and Biologically Inspired Computing*, 2009, NaBIC 2009, pp. 210–214. IEEE (2009)
26. Yelbay, B., Birbil, Ş.İ., Bülbül, K.: The set covering problem revisited: an empirical study of the value of dual information. *Eur. J. Oper. Res.* (2012)

An Improved Hybrid Firefly Algorithm for Solving Optimization Problems

Fazli Wahid^{1(✉)}, Rozaida Ghazali¹, and Habib Shah²

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
wahid_uomian@hotmail.com, rozaida@uthm.edu.my

² Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia
hurrahman@kku.edu.sa

Abstract. The standard firefly algorithm is suffered from three major drawbacks. Firstly, imbalanced exploration and exploitation due to random initial solution generation. Secondly, the local convergence rate is low when the randomization factor is large. Thirdly, low quality local and global search capability at termination stage that result in failing to get the most optimal solution. To overcome all these drawbacks, a new approach is introduced which has been named GA-FA-PS algorithm in which genetic algorithm (GA) has been applied to generate the initial solution for balancing the exploration and exploitation at the initial stage. In the second stage, crossed over operator is embedded in firefly changing position to improve local search which ultimately enhances local convergence. To further improve the local and global convergence rate, pattern search (PS) is introduced which is used to obtain the most optimal solution or at least the solution better than the solution provided by the standard firefly algorithm. The performance of the proposed approach has been compared with standard FA and GA and the proposed method outperforms both of these approaches in terms solution quality.

Keywords: Firefly algorithm · Swarm intelligence · Genetic algorithm
Pattern search · GA-FA-PS

1 Introduction

1.1 Background

Swarm intelligence is a field of artificial intelligence (AI) that gained high popularity over last couple of years [1]. Swarm intelligence gets inspiration from the collective behavior of social swarms of bees, ants, worms, termites, schools of fish and flock of birds. Although the main constituent of these swarms is the individuals, a coordinated behavior exists among these individuals that assist them in obtaining their desired goals. A self-organizing behavior exists inside the whole system which is the result of this coordinated behavior and swarm intelligence or collective intelligence and a simple rule based interaction of multi-agent systems. A coordinated interaction among different individuals in the swarm performs this well-managed behavior.

A trade-off is required between the collection of new information (Exploration) and the use of existing information (exploitation) during the search of new food sources [2]. The process of exploration and exploitation is used by the bee colony for maximizing the nectar amount and minimizing the foraging efforts. A collective behavior is shown by the swarm of individuals for foraging, reproduction, living and division of important tasks among the available individuals. In fact, a decentralized manner by individuals based on local information collected from the environment is used for decision making.

Swarm intelligence is a research field which is concentrated on the collective behavior inside a decentralized and self-organized system. Beni and Wang [3] used this term for the first time in the sense of cellular robotic system that consists of simple agents organizing themselves using neighborhood interaction system. Recently, the swarm intelligence methods have been applied in the robot control systems, solving optimization problems, load balancing and routing in new mobile telecommunication networks, demanding flexibility and robustness. Examples of swarm intelligence methods used for solving optimization problems include PSO (Particle swarm optimization) [4], ant colony optimization (ACO) [5], artificial bee colony optimization (ABC) [6–8]. Recently, some more advanced swarm intelligence optimization algorithms e.g. cuckoo search [9], firefly algorithm (FA) [10], bat algorithm [11], krill herd bio-inspired optimization algorithm [12] and clustering algorithms [13] have been introduced.

1.2 Firefly Algorithm (FA)

Firefly algorithm is one of the most recently introduced swarm intelligence technique developed by Yang [10] by taking inspiration from flash light signals which is the source of attraction among fireflies for potential mates. All the fireflies are unisexual and attract each other according to the intensities of their flash lights. Higher the flash light intensity, higher is the power of attraction and vice versa. For solving optimization problem, the brightness of flash is associated with the fitness function to be optimized. The light intensity $I(r)$ of a firefly at distance r is given by Eq. (1):

$$I(r) = \frac{I_0}{r^2} \quad (1)$$

where I_0 is the intensity of light at source. With the fixed light absorption coefficient γ of the medium, the intensity of light I at distance r is given by Eq. (2):

$$I = I_0 \exp(-\gamma r^2) \quad (2)$$

where r represents the distance between the fireflies in the firefly algorithm. Since the attractiveness of the fireflies by the adjacent fireflies is directly proportional to the intensity of light of the fireflies, the attractiveness of a firefly by another firefly is given by the Eq. (3):

$$B = \beta_0 \exp(-\gamma r^m) \quad (m \geq 1) \quad (3)$$

where β_0 represents the attractiveness at distance $r = 0$. The distance between two fireflies' x_i and y_j called as Euclidian distance is given by Eq. (4):

$$r_{ij} = |x_i - y_j| = \sqrt{\sum_{k=1}^d (x_{i,k} - y_{j,k})^2} \quad (4)$$

In each generation, the position of the firefly can be changed according to the following Eq. (5):

$$X_i = x_i + \beta_0 \exp(-\gamma r_{ij}^2) (x_i - y_j) + \alpha \varepsilon \quad (5)$$

where α represents the parameter of randomization, ε represents random number generated in Gaussian distribution.

To overcome the problem of balancing between the exploration and exploitation, many researchers introduced modifications in the standard FA to improve its efficiency. The improved versions of FA were introduced by the authors in [14–16]. All these authors introduced a single point modifications in some parts of the FA e.g. some authors modified the random movements of the FA, some authors brought some changes in the light absorption coefficient factor whereas other authors modified the step size of the firefly changing position equation. All these authors have addressed one or other issue mentioned earlier but failed to address all these problems identified in the FA. To balance the exploration and exploitation of the FA, many researchers have hybridized the FA with other optimization algorithms. FA has been used by the authors in [17–19] in hybrid mechanism with other optimization algorithms. In these approaches, the authors have used FA for exploration and other optimization algorithms for exploitation purposes or the other algorithms for exploration and FA for exploitation purposes but all these authors have not properly addressed the issues raised in the standard FA. In our work, we will develop a new hybrid firefly algorithm which will eliminate the major drawbacks for getting the most optimal value or at least the values better than the standard firefly algorithm when solving optimization problems.

2 Proposed Approach

The standard firefly algorithm, different modifications introduced in different parts of firefly algorithm and its different hybrid approaches suffer from three major drawbacks. The initial solution is generated randomly that results in imbalanced exploration and exploitation at this stage. The solution is trapped into local optima if the randomization factor in firefly changing position is very small and may skip the best solution if this is very large. Thirdly, if the maximum generation reaches and the solution obtained is not the most optimal solution. The improved hybrid approach is based on hybridization of three algorithms namely genetic algorithm, firefly algorithm and pattern search which will overcome all these three drawbacks. The newly introduced approach is called GA-FA-PS. The proposed approach is developed using following steps.

Step 1: Initial Population Generation

The initial population is generated randomly using the dependent and independent variables. The independent variables represent the variables to be optimized. The dependent variable represents the function to be optimized.

Step 2: Applying Genetic Algorithm

After the initial population is generated randomly, the independent and dependent variables are given as input to the genetic algorithm to represent genes and chromosomes of genetic algorithm. Genetic algorithm is a bio-inspired optimization approach that is inspired from the behavior transferring phenomenon of parents into children using heredity characteristics. Different types of characters are transferred from parents into children using various types of operators e.g. selection, mutation and cross over operators.

Step 3: Fitness function evaluation

After the assignment of independent and dependent variables, the fitness of each individual which represents the fitness value of the function to be optimized is evaluated.

Step 4: Generate the solution set for firefly algorithm

The genetic algorithm is run for a fixed number of iteration to generate the initial solution set for the firefly algorithm.

Step 5: Specification of fireflies and light intensities

From the solution set generated in step 4, assign the values of independent variables to fireflies and the dependent variable to light intensity associated with each firefly.

Step 6: Firefly changing position

During running the firefly algorithm, if the solution is not improved after few generations, modification is brought by introducing the cross over operator to enhance the exploitation at this stage.

Step 7: Firefly algorithm termination

The firefly algorithm is run for maximum number of generation and then terminated.

Step 8: Embed pattern search

If the maximum number of generations of firefly algorithm reaches and the solution obtained is not the most optimal solution, then the pattern search algorithm is introduced to further enhance the exploration and exploitation of the solution search space to obtain the most optimal solution or at least the solution better than the so far obtained solution. Pattern search is a kind of optimization algorithm having strong capability of solving various kinds of optimization problems where other standard optimization algorithms face failure in finding the most optimal solution. Its easy implementation,

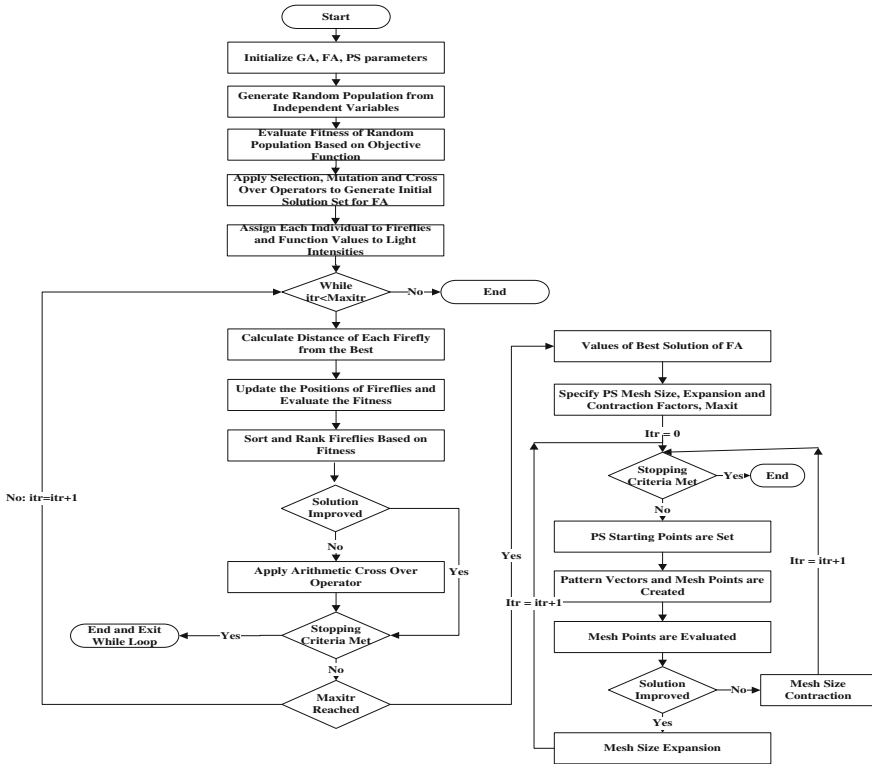


Fig. 1. Proposed approach

conceptual simplicity and computational efficiency make it more applicable than the other optimization techniques.

The proposed approach is shown in Fig. 1.

3 Experimental Setup

All the experiments were carried out using Core i3 with Windows 7 and Matlab 2015a installed on it. The detailed description of the whole experimental setup is given in this section. Different parameters of GA, FA and the proposed GA-FA-PS approaches are shown in Table 1.

GA parameters		FA parameters		GA-FA-PS parameters	
Maximum generations	200	MaxIt	200	MaxIt	200 (GA = 60, FA = 70, PS = 70)
Population	80	nPop	30	nPop	30
Cross over		Gamma	1	Gamma	1

(continued)

(continued)

GA parameters		FA parameters		GA-FA-PS parameters	
Maximum generations	200	MaxIt	200	MaxIt	200 (GA = 60, FA = 70, PS = 70)
	One point cross over				
Cross over probability	0.9	Beta	2	Beta	2
Mutation rate	0.1	Alpha	0.2	Alpha	0.2
Selection method	Rank based selection	Alpha_damp	0.98	Alpha_damp	0.98

Table 1. Comparative analysis of the proposed approach with other algorithms

Test function	Algorithm	Best case	Average best case	Worst case	Average worst case
Ackley	FA	0.7392	1.2174	8.0193	3.9371
	GA	0.5826	0.8917	6.8310	2.8649
	GA-FA-PS	0.0438	0.2516	0.9758	0.6192
Rosenbrock	FA	4.2832	8.7210	15.6209	10.2019
	GA	3.1271	6.9018	12.9047	9.4922
	GA-FA-PS	-5.2916	-1.9201	3.0154	2.0618
Sphere	FA	0.4173	1.3073	6.3361	2.1198
	GA	0.1709	0.7308	5.0193	1.8922
	GA-FA-PS	-7.0293	-3.0089	1.6911	-2.1015

3.1 Test Functions

For testing the performance of the proposed GA-FA-PS approach, three numerical benchmark functions namely Ackley's function, Rosenbrock's function and sphere functions were used in the experiments [20]. The global optimal value and the range of parameters are described as follow.

3.1.1 Ackley's Function

Ackley's function is a type of multimodal function mainly applied for testing the performance of the optimization algorithms and expressed by the following mathematical Eq. (6):

$$f_1(x) = -20 * \exp[-0.2\sqrt{\frac{1}{d}\sum_{i=1}^d x_i^2}] - \exp * [\frac{1}{d}\sum_{i=1}^d \cos(2\pi * x_i)] + (20 + e) \quad (6)$$

where the global minima for this function has been set to 0 in the range of $[-50, 50]^d$ where d represents the dimensions of problem and $i = (1, 2, \dots, d)$.

3.1.2 Rosenbrock's Function

Rosenbrock is a type of unimodal optimization function with mathematical representation by Eq. (7):

$$f_2(x) = \sum_{i=1}^{d-1} \left[(1 - x_i)^2 + 100 * (x_{i+1} - x_i^2)^2 \right] \quad (7)$$

The initial range for this function has been set to $[-120, 120]^d$ whereas the global minima for this function has been set to -5 to test the efficiency of the proposed optimization approach.

3.1.3 Sphere Function

The sphere function is given by Eq. (8):

$$f_3(x) = \sum_{i=1}^d x_i^2 \quad (8)$$

The initial range for this function has been set to $[-100, 100]^d$ whereas the global minima has been set to -10 .

3.2 Results

The proposed approach has been compared with the standard firefly algorithm and the standard genetic algorithm. For making proper comparison and meaningful analysis, extensive experimentation has been carried out. All of the three algorithms namely genetic algorithm, standard firefly algorithm and the proposed GA-FA-PS algorithms have been run 20 times and the results are shown in Table 1. The proposed approach was compared with the standard FA and GA by computing their best case, average best case, worst case and average worst case as shown in Table 1.

It is evident from the above table that the proposed approach outperforms both the FA and GA algorithms for all the three benchmark functions namely Ackley, Rosenbrock and Sphere functions.

3.3 Convergence

The convergence rate of the three algorithms is shown in Figs. 2, 3 and 4 respectively.

Figure 2 shows the convergence rates of genetic algorithm, standard firefly algorithm and the proposed approach.

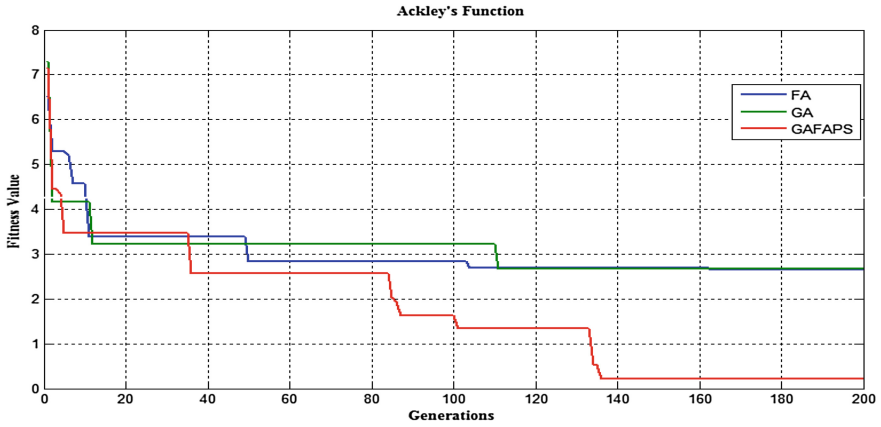


Fig. 2. Convergence rates of GA, FA and GA-FA-PS for Ackley's function

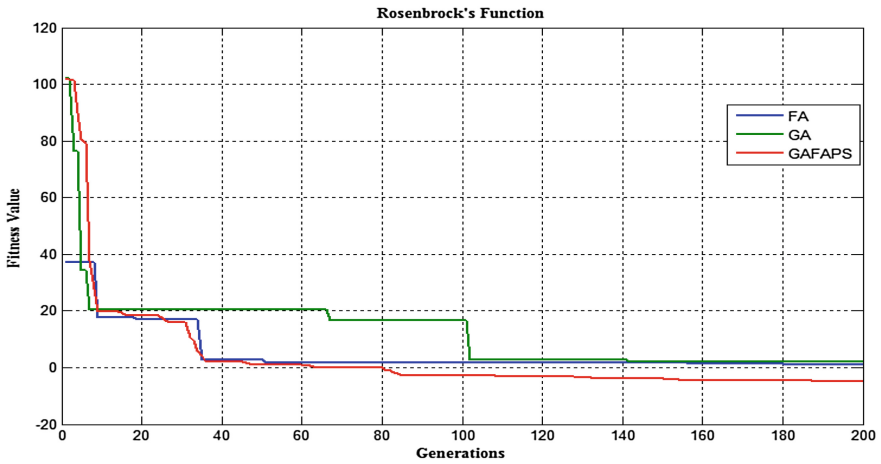


Fig. 3. Convergence rates of GA, FA and GA-FA-PS for Rosenbrock's function

In the convergence of the three algorithms, up to iteration 60, the lowest convergence has been observed for genetic algorithm followed by firefly algorithm and the proposed approach, respectively. After iteration 80, there is continuous improvement in the convergence rate of the proposed algorithm with sharp improvement after iteration 135 with the introduction of pattern search for bringing improvement.

Figure 4 shows the convergence rate of the genetic algorithm, firefly algorithm and the proposed approach. For the sphere function, the proposed approach has sudden changes in the convergence rate for getting the most optimal value.

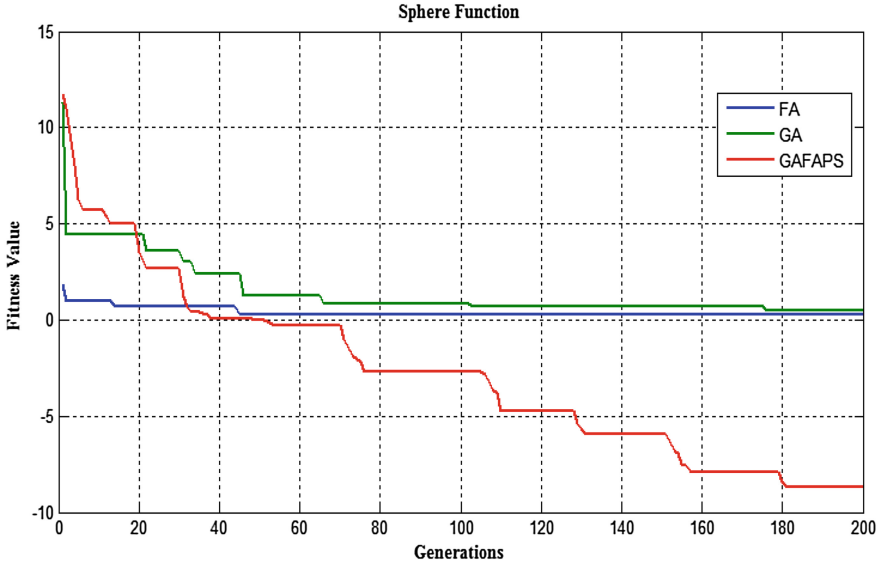


Fig. 4. Convergence rates of GA, FA and GA-FA-PS for sphere function

4 Conclusion

In this paper, a new hybrid algorithm of firefly algorithm in combination with the genetic algorithm and pattern search has been developed. In the proposed approach, the initial solution set was generated by genetic algorithm. To further enhance the convergence rate of the firefly algorithm, arithmetic cross over operator was introduced in firefly changing position. Pattern search was applied to further improve the solution quality of the firefly algorithm. The proposed approach was tested on three benchmark optimization functions namely Ackley's function, Rosenbrock and sphere function. The performance of the proposed approach was compared with the genetic algorithm and standard firefly algorithm. The developed algorithm outperforms both of these algorithms in terms of local and global convergence rate that results into better solution quality.

Acknowledgements. The authors would like to thank King Khalid University to provide the International Research Grant with Grant number A134 for supporting this research.

References

1. Blum, C., Li, X.: Swarm intelligence in optimization. In: Swarm Intelligence, pp. 43–85. Springer, Berlin, Heidelberg (2008)
2. Beekman, M., Sword, G.A., Simpson, S.J.: Biological foundations of swarm intelligence. In: Swarm intelligence, pp. 3–41. Springer, Berlin, Heidelberg (2008)

3. Beni, G., Wang, J.: Swarm intelligence in cellular robotic systems. In: *Robots and Biological Systems: Towards a New Bionics*, pp. 703–712. Springer, Berlin, Heidelberg (1993)
4. Kennedy, J., Eberhart, R.C.: The particle swarm: social adaptation in information-processing systems. In: *New Ideas in Optimization*, pp. 379–388. McGraw-Hill Ltd., UK (1999)
5. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. *IEEE Comput. Intell. Mag.* **4**(1) 28–39 (2006)
6. Shah, H., Ghazali, R.: Prediction of earthquake magnitude by an improved ABC-MLP. In: *Developments in E-systems Engineering (DeSE)*, pp. 312–317. IEEE (2011)
7. Shah, H., Ghazali, R., Nawawi, N.M.: Global artificial bee colony algorithm for boolean function classification. In: *Asian Conference on Intelligent Information and Database Systems*, pp. 12–20. Springer, Berlin, Heidelberg (2013)
8. Wahid, F., Kim, D.H.: An efficient approach for energy consumption optimization and management in residential building using artificial bee colony and fuzzy logic. *Math. Probl. Eng.* 1–13 (2016)
9. Yang, X.S., Deb, S.: Cuckoo search via Lévy flights. In: *IEEE World Congress on Nature & Biologically Inspired Computing, NaBIC*, pp. 210–214 (2009)
10. Yang, X.S.: Firefly algorithms for multimodal optimization. In: *International Symposium on Stochastic Algorithms*, pp. 169–178. Springer, Berlin, Heidelberg (2009)
11. Yang, X.S.: A new metaheuristic bat-inspired algorithm. *Nat. Inspir. Coop. Strateg. Optim. NISCO* 65–74 (2010)
12. Gandomi, A.H., Alavi, A.H.: Krill herd: a new bio-inspired optimization algorithm. *Commun. Nonli. Sci. Num. Simul.* **17**, 4831–4845 (2012)
13. Hatamlou, A., Abdullah, S., Nezamabadi-Pour, H.: A combined approach for clustering based on K-means and gravitational search algorithms. *Swarm Evol. Comput.* **6**, 47–52 (2012)
14. Yu, S., Yang, S., Su, S.: Self-adaptive step firefly algorithm. *J. Appl. Math.* (2013)
15. Gupta, A., Padhy, P.K.: Modified Firefly Algorithm based controller design for integrating and unstable delay processes. *Eng. Sci. Technol. Int. J.* **19**, 548–558 (2016)
16. Sundari, M.G., Rajaram, M., Balaraman, S.: Application of improved firefly algorithm for programmed PWM in multilevel inverter with adjustable DC sources. *Appl. Soft Comput.* **41**, 169–179 (2016)
17. Kaushik, K., Arora, V.: A hybrid data clustering using firefly algorithm based improved genetic algorithm. *Proced. Comput. Sci.* **58**, 249–256 (2015)
18. Farook, S.: Regulating LFC regulations in a deregulated power system using Hybrid Genetic-Firefly algorithm. In: *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–7. IEEE (2015)
19. Sur, U., Gautam, S.: Hybrid firefly algorithm based distribution state estimation with regard to renewable energy sources. In: *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, pp. 1–6. IEEE (2016)
20. Zhu, G., Kwong, S.: Gbest-guided artificial bee colony algorithm for numerical function optimization. *Appl. Math. Comput.* **217**, 3166–3173 (2010)

Exploration and Exploitation Measurement in Swarm-Based Metaheuristic Algorithms: An Empirical Analysis

Mohd Najib Mohd Salleh¹(✉), Kashif Hussain¹, Shi Cheng², Yuhui Shi³,
Arshad Muhammad⁴, Ghufraan Ullah⁵, and Rashid Naseem⁵

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
najib@uthm.edu.my

² School of Computer Science, Shaanxi Normal University, Xian, China

³ Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

⁴ Faculty of Computing and Information Technology, Sohar University, Sohar, Oman

⁵ Department of Computer Science, City University of Science and Information Technology, Peshawar, Pakistan

Abstract. Swarm-based metaheuristics, inspired from intelligent social behaviors in nature, have achieved wider acceptance among researchers as compared to other population-based methods. The success of any swarm-based algorithm highly depends upon the mechanism of social interaction which maintains the balance between exploration and exploitation. This research examines these two significant cornerstones of top five swarm-based metaheuristics using diversity measurement. The results show that ACO and FA maintained balance between exploration and exploitation throughout iterations thus achieved better results as compared to counterparts taken in this study.

Keywords: Swarm intelligence · Metaheuristics · Optimization
Exploration and exploitation

1 Introduction

In operations research, metaheuristic algorithms have an edge over deterministic methods for solving hard optimization problems. There exists plenty of literature proving outstanding results of applications on science and engineering optimization problems e.g. [1–3], etc. Furthermore, in metaheuristic research, swarm-based algorithms have earned more popularity over other population and natural evolution based methods [4, 5]. Moreover, the popularity of particle swarm optimization (PSO) [6], artificial bee colony (ABC) [7], and ant colony optimization (ACO) [8] motivated emergence of more and more swarm-based algorithms.

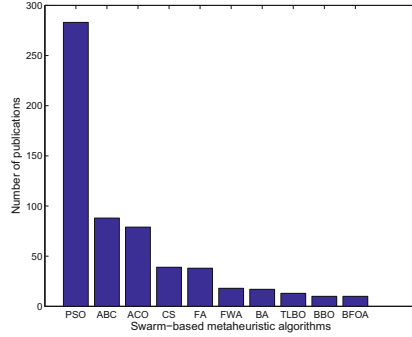


Fig. 1. Popular swarm-based metaheuristic algorithms

According to a preliminary survey performed for this paper, there appear at least more than 50 swarm-based metaheuristics which present a vast collection for practitioner to choose the best suitable algorithm for the problem in hand. Among these algorithms, PSO, ABC, ACO, cuckoo search (CS) [9], and firefly algorithm (FA) [10] are the most popular metaheuristics under the category of swarm intelligence (Fig. 1). Rest of the algorithms among top ten, according to the number of publications between 1995 and 2015, are FWA [11], BA [12], TBLO [13], BBO [14], BFOA [15].

The critics emphasize metaheuristic research to be directed more towards insightful analysis instead of solely raising the pile of methods that beat one or the other algorithms in one or the other optimization problems [16, 17]. Even though, “no-free-lunch” theorem is adequately discussed as evidence, yet in-depth analysis will reveal those relatively unknown reasons of varying metaheuristic performances on different applications. The measurement of exploration and exploitation, the corner-stone of any metaheuristic, opens the knots of questions that merely convergence graph and end results may not convey the point. Therefore, this research is aimed at analyzing the performances of top five popular metaheuristics, mentioned above, with the help of dimension-wise diversity measurement proposed by [18]. For optimization problems, this experimental study considered two popular benchmark test functions.

Rest of the paper is organized as follows. The following section briefs about swarm-based metaheuristic algorithms taken in this study. Section 3 explains the method adopted to measure diversity-based measurement of exploration and exploitation in order to observe the reasons behind difference in metaheuristic performances. Followed by Sect. 4 which describes the experimental settings, are the observations reported and further analyzed in Sect. 5. Lastly, but importantly, this empirical study is concluded by Sect. 6.

2 Swarm-Based Metaheuristic Algorithms

Besides other population-based metaheuristic algorithms, swarm-based methods employ swarm agents which socially interact to make collective decision about the search moves. Inspired from various highly intelligent swarm behaviors in nature, e.g. flocks of birds, school of fish, and colonies of ants, etc., researchers have developed dozens of optimization algorithms. Among those algorithms, some have significantly achieved acceptance in research community. Such popular algorithms are taken into observations in this experimental study. Due to limited space, following is a brief introduction of these algorithms, since the greater detail can be found in the cited papers.

Particle Swarm Optimization (PSO) [6] uses particles, representing a flock of birds or school of fish, to observe the search environment based on cognitive and social intelligence for searching the best food location. Particles in PSO have velocity and position. Initially, the swarm individuals are randomly placed, and as the iterations proceed, the next move is decided on the basis of current position and new velocity which is calculated with respect to personal best position and the globally best position. To control the explorative and exploitative capability of the swarm, PSO uses a parameter called inertia weight. Two more parameters, cognitive factor and social factor, are also tuned for optimum performance.

Artificial Bee Colony (ABC) [7] mimics the swarm behavior of bees that fly in search of best location for collecting nectar before returning to the hive. The swarm agents in this algorithm are employed bees, onlooker bees, and scout bees. Employed bees are the first to sightsee and discover food sources, followed by onlooker bees which pursue the potential locations shared by employed bees. This is based on the probability of selection of employed bees using Roulette wheel selection method. After certain number of attempts (controlled by parameter *Limit*), when some of the bees that are unable to produce better results, are replaced with scout bees that randomly explore the search space.

Ant Colony Optimization (ACO) [8] metaphorizes the foraging behavior of ants that follow the shortest path with maximum pheromone representing optimum food source. Initially ants search food source randomly, and while returning to the nest, deposit pheromone along the path which is, later on, gauged and reinforced by other ants through further depositing the pheromone. The path that is the shortest to food source and contains dense pheromone is considered by most of the ants, hence converging to optimum location. Since ACO is mainly designed for combinatorial optimization problems, we chose $ACO_{\mathbb{R}}$ [19] for solving continuous optimization problems in this work.

Cuckoo Search (CS) [9] algorithm follows the way Cuckoo birds search the most suitable host nests where the laid eggs can survive maximum. The eggs that survive and hatch successfully develop further societies. CS starts with initial random solutions in terms of habitats (host nests) where cuckoos lay eggs. Each habitat has fitness value representing suitability of eggs to survive. After laying new eggs in randomly chosen host nests in the predefined radius, certain percentage of eggs with worst fitness value are destroyed. CS uses levy

flight random walk to decide the next move. There is only one parameter which is discovery rate of poor eggs to be destroyed and replaced with new ones.

Firefly Algorithm (FA) [10] mimics the way fireflies glow for attracting other fireflies. The light intensity increases and decreases with respect to distance from other counterparts. FA starts with initial random population with random light intensity which, in later iterations, varies based on light absorption concept. In FA, the best firefly that converges others is the one with highest light intensity. For exploring the unseen neighborhoods, FA employs light absorption parameter.

3 Exploration and Exploitation Measurement

Exploration and exploitation are the omnipresent two cornerstones of any metaheuristic algorithm, if balanced adequately, drive any algorithm towards successful convergence. Exploration is visiting entire search space for discovering unseen locations, whereas exploitation refers to furthering search towards potential neighborhoods already visited [20]. The tradeoff between the two features is highly dependent on search philosophy adopted by a metaheuristic algorithm, that is searching for highly potential regions and avoiding unnecessary traps in local minima [21].

Since, merely observing convergence graph and end results does not explain these two features of an algorithm, this research adopts dimension-wise diversity measurement to perform in-depth analysis. That said, for swarm-base metaheuristic algorithms, it is significantly important to analyze the behavior of each individual in a swarm, as well as, swarm as a whole. In any metaheuristic algorithm, the parameter values of D dimensional population individuals converge towards values obtained by best individual in a population. Hence, in case of exploration the distance between values increases, and otherwise, it decreases while converging during exploitation phase. This motivated the research to adopt dimension-wise diversity measurement proposed by [18] with modification where mean is replaced with median in (1); as it reflects center of the dimension j in population more effectively.

$$\begin{aligned} Div_j &= \frac{1}{n} \sum_{i=1}^n median(x^j) - x_i^j; \\ Div &= \frac{1}{D} \sum_{j=1}^D Div_j \end{aligned} \tag{1}$$

where $median(x^j)$ is the median of dimension j in whole swarm, whereas x_i^j is the dimension j of swarm individual i , and n is the size of swarm. After taking dimension-wise distance of each swarm-individual i from the median of dimension j , we take average Div_j for all the individuals. Later on, Div_j is summed and averaged for diversity measurement Div of the swarm.

Furthermore, with the help of diversity measurement in hand, we calculated percentage of exploration and exploitation during each iteration using (2):

$$\begin{aligned} Xpl\% &= \left(\frac{Div}{Div_{max}} \right) \times 100; \\ Xpt\% &= \left(\frac{|Div - Div_{max}|}{Div_{max}} \right) \times 100 \end{aligned} \quad (2)$$

Table 1. Algorithm-specific parameter settings

Algorithm	Parameter settings
PSO	$\omega = [0.9 - 0.4], C_1 = C_2 = 2$ [22]
ABC	$Limit = SwarmSize \times D$ [23]
ACO _R	$\tau_0 = 1, \rho = 0.5, \omega = 0.5$ (Weight factor), $z = 1$ (Deviation–Distance Ratio) [19]
CS	$\rho = 0.25$ [24]
FA	$\beta_0 = 1, \gamma = 1, \alpha = 0.2$ [25]

where $Xpl\%$ and $Xpt\%$ are exploration and exploitation percentages respectively, while Div_{max} is the maximum diversity value of the swarm in an iteration.

4 Experimental Settings

In order to analyze the two highly influential factors (exploration and exploitation) of the metaheuristic algorithms, we examined the performances on numerical optimization problems. Two popular continuous benchmark test functions chosen for this experimental work are: Sphere (unimodal and separable) and Ackley (multimodal and non-separable), with 30 dimensions. For all the algorithms, the population size was 50 and maximum iterations were 1500. Since, the purpose was to analyze explorative and exploitative capability of the algorithms instead of merely comparing best objective function values; the algorithms were run once with adequately sufficient number of iterations as few of the preliminary tests proved the point. Apart from common experimental settings, Table 1 presents algorithm specific parameter settings, taken from literature.

5 Results Discussion

In this research, we observed explorative and exploitative capabilities of top five swarm-based metaheuristic algorithms on two benchmark test functions. According to results depicted in Figs. 2, 3 and 4, ACO_R and FA achieved best objective function values (Table 2) in both the functions because of consistent

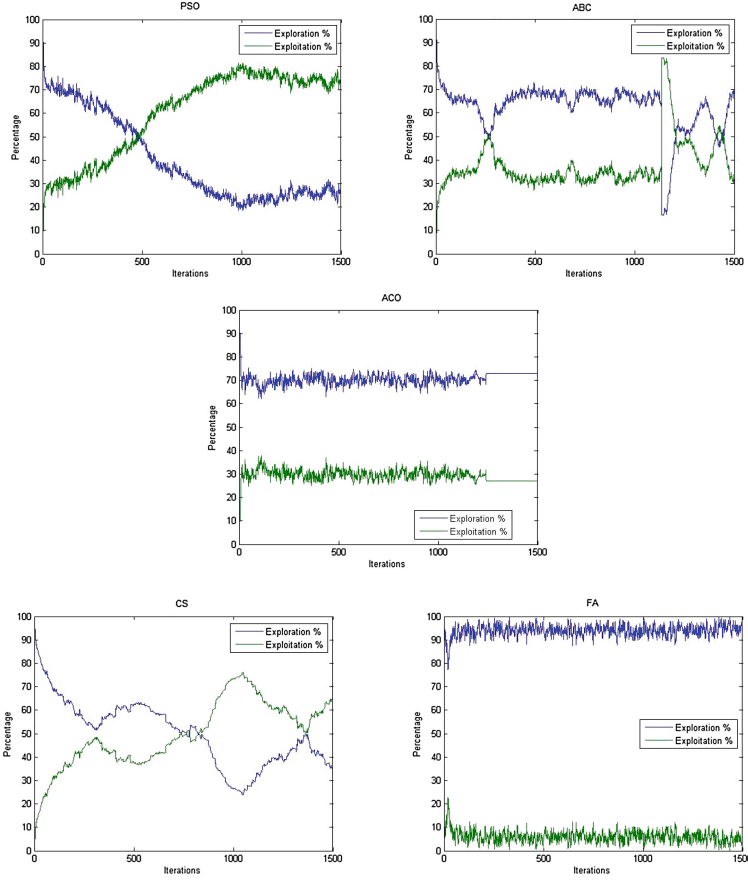


Fig. 2. Exploration and exploitation of swarm-based metaheuristics on Ackley function

exploration and exploitation ratio. In fact, $ACO_{\mathbb{R}}$ was more balanced as compared to FA which was highly explorative and considerably low at exploitation probably because of lévy flight to avoid trapping in local minima. The optimum ratio of exploration and exploitation was found around 70% : 30%, however it strongly depends on type of problem. In both the cases, Ackley and Sphere, PSO showed poor explorative capability, as it prematurely converged and retained within suboptimal locations persistently throughout majority of iterations. In case of ABC and CS, both the algorithms inconsistently maintained the ratio during the course of search. As mentioned earlier, convergence graph does not reveal much about algorithm performance as later part of iterations, no visible difference among the curves can be found in Fig. 4.

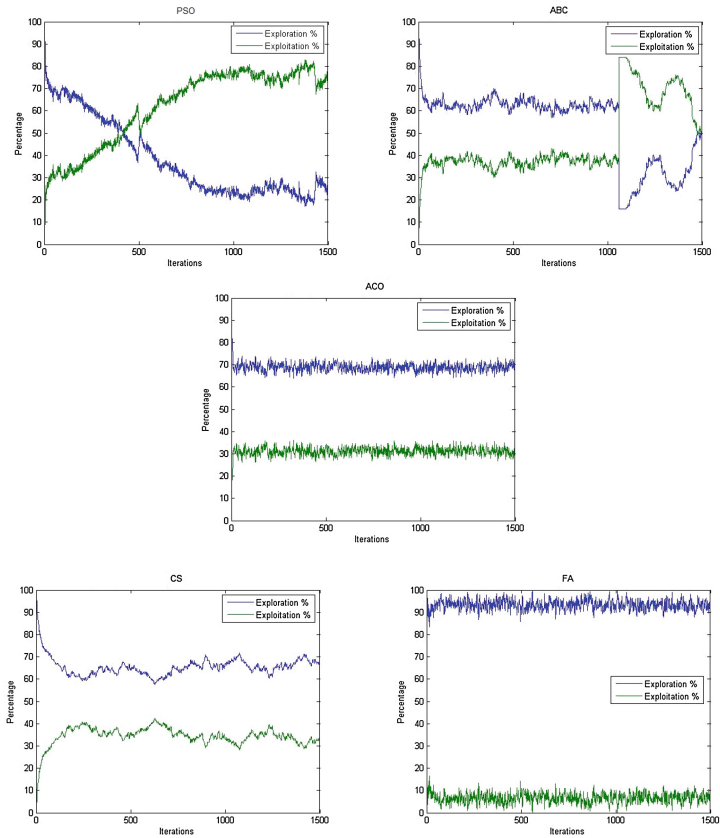


Fig. 3. Exploration and exploitation of swarm-based metaheuristics on Sphere function

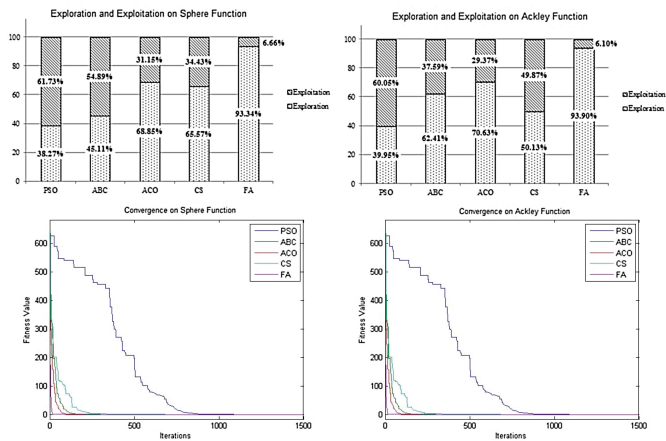


Fig. 4. Exploration:exploitation percentage and convergence on Ackley and Sphere functions

Table 2. Experimental results

	Function	PSO	ABC	ACO _R	CS	FA
Best solution	Ackley	1.66E-05	6.27E-08	6.22E-15	0.007807	3.82E-14
	Sphere	1.08E-09	1.70E-15	1.73E-37	3.77E-08	2.47E-27
Xpl%:Xpt%	Ackley	39.95:60.05	62.41:37.59	70.63:29.37	50.13:49.87	93.90:6.10
	Sphere	38.27:61.73	45.11:54.89	68.85:31.15	65.57:34.43	93.34:6.10
Diversity measurement	Ackley	116.8594	134.2424	156.4539	113.8922	163.5619
	Sphere	113.955	118.5775	155.1894	148.4064	163.8374
NFEs		75,000	112,600	120,050	150,050	1,837,550

6 Conclusion

In this experimental study, we measured the explorative and exploitative capabilities of top five swarm-based metaheuristic algorithms. According to diversity measurement and pictorial evidence provided in this paper, consistency and coherence among swarm individuals is the key factor for the success of any swarm-based metaheuristic algorithm. Other than ACO and FA, PSO, ABC, CS maintained inconsistent ratio of exploration and exploitation throughout iterations. Due to high variance of diversity among swarm individuals, the swarm is unable to improvise the available potential solutions. Convergence graph may provide answers to question about “what happened”, but “how and why it happened” is more related to in-depth analysis. The significantly useful methodology adopted in this work can be further extended in future to perform component-wise performance analysis in order to improve and develop metaheuristics more effectively. Moreover, the exploration and exploitation measurements can be performed on wider range of optimization problems, since difficulty in problem space defines the suitability of exploration and exploitation capabilities.

Acknowledgements. The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia for supporting this research under Postgraduate Incentive Research Grant, Vote No.U560.

References

1. Zheng, Yu-Jun, Chen, Sheng-Yong, Ling, Hai-Feng: Evolutionary optimization for disaster relief operations: a survey. *Appl. Soft Comput.* **27**, 553–566 (2015)
2. Hidalgo, I.G., de Barros, R.S., Fernandes, J., Estrócio, J.P., Correia, P.B.: Metaheuristic approaches for hydropower system scheduling. *J. Appl. Math.* **2015** (2015)
3. Duarte, A., Martí, R., Álvarez, A., Ángel-Bello, F.: Metaheuristics for the linear ordering problem with cumulative costs. *Eur. J. Oper. Res.* **216**(2), 270–277 (2012)

4. Yang, X.-S., Cui, Z., Xiao, R., Gandomi, A.H., Karamanoglu, M.: *Swarm Intelligence and Bio-Inspired Computation: Theory and Applications*. Newnes (2013)
5. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Technical report, Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
6. Kennedy, J., Eberhart, R.: Particle swarm optimization (ps). In: *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948. Perth, Australia (1995)
7. Tereshko, V., Loengarov, A.: Collective decision making in honey-bee foraging dynamics. *Comput. Inf. Syst.* **9**(3), 1 (2005)
8. Dorigo, M., Di Caro, G.: Ant colony optimization: a new meta-heuristic. In: *Evolutionary Computation, CEC 99. Proceedings of the 1999 Congress on*, vol. 2, pp. 1470–1477. IEEE (1999)
9. Yang, X.-S., Deb, S.: Cuckoo search via lévy flights. In: *Nature & Biologically Inspired Computing, 2009. NaBIC World Congress on*, pp. 210–214. IEEE (2009)
10. Yang, X.-S.: *Engineering Optimization. Firefly algorithm*, pp. 221–230 (2010)
11. Tan, Y., Zhu, Y.: Fireworks algorithm for optimization. *Advances in Swarm Intelligence*, pp. 355–364 (2010)
12. Yang, X.-S.: A new metaheuristic bat-inspired algorithm. *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, pp. 65–74 (2010)
13. Rao, R.V., Savsani, V.J., Vakharia, D.P.: Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. *Comput.-Aided Design* **43**(3), 303–315 (2011)
14. Simon, Dan: Biogeography-based optimization. *IEEE Trans. Evol. Comput.* **12**(6), 702–713 (2008)
15. Passino, K.M.: Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Syst.* **22**(3), 52–67 (2002)
16. Sorensen, K., Sevaux, M., Glover, F.: A History of Metaheuristics (2017). [arXiv:1704.00853](https://arxiv.org/abs/1704.00853), arXiv preprint
17. Yang, X.-S.: Nature-inspired metaheuristic algorithms: Success and new challenges (2012). [arXiv:1211.6658](https://arxiv.org/abs/1211.6658), arXiv preprint
18. Cheng, S., Shi, Y., Qin, Q., Zhang, Q., Bai, R.: Population diversity maintenance in brain storm optimization algorithm. *J. Artif. Intell. Soft Comput. Res.* **4**(2), 83–97 (2014)
19. Leguizamón, G., Coello Coello, C.A.: An alternative ACO_R algorithm for continuous optimization problems. In: *ANTS Conference*, pp. 48–59. Springer (2010)
20. Črepinšek, M., Liu, S.-H., Mernik, M.: Exploration and exploitation in evolutionary algorithms: A survey. *ACM Comput. Surv. (CSUR)* **45**(3), 35 (2013)
21. Jr, I.F., Yang, X.-S., Fister, I., Brest, J., Fister, D.: A brief review of nature-inspired algorithms for optimization (2013). [arXiv:1307.4186](https://arxiv.org/abs/1307.4186), arXiv preprint
22. Zhan, Z.-H., Zhang, J., Shi, Y.-H., Liu, H.-L.: A modified brain storm optimization. In: *Evolutionary Computation (CEC), IEEE Congress on*, pp. 1–8. IEEE (2012)
23. Karaboga, D., Akay, B.: A comparative study of artificial bee colony algorithm. *Appl. Math. Comput.* **214**(1), 108–132 (2009)
24. Nawi, N.M., Rehman, M.Z., Khan, A., Chirima, H., Herawan, T.: A modified bat algorithm based on gaussian distribution for solving optimization problem. *J. Comput. Theor. Nanosci.* **13**(1), 706–714 (2016)
25. Zhang, L., Liu, L., Yang, X.-S., Dai, Y.: A novel hybrid firefly algorithm for global optimization. *PloS one* **11**(9), e0163230 (2016)

Classification of JPEG Files by Using Extreme Learning Machine

Rabei Raad Ali^(✉), Kamaruddin Malik Mohamad, Sapiee Jamel,
and Shamsul Kamal Ahmad Khalid

Faculty of Computer Science and Information Technology, Information Security
Interest Group (ISIG), Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja,
Johor, Malaysia

rabei.aljawary@gmail.com, {malik, sapiee, shamsulk}
@uthm.edu.my

Abstract. Recovery of data files when their system information missing is a challenging research issue. The recovery process entails methods that analyze the structure and contents of each individual file clusters. A primary and important process of files' recovery is determining the files' types including JPEG, DOC or HTML. This paper proposes an Extreme Learning Machine (ELM) algorithm to assign a class label of JPEG or Non-JPEG image for files in a continuous series of data clusters. The algorithm automatically classifies the files based on evaluation measures of three methods Entropy, Byte Frequency Distribution and Rate of Change. The ELM algorithm is applied to RABEI-2017 and DFRWS-2006 datasets. The experimental results show that the ELM algorithm is able to identify JPEG files of fragmented clusters with high accuracy rate. The classification accuracy of the RABEI-2017 dataset is 90.15% and the DFRWS-2006 is 93.46%. The DFRWS-2006 has more classes than the RABEI-2017 which improves the ELM classifier fitting.

Keywords: Multimedia clusters · JPEG image · Classification · Extreme learning machine (ELM)

1 Introduction

Multimedia files of digital images and documents are the current trends in retaining important information or memories [1]. The digital devices such as computers and smartphones can deal with the huge number of files and in different file formats [2]. The files are exposed to deformation or damage due to many reasons including device failure, deliberate destruction or human errors. The recovery of the damaged files is a very important issue. Identifying the files' type is an essential step in the files' management and recovery functionalities [3].

Storages of computer devices are divided into digital spaces of blocks and clusters. A cluster is a smallest allocation of a device storage that carries a data file [1]. The clusters of a file have marker contains that include a header and footer data known as a signature. The signature describes the file's system information of type and contents. Applying file carving methods is one way to recover files [4]. However, in many cases,

this signature cluster is damaged or disconnected due to storage damage or system fragmentation process [5]. This issue entails investigating advanced carving methods. Identifying the files' type is an essential step to recover files with missing or damaged file system information [3]. There are three categories of files' type detection methods: extension-based method, magic bytes-based method and content-based methods. Each method has a number of advantages and disadvantages. Therefore, none of the methods are perfect and can provide comprehensive solutions [4]. The content-based method is an active research area to recover files with damage or missing marker cluster. It is used to extract features that identify the file types and contents. For example, McDaniel and Heydari [6], proposed Byte Frequency Analysis (BFA), File Header/Trailer (FHT) and Byte Frequency Cross-correlation (BFC) methods to analyze files' contents. Karresand and Shahmehri [7] improve the Oscar method through measuring the Rate of Change (ROC) of the byte contents. However, statistical or non-statistical methods are needed to analyze the extracted features [8]. Therefore, effective and efficient classification algorithms need to be investigated in order to improve the accuracy of the files identification and ultimately the recovery process.

The main focus of this paper is applying the Extreme Learning Machine (ELM) neural network classifier to detect class label of files in patterns of fragmented clusters. The ELM is emerged as a powerful supervised learning technique for data classification. It is found to be better than other methods in solving similar problems as in Zhang et al. [10] for instance. The ELM algorithm classifies the files to JPEG or Non-JPEG image in a continuous series of fragmented data clusters. The Entropy, BFD and ROC methods are used to extract the features of the file types from their corresponding clusters. The rest of the paper is organized such that Sect. 2 presents the related work. Section 3 describes the used datasets and methods in this work. Section 4 presents the implementation and results. Finally, Sect. 5 addresses the conclusion and the suggested future work.

2 Related Work

The literature offers many classification algorithms that solve the classification problems of binary and multi-classes such as decision trees, neural network, k-nearest neighbor, naive Bayes, and support vector machines [8–11]. However, there are a few attempts to classify high entropy file fragments in the literature [3, 4]. This section presents a review of the related work of solving files recovery problems. Qiu et al. [4], propose a new multimedia files carving method to enhance the recovery accuracy of fragmented files. The carving method includes extracting BFD and ROC features. They use SVM for data files classification. The method is able to recover fragmented files and the JPEG files show higher recovery accuracy.

Veenman [5], propose a classification method based on a statistical learning approach to classify fragmented clusters of different file types. The method collects a number of evidences from the neighboring clusters to form a set of characteristic features. Then apply the statistical learning classifier on the extracted features to recognize files' patterns of relevant file types. The method targets files that exposed to

fragmented cluster in which both the header and footer of a cluster are not available. The method successively identifies file types of fragmented clusters.

Amirani et al. [3], propose a context-based file type identification method. The method includes a Principle Component Analysis (PCA) technique for feature extraction and an unsupervised Multi-Layer Perceptron (MLP) neural network for classification. It applies the extracted features to obtain a fileprint of each file of a type. The fileprint are used to identify unknown files. The application of the method shows good classification accuracy results and fast performance. The method can be further improved to cover the wide range of file types and different file size of types.

Zhang et al. [10], compare between the performance of SVM and ELM with kernel functions when solving images classification problem. The comparison variable is the quality of the images' features. The features are extracted from ImageNet dataset. The comparison results show that the ELM outperforms the SVM about 4% of the average accuracy. However, both the SVM and ELM are used for objects recognition purposes and not for recovering images of fragmented clusters. Hence, this work includes applying the ELM to classify JPEG or Non-JPEG image in a continuous series of fragmented data clusters.

3 Methods

This work includes applying ELM algorithm to assign a class label to multimedia files. In this paper, ELM is used to classify the files to JPEG and Non-JPEG images. The files are extracted from a continuous series of fragmented data clusters. The main steps to complete this work are summarized in Fig. 1. Each step is detailed in the following subsections.

3.1 Datasets Description

The first step of the work is to prepare a dataset that serves in testing and validating the proposed method. These datasets are used to check the ability of the method to identify JPEG clusters from other file types. The datasets are described in the following subsections.

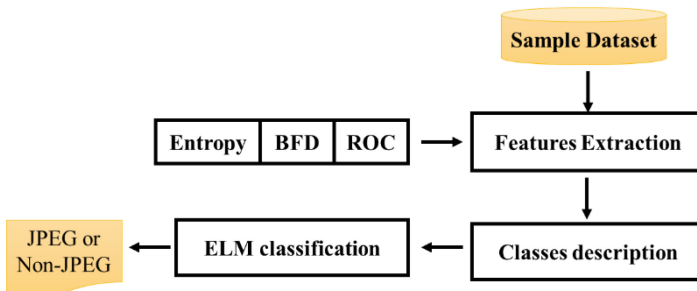


Fig. 1. The flow diagram of the research methods

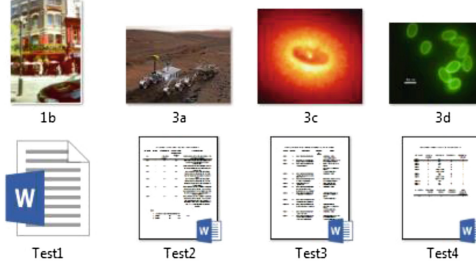


Fig. 2. The data sample of the RABEI-2017 dataset

3.1.1 The RABEI-2017

The RABEI-2017 dataset contains eight files selected from DFRWS-2006 (four JPEG images and four Microsoft Word documents). This dataset is created with the purpose of testing and verification. It facilitates the testing as it contains number of selected cases of data that closely related to the problem scope. The chosen clusters of images are extracted from scan segment where the first cluster exists after the image information cluster. The last cluster of an image is identified after the last incomplete cluster and ignore the incomplete cluster. The Non-JPEG images have been chosen from the DFRWS-2006 dataset. These files occupy 400 clusters and each cluster has a size of 512 bytes. The clusters are divided such that four different JPEG images occupy 200 clusters (each image occupies 50 clusters) and four different Microsoft Word documents occupy 200 clusters (each document occupies 50 clusters). Figure 2 shows a sample data of the RABEI-2017 dataset.

3.1.2 DFRWS-2006

DFRWS-2006 is a 50 MB multimedia files of JPEG, ZIP, HTML, Text, and Microsoft word files [11]. The clusters have the size of 512 bytes, each cluster is represented by 513 features and the final column (514) of each features row represents the class of the cluster. The number of clusters of JPEG images is 23639 and the number of clusters of non-JPEG images is 23639. Figure 3. shows the distribution of testing files. There are



Fig. 3. The data sample of the DFRWS-2006 dataset

13 JPEG image files and 23 of other file types. Figure 3 shows a sample data of the DFRWS-2006 dataset. The 3i.jpg image is excluded to equalize the image and non-image classes.

3.2 Features Extraction

The features extraction step is performed in three processes. Firstly, it is taken into distinguishing account the entropy. The computer byte value in the class is between 0 and 255. The entropy value of the class is considered ($1 \leq \text{entropy value} < 0$). It is used to distinguish JPEG files from other files [12]. Second, the BFD feature is represented as feature vector and used in cluster class classification. The feature vector consists of 256 basic features. The BFD feature neglects the order of the bytes and only considers the bytes' value. Thirdly, the ROC indexes the order of the bytes to track the relevant bytes. The feature represents the absolute value of the difference between two consecutive byte values in a data class [7]. Table 1 shows the feature vector description.

Table 1. Features description

Number	Features	Type	Range
1	Entropy	Float	[0, 1]
$2 \rightarrow 257(256)$	BFD	Float	[0, 1]
$258 \rightarrow 513(256)$	ROC	Float	[0, 1]

3.3 Classes Description

Above step results features that contain relevant information of the dataset. The classification task can be performed by using this reduced features' representation instead of the complete initial data [14, 15]. The classes are indexed by a class label of a column number. Binary classifications distinguish between two classes of a type. The JPEG clusters are tested against file clusters of different file types that have high entropies including Zip, Text and Word documents. However, only two cases are considered to be under investigation, JPEG and the non-JPEG classes. The two datasets are split equally into 50% training and 50% testing data. The feature vectors of the training data are linearly scaled within the range of [0, 1]. The same scaling is applied to the testing data.

3.4 Extreme Learning Machine Classification

Extreme Learning Machine (ELM) classification algorithms goal to assign a class label for each input of a binary or multi-class classification problem [10]. The classification process entails associating a number of data samples to specific labels or class.

The ELM predicts the classes by looking a set of interrelated features or attributes. The ELM is a feedforward neural network with one hidden layer. The weights of the hidden layer are chosen randomly and never updated. The weights of the output layer are learned in one iteration. The ELM architecture is shown in Fig. 4.

Let suppose that a set of stochastic samples, (x_i, t_i) where $x_i = [x_1^i, x_2^i, \dots, x_n^i]^T$ is the i th training sample of a n -dimensional vector quantity, $t_i = [t_1^i, t_2^i, \dots, t_l^i]^T$ is the

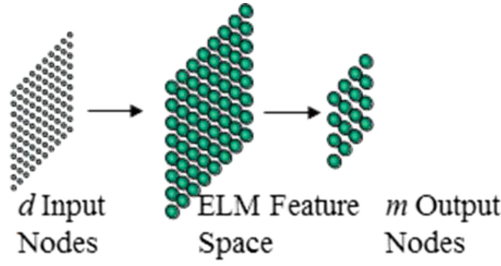


Fig. 4. The architecture of the ELM [13]

target vector. Here, the input-weights, $W_{M \times n}$, the bias of hidden layer is $b_{M \times 1}$, and the output-weights are $\beta_{l \times M}$, where M delegates the number of hidden nodes. The matrix $W_{M \times n}$ and $b_{M \times 1}$ are generated randomly. The target output of the ELM can be calculated by the follow equation [13]:

$$t_k^i = \sum_{j=1}^m \beta_{kjg_j}(w, b, x), \quad k = 1, 2, \dots, l \quad (1)$$

The above equation could be written as follows. Where $g_i(\cdot)$ is the activation function of a hidden layer.

$$T = H\beta \quad (2)$$

where H is defined as the output matrix of a hidden layer and determined as:

$$H(W, B, X) = \begin{bmatrix} g(w_1, x_1 = b_1) \dots g(w_M, x_1 + b_M) & & \\ & \dots & \\ g(w_1, x_l = b_1) \dots g(w_M, x_N + b_N) \end{bmatrix} \quad (3)$$

Where $\beta = [\beta_1, \beta_2, \dots, \beta_M]_{l \times M}^T$ is matrix output weight can be determined systematically through the least norm minimum square solution:

$$\tilde{\beta} = \arg \min_{\beta} \|H\beta - T\| = H^+ T \quad (4)$$

where H is generalized inverse of H^+ . If H^+ is the Moore-Penrose and H nonsingular, the above Eq. (4) could be written as follows:

$$\tilde{\beta} = (H^T H)^{-1} H^T T \quad (5)$$

The theory and mathematics of the ELM algorithm is detailed in [13]. The ELM is found to be an efficient classifier for its fast training and high generalization.

Table 2. The ELM classifier details

Dataset	RABEI-2017	DFRWS-2006
Size	400×513	47278×513
Number of hidden neurons	691	600
Function	Sigmoid	Sigmoid
Learing approach	10-Fold cross validation	10-Fold cross validation

4 Implementation and Results

The ELM algorithm is applied to RABEI-2017 dataset and DFRWS-2006 forensic challenge dataset. The overall method is implemented using MATLAB 2016a. The dataset is split into training and testing data. The two-thirds of the data is used for training and the remaining is used for testing. Table 2 shows the ELM classification setting when it is applied to the RABEI-2017 and DFRWS-2006 datasets.

The confusion matrix results are summarized in Table 3.

Table 3. The ELM evaluation measures

RABEI-2017		DFRWS-2006	
JPEG	Non-JPEG	JPEG	Non-JPEG
0.8788	0.1212	0.8788	0.1212
0.0758	0.9242	0.0758	0.9242

Table 4 shows the classification accuracy, the average of classification accuracy and the evaluation of the performance.

Table 4. The ELM classification measures

RABEI-2017								
Measures	Accuracy		Precision		Recall		F measure	
Classes	0.9015	0.9015	0.9206	0.8841	0.9242	0.8788	0.9037	0.8992
Overall	0.9015		0.9015		0.9015		0.9015	
Measures	Accuracy		Precision		Recall		F measure	
Classes	0.9346	0.9346	0.9473	0.9225	0.9203	0.9489	0.9336	0.9355
Overall	0.9346		0.9346		0.9346		0.9346	

The experimental results show that the ELM algorithm is able to identify files of fragmented clusters with high accuracy rate. The classification analysis of the results for the two datasets is shown in Table 5. The ELM classification accuracy is 90.15% in the RABEI-2017 dataset and 93.46% in the DFRWS-2006 dataset.

Tables 5 show the average accuracies in actual and predicted for JPEG and Non-JPEG classifiers for the two datasets clusters with a less error classifiers. It further shows that the ELM classification provides feasible measures in file type detection accuracy. The deference between the two datasets accuracy results from the difference in the data complexity in which the DFRWS-2006 dataset is more complex than the RABEI-2017 dataset. The extracted features and the ELM can be farther improved to improve the accuracy and reduce the classification error.

Table 5. The classification analysis of the results

Dataset	RABEI-2017		DFRWS-2006	
The no. of file types	The no. of classified JPEG images	The no. of classified Non-JPEG images	The no. of classified JPEG images	The no. of classified Non-JPEG images
Actual	66	66	7879	7879
Predicted	58	61	7251	7476
Error classification	8	5	403	628
Total file cluster	132		15758	
Accuracy (%)	90.15		93.46	

5 Conclusion and Future Work

In this paper, a content-based file type classification method is proposed based on Extreme Learning Machine (ELM) algorithm. The ELM classifies JPEG and Non-JPEG files of fragmented clusters that their system information is missing. Entropy, Byte Frequency Distribution (BFD) and Rate of Change (ROC) features extraction methods are applied to extract the contents of the file clusters. With the aid of the ELM characteristics, the method manifests the capabilities of high accuracy of 93.46% that promotes it to be adopted in real scenarios. Furthermore, the method works in situations of missing headers or fragmented clusters positions; hence, it is invulnerable to scrambled or fragmented files. This work will be used in the carving process of fragmented files recovery.

Acknowledgements. This work was supported by the Universiti Tun Hussein Onn Malaysia, Ministry of Higher Education Malaysia under Grant Vote No. U495.

References

1. Abdullah, N., Ibrahim, R., Mohamad, K.: Cluster size determination using JPEG files. In: 2012 Computational Science and Its Applications–ICCSA, pp. 353–363 (2012)
2. Mohammed, M.A., Gani, M.K. A., Hamed, R.I., Mostafa, S.A., Ahmad, M.S., Ibrahim, D. A.: Solving vehicle routing problem by using improved genetic algorithm for optimal solution. *J. Comput. Sci.* (2017)
3. Amirani, M.C., Toorani, M., Mihandoost, S.: Feature-based type identification of file fragments. *Secur. Commun. Netw.* **6**(1), 115–128 (2013)
4. Qiu, W., Zhu, R., Guo, J., Tang, X., Liu, B., Huang, Z.: A new approach to multimedia files carving. In: 2014 IEEE International Conference on Bioinformatics and Bioengineering (BIBE), pp. 105–110. IEEE, Nov 2014
5. Veenman, C.J.: Statistical disk cluster classification for file carving. In: 2007 Third International Symposium on Information Assurance and Security, IAS 2007, pp. 393–398. IEEE, Aug 2007
6. McDaniel, M., Heydari, M.H.: Content based file type detection algorithms. In: 2003 Proceedings of the 36th Annual Hawaii International Conference on System Sciences, pp. 10–pp. IEEE, Jan 2003
7. Karresand, M., Shahmehri, N.: Oscar-file type identification of binary data in disk clusters and ram pages. *Secur. Priv. Dyn. Environ.* 413–424 (2006)
8. Li, Q., Ong, A., Suganthan, P., Thing, V.: A novel support vector machine approach to high entropy data fragment classification. In: Proceedings of the South African Information Security Multi-Conf (SAISMC), pp. 236–247 (2011)
9. Mehra, N., Gupta, S.: Survey on multiclass classification methods. *Int. J. Comput. Sci. Inf. Technol.* **4**(4), 572–576 (2013)
10. Zhang, L., Zhang, D., Tian, F.: SVM and ELM: Who Wins? Object recognition with deep convolutional features from ImageNet. In Proceedings of ELM-2015, vol. 1, pp. 249–263. Springer International Publishing (2016)
11. Data dump DFRWS2006. <http://old.dfrws.org/2006/challenge/dfrws-2006-challenge-files.zip>. Accessed 12 Mar 2017

12. Shannon, M.: Forensic relative strength scoring: ASCII and entropy scoring. *Int. J. Digit. Evid.* **2**(4), 1–19 (2004)
13. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. *Neurocomputing* **70**(1), 489–501 (2006)
14. Mohammed, M.A., Ghani, M. K.A., Hamed, R.I., Mostafa, S.A., Ibrahim, D.A., Jameel, H. K., Alallah, A.H.: Solving vehicle routing problem by using improved K-nearest neighbor algorithm for best solution. *J. Comput. Sci.* (2017).
15. Khaleefah, S.H., Nasrudin, M.F., Mostafa, S.A.: Fingerprinting of deformed paper images acquired by scanners. In: 2015 IEEE Student Conference on Research and Development (SCoReD), pp. 393–397. IEEE, Dec 2015

Evaluating the Performance of Three Classification Methods in Diagnosis of Parkinson's Disease

Salama A. Mostafa¹(✉), Aida Mustapha¹, Shihab Hamad Khaleefah²,
Mohd Sharifuddin Ahmad³, and Mazin Abed Mohammed⁴

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
{salama, aidam}@uthm.edu.my

² Faculty of Computer Science, Almaaref University College, Anbar, Iraq
shi90hab@gmail.com

³ College of Computer Science and Information Technology, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia
sharif@uniten.edu.my

⁴ Planning and Follow-up Department, University of Anbar, Anbar, Iraq
mazin_top_86@yahoo.com

Abstract. Accurate diagnosis of the Parkinson's disease is a challenging task that involves many physical, psychological and neurological examinations. The examinations include investigating a number of signs and symptoms, reviewing the medical history and checking the nervous system conditions of a patient. Recently, researchers use voice disorders to diagnose Parkinson's disease patients. They extract features of a recorded human voice and apply classification methods to diagnosis this disease. In this paper, we apply a Decision Tree, Naïve Bayes and Neural Network classification methods for the diagnosis of Parkinson's disease. The aim of this paper is to resolve the problem by evaluating the performance of the three methods. The objectives of the paper are to (i) implement three classification methods independently on a Parkinson's dataset, and (ii) determine the best method among the three. The classification results show that the Decision Tree produces the highest accuracy rate of 91.63%, followed by the Neural Network, 91.01% and the Naïve Bayes produces the lowest accuracy rate of 89.46%. The results recommend using the Decision Tree or the Neural Network over the Naïve Bayes for datasets with similar properties.

Keywords: Classification · Decision tree · Naïve bayes · Neural network

1 Introduction

Parkinson's disease is a degenerative disorder of the central nervous system. It is a common disease among people who are above 50 years old [1, 2]. The main cause of the disease is the death of dopamine-generating cells in the parts of the midbrain known as *substantia nigra* [3]. Its physical symptoms are movement difficulties including

rigidity, shaking, and slowness. The psychological symptoms include depression, dementia, and emotional problems. The neurological symptoms are beyond human direct observations and require advanced diagnostic technologies to the humans' nervous system and brain [4, 5]. Figure 1 shows a healthy and Parkinson's brain.

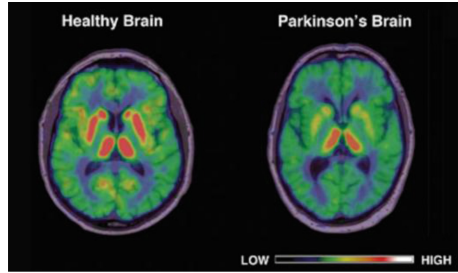


Fig. 1. Parkinson's disease symptoms in a human brain [3]

Little et al. [5] propose Dysphonia measures algorithms and speech analysis methods to study voice disorders effect on the Parkinson's disease patients. The main aim of the data is to discriminate healthy people from those with Parkinson's disease. They propose two features of recurrence and fractal scaling to differentiate normal and disordered voices. The research outcomes offer a wide range of voice disorder classes of Parkinson's dataset. Arjmandi and Pooyan [6] follow Little et al. [5] steps and propose an improved algorithm in pathological voice quality assessment. They identify different voice disorders of the vocal folds. They recommend using linear discriminant analysis over other detection approach of voice disorders. Recently, different classification methods such as a Decision Tree, Neural Network, Naïve Bayes, Support Vector Machine, and Regression are applied in diagnosing Parkinson's disease and many of which use the Parkinson's dataset of [5]. The methods investigate the patterns of the provided data and predict the corresponding class of the data [7].

In this paper, we evaluate the performance of three classification methods, which are a Decision Tree, Naïve Bayes, and Neural Network for the diagnosis of Parkinson's disease. The following section reviews the related work. Section 3 presents the research methodology and the activities that are implemented to complete this work. Section 4 outlines the experiments and the corresponding results and Sect. 5 analyzes and discusses the results. Section 6 presents the conclusion and proposes future work.

2 Related Work

A number of works attempt to increase the accuracy of the Parkinson's disease diagnosis. These include applying a number of classification and statistic methods. Tatu et al. [8], for example, propose an automated analysis and visualization technique in which, they apply class density measure, similarity measure, histogram density measure, overlap measure, rotating variance measure and Hough Space measure on the

Parkinson's dataset. The aim of their work is to find the linear and nonlinear correlations and clusters between the pairwise of the Parkinson's dataset.

Das [9] use the Parkinson's dataset to compare between a number of classification methods. The classification methods are a Decision Tree, DMneural, Neural Network, and Regression. The Neural Network yields the best classification result with a score of 92.90%. Table 1 shows the classification scores during the training and testing phases for each of the classification methods.

Table 1. The classification accuracy rates of [9]

Method	Training (%)	Testing (%)
Decision tree	93.60	84.30
DMneural	89.60	84.30
Neural network	100.00	92.90
Regression	89.00	88.60

Exarchos et al. [10] integrate a Partial Decision Tree and Association Rule Mining algorithms within a PERFORM system. The PERFORM system has a set of wearable sensors that extract Parkinson's disease symptoms during daily patient activities. The system online classifies users to Parkinson's or healthy based on analyzing initial information provided by the users and the extracted symptoms. The diagnosis accuracy depends on the provided information and it is found to be between 57.10 and 77.40%.

Can [11] applies Parallel Distributed Neural Network with backpropagation algorithm for Parkinson's disease diagnosis. The research aims to improve the classification or prediction accuracy rates via a majority voting scheme technique in Neural Network design. They consistently achieve 90.00% prediction accuracy rate.

3 Methods

This work evaluates the performance of a Decision Tree, Naïve Bayes, and Neural Network in the diagnosis of Parkinson's disease using the Parkinson's dataset. The dataset has been previously used in testing classification methods in which a relevant classifier identifies healthy and Parkinson's persons. The main activities that we conduct in this work are data preparation, methods setting, classification, and evaluation [12–14]. The activities are detailed in the following subsections and the sequence of the activities is shown in Fig. 2.

3.1 Parkinson's Dataset

The Parkinson's dataset is a real dataset of a human voice records. It contains voice disorder prediction features that are used for Parkinson's disease diagnosis. The data is collected using the Intel At-Home Testing Device (AHTD) at the home of volunteered patients [4]. The recorded data of the patients contains a number of flawed recordings, e.g., patients coughing but have been removed from the dataset. Sustained vowels

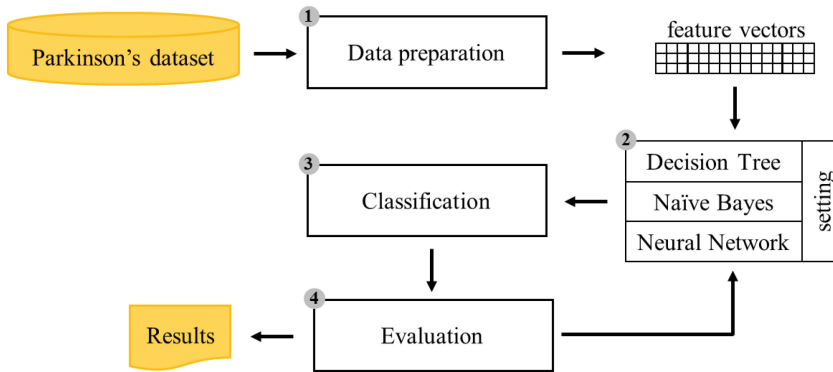


Fig. 2. The flow diagram of the research activities

method is used to simplify the signal analysis and reduce the confounding effects of the speech [6]. Then, the data is transmitted online via the Internet to a clinic to find the Unified Parkinson’s Disease Rating Scale (UPDRS) of the patients. The dataset composes of a range of biomedical voice measurements for 31 people, 23 of which are diagnosed with the Parkinson’s disease [5]. We obtain a copy of the Parkinson’s dataset from the University of California-Irvine (UCI) machine learning repository [4]. Table 2 shows the description of the Parkinson’s dataset.

Table 2. Parkinson’s dataset [4]

Data characteristic	Multivariate	Number of instances	197
Features characteristic	Real	Number of features	23
Purpose	Classification	Number of missing values	0

3.2 Data Preparation

From our observation of the data and the literature review, the Parkinson’s dataset does not have observable patterns. The features are numerical that contains only real numbers. The dataset is complete and has no missing values. Also, there is no noisy or non-numeric data found. We examine the dataset and perform a number of

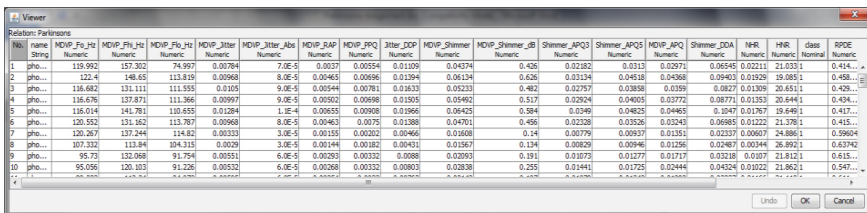


Fig. 3. The Parkinson's data examination [15]

preprocessing steps in order to check its quality and make it ready for running. The data pre-preparation includes data examination, cleaning, discretization, reformatting, and transformation. Figure 3 shows the examination of Parkinson's dataset.

3.3 Classification Methods

We use three well-known classification methods in Parkinson's disease classification which respectively are a Decision Tree, Naïve Bayes, and Neural Network. This section presents a briefing for each of the classification methods.

Decision Tree. A Decision Tree is a decision-making method that has a tree structure. It consists of four components which are a root, leaf nodes, branches and internal nodes [9]. The root connects the classes of a tree in which the leaf nodes represent the classes, the branches represent the outcomes and the internal leaves represent the processes. The classification rules are the paths from the root to the leaves [7]. The Decision Tree is represented by many algorithms and one of which is the Iterative Dichotomiser 3 (ID3) algorithm. The ID3 constructs a Decision Tree on a top-down manner [16]. It is used in the domains of classification, natural language processing, and machine learning.

Naïve Bayes. Naïve Bayes is a machine learning classifier that utilizes supervised learning or statistical approach. It is based on the Bayes probabilistic theorem and uses a conditional probability to determine the outcomes [17]:

$$posterior = \frac{prior * likelihood}{evidence} \quad (1)$$

The Naïve Bayes uses the traditional classification approach setting in which a problem instance is represented by vectors of feature values known as feature vectors. The feature vectors are classified by the method to certain classes.

$$probability(Class_i | feature_1, feature_2, \dots) \quad (2)$$

The features have independent relationships (naive) in which the evaluation a feature does not affect the value of other features. This assumption reduces the accuracy of the classification; however, it also reduces the required training samples to estimate successful classification and reduces the effects of the noise in the data too. Nonetheless, the Naïve Bayes classifiers have been found to work exceptionally well when the dataset contains plenty of input features but a small number of records.

Neural Network. An Artificial Neural Network is one of the numerical learning methods that simulates human biological neural networks [11]. It consists of many nonlinear computational elements that form the network nodes or artificial neurons. The neurons are linked by weighted interconnections. One of the well-known neurons is the sigmoid that has the following form:

$$\sigma(z) = \frac{1}{1 + \exp\left(-\sum_j weight_i * input_i - bias\right)} \quad (3)$$

Neural Networks are used in many research fields for classification, clustering, approximation, filtering, compression and blind source separation [9]. They are particularly useful in dealing with high complexity data that is difficult or impractical to be solved by traditional methods.

4 Experiments and Results

In this section, we detail the experimental setup and the results of the three classification methods. We apply the 10-fold cross-validation approach to evaluate the performance of the classifiers under study [18].

4.1 Decision Tree

In the Decision Tree experiment, we use an ID3 function to generate The decision tree from the dataset. We conduct 10 tests with different data allocations of cross-validation folds and split percentages [11, 19]. The test results show that the highest accuracy is found in test number seven where the data is divided into 30% training and 70% testing. The highest accuracy score is 91.63% when the Root Mean Square Error (RMSE) is the lowest, 0.2701. Subsequently, the lowest accuracy is found in test number one where the data is divided into 90% training and 10% testing. The lowest accuracy score is 80.00% when the RMSE is the highest, 0.4083. In general, the classification results of the Decision Tree are tending to be very high in which the average accuracy score is 86.42%. Table 3 shows the test results of the Decision Tree.

Table 3. The classification accuracy results of the decision tree

Test	Data allocation	Accuracy (%)	RMSE
1	90.10	80.0000	0.4083
2	80.20	82.3846	0.3789
3	70.30	86.7740	0.3549
4	60.40	86.4092	0.3448
5	50.50	88.0884	0.3235
6	40.60	91.0113	0.2920
7	30.70	91.6376	0.2701
8	20.80	90.8396	0.2948
9	10.90	81.6892	0.3435
10	66.34	85.4079	0.3560

4.2 Naïve Bayes

In the Naïve Bayes experiment, we use activation function to approximate the Naïve Bayes outputs. We conduct 10 tests with different data allocations of cross-validation folds and split percentages. The test results show that the highest accuracy is found in test number seven where the data is divided into 30% training and 70% testing. The

highest accuracy score is 89.46% when the RMSE is the lowest, 0.2668. Subsequently, the lowest accuracy is found in test number two where the data is divided into 80% training and 20% testing. The lowest accuracy score is 79.82% when the RMSE is the highest, 0.3668. In general, the classification results of the Naïve Bayes are tending to be slightly lower than the other methods in which the average accuracy score is 86.18%. Table 4 shows the test results of the Naïve Bayes.

Table 4. The classification accuracy results of the Naïve Bayes

Test	Data allocation	Accuracy (%)	RMSE
1	90.10	80.0000	0.3646
2	80.20	79.8205	0.3668
3	70.30	86.7740	0.3105
4	60.40	87.6751	0.3092
5	50.50	86.0476	0.3229
6	40.60	88.4689	0.2873
7	30.70	89.4637	0.2668
8	20.80	87.6751	0.3087
9	10.90	89.0339	0.2945
10	66.34	86.9005	0.3211

4.3 Neural Network

In the Neural Network experiment, we use multi-layer perceptron that has two hidden layers with a sigmoid function. We conduct 10 tests with different data allocations of cross-validation folds and split percentages. The Neural Network test results show that the highest accuracy is found in test number six where the data is divided into 40% training and 60% testing. The highest accuracy score is 91.01% when the RMSE is the lowest, 0.2871. Subsequently, the lowest accuracy is found in test number one where the data is divided into 90% training and 10% testing. The lowest accuracy score is

Table 5. The classification accuracy results of the neural network

Test	Data allocation	Accuracy (%)	RMSE
1	90.10	80.0000	0.4061
2	80.20	84.9487	0.3690
3	70.30	86.7740	0.3509
4	60.40	86.4092	0.3416
5	50.50	90.1292	0.3019
6	40.60	91.0113	0.2871
7	30.70	89.4637	0.2926
8	20.80	88.3080	0.3092
9	10.90	86.7740	0.3252
10	66.34	85.4079	0.3646

80.00% when the RMSE is the highest, 0.4061. In general, the classification results of the Neural Network are tending to be very high in which the average accuracy score is the highest among the three methods, 86.92%. Table 5 shows the test results of the Neural Network.

5 Discussion

The classification accuracy results of the 10-fold cross-validation are confined within 79.82–91.63%. The Decision Tree scores the highest accuracy percentage followed by the Neural Network and the Naïve Bayes. Table 6 shows the overall accuracy percentages for the classification methods based on the 10-fold cross-validation.

Table 6. The analysis of the results

Test	1. Decision tree accuracy (%)	2. Naïve bayes accuracy (%)	3. Neural network accuracy (%)	Highest accuracy (%)
1	80.0000	80.0000	80.0000	1 + 2 + 3
2	82.3846	79.8205	84.9487	3
3	86.7740	86.7740	86.7740	1 + 2 + 3
4	86.4092	87.6751	86.4092	2
5	88.0884	86.0476	90.1292	3
6	91.0113	88.4689	91.0113	1 + 3
7	91.6376	89.4637	89.4637	1
8	90.8396	87.6751	88.3080	1
9	81.6892	89.0339	86.7740	2
10	85.4079	86.9005	85.4079	2

Figure 4 shows the variations between the classification accuracy results of the Decision Tree, Naïve Bayes and Neural Network for diagnosis of Parkinson’s disease. It also shows the variations in their related accuracy linear measurements. However, we

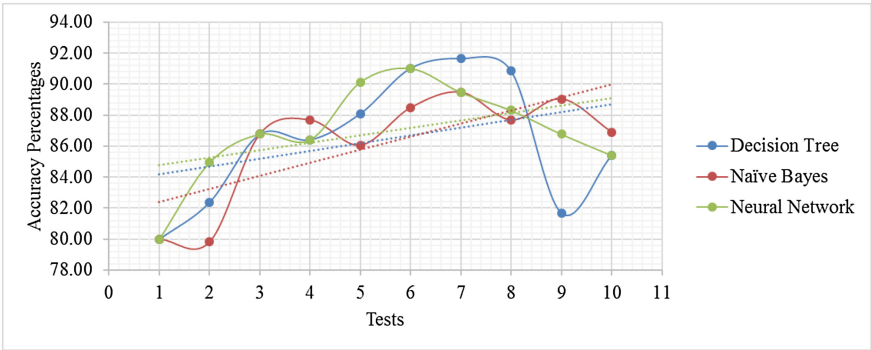


Fig. 4. The variation between the classification accuracy results

observe that the classification processing time of the Neural Network is dramatically longer than the other methods and the shortest time is scored by the Naïve Bayes.

6 Conclusion

The Parkinson's dataset has been used by many researchers in medical and classification researches. The dataset includes human voice disorder detection features. These features reflect voice disorders effect on the diagnosis of Parkinson's disease patients. In this paper, we use the Parkinson's dataset to experimentally evaluate the performance of three classification methods. The methods are a Decision Tree, Naïve Bayes, and Neural Network respectively. We implement 10 tests for each of the three methods and obtain the classification accuracy results. The classification results show that the Decision Tree produces higher accuracy rate of 91.63%, followed by the Neural Network with an accuracy rate of 91.01% and the Naïve Bases comes last with an accuracy rate of 89.46%.

In our future work, we will evaluate the features of the Parkinson's dataset and apply the three classifiers on the best-chosen features with the aim of improving the Parkinson's disease diagnosis.

Acknowledgements. This project is sponsored by Universiti Tun Hussein Onn Malaysia, ORICC, under Vot D004.

References

1. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981)
2. Davie, C.A.: A review of Parkinson's disease. *Br. Med. Bull.* **86**(1), 109–127 (2008)
3. Sutherland, M., Dean, P.: What is Parkinson's Disease? Neuro Challenges, Foundation for Parkinson's. <http://www.parkinsonsneurochallenge.org> (2017). Accessed 08 June 2017
4. Asuncion, A., Newman, D.: UCI machine learning repository. <https://archive.ics.uci.edu/ml/datasets/parkinsons> (2007)
5. Little, M.A., McSharry, P.E., Roberts, S.J., Costello, D.A., Moroz, I.M.: Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomed. Eng. Online* **6**(1), 23 (2007)
6. Arjmandi, M.K., Pooyan, M.: An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomed. Sign. Process. Control* **7**(1), 3–19 (2012)
7. Gnanapriya, S., Suganya, R., Devi, G.S., Kumar, M.S.: Data mining concepts and techniques. *Data Min. Knowl. Eng.* **2**(9), 256–263 (2010)
8. Tatu, A., Albuquerque, G., Eisemann, M., Schneidewind, J., Theisel, H., Magnor, M., Keim, D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In: 2009 IEEE Symposium on Visual Analytics Science and Technology VAST 2009, pp. 59–66. IEEE (2009)
9. Das, R.: A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Syst. Appl.* **37**(2), 1568–1572 (2010)

10. Exarchos, T.P., Tzallas, A.T., Baga, D., Chaloglou, D., Fotiadis, D.I., Tsouli, S., Konitsiotis, S.: Using partial decision trees to predict Parkinson's symptoms: a new approach for diagnosis and therapy in patients suffering from Parkinson's disease. *Comput. Biol. Med.* **42** (2), 195–204 (2012)
11. Can, M.: Neural networks to diagnose the Parkinson's disease. *SouthEast Eur. J. Soft Comput.* **2**(1) (2013)
12. Mohammed, M.A., Ghani, M.K.A., Hamed, R.I., Mostafa, S.A., Ibrahim, D.A., Jameel, H. K., Alallah, A.H.: Solving vehicle routing problem by using improved K-nearest neighbor algorithm for best solution. *J. Comput. Sci.* (2017)
13. Khaleefah, S.H., Nasrudin, M.F., Mostafa, S.A.: Fingerprinting of deformed paper images acquired by scanners. In: 2015 IEEE Student Conference on Research and Development (SCoReD), pp. 393–397. IEEE, Dec 2015
14. Mohammed, M.A., Gani, M.K.A., Hamed, R.I., Mostafa, S.A., Ahmad, M.S., Ibrahim, D.A.: Solving vehicle routing problem by using improved genetic algorithm for optimal solution. *J. Comput. Sci.* (2017)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009)
16. Kumar, S.A., Vijayalakshmi, M.: Efficiency of decision trees in predicting student's academic performance. In: The First International Conference on Computer Science, Engineering and Applications, CS and IT, vol. 2, pp. 335–343
17. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the International conference on Machine Learning ICML, Vol. 3, pp. 616–623 (2003)
18. Bahramirad, S., Mustapha, A., Eshraghi, M.: Classification of liver disease diagnosis: a comparative study. In: 2013 Second International Conference on Informatics and Applications (ICIA), pp. 42–46. IEEE, Sept 2013
19. Hossain, J., FazlidaMohdSani, N., Mustapha, A., SurianiAffendey, L.: Using feature selection as accuracy benchmarking in clinical data mining. *J. Comput. Sci.* **9**(7), 883 (2013)

Some New Results on the Stability of Fractional Integro-Differential Equations Under Uncertainty

A. Ahmadian¹, S. Salahshour^{2(✉)}, N. Senu¹, and F. Ismail¹

¹ Laboratory of Computational Sciences and Mathematical Physics, Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ahmadian.hosseini@gmail.com, norazak@upm.edu.my, fudziah@upm.edu.my

² Young Researchers and Elite Club, Mobarakeh Branch, Islamic Azad University, Mobarakeh, Iran
soheilsalahshour@yahoo.com

Abstract. In this research, we introduce the concept of E_α -type stability for fractional integro-differential equations with uncertainty. We propose different types of fuzzy E_α stabilities for some classes of fuzzy integro-differential equations of fractional order. Besides, we present some new findings on the existence and uniqueness of the solutions of fuzzy integro-differential equations of fractional order using the proposed new concept.

Keywords: Fractional integro-differential equations · Ulam stability · Existence and uniqueness of solution · Fuzzy settings theory · Uncertainty

1 Introduction

The investigation of stability of functional equations has been developed at a high rate in the most recent three decades. It was initially raised by S. M. Ulam in 1940 on a discussion given at Wisconsin University. The issue postured by Ulam was the accompanying: “Under what conditions does there exist an additive mapping near an approximately additive mapping?” [1]. In 1941, this issue was illuminated by Hyers [2] on account of Banach spaces. From that point, this sort of stability is known as the Hyers–Ulam stability. In 1978, Rassias [3] gave a striking speculation of the Hyers–Ulam stability of mappings by considering variables. The stability properties of a wide range of conditions have pulled in the consideration of several researchers. In specific, the Hyers–Ulam dependability and Hyers–Ulam–Rassias dependability have been taken up by various

mathematicians and the investigation of this field has become one of the focal subjects in the mathematical analysis field. For more points of interest on the latest progresses on the Hyers–Ulam stability and Hyers–Ulam–Rassias stability of differential equations, one can follow [4–12]. Moreover, the advancement of stability of non-integer differential equations is somewhat moderate. Li and Zhang [13] make a succinct review on the stability consequences of the fractional differential equations. Li et al. [14, 15] urged on the Mittag-Leffler stability and presented the fractional Lyapunov’s second technique. Be that as it may, there are just few chips away at Ulam type stability for non-integer order differential equations [16–19].

As of late, a few investigations of uncertain fractional differential and integral equations have been investigated in some research reports. Agarwal et al. [20] introduced the idea of solutions for non-integer differential equations with uncertainty. They considered the Riemann–Liouville differentiability notion based on the Hukuhara differentiability to solve uncertain differential equations of non-integer order. Allahviranloo et al. in [21–23] have concentrated on the ideas of the generalized Hukuhara Riemann–Liouville and Caputo differentiability of fuzzy-valued mappings. Mazandarani et al. [24] presented the approximate the solution for initial value problem under Caputo-type fuzzy fractional derivatives using a type of Euler’s technique. Additionally, they explored some outcomes on the existence and uniqueness of solution to differential equation under Caputo type-2 fuzzy derivative and the application of Laplace transform of type-2 fuzzy functions in [25]. Salahshour et al. [23] suggested some new findings toward existence and uniqueness of solution of differential equation with respect to fuzzy fractional calculus. The solutions of the later type of differential equations are examined by utilizing the fuzzy Laplace transforms in [26]. Regarding to the notion of Caputo-type derivative with respect to the generalized fuzzy differentiability, some investigations were implemented on the numerical simulations of fuzzy fractional differential equation [27, 28]. Besides, Malinowski proposed some new outcomes toward the solutions of random integral equations with fuzzy fractional derivative [29]. Ngo [30, 31] considered a few discourses on the existence and uniqueness solutions for functional integral and differential equations with the fuzzy fractional operators.

In recent years, there are very rare works were done to establish Ulam stability in the fuzzy normed spaces [32, 33]. Very recently, Shen [34] investigated the Ulam stability of some kinds of first order linear uncertain differential equations. Motivated by [17, 19], we will propose fuzzy E_α -Ulam type stability for the integro-differential equations based on the fuzzy fractional Caputo derivative. We present four new sorts of fuzzy E_α -Ulam stabilities: fuzzy E_α -Ulam–Hyers stability, fuzzy generalized E_α -Ulam–Hyers stability, fuzzy E_α -Ulam–Hyers–Rassias stability and fuzzy generalized E_α -Ulam–Hyers–Rassias stability. Next, we demonstrate fuzzy E_α -Ulam–Hyers–Rassias stability result for Eq. (3) on a compact interval $I = [0, b)$ via an integral inequality of singular Gronwall type. To the best of our knowledge, there is no report on the fuzzy Ulam stability of fuzzy fractional integro-differential equations, which perhaps give another path

to the researchers to examine the stability of differential equations in terms of fuzzy fractional notion from new viewpoints.

Whatever is left of this paper is composed as: Sect. 2 recalls some conceptions and definitions which are required for the mathematical background of our study in fuzzy setting theory and fractional calculus. In Sect. 3, the fuzzy Ulam stability problem is provided for fuzzy fractional integro-differential equations and different types of this stability are presented. The paper is closed with a few conclusion remarks.

2 Preliminaries

We review some fundamental definitions and concepts which are utilized as a part of the entire of the paper.

We signify the set of all real numbers by \mathbb{R} . Also, the set of all fuzzy numbers on \mathbb{R} is shown by \mathcal{P} .

Definition 1 (see, [35]) We describe a metric d on \mathcal{P} ($d : \mathcal{P} \times \mathcal{P} \longrightarrow \mathbb{R}_+ \cup 0$) by a distance which is so-called Hausdorff distance as follows:

$$d(g, h) = \sup_{r \in [0, 1]} \max\{|g_-^r - h_-^r|, |g_+^r - h_+^r|\}. \quad (1)$$

Definition 2 (see, [35]) Suppose $g, h \in \mathcal{P}$. If there exists $j \in \mathcal{P}$ such that $g = h \oplus j$, then j is called the Hukuhara-difference (H-difference) of g and h , and it is expressed by $g \ominus h$.

In the current research, the symbol “ \ominus ” arises for H-difference and consider that $g \ominus h \neq g + (-h)$. Additionally all through of paper is accepted the H-difference and generalized H-differentiability are existed. Let us to review the meaning of the strongly generalized H-differentiability suggested in [36].

Definition 3 (see, [36]) Let $z : (u, v) \rightarrow \mathcal{P}$ and $u_0 \in (u, v)$. z is called strongly generalized H-differentiable at u_0 , if there exists an element $z'(u_0) \in \mathcal{P}$, such that

(i) for all $\zeta > 0$ sufficiently small, $\exists z(u_0 + \zeta) \ominus z(u_0)$, $\exists z(u_0) \ominus z(u_0 - \zeta)$ and limits (in the metric space \mathfrak{D}):

$$\begin{aligned} \lim_{\zeta \searrow 0} \frac{z(u_0 + \zeta) \ominus z(u_0)}{\zeta} &= \lim_{\zeta \searrow 0} \frac{z(u_0) \ominus z(u_0 - \zeta)}{\zeta}, \\ &= z'(u_0). \end{aligned}$$

or

(ii) for all $\zeta > 0$ sufficiently small, $\exists z(u_0) \ominus z(u_0 + \zeta)$, $\exists z(u_0 - \zeta) \ominus z(u_0)$ and limits (in the metric space \mathfrak{D}):

$$\begin{aligned} \lim_{\zeta \searrow 0} \frac{z(u_0) \ominus z(u_0 + \zeta)}{-\zeta} &= \lim_{\zeta \searrow 0} \frac{z(u_0 - \zeta) \ominus z(u_0)}{-\zeta}, \\ &= z'(u_0). \end{aligned}$$

z is called (1)-differentiable on (u, v) if z is differentiable under description (i) of Definition 3 and likewise for (2)-differentiability in Definition 3, case (ii).

Definition 4 (see, [23]) Assume that $z \in C(I, \mathcal{P}) \cap L(I, \mathcal{P})$. The Riemann–Liouville integral of the fuzzy-valued function f is stated as:

$$({}^{RL}I_{a+}^{\kappa}z)(x) = \frac{1}{\Gamma(\kappa)} \int_a^x \frac{z(v)dv}{(x-v)^{1-\kappa}}, \quad x > a, \quad 0 < \kappa \leq 1.$$

Definition 5 (see, [23]) Suppose that $z \in C(I, \mathcal{P}) \cap L(I, \mathcal{P})$ is a fuzzy set-value function, then f is a Caputo fuzzy H-differentiable at x when:

$$({}^CD_{a+}^{\kappa}z)(x) = \frac{1}{\Gamma(1-\kappa)} \int_a^x \frac{fz'(v)}{(x-v)^{\kappa}} dv, \quad (2)$$

in which $0 < \kappa \leq 1$; then, z is ${}^C[(1-\kappa)]$ -differentiable if Eq. (2) holds while z is (1)-differentiable, and z is ${}^C[(2-\kappa)]$ -differentiable if Eq. (2) holds while z is (2)-differentiable.

3 Main Results

Consider the following fuzzy fractional integro-differential equation involving generalized Caputo differentiability:

$${}^cD^{\kappa}y(t) = f(t, y(t)) + \int_0^t K(x, t)y(x)dx, \quad (3)$$

where $K(x, t)$ is a kernel and $f(t, y(t))$ is a continuous fuzzy-valued function.

Definition 6 The problem (3) is fuzzy Hyers–Ulam stable if there exists a real number α such that for each ϵ and for each solution $y \in C^F(I, \mathcal{P})$ of the inequality $d({}^cD^{\kappa}y(t), f(t, y(t)) + \int_0^t K(x, t)y(x)dx) < \epsilon$, there exists a solution $\bar{y}(t)$ belongs to $C^F(I, \mathcal{P})$ of the Eq. (3) with

$$d(y(t), \bar{y}(t)) < \alpha\epsilon, \quad (4)$$

for $t \in [0, b)$.

Definition 7 The problem (3) is fuzzy generalized Hyers–Ulam stable, if there exists $\eta \in C(R_+, R_+)$, $\eta(0) = 0$ such that for each solution $y \in C^F(I, \mathcal{P})$ of the inequality $d({}^cD^{\kappa}y(t), f(t, y(t)) + \int_0^t K(x, t)y(x)dx) < \epsilon$, there exist a solution $\bar{y} \in C^F(I, \mathcal{P})$ of the Eq. (3) with

$$d(y(t), \bar{y}(t)) < \eta\epsilon, \quad (5)$$

for $t \in [0, b)$.

Definition 8 The problem (3) is fuzzy generalized Hyers–Ulam–Rassias stable w.r.t. θ , if there exists $q > 0$ such that for each $\epsilon > 0$ and for each solution $y \in C^F(I, \mathcal{P})$ of the inequality $d({}^c D^\kappa y(t), f(t, y(t)) + \int_0^t K(x, t)y(x)dx) < \epsilon\theta(t)$, there exist a solution $\bar{y} \in C^F(I, \mathcal{P})$ of the Eq. (3) with

$$d(y(t), \bar{y}(t)) < \epsilon q \theta(t), \quad (6)$$

for $t \in [0, b)$.

Definition 9 The problem (3) is fuzzy generalized Hyers–Ulam–Rassias stable w.r.t. θ , if there exists $q > 0$ such that for each $\epsilon > 0$ and for each solution $y \in C^F(I, \mathcal{P})$ of the inequality $d({}^c D^\kappa y(t), f(t, y(t)) + \int_0^t K(x, t)y(x)dx) < \theta(t)$, there exist a solution $\bar{y} \in C^F(I, \mathcal{P})$ of the Eq. (3) with

$$d(y(t), \bar{y}(t)) < q \theta(t), \quad (7)$$

for $t \in [0, b)$.

Note that we say $y \in C^F(I, \mathcal{P})$ is a solution of inequity $d({}^c D^\kappa y(t), f(t, y(t)) + \int_0^t K(x, t)y(x)dx) < \epsilon$ iff there exists a function $h(t) \in C^F(I, \mathcal{P})$ with the following properties:

- (1) $d(h(t), \tilde{0}) \leq \epsilon, \quad t \in [0, b)$,
- (2) ${}^c D^\kappa y(t) = f(t, y(t)) + \int_0^t K(x, t)y(x)dx + h(t)$.

Let us to define the right-hand side of Eq. (3) as follows:

$$M(t, K(x, t), y(t)) = f(t, y(t)) + \int_0^t K(x, t)y(x)dx,$$

then we have the following result.

Lemma 1 Let $y \in C^F(I, \mathcal{P})$ and $\kappa \in (0, 1)$, such that y is a solution of the inequality

$$d({}^c D^\kappa y(t), M(t, K(x, t), y(t))) \leq \epsilon,$$

then y is a solution of the following inequality:

$$d\left(y(t), y(0) + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} M(s, K(x, s), y(s))ds\right) \leq \frac{(t-0)^\kappa}{\Gamma(1+\kappa)} \epsilon, \quad (8)$$

for $t \in [0, b)$.

Proof Due to the meaning of the definition of the inequality, we have:

$${}^c D^\kappa y(t) = M(t, K(x, t), y(t)) + h(t), \quad (9)$$

then, using the fractional integral on the both sides of Eq. (9), we have:

$$y(t) = y(0) + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} [M(s, K(x, s), y(s)) + h(s)]ds, \quad (10)$$

for $t \in [0, b)$. Consequently,

$$\begin{aligned} d\left(y(t), y(0) + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} M(s, K(x, s), y(s)) ds\right) &\leq \\ \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} d(h(s), \tilde{0}) ds &\leq \\ \frac{\epsilon}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} ds &\leq \frac{\epsilon}{\Gamma(\kappa+1)} (t-0)^\kappa. \end{aligned}$$

3.1 Fuzzy Ulam Stability

Here, we intend to derive some new results and extensions regarding to the fuzzy Ulam stability.

Consider the following assumptions hold:

(A1) $f \in C^F([0, b), \mathcal{P})$,

(A2) (Lipschitz condition), there exists $\zeta > 0$ such that

$$d(M(t, K(x, t), y(t)), M(t, K(x, t), \bar{y}(t))) \leq \zeta d(y, \bar{y}).$$

Next, the following theorem proves that under the aforementioned assumptions, Eq. (3) is fuzzy Hyers–Ulam stable.

Theorem 1 *Let us consider that the assumptions (A1) and (A2) hold, then, Eq. (3) is fuzzy Hyers–Ulam stable.*

Proof Suppose that $y \in C^F(I, \mathcal{P})$ is a solution of the inequality $d({}^c D^\kappa y(t), M(t, K(x, t), y(t))) \leq \epsilon$. Set, $\bar{y}(t)$ is the solution of problem

$$\begin{cases} {}^c D^\kappa \bar{y}(t) = M(t, K(x, t), \bar{y}(t)), \\ \bar{y}(t) = y(0), \quad 0 < \kappa < 1, \quad t \in [0, b), \end{cases} \quad (11)$$

then, we have:

$$\bar{y}(t) = y(0) + \frac{1}{\Gamma(\kappa)} \int_a^t (t-s)^{\kappa-1} M(s, K(x, s), \bar{y}(s)) ds. \quad (12)$$

From the obtained results, we have:

$$d\left(y(t), y(0) + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} M(s, K(x, s), y(s)) ds\right) \leq \frac{(t-0)^\kappa}{\Gamma(\kappa+1)} \cdot \epsilon, \quad (13)$$

also, we have:

$$\begin{aligned} d(y(t), \bar{y}(t)) &= d\left(y(t), y(0) + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} M(s, K(x, s), \bar{y}(s)) ds\right) \leq \\ d\left(y(t), y(0) + \frac{1}{\Gamma(\kappa)} \int_a^t (t-s)^{\kappa-1} M(s, K(x, s), y(s)) ds\right) &+ \\ + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} d(M(s, K(x, s), y(s)), M(s, K(x, s), \bar{y}(s))) ds &\leq \\ \epsilon + \frac{\zeta}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} d(y(s), \bar{y}(s)) ds & \end{aligned}$$

Moreover, there exists $\alpha > 0$ such that:

$$d(y(t), \bar{y}(t)) \leq \alpha \epsilon,$$

and the proof is completed.

3.2 Fuzzy E_α -Ulam Stability

In this part, we proposed a new type of stability for fuzzy fractional integro-differential equations, so called fuzzy E_α -Ulam stability. In the fractional literature, Mittag-Leffler function plays a major role in the theory of fractional calculus and fractional differential equations which is presented in the following lemma.

Lemma 2 (see, [19]) E_κ , is the Mittag-Leffler function defined as:

$$E_\kappa(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(k\kappa + 1)}$$

Also, $E_{\kappa,\kappa}(z)$ is defined as:

$$E_{\kappa,\kappa}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(k\kappa + \kappa)}.$$

Moreover, For any $\lambda > 0$ and $t \in [0, T]$,

$$E_\kappa(-t^\kappa \lambda) \leq 1, \quad E_{\kappa,\kappa}(-t^\kappa \lambda) \leq \frac{1}{\Gamma(\kappa)}.$$

Furthermore,

$$E_\kappa(0) = 1, \quad E_{\kappa,\kappa}(0) = \frac{1}{\Gamma(\kappa)}.$$

Similar to Sect. 3.1, consider the following three inequalities:

$$d\left({}^c D^\kappa y(t), f(t, y(t)) + \int_0^t K(s, t)y(s)ds\right) \leq \epsilon, \quad (14)$$

$$d\left({}^c D^\kappa y(t), f(t, y(t)) + \int_0^t K(s, t)y(s)ds\right) \leq \phi(t), \quad (15)$$

$$d\left({}^c D^\kappa y(t), f(t, y(t)) + \int_0^t K(s, t)y(s)ds\right) \leq \epsilon\phi(t), \quad (16)$$

in which then, analogously to the demonstration of the definitions presented in Sect. 3, we extend them to the following cases:

(A) (*Fuzzy E_α -Ulam stability*); Eq. (3) is fuzzy E_α -Hyers-Ulam stable if there exists $c > 0$ such that for each $\epsilon > 0$ and for each solution $y \in C^F(I, \mathcal{P})$ of the inequality (14), there exists a solution $\bar{y} \in C^F(I, \mathcal{P})$ of Eq. (3) with

$$d(y(t), \bar{y}(t)) \leq cE_\kappa(qt^\kappa)\epsilon, \quad q \geq 0. \quad (17)$$

(B) (*Fuzzy E_α -Hyers-Ulam-Rassias stability*); Eq. (3) is fuzzy E_α -Hyers-Ulam-Rassias stable w.r.t. ϕ if there exists $c_\phi > 0$ such that for each ϵ and for each

solution $y \in C^F(I, \mathcal{P})$ of the inequality (16) there exists a solution $\bar{y} \in C^F(I, \mathcal{P})$ of Eq. (3) with

$$d(y(t), \bar{y}(t)) \leq c_\phi E_\kappa(qt^\kappa)\epsilon, \quad q \geq 0. \quad (18)$$

(C) (*Fuzzy generalized E_α -Hyers–Ulam–Rassias stability*); Eq. (3) is fuzzy generalized E_α -Hyers–Ulam–Rassias stable w.r.t. ϕ if there exists $c_\phi > 0$ such that for each solution $y \in C^F(I, \mathcal{P})$ of the inequality (15), there exists a solution $\bar{y} \in C^F(I, \mathcal{P})$ of Eq. (3) with

$$d(y(t), \bar{y}(t)) \leq c_\phi \phi(t) E_\kappa(qt^\kappa)\epsilon, \quad q \geq 0. \quad (19)$$

Lemma 3 Suppose that $y \in C^F(I, \mathcal{P})$ is a solution of Eq. (14), then we have

$$d\left(y(t), E_\kappa(t^\kappa)y(0) + \int_0^t (t-s)^{\kappa-1} E_{\kappa,\kappa}((t-s)^\kappa) f(s, y) ds\right) \leq \frac{(t-0)^\kappa}{\Gamma(\kappa+1)} \cdot \epsilon. \quad (20)$$

Proof Using the fact that $|E_{\kappa,\kappa}(\cdot)| \leq \frac{1}{\Gamma(\kappa)}$, the proof will be achieved completely similar to the proof of Lemma 1.

It is worth noting that the similar results can be obtained for other types of stability.

Theorem 2 Suppose that (A1) and (A2) hold, then Eq. (3) is fuzzy E_α -Hyers–Ulam stable.

Proof It is completely similar to the proof of Theorem 1. Since, let $y \in C^F(I, \mathcal{P})$ is a solution of inequality (14) and $\bar{y}(t)$ is the unique solution of the problem

$$\begin{cases} {}^c D^\kappa \bar{y}(t) = f(t, \bar{y}(t)) + \int_0^t K(s, t) \bar{y}(s) ds, \\ \bar{y}(0) = y(0), \end{cases} \quad (21)$$

then, we have

$$\bar{y}(t) = E_\kappa(t^\kappa)y(0) + \int_0^t (t-s)^{\kappa-1} E_{\kappa,\kappa}((t-s)^\kappa) M(s, K(x, s), \bar{y}(s)) ds. \quad (22)$$

Consequently, we obtain:

$$\begin{aligned} d(y(t), \bar{y}(t)) &= d(y(t), E_\kappa(t^\kappa)y(0) + \int_0^t (t-s)^{\kappa-1} E_{\kappa,\kappa}((t-s)^\kappa) M(s, K(x, s), \bar{y}(s)) ds) \leq \\ &\leq d\left(y(t), E_\kappa(t^\kappa)y(0) + \frac{1}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} E_{\kappa,\kappa}((t-s)^\kappa) M(s, K(x, s), y(s)) ds\right) \\ &+ \int_0^t (t-s)^{\kappa-1} E_{\kappa,\kappa}((t-s)^\kappa) d(f(s, y(s)), f(s, \bar{y}(s))) ds \\ &\leq \frac{(t-0)^\kappa}{\Gamma(\kappa+1)} \cdot \epsilon + \frac{\zeta}{\Gamma(\kappa)} \int_0^t (t-s)^{\kappa-1} d(y(s), \bar{y}(s)) ds. \end{aligned} \quad (23)$$

Then, using the generalized singular Gronwall inequality [37], we finally obtain:

$$d(y(t), \bar{y}(t)) \leq c E_\kappa(\zeta t^\kappa)\epsilon. \quad (24)$$

So, Eq. (3) is fuzzy E_α -Hyers–Ulam stable.

4 Conclusion

We offer some new concepts in stability of a class of fuzzy fractional integro-differential equations with Caputo fuzzy fractional derivative from different perspectives. By applying a Gronwall inequality in a complete metric space, we present the Hyers–Ulam–Rassias stability as well as Hyers–Ulam stability of fuzzy fractional integro-differential equations, which maybe provide a new way for the researchers to discuss such interesting problems in the fuzzy mathematical analysis area. This is quite useful in many applications such as numerical analysis, optimization, biology and economics, where finding the exact solution is quite difficult. Since, if we are studying Hyers–Ulam–Rassias stable (or Hyers–Ulam stable) system then one does not have to reach the exact solution. All what is required is to get a function which satisfies a suitable approximation inequality.

Acknowledgements. The authors acknowledge the financial support from Universiti Putra Malaysia under Putra-IPB grant: GP-IPB/2017/9542402.

References

1. Ulam, S.M.: A Collection of Mathematical Problems. Interscience, New York (1960)
2. Hyers, D.H.: On the stability of the linear functional equation. *Proc. Natl. Acad. Sci.* **27**, 222–224 (1941)
3. Rassias, Th.M.: On the stability of linear mappings in Banach spaces. *Proc. Am. Math. Soc.* **72**, 297–300 (1978)
4. Hyers, D.H., Isac, G., Rassias, Th.M.: *Stability of Functional Equations in Several Variables*. Birkhäuser, Basel (1998)
5. Jung, S.-M.: *Hyers–Ulam–Rassias Stability of Functional Equations in Mathematical Analysis*. Hadronic Press, Palm Harbor (2001)
6. Jung, S.-M.: Hyers–Ulam stability of linear differential equations of first order. *Appl. Math. Lett.* **17**, 1135–1140 (2004)
7. Wang, G., Zhou, M., Sun, L.: Hyers–Ulam stability of linear differential equations of first order. *Appl. Math. Lett.* **21**, 1024–1028 (2008)
8. Rezaei, H., Jung, S.-M., Rassias, Th.M.: Laplace transform and Hyers–Ulam stability of linear differential equations. *J. Math. Anal. Appl.* **403**, 244–251 (2013)
9. Abdollahpour, M.R., Aghayari, R., Rassias, Th.M.: HyersUlam stability of associated Laguerre differential equations in a subclass of analytic functions. *J. Math. Anal. Appl.* **437**, 605–612 (2016)
10. Huang, J., Li, Y.: Hyers-Ulam stability of linear functional differential equations. *J. Math. Anal. Appl.* **426**, 1192–1200 (2015)
11. Kim, B., Jung, S.M.: Bessel’s differential equation and its Hyers–Ulam stability. *J. Inequalities Appl.* **2007**(1), 1–8 (2007)
12. Kim, S.S., Cho, Y.J., Eshaghi Gordji, M.: On the generalized Hyers–Ulam–Rassias stability problem of radical functional equations. *J. Inequalities Appl.* **2012**(1), 1–13 (2012)
13. Li, C.P., Zhang, F.R.: A survey on the stability of fractional differential equations. *Eur. Phys. J. Spec. Top.* **193**, 27–47 (2011)

14. Li, Y., Chen, Y., Podlubny, I.: Mittag–Leffler stability of fractional order nonlinear dynamic systems. *Automatica* **45**, 1965–1969 (2009)
15. Li, Y., Chen, Y., Podlubny, I.: Stability of fractional-order nonlinear dynamic systems: Lyapunov direct method and generalized Mittag–Leffler stability. *Comput. Math. Appl.* **59**, 1810–1821 (2010)
16. Wang, J., Lv, L., Zhou, Y.: Ulam stability and data dependence for fractional differential equations with Caputo derivative. *Electron. J. Qual. Theory Differ. Equ.* **63**, 1–10 (2011)
17. Wang, J., Lv, L., Zhou, Y.: New concepts and results in stability of fractional differential equations. *Commun. Nonlinear Sci. Numer. Simul.* **17**, 2530–2538 (2012)
18. Wang, J., Zhou, Y., Fečkan, M.: Nonlinear impulsive problems for fractional differential equations and Ulam stability. *Comput. Math. Appl.* **64**, 3389–3405 (2012)
19. Wang, J., Li, X.: E_α -Ulam type stability of fractional order ordinary differential equations. *J. Appl. Math. Comput.* **45**, 449–459 (2014)
20. Agarwal, R.P., Lakshmikantham, V., Nieto, J.J.: On the concept of solution for fractional differential equations with uncertainty. *Nonlinear Anal.* **72**, 2859–2862 (2010)
21. Allahviranloo, T., Gouyandeh, Z., Armand, A.: Fuzzy fractional differential equations under generalized fuzzy Caputo derivative. *J. Intell. Fuzzy Syst.* **26**, 1481–90 (2014)
22. Allahviranloo, T., Salahshour, S., Abbasbandy, S.: Explicit solutions of fractional differential equations with uncertainty. *Soft Comput.* **16**, 297–302 (2012)
23. Salahshour, S., Allahviranloo, T., Abbasbandy, S., Baleanu, D.: Existence and uniqueness results for fractional differential equations with uncertainty. *Adv. Differ. Equ.* **2012**, 112 (2012)
24. Mazandarani, M., Vahidian Kamyad, A.: Modified fractional Euler method for solving fuzzy fractional initial value problem. *Commun. Nonlinear Sci. Numer. Simul.* **18**, 12–21 (2013)
25. Mazandarani, M., Najariyan, M.: Type-2 fuzzy fractional derivatives. *Commun. Nonlinear Sci. Numer. Simul.* **19**, 2354–2372 (2014)
26. Salahshour, S., Allahviranloo, T., Abbasbandy, S.: Solving fuzzy fractional differential equations by fuzzy Laplace transforms. *Commun. Nonlinear Sci. Numer. Simul.* **17**, 1372–1381 (2012)
27. Ahmadian, A., Chang, C.S., Salahshour, S.: Fuzzy approximate solutions to fractional differential equations under uncertainty: operational matrices approach. *IEEE Trans. Fuzzy Syst.* <https://doi.org/10.1109/TFUZZ.2016.2554156>
28. Ahmadian, A., Salahshour, S., Amirkhani, H., Baleanu, D., Yunus, R.: An efficient Tau method for numerical solution of a fuzzy fractional kinetic model and its application to oil palm frond as a promising source of xylose. *J. Comput. Phys.* **264**, 562–564 (2015)
29. Malinowski, M.T.: Random fuzzy fractional integral equations-theoretical foundations. *Fuzzy Sets Syst.* (In press). <https://doi.org/10.1016/j.fss.2014.09.019>
30. Hoa, N.V.: Fuzzy fractional functional differential equations under Caputo gH-differentiability. *Commun. Nonlinear Sci. Numer. Simul.* **22**, 1134–1157 (2015)
31. Hoa, N.V.: Fuzzy fractional functional integral and differential equations. *Fuzzy Sets Syst.* (In press). <https://doi.org/10.1016/j.fss.2015.01.009>
32. Mirmostafae, A.K., Moslehian, M.: Fuzzy versions of Hyers–Ulam–Rassias theorem. *Fuzzy Sets Syst.* **159**, 720–729 (2008)
33. Mirmostafae, A.K., Mirzavaziri, M., Moslehian, M.S.: Fuzzy stability of the Jensen functional equation. *Fuzzy Sets Syst.* **159**, 730–738 (2008)

34. Shen, Y.: On the Ulam stability of first order linear fuzzy differential equations under generalized differentiability. *Fuzzy Sets Syst.* (In press). <https://doi.org/10.1016/j.fss.2015.01.002>
35. Diamond, P., Kloeden, P.E.: *Metric Spaces of Fuzzy Sets: Theory and Applications*. World Scientific, Singapore (1994)
36. Bede, B., Rudas, I.J., Bencsik, A.L.: First order linear fuzzy differential equations under generalized differentiability. *Inform. Sci.* **177**, 1648–1662 (2007)
37. Ye, H., Gao, J., Ding, Y.: A generalized Gronwall inequality and its application to a fractional differential equation. *J. Math. Anal. Appl.* **328**, 1075–1081 (2007)

Semantic Approach for Web-Based Presentation Mining Based Ontology Support

Vinothini Kasinathan^{1(✉)}, Aida Mustapha², and Imran Medi¹

¹ Faculty of Computing Engineering and Technology, Asia Pacific University of Innovation and Technology, Technology Park Malaysia, 57000 Bukit Jalil, Kuala Lumpur, Malaysia

vinothini@apiit.edu.my, dr.imran.medi@apu.edu.my

² Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
aidam@uthm.edu.my

Abstract. This paper proposes a Presentation Mining system, the purpose of which is the automatic formulation of a visual mind map of keyphrases identified and extracted from a set of presentation slides. The paper empirically demonstrates that the use of ontology increases the effectiveness in evaluating the quality of mind maps generated by the system, which in turn implies that the users of this system are able to differentiate the keywords or keyphrases unique to the presentation source and those existing in the domain ontology.

Keywords: Presentation mining · Ontology · Keyword extraction
Information visualisation

1 Introduction and Related Work

Keywords articulate, in compacted form, a document's content and the topics of relevance that it covers [1]. They are thus used by information retrieval systems as a means to represent the basic idea and concepts in the originating document [2]. The increase in digital information and documents has escalated the attention given to the way in which keywords and phrases are extracted from a text, particularly given the often laborious effort of manually extracting keywords. A typical manual process involves the careful identification and selection of an appropriate set of keyphrases that best represent the content of the document, the precise nature of which will help audiences to gain better understanding of the document. The process usually necessitates a level of familiarity with the document only available to its author, thus an automatic extraction methodology is highly desirable. It should be noted that keyphrase extraction is not predicated upon a predefined vocabulary from which keywords and phrases are selected, rather the document itself represents the source from which selection takes place. Candidate keyphrases are examined on several conditions such as the frequency and length [3].

Automated keywords extraction can be distinguished based on method and can either be supervised or unsupervised in nature. Supervised techniques focus on classification, whilst unsupervised approaches focus on ranking information. Supervised

approaches expect a human-defined keyphrase list and machine learning algorithm [4, 5], with which the system is trained to retrieve keyphrases accurately using the defined list and the involvement of a classifier algorithm. The dependence upon a training list raises questions as to the potential effectiveness of the supervised approach, as inaccuracies in the list will have implications upon how well supervised automated keyword identification performs.

This in turn limits the viability and utility of such systems in domain specific contexts [2], where domain information is vital to the document. In order to mitigate potential keyphrase incoherence, keyphrase extraction algorithms should leverage domain information through the allocation of higher scores or weighting to keyphrases relevant to the domain information provided. Common supervised extraction algorithms which facilitate this include GenEx and KEA [6, 7]. Unsupervised extraction on the other hand is not dependent upon a keyphrase list for training purposes. Rather, it statistically extracts keyphrases on the basis of semantic relatedness or the extent to which one candidate relates to other candidates [6]. Unsupervised extraction approaches are underpinned by learning algorithms, the rules by which words are examined and ranked in a text were determined by [6]. The unsupervised approach saves more time in comparison to its supervised counterpart due to a predefined list not being a requirement. The most well known unsupervised extraction approaches include Term Frequency Inverse Document Frequency (TFxIDF), TextRank, and PageRank [8].

The focus of this paper is the extraction of keyphrases from presentation slides. In presentation mining, keyphrases are extracted from a large set of input presentation slides before being used to generate a mind map representing the content of the original slides. The paper proposes a web-based presentation mining system which has the ability to retrieve text from presentation slides, calculate and extract keyphrases from the texts, as well as be able to construct a mind-map displaying these keyphrases. The keyphrases which emerge at the end of the process are not necessarily ‘correct’ or comprehensible due to the probabilities associated with the coupling of the extensive number of keyphrases that exist within the domain. With this in mind, the objective of this paper is to evaluate the degree of effectiveness when using a keyphrase generator to generate the mind map. Ontology is known to support semantic capability among many systems such as search engines and question-answering systems.

The remainder of this paper proceeds as follows: Sect. 2 presents the web-based presentation mining system, Sect. 3 details out the experimental setup as well as presents the results. Finally, Sect. 4 concludes with some direction for future works.

2 Presentation Mining

Presentation mining systems are designed to extract keywords and keyphrases from a large group of input slides and automatically generate a single mind map that summarizes the entire slides. The motivation of such systems is to assist learners to reconstruct their understanding in radial form as opposed to the linear manner in which presentations are traditionally delivered. Figure 1 shows the framework for the proposed presentation mining system. The system consists of three main layers, input, core and output. As a whole, the system receives a PowerPoint file path to process and an

output file path. Contents of the PowerPoint file are retrieved and go through the first stage-text pre-processing. Text pre-processing is a mandatory step in text mining to standardize the text format so as to facilitate a better mining process.

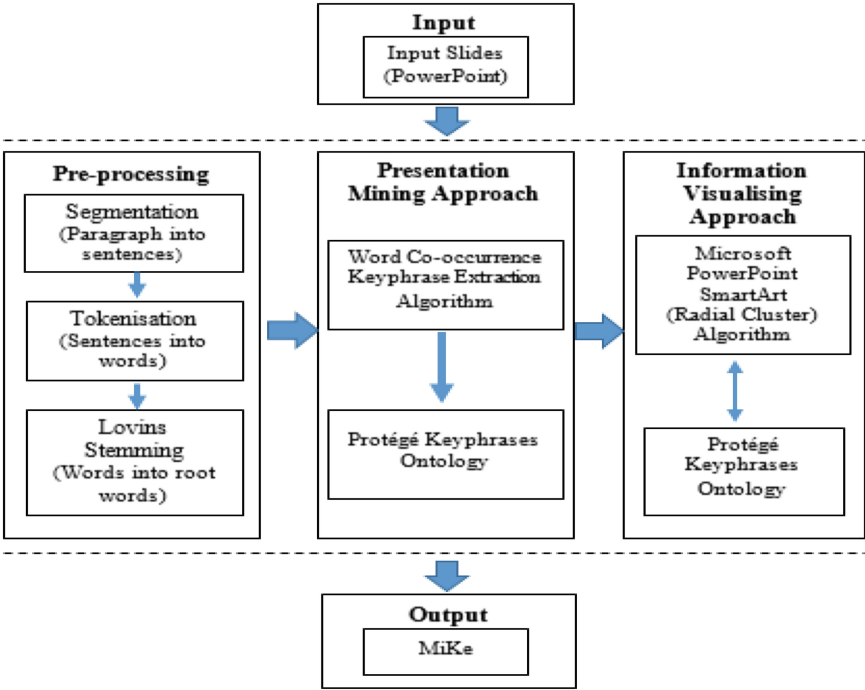


Fig. 1. Framework of the presentation mining system

Next, the processed text is sent to keyphrase extraction stage using the algorithms as shown in Tables 1 and 2. In Algorithm 1, standardisation involves performing ASCII-English transformation, converting single and double quotes, hyphen, cross-marking remaining unreadable ASCII characters, replacing tab with newline, and trim whitespaces. After standardization, pre- processing is carried out. Pre-processing involves sentence segmentation from individual slides, tokenization of sentences, stemming and lemmatization is to reduce inflectional forms to a common base form, and finally part-of-speech tagging to mark up the common base words into a particular part of speech such as verb and noun.

The keywords and keyphrases extracted are then generated into a mind map via the PowerPoint’s SmartArt provision. To improve the keyphrases relevancy in the mind map, the keyphrases will be compared against the domain ontology before the mind map is generated. The objective of using ontology is to verify whether the extracted keyphrase appears in the ontology. Nodes in the mind map are colored blue if the

Table 1. Algorithm 1: keyphrase extraction

```

for each input slide do
  standardization, pre-processing, phrase
  recognition and chunking
  if phrase matched with dictionary then
    select candidate phrase
  else
    discard candidate phrase
  end if
  for every candidate phrase do
    generate n-grams {Refer Algorithm 2}
    calculate c-value and weigh candidate
    phrases
  end for
end for

```

Table 2. Algorithm 2: generate N-grams

```

min n = (word count == 1) ? 1 : 2
max n = (word count >= 3) ? 3 : word count

for each n in max n do
  for each word in phrase do
    int gram count = n
    int pick index = word index
  end for
  while gram count > 0 do
    add picked word to phrase
    go to next word gram count
  end while
end for

```

keyphrase exists in the ontology. This reflects that the keyphrase is indeed part of the domain literature. Otherwise the node is colored red to indicate possibility of new keywords or keyphrases.

3 Experiments and Results

In order to measure the impact of ontology in evaluating the keyphrases extracted from the proposed keyphrase extraction system, an experiment is carried out to retrieve keywords and keyphrases. The latter two are retrieved from a set of input slides before generating a mind map using the keyphrases extracted. The resulting mind map is evaluated in terms of precision and recall, and then compared with the system supported by a domain ontology as opposed to the extraction system alone.

3.1 Figures

The datasets involved in the experiments include 500 pages of presentation slides for an Artificial Intelligence course at introductory level across different universities worldwide. Protégé was used to build a domain ontology for Artificial Intelligence (AI) based on the widely used AI textbook by Russell and Norvig [9]. This textbook was selected on the basis of its use in over 1,200 universities and 100 countries where it serves as the leading resource for Artificial Intelligence. An in depth number of related topics are covered hence keyphrases in the textbook are highly likely to appear in many universities slides. Figure 2 shows the AI ontology developed to support the web-based presentation mining system.

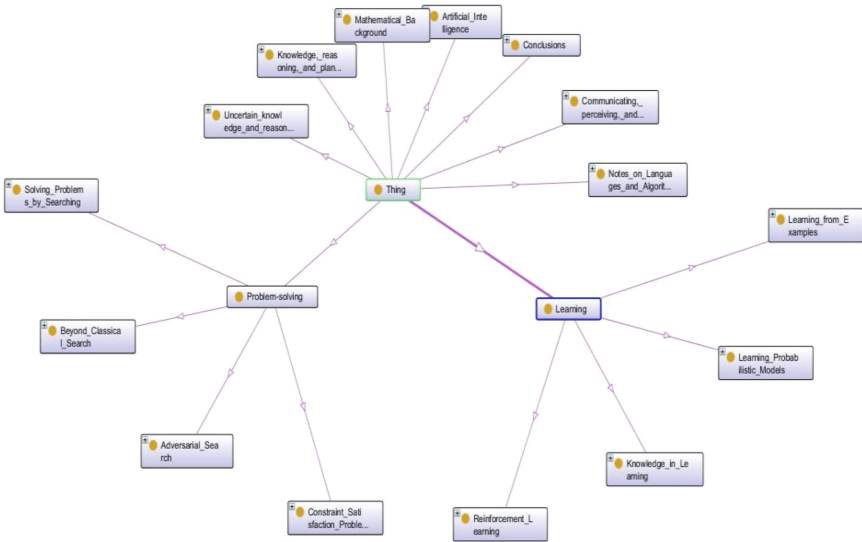


Fig. 2. Ontology for artificial intelligence based on [9]

Protégé IDE is an open-sourced ontology development tool developed and managed by the Stanford Center for Biomedical Informatics Research (BMIR) at the

Stanford University School of Medicine. It is widely used in constructing intelligent systems and knowledge-based solutions [10].

3.2 Evaluation Metrics

According to [11], the performance of keyphrase extraction system is measured by precision and recall criteria defined by the number of matches between the system-extracted keyphrases against the human-extracted keyphrases. The result of keyword extraction usually improves when the selected keywords get closer to those input by a user the ones suggested by a person. Since recall and precision have a mutual effect on each other, increase in precision leads to recall decrease and vice versa [12]. In this work, P is the number of correct keyphrases extracted divided by the sum of all of the keyphrases extracted. On the other hand, R is the number of relevant keyphrases retrieved divided by the total number of instances in the real class.

The formula for precision and recall are given in Eq. 1 and Eq. 2, respectively:

True = TP—True Positive (correct blue), TN True Negative (wrong in red)

False = FP—False Positive (wrong blue), FN False Negative (correct in red)

$$\text{Precision} = \frac{tp}{tp + fp} \quad (1)$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad (2)$$

where tp is a true positive (correct keyphrases extracted) and fp is a false negative (incorrect keyphrases not extracted). Since the goal of an extraction system is to have high precision, meaning high percentage of returned keyphrases are relevant; and high recall, meaning high percentage of relevant keyphrases are returned.

3.3 Results

In order to evaluate how well MiKe system performs, it was tested using 16 chapters from the AIMA text book, with a total 367 slides. The Table 3 shows the detail calculation of precision and recall for each chapter. In Chap. 3 where the total number of slides was 63, retrieves a small value for precision which is 0.696 and recall value 0.533. Based on the precision and recall calculation from the above, the more slides the smaller value of precision and recall. The smaller the sets of slides the more accurate keyphrases and keywords prove to be by value on precision and recall. Chapter 4b shows precision is 1 as there are only 27 slides.

Table 3. Results of all slides for precision and recall

Chapter 1	33	1	0.77
Chapter 2	16	1	0.5
Chapter 3	63	0.696	0.533
Chapter 4a	34	0.611	0.44
Chapter 4b	27	1	0.5
Chapter 5	26	0.6	0.53
Chapter 6	30	0.842	0.485
Chapter 7	22	0.938	0.429
Chapter 9a	16	0.688	0.55
Chapter 9b	13	1	0.393
Chapter 11	16	1	0.211
Chapter 13	15	1	1
Chapter 14	17	0.722	0.394
Chapter 15b	27	0.762	0.5
Chapter 16	22	0.857	0.857
Chapter 17	10	1	0.643

4 Conclusions

This paper has discussed how, through the use of keyword extraction, presentation slides can be synthesised into a visual mind map of the interlinked keyphrases. Not only that, this process can work in reverse, so as to revert back into slides based on the keywords and phrases present within the mind map. To the best of our knowledge this is the first time this tool/framework is presented. The current limitations of the system is that the ontology is domain specific. For future work, a domain agnostic system is proposed given the fact that mind maps are very useful in different fields or subjects in education.

As part of future work, further investigation will be given to making the algorithm more effective and just drawing one single mind map for each chapter. The development of presentation mining algorithm will be part of the system and outputs of this algorithm will be passed to an automatic mind map generating algorithm. The development of such system would contribute towards any area which uses presentation slides as an input and not just domain specific areas.

References

1. Gao, Y., Liu, J., Ma, P.: The hot keyphrase extraction based on tf^* pdf. In: 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1524–1528. IEEE (2011)
2. Mishra, A., Singh, G.: Improving keyphrase extraction by using document topic information. In: 2011 IEEE International Conference on Granular Computing (GrC), pp. 463–467. IEEE (2011)

3. Medelyan, O., Witten, I.H.: Thesaurus based automatic keyphrase indexing. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 296–297. ACM (2006)
4. Huang, H., Wang, H.: Keyphrases extraction research based on structure of document. In: 2010 2nd International Conference on Education Technology and Computer (ICETC), vol. 3, pp. V3–180. IEEE (2010)
5. Bellaachia, A., Al-Dhelaan, M.: Ne-rank: a novel graph-based keyphrase extraction in twitter. In: Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 372–379. IEEE Computer Society (2012)
6. Xie, F., Wu, X., Hu, X.: Keyphrase extraction based on semantic relatedness. In: 2010 9th IEEE International Conference on Cognitive Informatics (ICCI), pp. 308–312. IEEE (2010)
7. Lim, V.H., Wong, S.F., Lim, T.M.: Automatic keyphrase extraction techniques: a review. In: 2013 IEEE Symposium on Computers Informatics (ISCI), pp. 196–200, Apr 2013
8. Liu, Z., Chen, X., Zheng, Y., Sun, M.: Automatic keyphrase extraction by bridging vocabulary gap. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 135–144. Association for Computational Linguistics (2011)
9. Russell, S., Norvig, P.: Artificial intelligence: a modern approach. In: Artificial Intelligence. Prentice-Hall, Englewood Cliffs (2003)
10. Protégé IDE. <http://protege.stanford.edu/> (2013)
11. Turney, P.: Learning to extract keyphrases from text (1999)
12. Kian, H., Zahedi, M.: Improving precision in automatic keyword extraction using attention attractive strings. Arab. J. Sci. Eng. **38**(8), 2063–2068 (2013). <https://doi.org/10.1007/s13369-013-0573-6>

A Relative Tolerance Relation of Rough Set for Incomplete Information Systems

Rd. Rohmat Saedudin¹(✉), Hairulnizam Mahdin², Shahreen Kasim²,
Edi Sutoyo¹, Iwan Tri Riyadi Yanto³, and Rohayanti Hassan⁴

¹ School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia

{rdrohmat, edisutoyo}@telkomuniversity.ac.id

² Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

{hairuln, shahreen}@uthm.edu.my

³ Department of Information Systems, Universitas Ahmad Dahlan, 55161 Yogyakarta, Indonesia

yanto.itr@is.uad.ac.id

⁴ Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

rohayanti@utm.my

Abstract. Rough set theory is an effective approach to imprecision, vagueness, and uncertainty. This theory overlaps with many other theories such that fuzzy sets, evidence theory, and statistics. From a practical point of view, it is a good tool for data analysis. However, classical rough set theory cannot cope with the incomplete information systems where some attribute values are missing. There have been efforts in studying incomplete information systems for data classification which are based on the extensions of rough set theory. Moreover, the existing approaches have their weaknesses in terms of inflexible and imprecise in data classifications. To overcome these issues, we propose a relative tolerance relation of rough set (RTRS) to handling incomplete information systems, which it has flexibility and precisely for data classification. We compared RTRS with the existing approaches, the results show that our proposed method relatively achieves higher flexibility and precisely in data classification in incomplete information systems.

Keywords: Rough set theory · Limited tolerance relation · Relative precision
Incomplete information system

1 Introduction

In real world problems, it is common to find situations in which users are not able to provide all the preference values that are required, and then, we have to deal with missing value or incomplete information systems (IIS). Knowing how to handle missing values become important since the data insights or the performance of the predictive model could be impacted if the missing values are not appropriately handled.

There have been efforts in studying incomplete information systems. The simplest method to deal with incomplete information systems is to remove objects with unknown values [1] or to replace missing values with most common values [2]. Thus, this approach certainly reduces the sample size of data so that the objects with missing values may be different than the objects without missing values (e.g., missing values that are non-random). Recently, the extension of the classical rough set theory based on tolerance relation [3–7], non-symmetric similarity relation [8–10], and limited tolerance relation [11, 12] have also been proposed and studied to cope with incomplete information systems. In fact, this approach leads to poor results in terms of approximation. Consequently, Stefanowski and Tsoukias [8, 9] introduced non-symmetric similarity relation to refining the results obtained using tolerance relation approach. In Wang [12] and Yang et al. [11] proved that similarity relation will lost some information and proposed limited tolerance relation. Nevertheless, some information may also loss because the limited tolerance relation does not consider the similarity precision between two objects. Nguyen et al. [13] improved the tolerance relation by considering the probability matching between two objects. However, it needs to know the probability distribution of data in advance. In order to overcome their drawbacks, in this paper, we propose a relative tolerance relation of rough set (RTRS). The RTRS is based on limited tolerance relation by taking account into consideration the relative precision between two objects.

The rest of the paper is organized as follows. In Sect. 2, basic notions of rough set theory and rough set in an incomplete information system are presented. Section 3 describes the RTRS for handling incomplete information systems. Section 4 presents the results and compares them with existing approaches. Finally, the conclusion of this work is described in Sect. 5.

2 Theoretical Background

In this section, we recalled the notion of information systems, the idea of rough set theory and continued with the essential definitions of incomplete information system techniques based on rough set theory, namely Tolerance Relation, Non-Symmetric Similarity Relation, and Limited Tolerance Relation.

2.1 Information System

The idea of information system gives an appropriate tool for the representation of objects in terms of their attribute values. An information system is a 4-tuple (quadruple) $S = (U, A, V, f)$, where $U = \{u_1, u_2, \dots, u_{|U|}\}$ is a non-empty finite set of objects, $A = \{a_1, a_2, \dots, a_{|A|}\}$ is a non-empty finite set of attributes, $V = \cup_{a \in A} V_a$, V_a is the domain (value set) of attribute a , $f : U \times A \rightarrow V$ is an information function such that $f(u, a) \in V_a$, for every $(u, a) \in U \times A$, called information (knowledge) function [14]. If U in $S = (U, A, V, f)$ contains at least one object with an unknown or missing value, then S is called incomplete information system. The unknown or missing value is denoted by “*” in an incomplete information system. In this paper, we used the

quadruple $S^* = (U, A, V_*, f)$ to denote an incomplete information system. After the idea of an information system was presented as above, we recalled the notion of rough set theory in the following section.

2.2 Rough Set Theory

The idea of rough set theory was founded on the assumption that every object of the universe of discourse can be associated with some information (data, knowledge). Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible (similar) objects is called an elementary set, and form a basic granule (atom) of knowledge about the universe. Any union of some elementary sets is referred to as a crisp (precise) set—otherwise, the set is rough (imprecise, vague). Each rough set has boundary-line cases, i.e. objects of which their certainty cannot be classified, by employing the available knowledge, as members of the set or its complement.

Foremost, we recalled some fundamental definitions of rough set theory. Formal definitions and detailed description of rough set theory are originated from [15]. The concept of an information table is a quadruple $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, V is the union of attribute domains such that $V = \bigcup_{a \in A} V_a$ for V_a denotes the value domain of attribute a , any $a \in A$ determines a function $f_a : U \rightarrow V_a$ where V_a is the set of values of a .

Two elements $x, y \in U$ in $S = (U, A, V, f)$ are said to be B-indiscernible (indiscernible by the set of the attribute $B \subseteq A$ in S) if and only if $f(x, a) = f(y, a)$, for every $a \in B$ [14]. An indiscernible relation induced by the set of attribute B , denoted by $IND(B)$, is an equivalence relation. It is well known that an equivalence relation can induce a unique partition. The partition of U induced by $IND(B)$ in $S = (U, A, V, f)$ is denoted by U/B and the equivalence class in the partition U/B contains $x \in U$ and denoted by $[x]_B$. Let B be any subset of A in S and let X be any subset of U , the B-lower approximation of X , denoted by $\underline{B}(X)$ and B-upper approximation of X , denoted by $\bar{B}(X)$ respectively, are defined by

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \bar{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\}.$$

The accuracy of approximation of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted by $\alpha_B(X)$ is measured by $\alpha_B(X) = |\underline{B}(X)|/|\bar{B}(X)|$, where $|X|$ denotes the cardinality of X . For the empty set \emptyset , it is defined as $\alpha_B(\emptyset) = 1$ [16]. Clearly, $0 \leq \alpha_B(X) \leq 1$. If X is a union of some equivalence classes of U , then $\alpha_B(X) = 1$. Thus, the set X is crisp (precise) with respect to B . And, if X is not a union of some equivalence classes of U , then $\alpha_B(X) < 1$. Thus, the set X is rough (imprecise) with respect to B . This means that the higher the accuracy of approximation of any subset $X \subseteq U$, the more precise (the less imprecise) it would be [17].

2.3 Tolerance Relation

Given a complete decision system $S = (U, A, V_*, f)$, where $A = C\{d\}$, C is a set of condition attributes and d is the decision attribute, such that $f : U \times A \rightarrow V$, for any $a \in A$, where V_a is called domain of attribute a . An incomplete information system $S^* = (U, A, V_*, f)$, for any subset $B \subseteq C$, the tolerance relation T is determined by the following definition.

Definition 1 Let $S^* = (U, A, V_*, f)$ be an incomplete information system. A tolerance relation TR is defined as

$$\forall_{x,y \in U} TR(x, y) \Leftrightarrow \forall_{c_j \in B} (c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *) \quad (1)$$

Thus,

$$TR = \{(x, y) | x \in U \wedge y \in U \wedge \forall_{c_j \in B} (c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *)\}$$

Obviously, TR is reflexive and symmetric but does not need to be transitive. The tolerance class $I_B^T(x)$ of an object x with reference to an attribute set B is defined as $I_B^T(x) = \{y | y \in U \wedge TR_B(x, y)\}$.

Definition 2 Let $S^* = (U, A, V_*, f)$ be an incomplete information. The lower approximation x_B^T and upper approximation x_B^B of an object set X with reference to attribute set B respectively can be defined as follows:

$$x_B^T = \{x | x \in U \wedge I_B^T(x) \subseteq X\} \text{ and } x_B^B = \{x | x \in U \wedge I_B^T(x) \cap X \neq \emptyset\} \quad (2)$$

2.4 Non-symmetric Similarity Relation

An object x is considered to be similar to object y only if all their known attribute values are the same. Thus, one object may have more complete description than the other, the inverse relation does not hold [5]. The notion of a non-symmetric similarity relation is given as follows:

Definition 3 Let $S^* = (U, A, V_*, f)$ be an incomplete information system. A non-symmetric similarity relation S is defined as

$$\forall_{x,y \in U} (S_B(x, y) \Leftrightarrow \forall_{c_j \in B} (c_j(x) = c_j(y) \vee c_j(x) = *)) \quad (3)$$

It is clear that S is transitive and reflexive but not symmetric. From Definition 3, we can induce two similarity sets as given in Definitions 4 and 5.

Definition 4 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$. The set of objects similar to object x denoted by $\text{Sim}_B(x)$ is defined as

$$\text{Sim}_B(x) = \{y | y \in U \wedge S_B(y, x)\} \quad (4)$$

Definition 5 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$. The set of objects which x is similar to Sim_B^{-1} is defined as

$$\text{Sim}_B^{-1}(x) = \{y | y \in U \wedge S_B(x, y)\} \quad (5)$$

Obviously, $\text{Sim}_B(x)$ and $\text{Sim}_B^{-1}(x)$ are two different sets.

Furthermore, from Definitions 4 and 5, the lower approximation and upper approximation of objects set X can be defined as follows:

Definition 6 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$. The lower-approximation X_B^S and the upper-approximation X_S^B of an object set X with respect to an attribute set $B \subseteq A$ are respectively defined as

$$X_B^S = \{x | x \in U \wedge \text{Sim}_B^{-1}(x) \subseteq X\} \text{ and } X_S^B = \cup \{\text{Sim}_B(x) | x \in X\} \quad (6)$$

The approximations showed by the non-symmetric similarity relation are more informative than those resulted by tolerance relation.

2.5 Limited Tolerance Relation

In an information system, two objects may be distinct because of a little missing information. For example, two objects $a = \{x^*, y, z, w\}$ and $b = \{^*, v, y, z, w\}$ are similar, but they do not satisfy the non-symmetric similarity relation. To avoid such problem, Wang [12] developed a limited tolerance relation based on the following definition.

Definition 7 Let $S^* = (U, A, V_*, f)$ be an incomplete information system, a subset $B \subseteq A$, and $P_B(x) = \{b | b \in B \wedge b(x) \neq *\}$. A binary relation L (limited tolerance relation) defined on U is given by

$$\forall_{x,y \in U} (L_B(x, y) \Leftrightarrow \forall_{b \in B} (b(x) = b(y) = *) \vee ((P_B(x) \cap P_B(y) \neq \emptyset) \wedge \forall_{b \in B} ((b(x) \neq *) \wedge (b(y) \neq *) \rightarrow (b(x) = b(y)))))) \quad (7)$$

Obviously, the limited tolerance relation is symmetric and reflexive but not transitive. In Definition 8, the condition that $(b(x) \neq *) \wedge (b(y) \neq *) \rightarrow (b(x) = b(y))$ is equivalent to $(b(x) = *) \vee (b(y) = *) \vee (b(x) = b(y))$. Thus, two objects that satisfy the tolerance relation but not limited tolerance relation are only those with the formula $P_B(x) \cap P_B(y) = \emptyset$.

In other words, two objects are in limited tolerance relation if they are in one of the two cases. The first case is that all attribute values of the two objects are missing. Meanwhile the second case is where there is at least an attribute having an ordinary value for both objects and the two objects have the same value for those attributes. The notion of limited tolerance class is given as follows:

Definition 8 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and a subset $B \subseteq A$. The limited tolerance class is defined as

$$I_B^L(x) = \{y | y \in U \wedge L_B(x, y)\}. \quad (8)$$

From Definition 8, the notions of lower approximation and upper approximation of an object x based on the limited tolerance class are given in the following definition.

Definition 9 The lower approximation and the upper approximation of an object x based on the limited tolerance class $I_B^L(x)$ are respectively defined as:

$$D_L^B = \{x | x \in U \wedge I_B^L(x) \cap D \neq \phi\} \text{ and } D_B^L = \{x | x \in U \wedge I_B^L(x) \subseteq D\}. \quad (9)$$

3 Proposed Method

We introduce the concept of relative precision between objects x and y in order to determine both objects are tolerant.

3.1 Similarity Precision

Given an incomplete information system $S = (U, A, V_*, f)$, where $A = C \cup \{d\}$, C is a set of condition attributes and d the decision attribute, such that $f : U \times A \rightarrow V_*$. For any $a \in A$, where V_a is called domain of an attribute a and a subset $B \subseteq C$, the *similarity precision* is defined as follows.

Definition 10 Let $P_B(x) = \{b | b \in B \wedge b(x) \neq *\}$, the relative precision δ , is defined as

$$\delta(x, y) = \frac{|P_B(x) \cap P_B(y)|}{|P_B(x) \cup P_B(y)|}, \quad (10)$$

where $|\bullet|$ represents the cardinality of the set.

From Definition 10, it is clear that $0 < \alpha(x, y) \leq 1$. From Definition 10, the relative tolerance relation with similarity precision is given as follow:

Definition 11 Let given an incomplete information system $S = (U, A, V_*, f)$. The relative tolerance relation with similarity precision $L \delta$ is defined as follows

$$\forall_{x, y \in U \times U} (L\delta_B(x, y) \Leftrightarrow \forall_{b \in B} (b(x) = b(y) = *) \vee ((\alpha(x, y)) \geq \delta) \wedge \forall_{b \in B} (((b(x) \neq *) \wedge (b(y) \neq *)) \rightarrow (b(x) = b(y)))) \quad (11)$$

where $\delta \in (0, 1]$ is a threshold value.

Since $\delta \in (0, 1]$, then $0 < \alpha(x, y) \leq 1$ which implies that $P_B(x) \cap P_B(y) \neq \phi$ holds, but not vice versa if certain threshold value of the similarity is given.

Now, we define the extended tolerance relation by using similarity precision with a threshold.

Definition 12 Let given an incomplete information system $S = (U, A, V_*, f)$, a subset $B \subseteq C$ and a threshold δ . The relative tolerance relation with similarity precision is defined as follows

$$L\delta_B(x, y) \Leftrightarrow \alpha_B(x, y) \geq \delta \quad (12)$$

It is easy to observe that the above relation is reflexive and symmetric but not necessarily transitive.

Definition 13 Let given an incomplete information system $S = (U, A, V_*, f)$ and $B \subseteq C$. The limited tolerance class is defined as

$$I_B^{L\delta}(x) = \{y | y \in U \wedge L\delta_B(x, y)\} \quad (13)$$

Thus, the notions of lower approximation and upper approximation of an object x based on the limited tolerance class are given in the Definition 14.

Definition 14 Let given an incomplete information system $S = (U, A, V_*, f)$. The lower approximation and the upper approximation of an object x based on the limited tolerance class $I_B^{L\delta}(x)$ denoted as $D_{L\delta}^B(x)$ and $D_B^{L\delta}(x)$ respectively are defined as

$$D_B^{L\delta} = \{x | x \in U \wedge I_B^{L\delta}(x) \subseteq D\} \text{ and } D_{L\delta}^B = \{x | x \in U \wedge I_B^{L\delta}(x) \cap D \neq \emptyset\} \quad (14)$$

Obviously, if there is $0 \leq \delta_1 < \delta_2 \leq 1$. Thus, for every $a \in I_B^{L\delta_2}(x)$, if $\alpha_B(x, y) \geq \delta_2$. Since $\delta_2 > \delta_1$, then $\alpha_B(x, y) \geq \delta_1$, that is $\forall a \in I_B^{L\delta_1}(x)$ which implies $I_B^{L\delta_2}(x) = I_B^{L\delta_1}(x)$. Otherwise, $\alpha_B(x, y) \geq \delta_1$ then it does not necessarily $\alpha_B(x, y) \geq \delta_2$. Hence $I_B^{L\delta_2} \subseteq I_B^{L\delta_1}$.

Therefore, the proposed relative tolerance relation with similarity precision is an improved approach of limited tolerance relation in incomplete information systems. In the following section, we make simulation on benchmark datasets and compare the results obtained with other techniques.

4 Result and Discussion

In this section, we compare the proposed based on accuracy in term of flexibility. We will first describe 7 datasets used for simulation as follows.

4.1 Dataset

We elaborate the four approaches through the UCI benchmark datasets [18] as follow

1. The Zoo data set is comprised of 101 objects, where each data point represents information of an animal in terms of 18 categorical attributes. Each animal data point is classified into 7 classes.

2. Soybean data set contains 47 instances and 35 categorical attributes.
3. Tic-tac-toe dataset contains 958 instances and 9 attributes.
4. Monk dataset contains 432 instances and 6 attributes.
5. Spect dataset contains 187 instances and 922 attributes.
6. Car dataset contains 1728 instances and 6 attributes.
7. United States Congressional Voting Records Dataset contains 435 instances and 16 attributes.

4.2 Results

In this section, we compare the proposed Relative Tolerance Relation of Rough Set (RTRS) with the existing baseline approaches i.e. Tolerance Relation (TR), Limited Tolerance Relation (LTR), and Non-Symmetric Similarity Relation (NSSR) approaches based on accuracy in term of flexibility. We will first recall the notion of accuracy. The accuracy in term of is defined as follows (Table 1).

Table 1. The accuracy measurement of each approach

Approaches	Accuracy measurement	Description
Tolerance relation	$\frac{x_B^T = \{x x \in U \wedge I_B^T(x) \subseteq X\}}{x_T^B = \{x x \in U \wedge I_B^T(x) \cap X \neq \phi\}}$	Definition 2
Limited tolerance relation	$\frac{X_B^S = \{x x \in U \wedge \text{Sim}_B^{-1}(x) \subseteq X\}}{X_S^B = \cup \{\text{Sim}_B(x) x \in X\}}$	Definition 6
Non-symmetric similarity relation	$\frac{D_L^B = \{x x \in U \wedge I_B^L(x) \cap D \neq \phi\}}{D_B^L = \{x x \in U \wedge I_B^L(x) \subseteq D\}}$	Definition 9
Relative tolerance relation of rough set	$\frac{D_{L\delta}^B = \{x x \in U \wedge I_B^{L\delta}(x) \subseteq D\}}{D_{L\delta}^B = \{x x \in U \wedge I_B^{L\delta}(x) \cap D \neq \phi\}}$	Definition 14

In the experimentation, the proposed approach and other three baseline approaches are implemented in MATLAB version 8.3.0.532 (R2014a). They are executed sequentially on a processor Intel Core i5-6200U Processor 2.30 GHz CPUs. The total main memory is 4 GB and the operating system is Windows 10.

Table 2. Accuracy in terms of flexibility and precise

Dataset	Tolerance relation	Limited tolerance relation	Non-symmetric similarity relation	Proposed relative precision tolerance relation of rough set	Improvement (%)
Spect	0.5889	0.5889	0.8831	1	11.69
Monk	0.4068	0.4068	0.8112	1	18.88
Soybean	0.8229	0.8829	0.9368	1	6.32
US Voting	0.1701	0.4637	0.9760	0.9819	0.59
Zoo	0.7908	0.7908	1	1	0.00
Car	0.1655	0.1665	0.5565	1	44.35
Tic-tac-toe	0.6315	0.6315	0.9594	0.9954	3.60
Improvement					12.20

The proposed and existing approaches are implemented to seven incomplete data set, i.e. Spect, Monk, Soybean, US voting, Zoo, Car, and tic tac toe dataset from UCI learning machine. The results show that the proposed approach outperforms as compared to the existing approaches in term of flexibility and precise of accuracy. The results are summarized in Table 2, and the proposed technique significantly improves the accuracy up to 12.20%.

5 Conclusion

Missing values in data is a common phenomenon in real world problems. Knowing how to handle missing values become important since the data insights or the performance of the predictive model could be impacted if the missing values are not appropriately handled. In this paper, we proposed a new approach based on limited tolerance relation by taking into account the similarity precision between two objects. The results showed that the proposed method is more flexible and precise as compared with three existing approaches.

References

1. Bunting, B.P., Adamson, G., Mulhall, P.K.: A Monte Carlo examination of an MTMM model with planned incomplete data structures. *Struct. Equ. Model.* **9**, 369–389 (2002)
2. Chmielewski, M.R., Grzymala-Busse, J.W., Peterson, N.W., Than, S.: The rule induction system LERS-a version for personal computers. *Found. Comput. Decis. Sci.* **18**, 181–212 (1993)
3. Kryszkiewicz, M.: Rough set approach to incomplete information systems. *Inf. Sci.* **112**, 39–49 (1998)
4. Kryszkiewicz, M.: Rules in incomplete information systems. *Inf. Sci.* **113**, 271–292 (1999)
5. Zhou, J., Yang, X.: Rough set model based on hybrid tolerance relation. In: *International Conference on Rough Sets and Knowledge Technology*, pp. 28–33. Springer (2012)
6. Zhou, Q.: Research on tolerance-based rough set models. In: *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM)*, pp. 137–139. IEEE (2010)
7. Yang, X.: An improved model of rough sets on incomplete information systems. In: *International Conference on Management of e-Commerce and e-Government*, 2009. ICMECG'09, pp. 193–196. IEEE (2009)
8. Stefanowski, J., Tsoukiàs, A.: On the extension of rough sets under incomplete information. In: *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pp. 73–81. Springer (1999)
9. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. *Comput. Intell.* **17**, 545–566 (2001)
10. Wu, Y., Guo, Q.: An extension model of rough set in incomplete information system. In: *2010 2nd International Conference on Future Computer and Communication (ICFCC)*, pp. 2–434. IEEE (2010)
11. Yang, X., Song, X., Hu, X.: Generalisation of rough set for rule induction in incomplete system. *Int. J. Granul. Comput. Rough Sets Intell. Syst.* **2**, 37–50 (2011)

12. Wang, G.: Extension of rough set under incomplete information systems. In: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, 2002. FUZZ-IEEE'02, pp. 1098–1103. IEEE (2002)
13. Van Nguyen, D., Yamada, K., Unehara, M.: Extended tolerance relation to define a new rough set model in incomplete information systems. *Adv. Fuzzy Syst.* **2013** 9 (2013)
14. Dai, J., Wang, W., Xu, Q., Tian, H.: Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. *Knowl.-Based Syst.* **27**, 443–450 (2012)
15. Pawlak, Z.: Rough sets. *Int. J. Comput. Inform. Sci.* **11**, 341–356 (1982)
16. Parmar, D., Wu, T., Blackhurst, J.: MMR: an algorithm for clustering categorical data using rough set theory. *Data Knowl. Eng.* **63**, 879–893 (2007)
17. Herawan, T., Deris, M.M., Abawajy, J.H.: A rough set approach for selecting clustering attribute. *Knowl.-Based Syst.* **23**, 220–231 (2010)
18. Lichman, M.: UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml> (2013)

Fuzzy Evaluation Scheme for KDF Based on Stream Ciphers

Hamijah Mohd. Rahman^{1,2(✉)}, Nureize Arbaiy¹,
and Chuah Chai Wen²

¹ Soft Computing and Data Mining (SMC), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

hamijahrahman@gmail.com, nureize@uthm.edu.my

² Information Security Interest Group (ISIG), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
cwchuah@uthm.edu.my

Abstract. Cryptography is a practice of technique to ensure security by using the cryptography keys. Key derivation function (KDF) is a standard algorithm to generate these cryptographic keys. Stream ciphers are one of the cryptographic primitives that are used to construct the key derivation function namely key derivation function based on stream ciphers. Though the key derivation function based on stream ciphers have a great role in security, it is necessary to have a framework which can evaluate the security level of the different types of key derivation function based on stream ciphers. Random oracle model (ROM) is the current procedure to proofs the security of KDF. However, the security evaluation of ROM did not evaluate the degree of secureness of KDF as it can only proof either the KDF is theoretically secure or insecure. Hence, this research applies fuzzy evaluation method to form a framework to evaluate the degree of secureness of the KDF for different types of key derivation function based on stream ciphers. Key sizes and complexity attacks are two main variables which are considered in the design of fuzzy rule. The proposed method introduces the information extraction to construct fuzzy membership function and rules. The result from this proposal is effective to approximate the security aspect in the computer system as well as network system.

Keywords: Key derivation function · Stream cipher · Fuzzy logic
Membership function · Fuzzy evaluation

1 Introduction

Cryptography is a method used to protect information especially in electronic communication. The cryptography transforms the information into an unreadable format using cryptography keys. The cryptographic keys are essential for safeguarding all the security application in computer system. These cryptographic keys may province confidentiality to the private message. One of the standard algorithms used to generate

these cryptographic keys is key derivation functions (KDF). KDFs are used to generate one or more cryptographic keys from a private string such as password, Diffie-Helman (DH) shared secret or some non-uniformly random source materials [1]. Block ciphers, hash functions and stream ciphers are three cryptographic primitives that used to construct these KDF proposals.

Stream cipher is a symmetric key cipher that provides confidentiality protection for data [2, 3]. To provide the confidentiality, the data is XORed with pseudorandom keystream bit by bit to form the ciphertext. This process is known as encryption. To recover the data from the ciphertext, the reverse process of encryption is performed, namely decryption. Stream cipher is simple and fast [4]. Therefore, the researchers had designed the KDF using these stream ciphers, namely stream cipher based KDF (SCKDF) [5].

To date, the security proofs of KDF are based on random oracle model (ROM) either the KDF is theoretically identified as secure or insecure [6]. The current procedure of ROM evaluates the security of KDF based on security model of KDF namely Adaptive Chosen Public Input Model with Multiple Salts [7] is either secure or insecure. For the proofs, it is assumed that the stream ciphers are ideal ciphers that perform well in ROM. Such evaluation did not evaluate the degree of secureness of KDF. As it is known that not all stream ciphers perform well in ROM due to other factors. Some of the stream ciphers are vulnerable to attacks such as correlation attack, related-key chosen IV, advanced algebraic attack, distinguish attack and so on [8–11]. It makes the existing security evaluation is incapable to measure the degree of secureness of the KDF. Hence, there is a need to evaluate the degree of secureness of the KDF when it is proofed as either secure or insecure by considering the attacks happen to that particular stream ciphers.

Furthermore, the security evaluation by ROM is based on standard complexity-theoretic assumption [7]. Practically, the specification of the procedure is complex that only known by the expert. In real world, this complex model can be defined with some simple and understandable specification using Fuzzy Logic approach. The fuzzy set and membership to a certain parameter can be used to describe the pseudorandom sequence used in security evaluation of KDF [12]. In addition, the fuzzy tools provide a simplified platform to reduce development time in the analysis model [13].

Hence, motivated by this situation, this paper introduced fuzzy evaluation system for KDF. Mamdani inference engine is used to evaluate the security strength based on KDF algorithm. This study incorporates information extraction steps to acquire knowledge from security experts. This information is crucial to construct fuzzy membership function. Fuzzy rules are acquired to construct fuzzy inferencing for the security evaluation. The designed fuzzy rules in this model consider two main variables which are the key length and the complexity of the attack. The linguistic term for security level such as ‘*strong*’, ‘*moderate*’ and ‘*weak*’ are included in the model to represent the degree of secureness using linguistic notation. The uncertain and imprecise information could be handled through integration of fuzzy logic concept. The result of security strength which will produce from this proposal can be used to approximate the security aspect in the network and computer system.

The remainder of this paper is organized as follows. Section 2 describes the overview of related works. Section 3 explains the evaluation of KDF using Fuzzy Logic. Section 4 explains the experiment of the fuzzy evaluation. Section 5 describes the result and discussion. Section 6 draws the conclusions.

2 Overview of Related Works

This section discusses about the related work to this research which are KDFs, stream cipher algorithm, types of stream cipher, and Fuzzy system.

2.1 Key Derivation Functions

KDF is the standard algorithm used to derive cryptographic keys from keying materials based on private string and public strings. KDF transform the inputs into n bit pseudorandom string which can be used as cryptographic key [7]. In cryptography, the algorithm to build the KDF usually called as cipher. The standard ciphers which are used to construct the KDF are block ciphers and stream ciphers. Both ciphers are categorized as symmetric keys encryption. However, stream cipher is faster than block as it works on a few bits of plaintext at a time.

2.2 Stream Cipher Algorithms

Stream cipher is a kind of symmetric encryption that encrypts plaintext one bit at a time. As such, this cipher tends to be simple and fast. Hence, this cipher widely applied in little computational resource such as cell phones or other small embedded devices [3]. As technology arises, a development to upgrade stream cipher efficiency was developed to provide confidentiality of the cipher [2]. In communication application, a lighter and faster stream cipher was developed to enhance the smartphone performance [4]. The stream cipher also conducted to the performance of resistive RAM due to the data security of embedded system is becoming more important [3].

The process of encryption in stream cipher uses a bitwise XOR with the corresponding keystream character to produce ciphertext. The decryption process uses XORed ciphertext with an identical keystream. The keystream in encryption and decryption process is generated by keystream generator. Secret key and initialization vector are two inputs provided for keystream generator. While the output is the longer pseudorandom binary sequence that may be in bits, bytes or words. Different types of stream ciphers have been developed despite on the advantages of its flexibility and speed. Four common types of stream ciphers are A5/1, Grain, Trivium and Rabbit.

2.3 Types of Stream Ciphers

A5/1, Grain, Trivium and Rabbit are types of stream ciphers that commonly used. A5/1 is an earlier stream cipher designed in 1989. This kind of stream cipher is used in the GSM cellular telephone standard. It uses a key of 64 bits, initial vector 22 bits and internal state of A5/1 is 64 bits [8]. Grain is a stream cipher developed in 2004. Grain is

designed for restricted hardware environment. The key size is 80 bits, initial vector is specified to be 64 bits and internal state of Grain is 160 bits [9]. Trivium is a synchronous stream cipher designed in 2004. It provides flexible trade-off between speed and gate count in hardware. Trivium generates output from 80-bit key, 80-bit initial vector and internal state for the Trivium is 288 bits [10]. In 2003, Rabbit was designed with high performance and compact in hardware. It uses 128-bit key, 64-bit initial vector and internal state of Rabbit is 513 bits [11]. These four types of stream cipher obviously different according to their key length, initial vector and internal state. Despite on the differences between those ciphers, they have been implemented in different computer system applications.

2.4 Fuzzy Inference System

The fuzzy concept has demonstrated its ability in many different applications including in network security domain. Fuzzy logic offers a nature way to model expert's knowledge which is often imprecise in nature. The fuzzy rule was used in cyber security to protect system administrator and prevent harmful behaviour [14–16]. However, there is still a few researches discussed on the security evaluation in term of strength level using fuzzy logic approach. One of the current literatures discussed on the subject matter was proposed by Mohammed and Sadkhan [17] to evaluate the security of block cipher by using the fuzzy logic as to protect the communication system in wireless network. Another application of fuzzy logic in security evaluation was presented in [12] whereby the method was applied to a type on non-linear feedback shift register. The results in these studies showed an effective way of carrying out security evaluation model by using fuzzy logic method.

Fuzzy inference system (FIS) is a major unit of a fuzzy logic system [18]. The FIS are also known as fuzzy model, fuzzy rule based system and fuzzy associative memory (FAM). Fuzzy membership functions, rules and reasoning are the main structures implied in fuzzy inference system. The membership function is a generalization of the indicator function to represents the degree of truth as an extension of valuation [19]. It is mapped to a membership value (or degree of membership) between 0 and 1. It can be expressed in the form of a curve such as trapezoidal.

The set of rules consist of a set 'IF' condition and one 'THEN' conclusion and an optional 'ELSE' conclusion [18]. 'AND', 'OR' and 'NOT' are connectives to join multiple conditions depending on the problem situation. Mamdani is a common method in fuzzy inference system due its simple min-max structure. The model was proposed by Mamdani and Assilian [20] as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators [21]. Mamdani-type FIS uses the technique of defuzzification of a fuzzy output. The output from Mamdani FIS can be easily transformed to a linguistic form as the inference result before defuzzification. The fuzzy inference system is a kind of tool that can control the uncertain input. Hence, this approach has been used recently for the evaluation of security in different situation.

3 Fuzzy Evaluation of KDF

Considering aforementioned concept, the existing KDF process is enhanced using a fuzzy method. That is the new F-KDF process contains a method to identify parameters of fuzzy variable, and find the optimal setting to evaluate the level of security.

3.1 Fuzzy KDF Model

The fuzzy evaluation based on KDF algorithm consists of four stages:

Step 1: Problem Description

Determine types of stream cipher to be evaluated.

Step 2: Information Extraction

Determine types of stream cipher to be evaluated.

Those factors are needed to develop the membership functions and fuzzy rules in the fuzzy model. The factors for the fuzzy evaluation model are tabulated in the Table 1 respectively.

Table 1. Elements required for fuzzy evaluation model

Cipher	Factor	Security level
C_1, C_2, \dots, C_n	(F_1, F_2, \dots, F_n)	S_1, S_2, \dots, S_n

Step 3: Fuzzy evaluation model

Three major steps included in the fuzzy model which are fuzzification, fuzzy inferencing and defuzzification.

i. Fuzzification

Fuzzification is a process of converting crisp inputs data and determines the degree to which these inputs belong to appropriate fuzzy set. The process of fuzzification is using fuzzy variables, fuzzy linguistic variables, fuzzy linguistic term and membership function.

a. Identify fuzzy variable (FV)

Fuzzy variables are the input or output variables of a system where the values are from natural language.

b. Identify linguistic variable for each FV

In real life, term such as “fast” and “slow” are used to qualify the complexity of attack, these words are called the linguistic values for the complexity of attack. Then $(A) = \{Fast, SlightlyFast, Moderate, SlightlySlow, Slow\}$ can be the set of decomposition for the linguistic variable attack. Each member of this decomposition is called a linguistic term.

c. Draw membership function graph

Membership functions are used to map the crisp input values to fuzzy linguistic terms and vice versa in fuzzification and

defuzzification step. Triangular, Trapezoidal and Gaussian shapes are the common types of membership function [22].

ii. Fuzzy Inferencing

The fuzzy rules are used to obtain output from crisp input in the fuzzification and defuzzification process. The fuzzy rule is also known as fuzzy if-then rules, fuzzy conditional statement or fuzzy implication [19]. The general form of the if-then rule is shown in Eq. (1) respectively.

$$IF (x \text{ is } A) \text{ AND } (y \text{ is } B) \text{ THEN } (z \text{ is } Z) \quad (1)$$

where x, y, z represent the variables and A, B, C are the linguistic values in the universe of discourse. Here, the IF part is referred as antecedent and THEN part is the consequent [19].

The fuzzy operator (AND) is used as the rule has multiple ancestry. The used of AND operator is to obtain a single number that represent the result of the predecessor evaluation. The fuzzy set operations are used to evaluate the fuzzy rules and combine the results of the individual rules. The common used operation for AND and OR are *max* and *min*. Equations (2) and (3) shows the operation for AND and OR respectively.

$$\begin{aligned} &AND(\text{intersection}) \\ &MIN = \min\{\mu_A(x), \mu_B(x)\} \end{aligned} \quad (2)$$

$$\begin{aligned} &OR(\text{union}) \\ &MAX = \max\{\mu_A(x), \mu_B(x)\} \end{aligned} \quad (3)$$

iii. Defuzzification

This is the last step of fuzzy process whereby the defuzzification input is the aggregate output fuzzy set. The defuzzification task will transforms the linguistic terms aggregate output fuzzy set back into crisp value as for the system output. Defuzzification is performed according to the membership function of output variables. Center of Gravity (COG), Weighted Average, Mean-Max and Center of Largest Area (COA) are the common method used in defuzzification process [19].

Step 4: Decision Making (Determination of security level)

The security level can be presented as “*critical, severe, substantial, moderate and high*”. For instance, if the security level is *high*, the combination of parameters and cipher selection are good to build the KDF.

The flow chart of the Fuzzy evaluation model based on KDF algorithm is shown in Fig. 1.

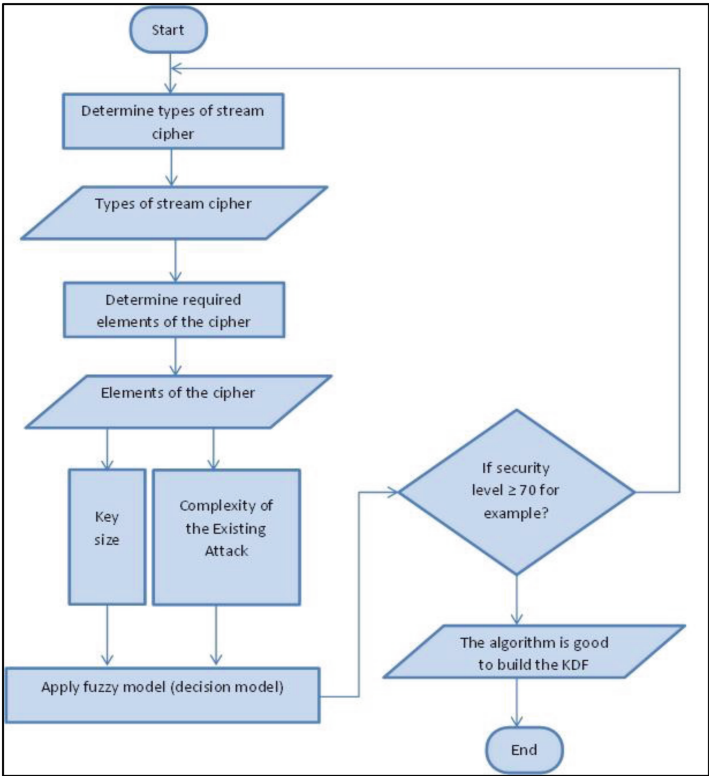


Fig. 1. Stream cipher based Fuzzy-KDF framework

4 Experiments of the Fuzzy Model

This section explains the result obtained from Fuzzy Evaluation Model to evaluate the security level respectively. The proposed solution model is implemented using the MATLAB fuzzy logic toolbox. The steps of the Fuzzy Evaluation implementation are as follows:

i. Fuzzification

In this experiment, *KeyLength* and *Attack* are used as two variable inputs provided in this fuzzy evaluation system. The output variable of the system is *SecurityLevel*. The fuzzy variables of *KeyLength* and *Attack* are tabulated in Table 2 respectively.

ii. Membership functions

The membership function editor displays the graph of membership function which is generated in the model. Membership functions is a process to generate the membership values to determine the degree to which these *KeyLength*, *Attacks* and *SecurityLevel* variables belong to each appropriate fuzzy set. The membership functions for

Table 2. Fuzzy evaluation elements

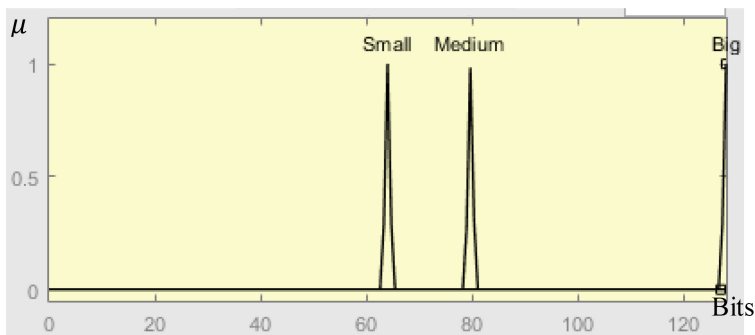
Stream Cipher	Factor (Bits)		Citation
	Key length, F_1	Attack, F_2	
A5/1-1989	64	Correlation attack • Attack needs 40 first bits from about 2^{16} (possible non-correlation frames)	[8]
Grain-Pre-2004	80	Related-Key Chosen IV • Attack recovers secret key on Grain-v1 with $2^{22.59}$ chosen IVs, $2^{26.29}$ bits keystream sequence and $2^{22.90}$ computational complexity	[9]
Trivium-Pre-2004	80	Advanced Algebraic attack • Complexity $2^{42.2}$ Trivium computations and data complexity of 2^{12}	[10]
Rabbit-2003	128	Distinguish attack based on multi cube tester • Using one iteration of next state function with complexity 2^{25}	[11]

KeyLength, *Attacks* and *SecurityLevel* are shown in Fig. 2, Fig. 3 and Fig. 4 respectively.

iii. Inferencing

In Table 3, fuzzy rules for the evaluation model are listed.

Rules in Table 4 are extracted from fuzzy inferencing process. If A5/1 is selected as the stream cipher with *KeyLength* 64, and the *attack* is Fast which less than 2^{31} , it indicates that the *SecurityLevel* states is Critical. Another example is if Trivium is chosen with *KeyLength* 80, and the attack is SlightlySlow which in between 2^{64} and 2^{79} , then *SecurityLevel* is Substantial.

**Fig. 2.** Membership function for *KeyLength*, F_1

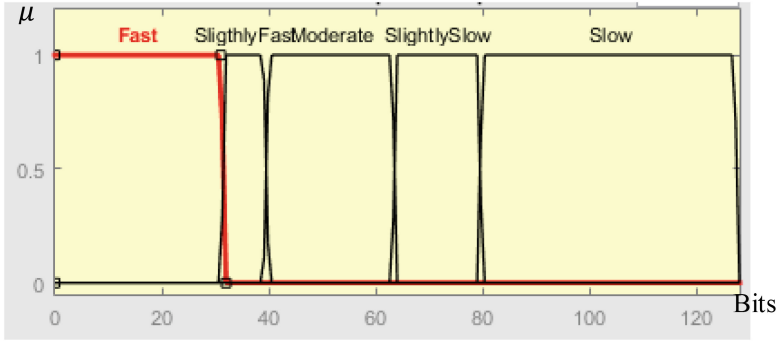


Fig. 3. Membership function for *Attack, F₂*

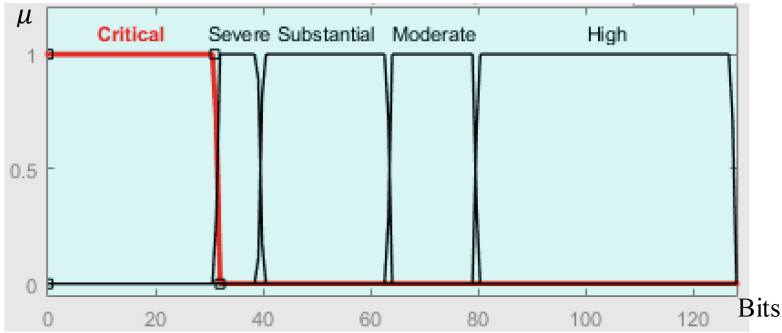


Fig. 4. Membership function for *SecurityLevel, S*

iv. Defuzzification

In this experiment, the Centroid of the Area method is used due its capability to produce an accurate result. Equation (4) shows the Centroid of the Area [19] respectively.

$$z^* = \frac{\int \mu_C(z) \cdot z dz}{\int \mu_C(z) dz} \quad (4)$$

5 Results and Discussion

This section explains the result obtained from fuzzy evaluation model respectively. The comparison was made by the result with those obtained by the fuzzy evaluation model approach.

Table 3. Fuzzy rules for evaluation model

Fuzzy rules

1. *If (keylength is A5/1) and (Attack is fast) then (security_level is critical)*
2. *If (keylength is A5/1) and (Attack is slightly_fast) then (security_level is severe)*
3. *If (keylength is A5/1) and (Attack is moderate) then (security_level is moderate)*
4. *If (keylength is Trivium_Grain) and (Attack is fast) then (security_level is critical)*
5. *If (keylength is Trivium_Grain) and (Attack is slightly_fast) then (security_level is critical)*
6. *If (keylength is Trivium_Grain) and (Attack is moderate) then (security_level is substantial)*
7. *If (keylength is Trivium_Grain) and (Attack is slightly_slow) then (security_level is moderate)*
8. *If (keylength is Rabbit) and (Attack is fast) then (security_level is critical)*
9. *If (keylength is Rabbit) and (Attack is slightly_fast) then (security_level is critical)*
10. *If (keylength is Rabbit) and (Attack is moderate) then (security_level is critical)*
11. *If (keylength is Rabbit) and (Attack is slightly_slow) then (security_level is substantial)*
12. *If (keylength is Rabbit) and (Attack is slow) then (security_level is moderate)*

Table 4. Results of the fuzzy evaluation model using real attacks

Stream cipher	Key length	Name of attack	Attack (2^n)	Security level	Level description
A5/1	64	Correlation attack [8]	16	15.4	Critical
Grain	80	Related-key chosen IV [9]	22.9	15.4	Critical
Trivium	80	Algebraic attack [10]	42.2	51.2	Substantial
Rabbit	128	Distinguish attack based on multi cube tester [11]	25	15.4	Critical

5.1 Experimental Result

The fuzzy evaluation model is developed based on key bits length and attack's complexity of the stream ciphers. Table 3 shows the security level obtained from the fuzzy evaluation.

The evaluation result as tabulated in Table 2 shows the security level values from four types of stream cipher which are A5/1, Grain, Trivium and Rabbit. From the result, we can see that the correlation attack can break A5/1 within 2^{16} bits. The attack is faster than standard attacks which are brute force and birthday attack. Hence, the attack's complexity is 2^{16} , the security level for A5/1 is classified as critical. Critical means that attack is expected imminently.

Grain has key length longer than A5/1. Generally, attack needs longer time to break the algorithm. However, the existences of related-key chosen IV influence the security level of Grain. Though that Grain has 80 bits key, but related-key chosen IV attack can break Grain within $2^{22.9}$ bits. The attack is faster than brute force and birthday attack. Therefore, the attack's complexity is $2^{22.9}$, the security level for Grain is classified as critical.

Trivium is a kind of stream cipher that also has key length 80 bits key. From the result, we can see that the advance algebraic attack can break Trivium within $2^{42.2}$ bits.

Hence, the attack’s complexity is $2^{42.2}$, the security level for Tivium is classified as substantial. Substantial means that the attacks have a strong possibility to break the algorithm (Table 4).

As for Rabbit, the key length is 128 bits. In this experimental, Rabbit has the longest key length compared to A5/1, Grain and Trivium. However, the distinguish attack based on multi cube tester can break Rabbit within 2^{25} bits. The attack is much faster than brute force and birthday attack. Hence, the attack’s complexity is 2^{25} , the security level for Rabbit is classified as critical also.

The security level for these stream ciphers can be represented in a graph. Figure 5 demonstrates the graph of security level for the stream ciphers.

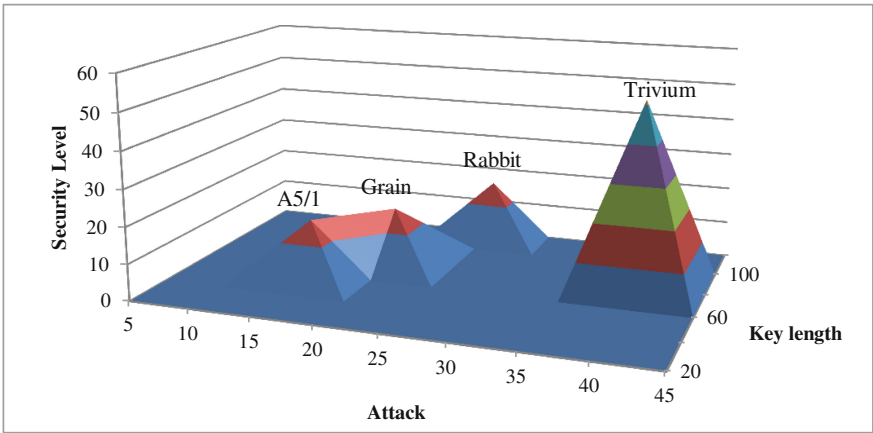


Fig. 5. Security level for stream ciphers

The graph shows the security level of A5/1, Grain, Trivium and Rabbit. Referring the result from the graph, it is shown that the security level for A5/1, Grain and Rabbit are lower than Trivium with value 15.4. This value is in the range of security level between 10 and 20. The graph also shows the higher level of security for stream cipher belongs to Trivium which is 51.2. This value is in the range of security level between 50 and 60. Typically, 128 bits of key length is used as baseline to achieve minimum security. However, it is not necessary mean that longer key length will guarantee higher security. From the graph, it shown that Rabbit with key length 128 bits can be critical in shorter time.

Hence, from this result, Trivium can be considered as more secure as compared to A5/1, Grain and Rabbit. This situation may due to Trivium’s algorithm structure itself. Trivium was designed to be both efficient and secure, though Trivium has the simplest structure among all the ciphers in eSTREAM [23]. The development of Trivium is a benchmark to explore how far a stream cipher can be simplified without sacrificing its security, speed and flexibility. From the result demonstration, it shows that the Trivium is good to build the KDF.

6 Conclusions

In this study, we demonstrate the use of fuzzy logic as an evaluation method to evaluate the degree of secureness of KDF. A5/1, Grain, Trivium and Rabbit are four stream ciphers used in this demonstration. The proposed method suggests the use of fuzzy evaluation model to evaluate the security level based on these four stream ciphers. The fuzzy evaluation model also considering the relation between the key length and attack of the stream ciphers. The property of fuzzy evaluation model is used to take consideration the existence of attack to the stream cipher. The experimental result demonstrated that the proposed method using the fuzzy logic tool can achieve security level of stream ciphers. As the result, it is shown that the fuzzy evaluation model effectively determined the security level. This fuzzy evaluation model also can help user to decide which cipher is better to build the KDF.

Acknowledgements. This research was supported by FRGS Vot 1558, RMC UTHM, and Gates IT Solution Sdn.Bhd.

References

1. Krawczyk, H.: Cryptographic extraction and key derivation: the HKDF scheme. In: Annual Cryptology Conference, pp. 631–648. Springer, Berlin, Heidelberg (2010)
2. Bakhtiari, M., Maarof, M.A.: An efficient stream cipher algorithm for data encryption. *Int. J. Comput. Sci. Issues* **8**(3) (2011)
3. Yun, J., Park, K.W., Shin, Y., Kim, H.D.: An efficient stream cipher for resistive RAM. *IEICE Electron. Express* **14**(7), 20170179–20170179 (2017)
4. Vidal, G., Baptista, M.S., Mancini, H.: A fast and light stream cipher for smartphones. *Eur. Phys. J. Spec. Top.* **223**(8), 1601–1610 (2014)
5. Chuah, C.W., Dawson, E., Simpson, L.: Key derivation function: the SCKDF scheme. In: IFIP International Information Security Conference, pp. 125–138. Springer, Berlin, Heidelberg (2013)
6. Bellare, M., Rogaway, P.: Random oracles are practical: a paradigm for designing efficient protocols. In: CCS '93, pp. 62–73. ACM Press (1993)
7. Chuah, C.W., Dawson, E., Nieto, J.M.G., Simpson, L.: A framework for security analysis of key derivation functions. In: International Conference on Information Security Practice and Experience, pp. 199–216. Springer, Berlin, Heidelberg (2012)
8. Ekdahl, P., Johansson, T.: Another attack on A5/1. *IEEE Trans. Inf. Theory* **49**(1), 284–289 (2003)
9. Lee, Y., Jeong, K., Sung, J., Hong, S.: Related-key chosen IV attacks on Grain-v1 and Grain-128. In: Australasian Conference on Information Security and Privacy, pp. 321–335. Springer, Berlin, Heidelberg (2008)
10. Quedenfeld, F.M., Wolf, C.: Advanced algebraic attack on Trivium. In: International Conference on Mathematical Aspects of Computer and Information Sciences, pp. 268–282. Springer International Publishing (2015)
11. A Distinguish attack on Rabbit Stream Cipher Based on Multiple Cube Tester. *IACR Cryptol. ePrint Archive* **780** (2013)

12. Al Maliky, S.B.S., Jawad, S.F.: Fuzzy logic-based security evaluation of stream cipher. In: *Multidisciplinary Perspectives in Cryptology and Information Security*, pp. 157–178. IGI Global (2014)
13. Azadegan, A., Porobic, L., Ghazinoory, S., Samouei, P., Kheirkhah, A.S.: Fuzzy logic in manufacturing: a review of literature and a specialized application. *Int. J. Prod. Econ.* **132**(2), 258–270 (2011)
14. Goztepe, K.: Designing fuzzy rule based expert system for cyber security. *Int. J. Inf. Secur. Sci.* **1**(1), 13–19 (2012)
15. Sallam, H.: Cyber security risk assessment using multi fuzzy inference system. *IJEIT* **4**(8), 13–19 (2015)
16. Bhusari, K.P., Kale, S.G.: Intrusion detection in wireless network using fuzzy rules. *Virus* **10** (11)
17. Mohammed, S.A., Sadkhan, S.B.: Block cipher security evaluation based on fuzzy logic. In: *2013 International Conference on Electrical, Communication, Computer, Power, and Control Engineering (ICECCPCE)*, pp. 169–173. IEEE (2013)
18. Klir, G.J., Yuan, B.: *Fuzzy Sets and Fuzzy Logic: Theory and Applications* (1996)
19. Sumathi, S., Paneerselvam, S.: *Computational Intelligence Paradigms: Theory and Application Using MATLAB*. CRC Press (2010)
20. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Man Mach. Stud.* **7**(1), 1–13 (1975)
21. Iancu, I.: *A Mamdani Type Fuzzy Logic Controller*. INTECH Open Access Publisher, Rijeka (2012)
22. Mendel, J.M.: Fuzzy logic system for engineering: a tutorial. *IEEE Trans. Fuzzy Syst.* (1995)
23. Mukherjee, P.: *An Overview of eSTREAM Ciphers*. Centre of Excellence in Cryptology, Indian Statistical Institute, Kolkata, India (2013)

A Similarity Precision for Selecting Ontology Component in an Incomplete Sentence

Fatin Nabila Rafei Heng¹(✉), Mustafa Mat Deris², and Nurlida Basir¹

¹ Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia

{fatin, nurlida}@usim.edu.my

² Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia
mmustafa@uthm.edu.my

Abstract. Most of the existing methods focus on extracting concepts and identifying the hierarchy of concepts. However, in order to provide the whole view of the domain, the non-taxonomic relationships between concepts are also needed. Most of extracting techniques for non-taxonomic relation only identify concepts and relations in a complete sentence. However, the domain texts may not be properly presented as some sentences in domain text have missing or unsure term of concepts. This paper proposes a technique to overcome the issue of missing concepts in incomplete sentence. The proposed technique is based on the similarity precision for selecting missing concept in incomplete sentence. The approach has been tested with Science corpus. The experiment results were compared with the results that have been evaluated by the domain experts manually. The result shows that the proposed method has increased the relationships of domain texts thus providing better results compared to several existing method.

Keywords: Ontology · Non-taxonomic relation · Similarity precision

1 Introduction

Ontology is a representation of knowledge with a pair of concepts and relationships within a particular domain. Ontology have been used in many areas such as information retrieval and semantic web. According to [16], an ontology can be generated from different sources such as set of documents, existing ontologies, or from a combination of both sources. It can also be generated from scratch. The general process of ontology construction from texts consists of extracting ontological components (such as concept, relation) from text, which then creates an ontology. For example, [8] proposed ontology construction using textual data in order to reduce the time and effort required for manual ontology construction.

The relations between concepts are known as taxonomic and non-taxonomic relations. A relationship between concepts is needed to explain more about the domain. A taxonomic relation represents a hierarchy of concepts, i.e. is-a relation. For example, mother is-a person. A non-taxonomic relation represents relation other than an “is-a” relation that exists in texts. The extraction of non-taxonomic relationships is considered

as one of the most challenging and important tasks [11, 12]. Many existing techniques [11–13] focused on extracting relationships between two concepts focus on terms that appear as subject and object in a single sentence. Villaverde et al. [17] proposed a technique that identifies nouns as concepts and verbs that hold a place between two nouns that occur in a single sentence, as a relationship. All nouns and verbs are identified by using parts-of-speech (POS) taggers that were applied to each sentence from the documents collection to fulfill the pattern: $\langle \text{term} \rangle \langle \text{verb} \rangle \langle \text{term} \rangle$, where the terms are identified nouns that also exist in ontology concepts. In both works, if the concepts do not exist in ontology concepts, then the relation is not identified. Punuru and Chen [11] and Serra and Girardi [13] also used the predicate as a relation between two concepts. In contrast to [17], this work do not refer to ontology concepts to find the relevant concepts. Punuru and Chen [11] proposed the Subject-Verb-Object (SVO) Triples method to identify non-taxonomic relations between two concepts, where the concepts must appear as the subject and the object of a sentence. They used MINIPAR dependency parser to determine the appearance of concepts. Then, the verb that occurred together with the concept pair was identified. Serra and Girardi [13] proposed a technique that used an NLP approach and data mining technique to identify potential non-taxonomic relationships from textual sources. Serra et al. [14] proposed a semi-automatic method called PARNT to extract non-taxonomic relations from texts. Ribeiro [12] proposed a framework that used A Nearly New Information Extraction System (ANNIE), Stanford dependency parser and association technique, to enrich ontology relations. The framework extracted relations between two concepts, where the concepts must appear as the subject and the object of a sentence, as similar to [11]. This work extracted non-taxonomic relationships of the domain of tennis sport collected from various sources. Muzaffar et al. [7] proposed a hybrid framework that used four features such as the bag of word model, NLP approach, Lexical and semantic based UMLS, to extract relation from biomedical datasets collected from MEDLINE database. This work extracted all verb phrases that occur between treatments and disease entities in the sentence. All these techniques only able to extract non-taxonomic relations between two concepts i.e. subject and object that appear in the same sentence. However, the existing methods do not cover if the subject or the object of a sentence is “unclear” or “missing”. Therefore the domain texts may improperly presented, as some concepts and relations cannot be identified. Hence, this paper proposed a technique that calculates the similarity precision for suggesting the relevant concept for missing concept in incomplete sentence.

The paper is organized as follows: Sect. 2 introduces the basic definition of formula. Section 3, a proposed approach is presented. Section 4 presents the experiments. Finally, Sect. 4 presents the conclusions and future work.

2 Definition of Formula

The basic concept of complete system as described in [9] will be explained as follow:

An information systems [9] is a quadruple, $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attribute, where $A = C \cup \{d\}$, C is a set of condition attributes and d the decision attribute, such that

$f : U \times A \rightarrow V$ for any $a \in A$, where V_a is called domain of an attribute a . If U contains at least one object with an unknown or missing value, the S is called as an incomplete information system [4, 5]. The unknown or missing value is denoted as “*” in incomplete information system.

From this definition, U represents a set of sentences in texts. A is a set of terms, where C is a term (i.e. concept) that appear as subject or object in texts and d is a term (i.e. predicate) that appears together with the subject and object in a sentence. If the sentence is incomplete i.e., does not contain object or subject, the missing value will be denoted as “*”.

In this work, we use voting machine text as a case study. We extract all sentences and present it in the information table as shown in Table 1. In Table 2, u_1, u_2, \dots, u_{18} represent sentences. C is a set of attributes that consist of subject, object and predicate of each sentence.

Table 1. A part of sentences in voting machine dataset

SentenceID	Subject	Object	Predicate
u_1	concern	reliability	grow
u_2	*	machine	supply
u_3	machine	paper	produce
u_4	voter	*	check
u_5	machine	record	produce
u_6	*	Rule	draft
u_7	company	record	produce
u_8	voter	*	trust
u_9	*	software	trust
u_{10}	machine	paper	calculate
u_{11}	machine	paper	produce
u_{12}	company	security	increase
u_{13}	company	machine	provide
u_{14}	*	machine	produce
u_{15}	official	election	mention
u_{16}	company	*	produce
u_{17}	machine	*	produce
u_{18}	voter	machine	trust

In non-taxonomic relation extraction, most techniques in [7, 11–13, 17] extract the predicate that appear together with the subject and object in a single sentence. For example, based on Table 1, only nine sentences ($u_1, u_3, u_5, u_{10}, u_{11}, u_{12}, u_{13}, u_{15}$ and u_{18}), which have predicate that co-occur with the subject and object in the same sentence. The remaining sentences (i.e. $u_2, u_4, u_6, u_7, u_8, u_9, u_{14}, u_{16}, u_{17}$) are considered as incomplete sentences. Currently, the existing techniques can only extract relations of sentences that have proper pattern i.e., subject-predicate-object (S-P-O). However, these do not properly represent the domain text, as the incomplete sentences

are not extracted due to the missing or unsure value on the subject or object in the sentence. Therefore, in this work, we propose an approach to overcome the issue of missing potential relations and suggest the most dominant term to replace the missing concepts.

3 Similarity Extraction Method

In this work, we will present the proposed method, which we refer to as similarity extraction method (SEM). The proposed method will identify the missing or unsure value (i.e. subject or object of a sentence that does not exist in a sentence) by giving a suggestion to replace the value based on the highest similarity value been calculated. The suggested value (i.e., object or subject) can be considered as the most potential value of all. Figure 1 shows the pseudo-code of the SEM algorithm.

Algorithm: proposed method	
Input: Domain texts documents	
Output: subject-object-predicate	
Begin	
Step 1.	Identify the subject, object and predicate of sentences by using pre-processing.
Step 2.	Compute the similarity precision of attribute.
Step 3.	Compute the most dominant term to replace the missing value of subject or object.
Step 4.	Compute and select the most relevance relationship for the subject-object pair.
End	

Fig. 1. The SEM algorithm

Details of the pseudo are described below.

3.1 Identify the Subject, Object, and Predicate of Sentences by Using Pre-processing

In this section, three main text-preprocessing steps i.e. part-of-speech (POS) tagging, stop word removal, and morphological analysis are used to extract terms from texts.

Part-of-speech (POS) tagging is to assign parts of speech (tag) to each word such as noun, verb, adverbs, adjective, etc. Hence, the noun can be determined by selecting the word that has the NN (i.e. Noun, singular), NNS (i.e. Noun, plural) or NNP (i.e. Proper noun) tag. Next, Stop word removal, i.e. Porter’s stop word list [10] is applied to remove the frequent words that do not give any information about the domain corpus, such as the, and, was, a, of, with, to, is, that, etc. Then, the morphological analysis is used to tag each remaining noun to obtain its common base form. Once the process is done, a list of definite terms is identified. For example, “papers” to “paper”.

Statistical analysis is used to determine the relevant term as a concept in domain text. Each term is produced during a pre-processing step and calculated using a statistical measurement such as term frequency, to determine its relevancy to the domain. Then, the dependency pairs between these terms (i.e., grammatical relation between subject and/or object with the predicate) are identified using the Minipar shallow parser [6]. All identified terms and their relations are presented in the information table. In this table, all incomplete sentences with missing object or subject are highlighted by “*”. This “*” is known as a “unclear” value of the sentences.

3.2 Compute the Similarity Precision of Attribute

The second step of the technique is to obtain the set of similarity attributes between sentences by using similarity precision [2] in formula (1). Set of attributes with $\text{simP} \geq 0.667$ thresholds are selected (at least two attributes are similar).

Definition 1 Let $P_B(x) = \{b | b \in B \wedge b(x) \neq *\}$, the similarity precision simP , is defined as

$$\text{simP}(x, y) = \frac{|P_B(x) \cap P_B(y)|}{|C|}, \quad (1)$$

where

- $|C|$ represents the cardinality of the set.
- B represent the set of attributes.

In this step, WordNet will be used to identify the synonymous meaning between predicates. Examples 1 and 2 describe the process.

Example 1 For $u_{17} = \{\text{machine}, *, \text{produce}\}$, we will find set of similarity attributes (subject, object, predicate) with $\text{simP} \geq 0.667$, we can get the following result

- $u_3 = \{\text{machine}, \text{paper}, \text{produce}\}$, in u_3 , two attributes are similar with attributes in u_{17} , i.e. machine and produce. Therefore,
 $P(u_{17}) \cap P(u_3) = \{2\}$ or $\text{simP}(u_{17}, u_3) = 2/3 = 0.66$
- $u_5 = \{\text{machine}, \text{record}, \text{produce}\}$,
 $P(u_{17}) \cap P(u_5) = \{2\}$ or $\text{simP}(u_{17}, u_5) = 2/3 = 0.66$
- $u_{11} = \{\text{machine}, \text{paper}, \text{produce}\}$,
 $P(u_{17}) \cap P(u_{11}) = \{2\}$ or $\text{simP}(u_{17}, u_{11}) = 2/3 = 0.66$.

Example 2 For $u_2 = \{*, \text{machine}, \text{supply}\}$, we will find set of similarity attributes (subject, object, predicate) with $\text{simP} > 0.5$, we can get the following result

- $u_{13} = \{\text{company}, \text{machine}, \text{provide}\}$
 $P(u_2) \cap P(u_{13}) = \{2\}$ or $\text{simP}(u_2, u_{13}) = 2/3 = 0.66$

In this example, by using WordNet, the predicate supply is synonymous with predicate provide.

3.3 Compute the Most Dominant Term to Replace the Missing Value of Subject or Object

In step 3, the most dominant term is calculated to determine which subject/object is selected to replace the missing value of subject/object [2]. The most dominant is defined as below.

Definition 2 The most dominant of value X (i.e., subject/object) in set of similarity attributes denoted as $MD_C(x)$. This can be defined as

$$MD_C(x) = (|x|/|Cs|, x \in C), \quad (2)$$

In step 3, the most dominant term is calculated to determine which subject/object is selected to replace the missing value of subject/object. The most dominant is defined as below.

where

- $(| |)$ represents the cardinality of the sets,
- Cs represent the cardinality of the set of similarity attributes.

Example 3 shows the results of Step 3.

Example 3 Based on the result produced in Example 1, the most dominant of object is calculated as follows

Table 2. Incomplete information table after Step 1 and 2

SentenceID	Subject	Object	Predicate
u ₁	concern	reliability	grow
u ₂	* company	machine	supply
u ₃	machine	paper	produce
u ₄	voter	* –	check
u ₅	machine	record	produce
u ₆	* –	rule	draft
u ₇	company	record	produce
u ₈	voter	* machine	trust
u ₉	* –	software	trust
u ₁₀	machine	paper	calculate
u ₁₁	machine	paper	produce
u ₁₂	company	security	increase
u ₁₃	company	machine	provide
u ₁₄	* –	machine	produce
u ₁₅	official	election	mention
u ₁₆	company	* record	produce
u ₁₇	machine	* paper	produce
u ₁₈	voter	machine	trust

- $u_3 = \{\text{machine, paper, produce}\},$
- $u_5 = \{\text{machine, record, produce}\},$
- $u_{11} = \{\text{machine, paper, produce}\},$

$C_s = 3, MD(\text{paper}) = 2/3, MD(\text{record}) = 1/3.$

Thus, paper is selected to replace the missing value of object in u_{17} .

The information table after the replacement in Step 1 and 2 is shown in Table 2.

3.4 Compute and Select the Most Relevance Relationship for the Subject-Object Pair

In step 4, we compute the association between subject-object pair with predicate by using association rules [1]. The support for an association rule $X \rightarrow Y$ is the number of sentences in texts that contain $(X \cup Y)$. The formula is as follows:

$$\text{supp}(X \rightarrow Y) = |(X \cup Y)|/|U|$$

The confidence of an association, denoted by $\text{conf}(X \rightarrow Y)$, is the ratio of the number of sentences that contain $X \cup Y$ to the number of transaction that contain X , is defined as follows:

$$\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$$

In our work, X is referring to subject-object pair and Y is referring to predicate.

Example 4 From Table 2, the confidence values for each subject-object pair with predicate are as follow:

$\text{conf}((\text{concern, reliability}) \rightarrow \text{grow}) = 1/1 = 1, \text{conf}((\text{company, machine}) \rightarrow \text{supply}) = 1/2 = 0.5, \text{conf}((\text{machine, paper}) \rightarrow \text{produce}) = 3/4 = 0.75, \text{conf}((\text{machine, paper}) \rightarrow \text{calculate}) = 1/4 = 0.25, \text{conf}(\text{machine, record} \rightarrow \text{produce}) = 1/1 = 1, \text{conf}(\text{company, record} \rightarrow \text{produce}) = 2/2 = 1, \text{conf}(\text{voter, machine} \rightarrow \text{trust}) = 2/2 = 1, \text{conf}((\text{company, security}) \rightarrow \text{increase}) = 1/1 = 1, \text{conf}((\text{company, machine}) \rightarrow \text{provide}) = 1/2 = 0.5, \text{conf}((\text{official, election}) \rightarrow \text{mention}) = 1/1 = 1.$

In our work, predicate with the highest degree of confidence is considered as a suitable relation for the subject-object pair. However, if there exists more than one predicates that have the highest and similar degree of confidence, both predicate can be used.

4 Experiment

The goal of the proposed method is to improve the knowledge of domain texts. Since this paper aims to extract non-taxonomic relations in domain texts as many as possible, the information retrieval metrics of precision and recall can be used. The precision and recall are used to analyze the quality of the knowledge extracted. Recall is used to measure the quality of correctly extracted relations in domain texts. The percentage of quality of the correct extracted relation is calculated as:

$$\text{Recall} = \frac{\# \text{ of correct extracted relations}}{\# \text{ of relevance relations}} \times 100$$

Precision is used to calculate the accuracy of correctly extracted relations to the total number of relations extracted. The percentage of accuracy for correct extracted relations to the total number of relations extracted is calculated as shown below:

$$\text{Precision} = \frac{\# \text{ of correct extracted relations}}{\# \text{ of extracted relations}} \times 100$$

For this experiment, a set of science dataset was used. The science domain texts were collected from Wikipedia website and the Understanding Science website that consists of 24 documents with 10,480 words describing science. For evaluation, a prototype was developed using Java and JavaScript. The relations from the Science corpus that are extracted by the SEM, and two methods [11, 13], are compared with the results produced by domain experts. In this evaluation, the experts analyzed the given domain texts and manually identified all the relevant relations to be used as benchmarks for the system. Then, the experiments' results are analyzed using recall metric to measure the completeness and sufficiency of the extracted relations to represent the domain texts. The relations from the Science corpus that are extracted are shown in Fig. 2. In Fig. 2, the sentence ID highlighted in green indicates the complete sentences that have the subject, object and predicate co-occur in the same sentence. While the remaining lists are irregular sentences in raw texts, which have an uncertain term (i.e. subject/object). Based on the proposed method, it extracted the relations and determined the possible subject/object to replace the unclear term by using synonyms of predicate and probability.

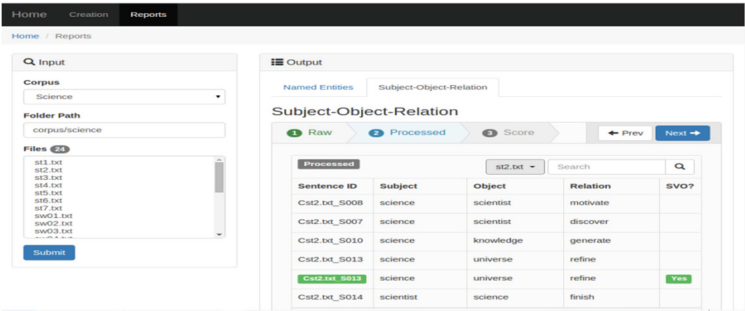


Fig. 2. The extracted relations for science corpus

Table 3 shows the results for the science domain texts. In Table 3, Expert 1 and Expert 2 have identified 74 and 102 relevant relations. Even though both are experts in the Science field, the varying results are because the terminology structuring itself varies from one terminologist to another [3]. They might define the term that may have

Table 3. The recall value for science domain texts

Method	# of extracted relations	Expert 1: 74 relevance relations			Expert 2: 102 relevant relations		
		# of correct relations	P (%)	R (%)	# of correct relations	P (%)	R (%)
SEM	122	65	53.28	87.84	115	94.26	89.15
SVO Method [11]	67	31	46.27	41.89	60	89.55	58.82
Serra Method [15]	75	36	48.00	48.65	65	86.67	63.72

P precision, *R* recall

similar meaning differently. For example, Expert 2 accepted more general terms (i.e. journalist, life, hypothesis, checklist, chemist, etc.) while expert 1 only considered specific and frequent terms that exists in the texts. In addition, inconsistencies of an expert are another issue. People are inconsistent in their judgments especially when dealing with large amount of data. Based on Expert 1, the recall value is 87.84% for SEM, 41.89% for the SVO method [11], and 48.65% for the method in [13]. Meanwhile, based on Expert 2, the recall value of SEM is 79.41%, which is higher than the SVO method [11] and the [13] method.

5 Conclusions

In conclusion, we have proposed a method that focused on extracting potential relations using similarity precision and can be used to solve the issue of an incomplete sentence in texts. Based on the experiment results, we can conclude that the proposed method managed to extract correct relations in domain texts and thus improve the knowledge for a domain text and help the ontology engineers to label the relations between concepts appropriately.

Acknowledgements. This work was supported under Grant Universiti Sains Islam Malaysia (USIM). Grants: PPP/USG-0116/FST/30/11616.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. *ACM SIGMOD Record* **22**(2), 207–216 (1993)
2. Deris, M.M., Abdullah, Z., Mamat, R., Yuan, Y.: A new limited tolerance relation for attribute selection in incomplete information systems. In: 2015 12th International

- Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 964–970. IEEE, Aug 2015
3. Hamon, T., Nazarenko, A.: Detection of synonymy links between terms: experiment and results. *Recent Adv. Comput. Terminol.* **2**, 185–208, page 13 (2001)
 4. Kryszkiewicz, M.: Rough set approach to incomplete information systems. *Inf. Sci.* **112**(1–4), 39–49 (1998)
 5. Kryszkiewicz, M.: Rules in incomplete information systems. *Inf. Sci.* **113**(3–4), 271–292
 6. Lin, D.: Dependency-based evaluation of MINIPAR. In: *Treebanks*, pp. 317–329. Springer, Netherlands (2003)
 7. Muzaffar, A.W., Azam, F., Qamar, U.: A relation extraction framework for biomedical text using hybrid feature set. *Comput. Math. Methods Med.* (2015)
 8. Navigli, R., Velardi, P., Gangemi, A.: Ontology learning and its application to automated terminology translation. *Intell. Syst. IEEE* **18**(1), 22–31 (2003)
 9. Pawlak, Z.: Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356
 10. Porter, M.F.: An algorithm for suffix stripping. *Program: Electron. Libr. Inf. Syst.* **14**(3), 130–137 (1980)
 11. Punuru, J., Chen, J.: Learning non-taxonomical semantic relations from domain texts. *J. Intell. Inf. Syst.* 191–207 (2012)
 12. Ribeiro: Extraction of non-taxonomic relations from texts to enrich a basic ontology (2014)
 13. Serra, I., Girardi, R.: A process for extracting non-taxonomic relationships of ontologies from text. *Intell. Inf. Manag.* **3**(4) (2011)
 14. Serra, I., Girardi, R., Novais, P.: PARNT: a statistic based approach to extract non-taxonomic relationships of ontologies from text. In: *2013 Tenth International Conference on Information Technology: New Generations (ITNG)*, pp. 561–566. IEEE (2013)
 15. Stewart, T.R.: Improving reliability of judgmental forecasts. In: *Principles of Forecasting*, pp. 81–106. Springer US (2001)
 16. Uschold, M.: Creating, integrating and maintaining local and global ontologies. In: *Proceedings of the First Workshop on Ontology Learning (OL-2000) in Conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000)* (2000)
 17. Villaverde, J., Persson, A., Godoy, D., Amandi, A.: Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. *Expert Syst. Appl.* **36**(7), 10288–10294 (2009)

Mitigating Manual Final Year Project (FYP) Management to Be Centralized Electronically

Noryusliza Abdullah¹(✉), Shahril Nazim Mohamed Salleh²,
Hairulnizam Mahdin¹, Rozanawati Darman¹, Basil David Daniel³,
and Ely Salwana Mat Surin⁴

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
{yusliza, hairuln, zana}@uthm.edu.my

² Information Technology Centre, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
shahril@uthm.edu.my

³ Faculty of Civil and Environmental Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
basil@uthm.edu.my

⁴ Institute of Visual Informatics, University Kebangsaan Malaysia, 43600 Bangi, Malaysia
elysalwana@ukm.edu.my

Abstract. Academic structure for bachelor degree in Malaysia universities consists of Final Year Project (FYP). Unfortunately, the management of the FYP is implemented manually or not fully utilizing an integrated computer system. It results in slower implementation of the FYP and inaccurate marks for the project's evaluation. This article proposes a new approach for solving the above stated problems named FYP Management System. The approach is tested using real dataset of University Tun Hussein Onn Malaysia's (UTHM) students and lecturers. As a user satisfaction measure, 92.3% users have acknowledge FYP Management System has assisted them in managing student's project while 95.6% users have recognized this approach has accelerated FYP grading process.

Keywords: Final Year Project (FYP) · Project management system effectiveness

1 Introduction

Final Year Project (FYP) is a requirement to complete a bachelor degree in Malaysia universities [1]. The subject is conducted in two last semesters carrying credit ranging from 6 to 8 credits depends on the program and universities. Normally, FYP administration is consists of registration, administration and evaluation. Most institutions are implementing FYP manually while some of them are using computerized system that are not fully integrated. For example, Faculty of Mechanical and Manufacturing

Engineering, UTHM has done Final Year Project System [2] using data that are non-centralized within UTHM. Those data may be inaccurate since they are not monitored by any party.

The drawbacks results in slower implementation of FYP especially for the project's evaluation. Some method is not consider evaluation from more than one examiner [3] but others are adopting that approach since it is essential to get comprehensive judgment. Manual system that lack of automatic average calculation on examiner's mark will give inaccurate result. It also time consuming to calculate the marks manually. Based on the stated problems, we proposed a computerized and centralized system that encapsulates all aspects of FYP administration, called FYP Management System. The usage of an automatic system will helps the administration of FYP, accelerate processes and reduce the error rate in evaluation.

Three target users for this application are coordinator, supervisor and examiner. The coordinator manage and administer data that used in the system while supervisor and examiner are responsible to evaluate the projects. Besides, supervisor is also monitor detail information regarding their student's FYP. Marks given by the supervisor and examiner are entered based on weightage and scoring scale that have been fully implemented on the system. This application is expected to reduce paper and time usage, prepare a hassle-free implementation on FYP and produce better outcome in terms of student's marks and grade.

The remainder of this paper is organized as follows. Section 2 discusses the related works. In this section, it is divided into sub sections that explain briefly on FYP management system in Faculty of Mechanical and Manufacturing Engineering (FKMP)—University Tun Hussein Onn Malaysia (UTHM) and Faculty of Computer Science and Information Technology (FSKTM)—University of Malaya (UM). Section 3 describes implementation of FYP Management System. The results and discussion of the proposed approach is explained in Sect. 4. Finally, Sect. 5 provides the conclusion and future work.

2 Related Works

This section discusses about the related works on managing project and student's system. For project management system, current researches on it are studied while for student's project, two systems are compared with the proposed approach. They are systems used for FKMP, UTHM and FSKTM, UM.

2.1 Project Management System

Research on project management is actively conducted. The purpose is to obtain data systematically and organize them in a structured manner. It is also used to transform the data into information that is useful in decision making. Nagar and Anouar [4] have proposed a computer implemented method to handle program, project and scheduling management. This approach concentrated on nested hierarchy of projects consists of multiple tiers structure. It also concentrated on weightage, aggregating the progress percentage and propagating the information.

On the other hand, Braglia and Frosolini [5] have highlighted an integrated method with extended enterprise in their approach. This research provided pertinent information and collaborative tools in web-based collaborative technologies that enhances the understanding of processes and helps detecting issues. In terms of student’s project management system, even though there are many developed applications, but not many are recorded. Hence, the curve of improvement is not clearly stated. Numerous researches on student information management system have been conducted. It covers student details, academic related reports, college details, course details, curriculum, batch details, placement details and other resource related details [6]. However, the module on Final Year Project is barely included in the system. The examples of FYP modules that have been developed in some institutions are explained in the next subsections. They are from FKMP, UTHM and FSKTM, UM.

2.2 FKMP, UTHM Final Year Project System

FKMP UTHM has their own FYP system to manage and administer FYP for their students [2]. The system consists of two main modules that are pre-FYP and FYP.

In pre-FYP module, supervisor and coordinator can suggest, approve and assign titles to the selected students. While in FYP module, supervisors and examiners (panels) are giving marks for the evaluation as shown in Fig. 1. Marks are different from one scheme (assessment) to another. The drawback of the differentiation is lecturers are tend to give incorrect marks due to the unstandardized scheme.

▼ Evaluation of PSM I report (50 %)

Scheme R1 *

0

1

2

3

4

5

Background [max 2.5%]

Scheme R2 *

Problem statement, objective development and research scope [max 3%]

Scheme R3 *

Literature review & theories [max 15%]

Scheme R4 *

Research methodology and planning (Gantt chart) [max 15%]

Scheme R5 *

Preliminary Work [max 5%]

Scheme R6 *

Grammar and usage of language [max 5%]

Scheme R7 *

0

1

2

3

4

5

References [max 2.5%]

▼ Evaluation of Log Book (15 %)

Scheme B1 *

0

1

2

3

4

5

Frequency of data collection [max 5%]

Scheme B2 *

0

1

2

3

4

5

Relevance of information [max 5%]

Scheme B3 *

0

1

2

3

4

5

Fig. 1. FYP evaluation in FKMP’s system

2.3 FSKTM, UM Academic Project Portal

Final Year Project or Academic Project in Faculty of Computer Science and Information Technology, UM is administered using a portal [7]. The portal includes assessment criteria, report format, announcement, general information, rules and regulation, list of titles and pro forma.

Management of the project for Supervisor and Examiner is done using several modules. They are Announcement, Project title, Assign student and Assign mark. In UM, project title and related information are updated by supervisors. They will then assign student for each project. Marks are given two times which are on week seven and 14 using. The marks are based on weightage for every item in the evaluation.

2.4 Comparison Between FKMP, FSKTM and the Proposed FYP System

Based on the criteria discussed in the previous sub-sections, there are several important features embedded in the proposed FYP System. The main feature is it using centralized data that obtained from organization's main database. These data are monitored regularly. Final marks are also calculated automatically based on the marks given by the supervisor and average marks from the examiners using specific weightage set by the coordinator. Supervisor and examiners are allowed to give comments and they are visible to each other to ensure students make improvement on any suggestion. Furthermore, the system is designed to give clear information to assist coordinator in detecting incomplete data. It also give notification through email regarding project's status.

3 Implementation

There is room for improvement for both systems explained in the previous section. Hence, Final Year Project Management System is proposed. It is implemented on JavaScript, HTML and PowerBuilder. Oracle 12.0 is chosen as its database. The application consists of two main modules, which are Registration and Evaluation. The Registration module is divided in sub-modules including Title registration, Assign supervisor, Supervisor approval and Notification. Flow chart on this module is shown in Fig. 2.

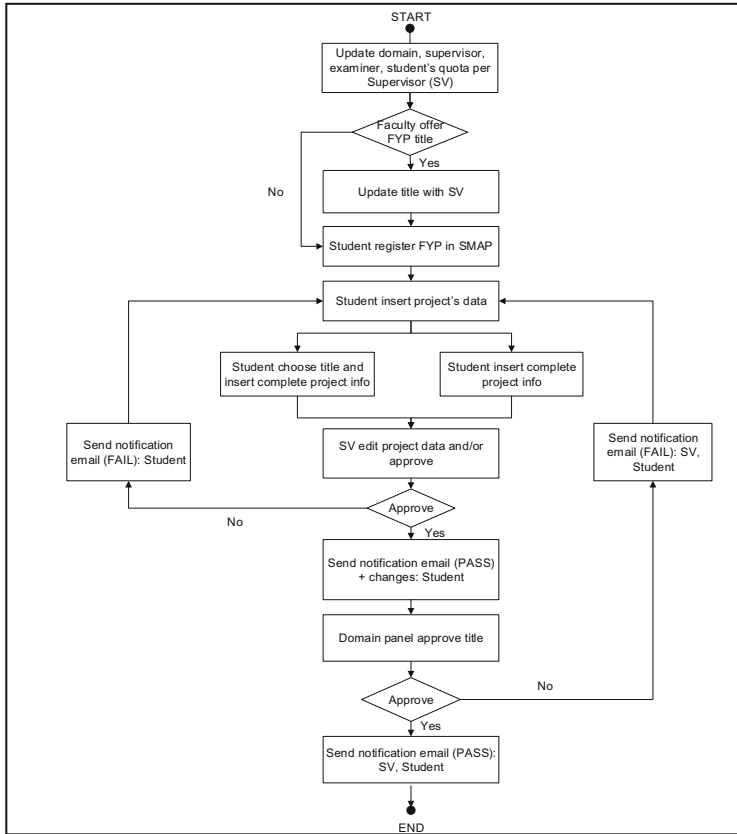


Fig. 2. Flow chart for registration module

In Title registration, it can be done by the student or supervisor using two different platform. Student's platform is using an application based on JavaScript and HTML while supervisor platform is using PowerBuilder. Supervisor is assigned in the second sub-module. Once supervisor is assigned, they can insert or change FYP data of their students. Approval of the title is also done by the supervisor. The supervision page will list all of their supervisees that will be used in the next module, Evaluation. Student's project info is shown in Fig. 3. Notification module will create email to notify students regarding approval or changes in project info. Second module of this application is evaluation with some sub-modules. They are Updating examiner, Weightage determination and Marks entry. The pseudo-code for this module is shown in Fig. 4.

MODUL PENYELIA PROJEK SARJANA MUDA

Input
Sesi / Semester: 20162017/2

Senarai Pelajar PSM Selaian Sesi 20162017/2 [Bilangan Rekod : 6]

BIL	NOMATRIK	NAMA PELAJAJAR	TAHUN / PROGRAM / FAKULTI	STATUS
1	CI140131	ARNI NORSYAHRAH BINTI MOHAMAD AYOB TAJUK : Sistem Pengurusan Surimas Catering	4 BW FSKTM	Lulus Ketua Panel
2	AI140228	LIM YU HAN TAJUK : Food Advisor	3 BW FSKTM	Lulus Ketua Panel
3	AI140113	NORHDAYU BINTI IWAN MOHD SAM TAJUK : Sistem e-perputakaan SMK Bander Baru Serting	3 BW FSKTM	Edit
4	AI140107	NUR SHAFATIN BINTI ABDUL AZIZ TAJUK : Sistem Senarai Semak Barang Runcit (grocery Checklist) di Part Raja, Batu Pahat	3 BW FSKTM	Edit
5	AI140144	NOOR HAFIZAH BINTI ZAINAL ABDIN	3 BW FSKTM	Edit

Info Pelajar
NO MATRIK: CI140131
NAMA: ARNI NORSYAHRAH BINTI MOHAMAD AYOB

Maklumat Permohonan Projek Sarjana Muda

BIDANG: Tiada Pilihan
KATEGORI: Tiada Pilihan
TAJUK: Sistem Pengurusan Surimas Catering

Senarai Tajuk: Komen

Status Permohonan
T.KEMAS: 13/02/2017 12:09:26
STATUS: KL - Lulus Ketua Panel
Kemaskini

ULASAN: Log

SINOPSIS: Surimas Catering adalah syarikat yang telah beroperasi selama enam tahun dan dimiliki oleh Md. Suri bin Rusman. Pada peringkat awal, Surimas Catering telah mendapat kontrak bagi mengusahakan kantin sekolah di Puchong Perdana tetapi tetap fokus pada perkhidmatan catering dan sajian. Surimas Catering menyediakan perkhidmatan untuk pelbagai jenis majlis seperti majlis perkahwinan, majlis-majlis korporat dan banyak lagi. Oleh kerana mendapat banyak permintaan daripada pelanggan, Surimas Catering memerlukan satu sistem

OBJEKTIF: Matlamat utama projek ini adalah untuk membangunkan Sistem Pengurusan Surimas Catering. Oleh yang demikian untuk mencapai matlamat yang didasarkan, beberapa objektif telah dibangunkan untuk memastikan objektif tercapai. Objektif-objektif tersebut ialah:
i. Untuk menaik taraf sistem yang sedia ada (sistem manual) iaitu menggunakan buku log kepada sistem komputer supaya lebih mudah sistematis.

SKOP: Sistem Pengurusan Surimas Catering ini dibangunkan untuk membantu pihak pengurusan Surimas Catering bagi menguruskan maklumat tempahan atau pesanan pelanggan-pelanggan untuk pelbagai majlis supaya lebih teratur dan sistematis. Selain itu, untuk memudahkan pihak pengurusan mengemaskini maklumat para pekerja supaya mudah untuk menyediakan butiran-butiran pekerja seperti gaji dan sebagainya.

Fig. 3. Information on student's FYP

```

START
Update examiner for each student
Insert assessment list, a
  For a > 0 do
    Insert weightage of assessment, w
  End for
  If user = supervisor then
    For a <> 0 do
      Insert supervisor marks, s1
      s1 * w
    End for
  Else if user = examiner then
    For a <> 0 do
      Insert examiner marks: e1, e2...en
      e1, e2...en * w
    End for
  Else
    View data
  End
Calculate marks
s1 + (e1, e2...en / n)
Monitor FYP marks
  If edit marks is needed then
    Update new marks
  Else
    Maintain marks
  End
Submit marks to main system
  If date > 0 then
    Submit marks
  Else
    Print "System closed"
  End if
Display grade for student
END

```

Fig. 4. Pseudo-code for evaluation module

The Updating Examiner module will ensure the right panels are assigned to each student. This is done by Coordinator. In order to assist the Coordinator, the proposed approach provides an indicator that shows number of Supervisor and Examiners. It will prevent cases of overlooked list of the Supervisor and Examiner.

Supervisors and Examiners are given a list of students with a set of evaluation form using Marks entry module. The marks are based on Weightage determination module set by Coordinator. Marks entered are based on 1–5 scales and the scales are multiplied with the weightage that have been set-up previously. Figure 5 shows evaluation form with embedded weightage.

Fig. 5. Evaluation

3.1 Dataset

This system uses real dataset of UTHM. The data is integrated from two main databases, Staff and Student. Development of the database for this research have to be analyzed carefully to ensure reliability of the data [7].

There are 3149 students from 8 faculties, who are each supervised by a supervisor and examined by at least two examiners. Managing student's project in FYP Management System involves several modules, as stated in the previous section. The division of student by faculty is listed in Table 1.

Table 1. Numbers of students from eight faculties in UTHM

Faculty	Student
Faculty of Computer Science and Information Technology	360
Faculty of Civil and Environmental Engineering	510
Faculty of Electrical and Electronic Engineering	508
Faculty of Mechanical and Manufacturing Engineering	417
Faculty of Technology Management and Business	429
Faculty of Technical and Vocational Education	441
Faculty of Science, Technology and Human Development	195
Faculty of Engineering Technology	289
Total	3149

Data as is 16 March 2017

4 Results and Discussion

The functionality of all modules are tested in terms of user acceptance, capability to implement task, time usage and error-free. Modules that are undergo testing phases are listed below:

- a. Title registration
- b. Assign supervisor
- c. Supervisor approval
- d. Notification
- e. Updating examiner
- f. Weightage determination
- g. Marks entry

To ensure the acceptance of this application, user's satisfaction is the significant factor to be considered. In line with it, a questionnaire was distributed to 91 active users of the proposed system. The questionnaire is divided into two categories of satisfaction:

- i. Assist in the execution of FYP
- ii. Accelerate FYP grading process

4.1 Assist in the Execution of FYP

This subsection explains user's satisfaction from various faculty in graph form. Respondents were asked for their level of satisfaction regarding the proposed approach assisting them in managing and evaluating their students in FYP. The result is shown in Fig. 6. From the graph, 41.8% users totally agree that FYP Management System has assisted them in managing FYP, 50.5% chose 'agree' while a small number of them, which is 7.7%, disagree with the system.

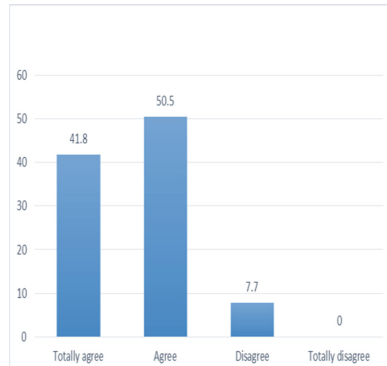


Fig. 6. Graph assist in execution of FYP

4.2 Accelerate FYP Grading Process

On the other hand, most of the users with 54.9% totally agree that FYP Management System has accelerated FYP grading process, 40.7% agree and 4.4% users disagree. The graph is shown in Fig. 7.

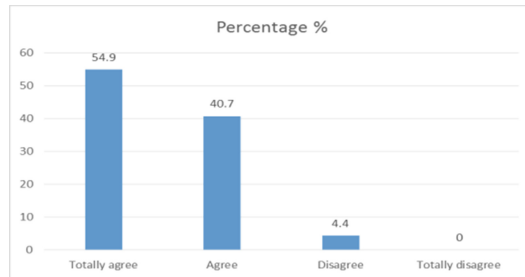


Fig. 7. Graph accelerate FYP grading process

Both graphs show the proposed approach is believed to benefit supervisors, examiners and students in managing the FYP.

5 Conclusion and Future Work

In this paper we have presented an approach to assist in FYP management. It is an essential move since current approaches have some insufficiencies in giving comprehensive outcome. The advantage of this approach is it encapsulates all aspects of FYP implementation from registration, administration and also evaluation. Furthermore, it integrates with other main systems to obtain or send data so that the processes are faster and reduces error in data entry. The future work for this research is to implement auto-submission every time marks are given by supervisor and examiner to eliminate

extra processes done by the Coordinator. The auto-submission will consider some constraints to prevent error.

Acknowledgements. This work is supported by Tier 1 Grant, Vote Number U897, Universiti Tun Hussein Onn Malaysia (UTHM).

References

1. Halim, M.A., Buniyamin, N., Imazawa, A., Naoe, N., Ito, M.: The role of Final Year Project and Capstone Project in undergraduate engineering education in Malaysia and Japan. In: 2014 IEEE 6th Conference on Engineering Education (ICEED), pp. 1–6. IEEE (2014)
2. Faculty of Mechanical and Manufacturing Engineering, UTHM. <http://fkmp.uthm.edu.my/mypsm> (2017)
3. Leung, C.H., Lai, C.L., Yuan, T.K., Pang, W.M., Tang, J.K., Ho, W.S., Huang, D.: The development of a final year project management system for information technology programmes. In: Technology in Education. Transforming Educational Practices with Technology: International Conference, ICTE 2014, Hong Kong, China, 2–4 July 2014. Revised Selected Papers, vol. 494, p. 86. Springer (2015)
4. Nagar, A.R., Anouar, J.: Method and system for handling program, project and asset scheduling management. U.S. Patent No. 8,738,414 (2014)
5. Braglia, M., Frosolini, M.: An integrated approach to implement project management information systems within the extended enterprise. *Int. J. Project Manag.* **32**(1), 18–29 (2014)
6. Callista, A., Fiona, F.: Development of a data management system for students' final year projects case study: Department of Information Systems. *Seminar Nasional Informatika (SEMNASIF)*, vol. 1, no. 5 (2015)
7. Faculty of Computer Science and Information Technology, UM. <http://ilmiah.fsktm.um.edu.my> (2017)

An Algorithm Design of Kansei Recommender System

Pei-Chun Lin^{1(✉)} and Nureize Arbaiz²

¹ Department of Information Engineering and Computer Science, Feng Chia University, No. 100, Wenhwa Rd., Seatwen 40724, Taichung, Taiwan
peiclin@fcu.edu.tw

² Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia
nureize@uthm.edu.my

Abstract. We propose an algorithm design for a Recommender System based on a Kansei model in this paper, we called this algorithm as Kansei Recommender System (hereafter, we denoted as KRS algorithm). The purpose of KRS algorithm is to support designers to pre-know the appearance feeling (Kansei) of products from consumers. To complete this algorithm, we divide the algorithm design into three parts: (1) Extract Kansei factors and evaluation factors from consumers' shopping items. (2) Determine a Kansei model for KRS algorithm. (3) Making decision by using KRS algorithm. We also give a concept map of paradigm by using KRS algorithm. In conclusion, we remain the future work to implement the KRS algorithm in real case studies with different fields of enterprises.

Keywords: Kansei Engineering · Fuzzy set theory · Statistical modeling
Recommender system · Classifier · Factor analysis · Algorithm design

1 Introduction

In traditional shopping process, consumers are able to feel and touch a product in a physical storefront before purchasing it. This enables consumers to quickly decide whether to buy this product or find other products that they would be more satisfied with them [1]. The consumers and sellers can also establish a good relationship during traditional shopping process. This relationship usually will increase the success rate of sales transactions [30]. Conversely, consumers can not touch or feel the product through online shopping. The image and narration of the product displayed through the website becomes key factors in their buying decision. Due to this situation, many manufacturers have begun to consider such subjective characteristics and develop their products in a way to convey their company's image. Therefore, a reliable recommendation system is an important tool for enterprises to encourage customers to make decision for buying their products.

Nowadays, product designing has been transformed from product-oriented to consumer-oriented. The consumer's needs and feeling are considered more valuable in product development than ever before [10]. The study of customer's subjective hidden

needs into specific products is critical to the enterprise in recent years. We call this research work as “emotional design” or “Kansei engineering”. Kansei is a Japanese word which means the consumer’s psychological feeling and image for considering a new product (see Refs. [21, 29]). Kansei engineering is an all-round human-centered technology to develop new products [5]. It has been widely used in construction machine, automobile, house construction, electric home appliance, textile industries, costume and so on (see Refs. [22, 23]).

There were numerous research works which were concerned about Kansei Engineering system (a system which is considered Kansei in the process) during the product design process (see Refs. [7, 24, 26, 28, 29, 31, 32]), but they still have some deficiencies. For example, the traditional Kansei image spend a lot of time and manpower when gathering data; the systems generally include only product image survey and lack of Kansei semantic works from product retrieval module for designers. Moreover, the systems are inconvenient for designers to search the matching products by inputting correct Kansei semantic words. Those systems are not common for a variety of products and they are only for one type of products such as modeling, color, material, etc. (see Refs. [2–4, 9, 33]).

In view of those considerations, a Recommender System based on Kansei engineering (hereafter, we denoted as KRS algorithm) is developed which is proposed to support designers to pre-know the appearance feeling (Kansei) of products from consumers. The research framework of KRS algorithm is arranged into three topics. (1) Extract Kansei factors and evaluation factors from consumers’ shopping items. (2) Determine a Kansei model for KRS algorithm. (3) Making decision by using KRS algorithm.

If the designers can master the consumers’ mind and explore the consumers’ feelings and needs, they will be able to successfully develop a better product [6]. Hence, in this paper, our objective is to implement KRS algorithm to identify the design trends that would attract to local designer. The KRS system could recommend the design elements consistent with consumer’s Kansei and the Kansei data is extracted out from consumer’s former behavior and the history of his purchase. This study also considers different analysis methods such as neural network, artificial intelligence and genetic algorithm as well as fuzzy logic to be utilized in the Kansei Model to construct the concerned databases and recommender system [25–27]. We believe that the KRS system could encourage the products of sales and reduces the excess products by pre-knowing consumers’ Kansei. Moreover, the KRS algorithm can serve in accordance with user’s Kansei immediately to increase customer’s satisfaction and consequently attract more sales in E-commerce.

The structure of this paper is organized as follows. Section 1 gives introduction of the study. Section 2 describes preliminary reviews of Kansei engineering system and statistical analysis. Section 3 establishes mathematical formulas of Kansei model. Section 4 explains a concept map of paradigm for using KRS Algorithm. Section 5 gives conclusions and future work.

2 Kansei Engineering System

Rapid development of science and technology in recent years makes traditional type of shopping in the store is changed. Before customers making a decision to buy goods, there are more choices for them and no longer through the way of shopping in the store. Designers are faced a new challenge to understand customers' preference and what kind of goods are more popular in their store. Therefore, to implement a recommender system based on consumer's Kansei becomes a crucial subject in E-commerce. In this section, we give the explanation of the process for modeling the consumers' Kansei. First, we give short explanation about Kansei Engineering System.

2.1 Kansei Engineering System (KES)

In our paper, we focus on the products from online shopping. The behavior of users is actively and passively observed, and the data from consumer's shopping behavior and history of purchase is also collected. By distinguishing the individual, we can accumulate the data from users, even if they do not intend to buy that product. To build the database of Kansei, data from online retailers' showcase is accumulated.

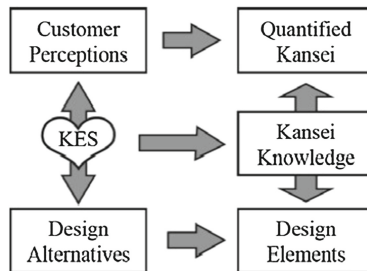


Fig. 1. The generalized Kansei Engineering System [35]

A common process of Kansei Engineering System (KES) is shown in Fig. 1 [35]. The customers' perceptions of a product are usually described in the form of Kansei semantic words and quantified by giving semantic scale ratings in system. On the other hand, the alternative design of the product can be decomposed into a set of different design elements. The Kansei knowledge can be established through advanced computer technologies and appropriate statistical analysis. Lin et al. have developed a series of statistical model based on fuzzy data [11–20]. Those concepts are closer to KES and suitable for modeling the Kansei from consumers. We will consequently use this concept in building the KRS in this paper.

In next subsection, we explain how to survey Kansei factors through statistical analysis from target products.

2.2 Extract Kansei Factors by Statistical Analysis

In this paper, the Kansei Engineering (KE) type I [8] is used. KE type I is used in the form of Kansei semantic words to distinguish emotional appeal and product category. The characteristics of commodities are added in the KES to determine the association of consumers' shopping behavior. Our goal is online shopping in E-commerce. The flowchart with statistical analysis through Kansei Engineering Type I is illustrated in Fig. 2.

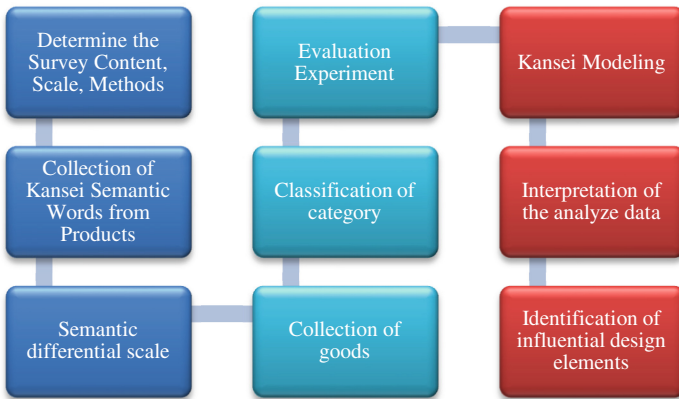


Fig. 2. Kansei Engineering Type I in KRS

In the procedure of statistical analysis, we will consider the cluster analysis in KES. Clustering analysis is a procedure to group a set of objects so that the elements in the same group are more similar to each other than those in other groups. The main task is to explore the data by using data mining or some basic statistical analysis. It is commonly used in many fields such as machine learning, retrieval system, bioinformatics, and image analysis.

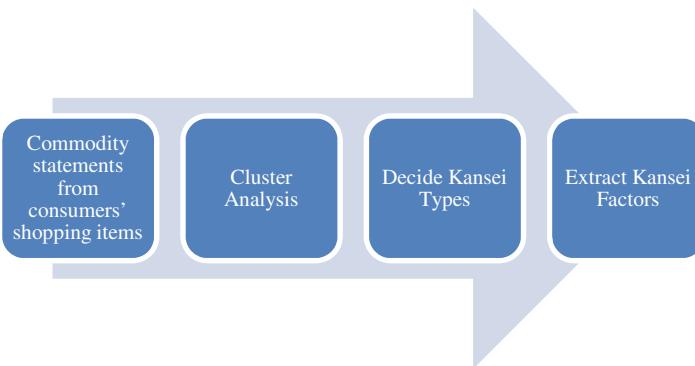


Fig. 3. Flowchart of extracting Kansei factors

To give an appropriate cluster algorithm, it is necessary to adjust the data processing and qualify the parameters until the result reaches our desired properties. We give a flowchart of extracting Kansei factors in Fig. 3. By using this statistical technique into machine learning, we could collect the consumers' Kansei and group the consumers into different types. Then, we take those Kansei factors into our Kansei model which will illustrate in next section.

3 Kansei Modeling

After we determined the Kansei factors, we need to give a score for the features (Kansei factors) that enable us to distinguish which are the most principal factors for consumers to buy this commodity. We also collect the evaluation sentences from commodities. We think that the evaluations from people who also buy this item will have the same type to buy those items they never bought before. We give a definition of mathematical formula to compute the Kansei factors into a score value in the following.

Definition 1 Scores of Kansei Factors and Evaluations Let x_i is denoted as consumer i , and \tilde{x}_i is the evaluations of the items from other consumers, for $i = 1, 2, 3, \dots$. We denote the $f(x_i)$ as the feature of commodity statement and $g(\tilde{x}_i)$ as the feature of evaluation statement which are extracted by cluster analysis. To decide the decision boundary, we give score functions for $f(x_i)$ and $g(\tilde{x}_i)$, separately, as follows:

$$\begin{aligned} Score_1(\mathbf{x}_i) &= \mathbf{W}^T f(\mathbf{x}_i) = \sum_{j=0}^{m-1} W_j f_j(x_i) \\ &= W_0 f_0(x_i) + W_1 f_1(x_i) + \dots + W_{m-1} f_{m-1}(x_i), \end{aligned}$$

and

$$\begin{aligned} Score_2(\tilde{\mathbf{x}}_i) &= \widetilde{\mathbf{W}}^T g(\tilde{\mathbf{x}}_i) = \sum_{k=0}^{n-1} \widetilde{\mathbf{W}}_k g_k(\tilde{x}_i) \\ &= \widetilde{\mathbf{W}}_0 g_0(\tilde{x}_i) + \widetilde{\mathbf{W}}_1 g_1(\tilde{x}_i) + \dots + \widetilde{\mathbf{W}}_{n-1} g_{n-1}(\tilde{x}_i), \end{aligned}$$

where \mathbf{W} is the effect factors (coefficients) of feature $f(x_i)$ and $\widetilde{\mathbf{W}}$ is the effect factors of feature $g(\tilde{x}_i)$, i.e. W_j is the weight value of feature $f_j(x_i)$ and $\widetilde{\mathbf{W}}_k$ is the weight value of feature $g_k(\tilde{x}_i)$, for $j = 0, 1, 2, \dots, m-1$ and $k = 0, 1, 2, \dots, n-1$.

It is not easy to deal with Kansei semantic words in statistical terms especially we have a series of Kansei factors in our database. Considering the vague data from consumers' Kansei, we will use the concept of fuzzy logic [34] to define the Kansei model. Also that Lin et al. have developed a series of statistical model based on fuzzy data [11–20]. We will consequently use those concepts in this paper.

We give the Definition of Kansei Model in the following that combine two scores which mentioning in Definition 1.

Definition 2 Kansei Model The Kansei model (KM function) is defined by an estimated function which is denoted as $\widehat{\text{KM}}$ function and given as follows:

$$\widehat{\text{KM}}(y = +1 | \mathbf{X}, \widehat{\mathbf{W}}) = \frac{1}{1 + e^{-\text{Score}_1(\mathbf{x})}} + \text{sign}(\text{Score}_2(\tilde{\mathbf{x}})),$$

where the values of Score_1 and Score_2 are calculated by using formulas in Definition 1, \mathbf{X} is a two-dimensional vector and denoted as $\mathbf{X} = [\mathbf{x}, \tilde{\mathbf{x}}]$. $\widehat{\mathbf{W}}$ is also a two-dimensional vector which is indicated the best solution of likelihood function $l(\widehat{\mathbf{W}})$ and denoted as $l(\widehat{\mathbf{W}}) = [l(\mathbf{W}), l(\widetilde{\mathbf{W}})]$.

Note that training a good classifier means learning the best coefficients in Machine Learning. We need to maximize the quality metric (likelihood $l(\mathbf{W})$ and $l(\widetilde{\mathbf{W}})$) over all possible W_0, W_1, \dots, W_{m-1} and $\widetilde{W}_0, \widetilde{W}_1, \dots, \widetilde{W}_{n-1}$.

We give two functions for Kansei Model to calculate an weight values for evaluating consumers' Kansei. One of the function is calculated by using logistic function which is in order to make decision in the values between 0 and 1 that will be closer consumer's Kansei. The other function is calculated by using sign function that will be given positive (+1) and negative (-1) evaluations only.

The concept of this KM function comes from the definition of combining two factors into one decision making formula which defined in Lin et al.'s paper [13]. They also give a serious proof to show that it is more significant to use exponential function in making decision.

To evaluate our $\widehat{\text{KM}}$ function, we calculate the accuracy by using following function.

$$\text{Accuracy}(\text{KM}, \widehat{\text{KM}}) = \frac{1}{q} \sum_{i=0}^{q-1} 1(\widehat{\text{KM}} = \text{KM}),$$

where $q = m + n, \forall m, n \in \mathbb{R}$, and $1(x)$ is the indicator function.

The procedure of training a good classifier for KRS algorithm is given in Fig. 4.

In the following section, we give a concept map of paradigm for using KRS Algorithm.

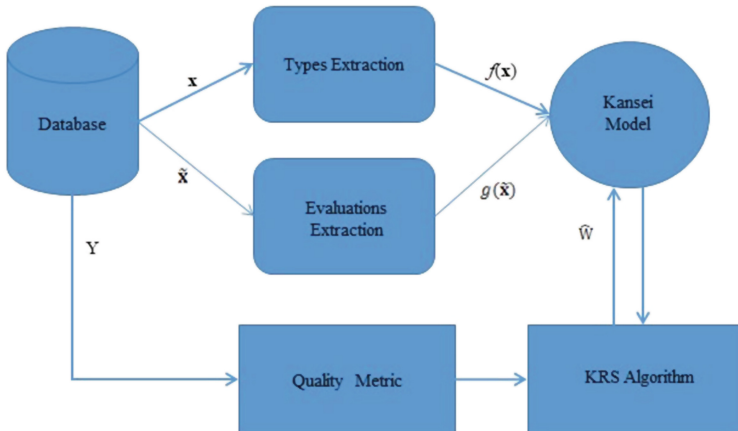


Fig. 4. Procedure of training classifier by KM function

4 A Concept Map of Paradigm by Using KRS Algorithm

We demonstrate a paradigm of online shopping market by using KRS Algorithm in Fig. 5.

The process of implementing KRS algorithm is given as follows:

- (1) Collect the shopping market data from database.
- (2) Training data by cluster analysis.
 - (i) Extract Kansei Factors from consumers' shopping items.
 - (ii) Extract evaluation factors of the items from the populations.
- (3) Decide a good classifier by KM function.
- (4) Implement KRS algorithm.
- (5) Calculate the accuracy.
- (6) Making decision (Recommending proper goods for consumers).

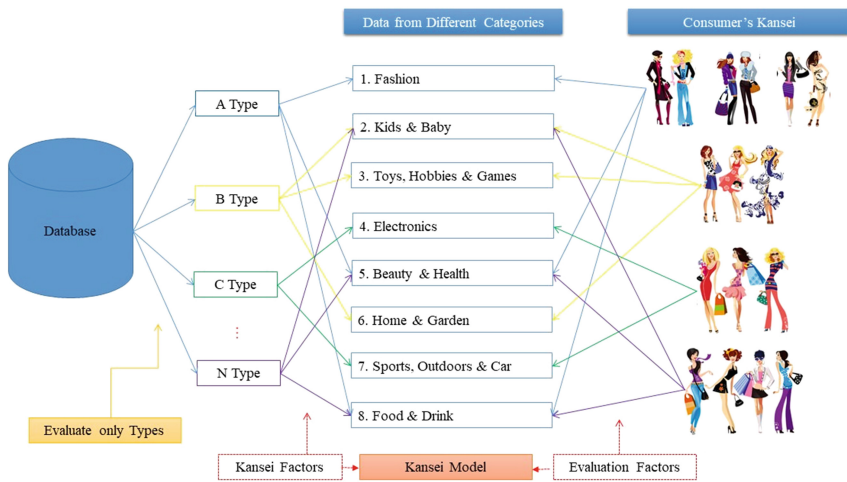


Fig. 5. A concept map of paradigm for using KRS algorithm

5 Conclusions and Future Work

The building of Recommender System based on Kansei Engineering (KRS Algorithm) is indispensable for designers because of on-line shopping will become a popular way for consumers and data analysis becomes a huge topic of internet of thing (IoT) in this word. We expect that the KRS algorithm could encourage the products of sales and reduces the excess products by pre-knowing consumers' Kansei. Furthermore, the KRS algorithm can serve in accordance with users' Kansei immediately. In conclusions of this paper, we demonstrate a real-word application by using KRS algorithm. We believe that using KRS algorithm would increase customer's satisfaction and consequently attract more sales in E-commerce. In future work, we will cooperate with enterprise for real case studies.

Acknowledgements. The authors express her appreciation to the University Tun Hussein Onn Malaysia (UTHM). This research also supported by GATES IT Solution Sdn. Bhd. Under its publication scheme.

References

1. Chuan, N.K., Sivaji, A., Shahimin, M.M., Saad, N.: Kansei Engineering for e-commerce sunglasses selection in Malaysia. *Proc. Soc. Behav. Sci.* **97**, 707–714 (2013)
2. Cao, Y., Li, Y.: An intelligent fuzzy-based recommendation system for consumer electronic products. *Expert Syst. Appl.* **33**, 230–240 (2007)
3. Das, A., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In: *Proceedings of the 16th IEEE*, pp. 271–280 (2007)
4. Hotta, H., Hagiwara, M.: A fuzzy rule based personal Kansei modeling system. In: *2006 IEEE International Conference on Fuzzy Systems*, Vancouver, Canada, 16–21 July, pp. 1031–1037 (2006)
5. Huang, M.S., Tsai, H.C., Lai, W.W.: Kansei Engineering applied to the form design of injection molding machines. *Open J. Appl. Sci.* **2**, 198–208 (2012)
6. He, Z.X., Wang, S.: Mapping customer requirements to product performance index based on data fusion by vague set. *J. Comput. Inf. Syst.* **6**, 1679–1686 (2009)
7. Jindo, T., Hirasago, K., Nagamachi, M.: Development of a design support system for office chairs using 3-D graphics. *Int. J. Ind. Ergon.* **15**(1), 49–62 (1995)
8. Lokman, A.M., Nagamachi, M.: Kansei Engineering: a beginner's perspective. UPENA, Malaysia (2010)
9. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. *Internet Comput. IEEE* **7**, 76–80 (2003)
10. Lu, H., Yan, C., Du, J.: An interactive system based on Kansei Engineering to support clothing design process. *Res. J. Appl. Sci. Eng. Technol.* **6**(24), 4531–4535 (2013)
11. Lin, P.-C., Nureize, A., Hsiao, Y.-C.: Hypothesis test for identifying the vague factors from consolidated income. In: *2017 IEEE International Conference on Fuzzy Systems*, Naples, Italy, 9–12 July 2017
12. Lin, P.-C., Nureize, A.: One-way ANOVA model with fuzzy data for distinguishing factors from consumer demand. In: *Recent Advances on Soft Computing and Data Mining. Advances in Intelligent Systems and Computing*, vol. 549, pp. 111–121 (2017)
13. Lin, P.-C., Watada, J., Wu, B.: A parametric assessment approach to solving facility location problems with fuzzy demands. *IEEJ Trans. Electron. Inf. Syst.* **9**(5), 484–493 (2014)
14. Lin, P.-C., Nureize, A.: Two-echelon logistic model based on game theory with fuzzy variable. In: *Recent Advances on Soft Computing and Data Mining. Advances in Intelligent Systems and Computing*, vol. 287, pp. 325–334 (2014)
15. Lin, P.-C., Watada, J., Wu, B.: Risk assessment of a portfolio selection model based on a fuzzy statistical test. *IEICE Trans. Inf. Syst.* **E96-D**(3), 579–588 (2013)
16. Lin, P.-C., Watada, J., Wu, B.: Identifying the distribution difference between two populations of fuzzy data based on a nonparametric statistical method. *IEEJ Trans. Electron. Inf. Syst.* **8**(6), 591–598 (2013)
17. Lin, P.-C., Watada, J., Wu, B.: Portfolio selection model with interval values base on fuzzy probability distribution functions. *Int. J. Innov. Comput. Inf. Control* **8**(8), 5935–5944 (2012)
18. Lin, P.-C., Wu, B., Watada, J.: Goodness-of-fit test for membership functions with fuzzy data. *Int. J. Innov. Comput. Inf. Control* **8**(10), 7437–7450 (2012)

19. Lin, P.-C., Watada, J., Wu, B.: A database for a new fuzzy probability distribution function and its application. *Int. J. Innov. Manag. Inf. Prod.* **2**(2), 1–7 (2011)
20. Lin, P.-C., Wu, B., Watada, J.: Kolmogorov-Smirnov two sample test with continuous fuzzy data. *Integr. Uncertain. Manag. Appl.* **68**, 175–186 (2010)
21. Nagamachi, M.: Introduction of Kansei Engineering. Standard Association, Tokyo, Japan (1996)
22. Nagamachi, M.: Kansei Engineering as a powerful consumer-oriented technology for product development. *Appl. Ergon.* **33**(3), 289–294 (2002)
23. Nagamachi, M., Matsubara, Y.: Hybrid Kansei Engineering system and design support. *Int. J. Ind. Ergon.* **19**(2), 81–92 (1997)
24. Nureize, A., Lin, P.-C.: Weighted value assessment of linear fractional programming for possibilistic multi-objective problem. *Int. J. Adv. Intell. Paradig.* **8**(1), 42–58 (2016)
25. Nureize, A., Watada, J., Lin, P.-C.: Fuzzy random regression-based modeling in uncertain environment. In: *Sustaining Power Resources through Energy Optimization and Engineering*, pp. 127–146. IGI Global (2016)
26. Nureize, A., Watada, J.: Building multi-attribute decision model based on Kansei Information in environment with hybrid uncertainty. In: *Intelligent Decision Technologies*, pp. 103–112. Springer, Berlin, Heidelberg (2011)
27. Nagamachi, M.: An image technology expert system and its application to design consultation. *Int. J. Hum.-Comput. Interact.* **3**(3), 267–279 (1991)
28. Nagamachi, M.: Kansei Engineering: a new ergonomic consumer-oriented technology for product development. *Int. J. Ind. Ergon.* **15**(1), 3–11 (1995)
29. Ozaki, S., Hisano, S., Iwamoto, Y.: Potency of animal models in Kansei Engineering. *Kansei Eng. Int. J.* **11**, 127–132 (2012)
30. Sivaji, A., Downe, A.G., Mazlan, M.F., Soo, S., Abdullah, A.: Importance of incorporating fundamental usability with social and trust elements for e-commerce website. In: *Proceedings of the Business, Engineering and Industrial Applications (ICBEIA)*, pp. 221–6, Kuala Lumpur, Malaysia, 5–7 June 2011
31. Tanaka, M., Miyaji, M., Yamamoto, U., Hiroyasu, T., Miki, M.: Interactive recommender system to estimate personal user's Kansei Model. *Int. J. Comput. Sci. Eng. (IJCSE)* **5**(11), 904–913 (2013)
32. Tan, Z.Y., Sun, S.Q.: Image retrieval technology based on imagery cognition model. *Eng. Sci.* **42**(5), 763–767 (2008)
33. Tang, Z., Sun, S., Zeng, X., Cao, H., Xing, B., Yang, Z.: Researching on Kansei Engineering system for product image survey and retrieval. *J. Comput. Inf. Syst.* **10**(10), 4029–4038 (2014)
34. Yoshiki, N.: *Kansei Data Analysis*. Morikita Shuppan (2000)
35. Zhai, L.Y., Khoo, L.P., Zhao, W.Z.: A dominance-based rough set approach to Kansei Engineering in product development. *Expert Syst. Appl.* **6**, 393–402 (2009)

Warehouse Picking Model for Single Picker Routing Problem in Multi Dimensional Warehouse Layout Using Genetic Algorithm Approach to Minimize Delay

Dida Diah Damayanti^(✉), Erlangga Bayu Setyawan,
Luciana Andrawina, and Budi Santosa

Faculty of Industrial and System Engineering, Telkom University, Bandung
40257, Indonesia
{didadiiah,erlangga.setyawan,lucianawina,bschulasoh}
@gmail.com

Abstract. Order picking process is one of the most time-consuming activities at a 3PL's warehouse system. This paper aims to determine picking process and routing method in finding the optimal order picking by using Single Picker Routing Problem in a multidimensional warehouse. Single Picker Routing Problem (SPRP) is method development of general Traveling Salesman Problem (TSP). SPRP is used to determine the minimum route in picking process to several points including depot areas and choosing a concerned location. In this paper, we develop a model to minimize travel time considering "x-axis", "y-axis" and "z-axis" (height of racking system) using Genetic Algorithm (GA). Genetic Algorithm as the chosen method is carried out to calculate process quickly because the number of variables in the case study used in. The result of the model is the picking sequence not only considering 2D layout, but also the height of the racking system. By applying the model to the case study, the improvement of picking time is 60%.

Keywords: Picking and routing method · Traveling salesman problem (TSP)
Single picker routing problem (SPRP)

1 Introduction

Warehousing is one of the activities in the logistics is very important and critical in industrial systems and services. Warehouse has an important role to enhance the success of business in the level of costs and customer service. Thus, in the activities of the warehouse is required allocation of products with classification based on the

Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

characteristics and speed of each product as well as to the arrangement and preparation of the number of products on every slot on every shelf in the warehouse [1].

One way to improve warehouse activity is by product allocation. The process used to solve these problems by making the process of classifying products based on Class-Based Storage using analysis of FSN. After that, the process of storage allocation by ZABRLS (zone, aisle, bay, row, level, and slot). In this research, will be the development of allocating goods stored by the optimal route using the Single Picker Routing Problem (SPRP) and supported by the WMS application.

Picking Activity is one of the largest contributions in the warehouse activity [2]. In picking activity, Genetic Algorithm (GA) selected in the determination of the routing method in the case of Traveling Salesman Problem (TSP), which serves to determine the optimal order picking. Another research aims to determine picking and routing method which result in two dimensions picking orders by ignoring racking heights 4. The method can not be applied to the most 3PL company as a supply chain because it does not pay attention to high-racking and time the process used to obtain the order picking optimal results take a long time because the value of outbound throughput at the facility is very high.

Single picker routing problem considering multidimensional warehouse is needed to be developed because picking time not only consider 2D warehouse (x-axis and y-axis) but also consider 3D (vertical racking height).

2 GA for Single Picker Routing Problem

2.1 Existing Model for Single Picker Routing Problem

In conducting warehouse activities necessary to optimize the location of the warehouse which aims for determining exact location and suitable for products in a storage area and aims to minimize time and costs, both in product storage process (storing) and picking process.

SPRP (Single Picker Routing Problem) is a routing problem in warehouse at the time of picking activity. SPRP is used to determine the minimum route in picking process to several points including depot area and choosing a concerned location. Thus, a special case of the TSP is able to represent a general TSP formulation which means that the general TSP could be eligible for modeling the SPRP [3].

Model development of Single Picker Routing Problem in this research was based on previous studies with basic concepts of the Traveling Salesman Problem 5. Traveling Salesman Problem basic model used to solve Single Picker Routing Problem has an objective function to minimize the total distance from the whole point of picking by ensuring that each vertex can only be visited once.

In resolving Single Picker Routing Problem using the model cannot be applied optimally. That is because the model does not consider the number of goods carried by the material handling equipment. In the process of picking the warehouse, material handling equipment will depart from the starting point (depot), then to the point of picking to take the goods, and the goods are brought to a certain point, both to staggering and back to the depot.

Another research has developed a basic TSP model by adding variables to consider material handling equipment in the carriage of goods (commodities). The goods carried are palletized, so in one way material handling equipment can only be transported per pallet. The model ensures that the path is located at the end of the aisle are not included in the calculation, because the material handling equipment to maneuver between aisle [4].

Later research has developed a problem-solving in a warehouse for Single Picker Routing Problem using mathematical models. The objective function of the model has some functional purpose of determining the shortest route from basic TSP into several factors, i.e. [5]:

1. The distance between front cross aisle to cross aisle behind.
2. The distance between front cross aisle to cross aisle next (back cross aisle).
3. The distance between vertex distance between slots in racking).
4. The distance between nearest aisle.
5. The distance between depot with cross aisle.

The model has a lot of functionality for the purpose of storage has a very large spacious and pallet position of 145,000 pallet positions in a separate location block 124, so as to generate optimal repair requires a lot of research factors [5].

2.2 Improved Model for Single Picker Routing Problem Considering Rack Height

By considering the height of racks in the warehouse, an improved model for single picker routing problem is developed, referring to third in the journal by adding the variable of storage position vertically, which is symbolized by (v).

The proposed model has also the addition of a variable, i.e., a variable that shows time of the vertical p smaller than vertical position q, where t_v is:

$$t_{vertical} = \frac{\text{level height}}{\text{lifting speed}} \quad (1)$$

Lifting speed is the speed of material handling when lift pallet will be allocated on a shelf. Calculation in speed lifting using data specification such as maximum lifting speed and lifting speed of at least the amount of material handling certain weight. Data specifications are then compared with the actual weight of the product is lifted in one pallet using interpolation formula. To obtain the lifting speed in the Eq. (1), using the following calculation:

$$\text{lifting speed} = v_{min} \frac{v_{max} - v_{min}}{\text{weight}_{max}} \cdot \text{actual weight} \quad (2)$$

The formulation of the model is constructed as follows:

$$\min \sum_{(p,q) \in E} c_{pq} \cdot x_{pq} \quad (3)$$

$$\sum_{p \in V} x_{pq} = 1 \quad \forall q \in V \quad (4)$$

$$\sum_{q \in V} x_{pq} = 1 \quad \forall p \in V \quad (5)$$

$$\sum_{q \in V} g_{pq} - \sum_{q \in V \setminus \{0\}} g_{pq} = 1 \quad \forall p \in V \setminus \{0\} \quad (6)$$

$$h_p - h_q + (n+1)x_{pq} \leq n \quad \forall_{pq} \in E; p, q \neq 0 \quad (7)$$

$$tv_p - tv_q + (n+1)x_{pq} \leq n \quad \forall_{pq} \in E; p, q \neq 0 \quad (8)$$

$$g_{pq} \leq nx_{pq} \quad \forall_{pq} \in E; p, q \neq 0 \quad (9)$$

$$x_{pq} \in \{0, 1\} \quad \forall_{pq} \in E \quad (10)$$

$$g_{pq} \geq 0 \quad \forall_{pq} \in E : q \neq 0 \quad (11)$$

Objective function (3) is to determine the shortest route (c) in the case of single picker routing problem from node (p) to node (q). Equations (4) and (5) ensure that each vertex p and q can only be visited once. Equation (6) which ensures there is only one commodity or goods (pallet) are brought in from vertex p to vertex q . Equation (7) ensuring that the position of the x-axis of a vertex p is smaller than q vertices. Equation (8) ensuring that the position of z-axis (racking height) of a vertex p is smaller than q vertices. Equation (9) ensures that the path is located at the end of the aisle are not included in the calculation, because the material handling equipment to maneuver between aisle.

For the model to be used, the process should be done in advance of the allocation based on the decision-making process of the fastest. The model is used to make process of slotting based on the time of the product is the fastest [6]. In the paper it is discussed slotting process with the objective function is to minimize the distance. With Material Handling Specification data, the distance is converted into a process.

The weakness of the Single Picker Routing Problem model developed with TSP concept is that the calculation takes a long time when picking a point (p, q) and many of the high throughput. The project will be carried out this great solution model using genetic algorithm approach. Genetic Algorithm is applied for Single Picker Routing Problem [7].

In modeling of the GA model for the case, picking points are interpreted as chromosomes. In Table 1 is shown the example of six picking points within a generation for five location points:

Table 1. Example of cumulatif probability calculation

Chro no.	Variat random uniform (<i>i</i>)				
	A	B	C	D	E
1	0.13	0.63	0.28	0.72	0.36
2	0.45	0.29	0.64	0.21	0.47
3	0.51	0.38	0.11	0.27	0.68
4	0.17	0.08	0.36	0.81	0.49
5	0.37	0.31	0.42	0.39	0.74
6	0.26	0.28	0.67	0.21	0.33

From the random number table above, the position *i* indicates the location (picking point) that must be visited. While the random value that is in position *i* determine the visit sequence of each point for the process of picking.

With the above random number parameter, it can be determined that on the second chromosome, the value of 0.21 is the smallest random number. So point D (point four) on the second chromosome is the first point to be visited (Table 2).

Table 2. Result of encoding

Chro no.	Result of encoding				
	A	C	E	B	D
1	A	C	E	B	D
2	D	B	A	E	C
3	C	D	B	A	E
4	B	A	C	E	D
5	B	A	D	C	E
6	D	A	B	E	C

Fitness value of a chromosome in this case, is calculated from the total travel distance of the picking list. Chromosomes that have a high fitness value will have a greater chance of surviving in the next generation. After getting the fitness value, the next step is to calculate the fitness value evaluation to get cumulative value of chromosomal probability, then the selection process is done by using the roulette wheel method.

To have some good alternative solutions, cross over is done by using cut-point method, and the final chromosome is obtained by doing the mutation processes.

3 Discussion

As described in the background problems, Single Picker Routing Problem in the situation of the warehouse XYZ can not be solved using the model developed before [3–5]. All models had not considered the height of the racking system for storing goods. The next discussion will explain the analysis of model development and the application result of the model to the case study in XYZ, and the validation process by Monte Carlo simulation.

3.1 Total Picking Time

In this study will discuss about an improved model for single picker routing problem, which is consider racking height. Single picker routing problem is an activity performed for the picking (dropping) process on a pallet using the shortest route in executing a picking list to minimize the delay time.

In conducting picking activity at XYZ, there are four variables that influence the picking time: horizontal travel time, vertical travel time, picking time and double handling time. So to know the total picking time can be formulated as follows:

$$\text{Total Picking Time} = \text{HT} + \text{VT} + \text{PT} + \text{DDT}$$

Notation:

HT	Horizontal Travel Time
VT	Vertical Travel Time
PT	Picking Time
DDT	Double Handling Time

Horizontal Travel Time

Horizontal travel time is the time required to perform picking activity from zero or central point (location of MHE) to the nearest picking location. The distance from the zero point or the center point to the nearest picking location is called the horizontal distance based on rectilinear. Rectilinear is a measured distance following a perpendicular path from a central point of a location to another central location point. Horizontal travel time is obtained from the horizontal distance division that is reached by the average speed of material handling.

If known the point of the location of material handling to the closest rack that is equal to 3.375 m, travel speed of 2.5 m/s. It is known horizontal travel time of 1.35 s.

Vertical Travel Time

Vertical travel time is the time it takes to do the rack picking process at a level above the zero level (level one, level two and so on). Vertical travel time is obtained by dividing the vertical distance obtained from the height of the rack specification with the specification of lifting speed material handling used per unit of total product weight on a pallet.

If it is known that the height of the rack is 1.67 m, then at level 1 has a vertical distance of 0, at level 2 has a vertical distance of 1.67 m, and on the next level has a vertical distance of distance at level 2 plus the height of the rack, and so on, as the number of levels. So it can be seen, the vertical travel time at level 1 is 0 s, where the vertical distance of level 1 is 0 m divided by lifting speed of 0.27 m/s.

Picking Time

Picking time is the time spent in the process of moving goods from storage to meet a specific need at each level. Picking time is obtained from observations in XYZ warehouse. Observations were made 30 times. The average picking time at level 1 is 11 s, level 2 is 23.47 s, and level 3 is 43.83 s.

Double Handling Time

Double handling time is the time spent for repetitive activities in handling one product. Double handling time has different time for each level and position.

If known a SKU is at location C at level 1, then double handling time equal to 33 s that is picking time multiply three. This happens because, when it comes to issuing

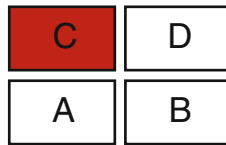


Fig. 1. Position of laying of goods on rack

SKUs from location C, the first thing to do is to issue another SKU at location A, then issue the destination SKU at location C and re-enter the previous SKU A location to location A again (Fig. 1).

3.2 Monte Carlo Simulation to Determine the Number of Picking Points

Simulation is a mathematical modeling technique that is used to mimic the real picking system. In this study, the real system to be emulated is the number of picking points on finished goods warehouse PT XYZ. The simulation model of stochastic simulation monte carlo, because the data demand affect picking point has a certain probability. Stochastic simulations are needed to compare the random variable is raised to determine the location of picking by sampling the probability of demand. Steps to perform Monte Carlo simulations are:

- a. Calculate the probability of demand for each SKU's

The first step taken to perform Monte Carlo simulations is to calculate the probability of demand (delivery order) for each SKU's within a certain period. The result of the calculation of the probability of demand can be seen in Table 2.

- b. Generating random numbers to determine the number of picking points on each trial picking list

In this simulation process, will use ten experiments picking list by generating a random variat uniform distribution to determine the number of points each picking list a number of twenty to fifty. Variate generated random uniform distribution, in accordance with the distribution of the demand of each SKU's.

H0: data distribution is uniform

H1: data distribution is not uniform

Because the test results for five pieces of SKU's has a value of greater significance than the alpha (0.05), then accept H0, so that it can be concluded that the distribution of data demand is uniform.

The generation of random variat for each picking list can be is then generated (Table 3).

- c. Generating random picking locations for each SKU's trial to determine which point would be picked and the amount uniform random numbers generated in accordance with the number of picking points each trial. Then random sampling will be done on demand probability distribution in accordance with the process of Monte Carlo (Table 4).

Table 3. Example of cumulatif probability calculation

SKU's	Probability	Cum. probability
G01D	3.62241E-07	3.62241E-07
V02FP	3.62241E-07	7.24482E-07
K09HL	7.24482E-07	1.44896E-06

Table 4. Random numbers number of picking points each trial

Picking list	Number of picking points
Picking 1	23
Picking 2	29
Picking 3	34
Picking 4	41
Picking 5	35
Picking 6	26
Picking 7	24
Picking 8	31
Picking 9	26
Picking 10	36

3.3 Genetic Algorithm Calculation for Single Picker Routing Problem

In this study, the use of Genetic Algorithm for Single Picker Routing Problem aims to obtain optimal results in determining the picking route. The reason for using Genetic Algorithm is because the model in this research can not be solved with simple mathematics solver. This happens because the SPRP model has a considerable limitation that can be seen in the above mathematical model. The steps in performing calculations using Genetic Algorithm as follows:

1. Encoding Chromosome

Encoding is the process of determining the value to be used as input of a genetic algorithm. Encoding process on single picker routing problem will use random generator technique in permutation representation. Random generator is one of the encoding techniques which generates chromosomes as the initial population by involving random numbers.

2. Evaluation and Selection of Chromosomes

The concept of genetic algorithm is that individuals who have a high fitness value will survive, whereas individuals who have low fitness values will die (eliminated). This process aims to get the best parent, because a good parent will produce good offspring as well. The value of fitness is a value that indicates the level of good or not an individual. Chromosomes that have a high fitness value will have a greater chance of surviving in the next generation.

3. Cross Over

Cross over process requires two parent chromosomes. This process serves to swap some of the information held by the first parent with the second parent and vice versa. The selected parent chromosome is randomly selected and influenced by cross over rate (pc). The higher cross-over rate will result in a more varied achievement of alternative solutions and reduce the likelihood of generating undesirable optimum value. However, if the value of cross over rate is too high, it will lead to a longer time wastage to calculate in a solution area that is not as promising as an optimal solution. Cross over process will use cut-point method.

4. Mutation

The mutation process is needed to replace the genes of a population during selection. In addition, mutations also serve to form genes that are not present in the initial population. In this process, a gene pair is selected in a random chromosome for gene exchange.

In the problem of traveling salesman problem, mutation scheme will use swapping mutation method (exchange mutation), that is mutation process by choosing two genes randomly and exchange them. The chromosome that undergoes a mutation in a population will be determined by the mutation rate (pm). The mutation rate should be applied with a small value, because if the mutation rate is too large it will produce too many mutations, resulting in a weak individual because the configuration of the superior genes is mutated.

3.4 Calculation Result

From the mathematical model of SPRP, the calculation using Genetic Algorithm which is processed using matlab application produces the image below which shows

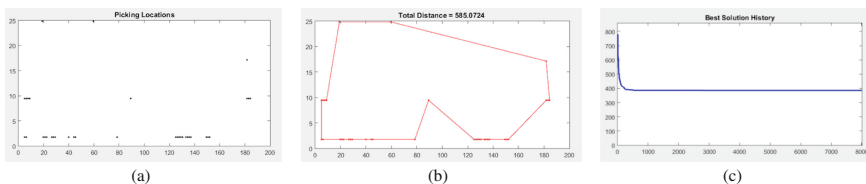


Fig. 2. Output MATLAB simulation

Table 5. Simulation result

Trial	1	2	3	4	5	6	7	8	9	10
Number picking location	23	29	34	41	35	26	24	31	26	36
Existing picking time (s)	466.9	588.7	690.2	832.3	710.5	527.8	487.2	629.3	527.8	730.8
Proposing picking time (s)	186.76	235.48	276.08	332.92	284.2	211.12	198.88	250.72	221.12	292.32

the points of picking location contained in the warehouse. These location points are used to perform the picking process.

In addition, the results of SPRS calculations using Genetic Algorithm resulted in a sequence of routes used for picking activity. In doing the picking activity is not determined from where to start to do picking, but in doing the picking activity must be in accordance with the order determined according to the results of the calculations performed.

To obtain the optimal distance results are iterated repeatedly so as to obtain steady state value, the optimal value where the graph does not increase or decrease shown in the existing graph (Fig. 2c).

Simulation result shown in Table 5. Number of picking location have been determined in Table 3 using Monte Carlo Simulation. The simulation result, our proposing algorithm can minimize the picking time about 60% from the existing.

4 Conclusion

SPRP (Single Picker Routing Problem) is a routing problem in the warehouse at the time of picking activity. SPRP relating to the determination of minimum service in doing picking to several points including depot picking and choosing the location in question. Thus, a special case of the TSP is able to represent a significant public TSP formulation TSP that can generally be feasible to do modeling SPRP. This paper has developed single picker routing problem consider racking height, because lot of logistics operation use racking system for the storage system.

References

1. Manzini, R.: Warehousing in the Global Supply Chain. Springer, Bologna, London, Dordrecht, Heidelberg (2012)
2. Frazelle, E.H.: World-Class Warehousing and Material Handling. McGraw-Hill, Singapore (2002)
3. Miller, C.E., Tucker, A.W., Zemlin, R.A.: Integer programming formulations and traveling salesman problems. *J. Assoc. Comput. Mach.* **7**, 326–329 (1960)
4. Gavish, B., Graves, S.C.: The traveling salesman problem and related problems (1978)
5. Henn, S., Scholz, A., Stuhlmann, M., Wascher, G.: A New Mathematical Programming Formulation for the Single-Picker Routing problem in a Single- Block Layout, Working Paper No. 5/2015, Bezug über den Herausgeber (2015). ISSN 1615–4274
6. Claus, A.: A new formulation for the traveling salesman problem. *SIAM J. Algebraic Discrete Methods* **5**, 21–25 (1984)
7. Bottani, E., et al.: Optimisation of storage allocation in order picking operations through a genetic algorithm (2012)

Relationship Between Angiotensin Converting Enzyme Gene and Cardiac Autonomic Neuropathy Among Australian Population

Ahmad Shaker Abdalrada^{1(✉)}, Jemal H. Abawajy¹, Morshed U. Chowdhury¹,
Sutharshan Rajasegarar¹, Tahsien Al-Quraishi¹, and Herbert F. Jelinek²

¹ Deakin University, Burwood, VIC, Australia
aabdalra@deakin.edu.au, jemal.abawajy@deakin.edu.au

² Charles Sturt University, Albury, NSW, Australia
hjjelinek@csu.edu.au

Abstract. Angiotensin Converting Enzyme (ACE) gene considers a risk factor for many pathologies such as hypertensive, and diabetic nephropathy. The objective of this study was to investigate the role of ACE genotype with Cardiovascular Autonomic Neuropathy (CAN). We used a data set for 299 participants with and without CAN, as well as with different ACE genotype. Various statistical tests have been considered. Logistic regression was applied to demonstrate the size effect of each predictor. The results revealed, there was no significant different between ACE genotype in the patients with and without CAN. Logistic regression demonstrated only Ewing battery tests as an effective predictive factors. Our investigation found ACE genotype was not a risk factor in Cardiovascular Autonomic Neuropathy in the population of our study.

Keywords: Angiotensin Converting Enzyme (ACE) gene
Cardiovascular Autonomic Neuropathy (CAN) · Logistic regression

1 Introduction

Cardiovascular Autonomic Neuropathy (CAN) is one of the most common complications of diabetes disease that is overlooked [1–3]. Patients with CAN are more likely to have a damage in their heart nerve fibers, innervate the heart and blood vessels [4]. It causes abnormality in the heart rate and vascular dynamics [5]. The prevalence of CAN has increased significantly in worldwide [6]. For example, recent statistics have shown there are 382 million individual with Diabetes Mellitus (DM). This number is expected to rise to 592 million by 2025 [7]. Most of those patients are with diabetes type2, also 22% of patients with

diabetes type2 are suffering from CAN. Patients with CAN are more likely to face the sudden cardiac death and the risk of arrhythmia [8,9].

The task of CAN diagnosis traditionally based on five common tests but these tests known as Ewing tests battery [9] are not conclusive. Nevertheless, the criteria to diagnosis CAN is still under debate [10]. The predictive ability of such tests is still limited due to many reasons [11]. For instance, the mobility challenges often prevent patients to perform the lying to stand test. Some patients also have breathing problems, they cannot perform the valsalva manoeuvre test. Ewing tests are also not suitable for patients who are suffering from comorbidities such respiratory or Cardiovascular disease [11]. Therefore, it is important to discover new tests to enhance the diagnosis of CAN.

The ACE gene encodes the angiotensin-converting enzyme, which is a key component of the reninangiotensin system (RAS). The main task of RAS is to regulate blood pressure as well as balances the fluids and salts in human body [13]. ACE has ability to cut(cleave)proteins [14–16]. By cutting proteins angiotensin I at specific location, the (RAS) will convert angiotensin I into angiotensin II, which causes narrow in the blood vessels [17] and increases the blood pressure [18]. Individuals with diabetes type1,2 and DD allele pattern are more likely to developing diabetic nephropathy [17], which contributes to kidney failure. ACE gene is also capable to produce the aldosterone hormone, which causes absorption for water and salt in kidney. The ACE also can cleave other protein so-called Bradykinin2, which widens blood vessels, and reduces blood pressure [14–16]. We therefore investigated the role of ACE gene with CAN in the population of our study.

Machine learning algorithms and statistical analysis tests have shown ability to analyse medical data collected for diabetes [12,19–21]. Applying such techniques in the medical field has many positive outcomes. For instance, it improves patients' well-being, and reduces the morbidity and mortality related with cardiac arrhythmias in diabetes. Thus, we will investigate the relation between angiotensin converting enzyme (ACE) genotypes, cholesterol blood biochemistry markers, demographic data and CAN using machine learning methods and the statistical analysis tests. Specifically, investigating whether the diagnosis of CAN corresponds with the ACE genotypes or not.

Due to the aforementioned challenges and motivations,the contributions of this study are:

- Investigate the role of ACE genotype with CAN disease.
- Examine the association of cholesterol blood biochemistry markers, demographic patient's data.
- Assess and identify the significant predictors of CAN.

This paper is organized as follow: Sect.2 review previous works. Section3 explains the materials and methods used in this study. Experiments is in Sect.4 and results and discuss are in Sect.5. Finally, draw the conclusion in Sect.6.

2 Related Work

A significant number of studies have investigated the relationship between ACE genotype and variety of pathologies. To the best of our knowledge, the relationship between ACE genotype and cardiovascular autonomic neuropathy (CAN) has not been investigated before (Table 1 presented number of related works). For instance, Bhaskar, Mahin [22] have investigated the role of Angiotensin-converting enzyme (ACE), and peroxisome proliferator activated receptor gamma (PPARG) in the diabetic nephropathy disease among population from south India. In their study, 54 patients with diabetic nephropathy and 67 control patients have been used. Hardy–Weinberg equilibrium test was used to evaluate the distribution of genotype. Chi-Square test was applied to find out the association and to analysis the distributions in genotype. Odds ratios have been calculated. The results showed that there is no significant difference in the frequencies of the genotype of the ACE and PPARG polymorphisms between diabetic nephropathy and control patients.

Jayapalan, Muniandy [23] have conducted a research on the relationship between ACE genes and diabetic nephropathy disease among multiethnic Malaysian subjects. In their research 137 control (healthy) patients and 256 patients with diabetic were considered. Patients with diabetes have divided into normoalbuminuric, microalbuminuria, and End Stage Renal Failure (ESRF) based on the ratio of urinary albumin creatinine (ACR) and the ratio of glomerular filtration (GFR). They also included demographic and clinical characteristics. A number of statistical analysis tests were used such as (ANOVA) test, Chi-square test, Kruskal–Wallis test, Levenes test of homogeneity and Tukeys HSD and Dunnetts T3 post hoc test. Furthermore, to evaluate the significant of each predictor, multinomial logistic regression has been used too. The results of their research revealed that there is no significant existence of relation between the ACE gene polymorphism and both T2DM and/or diabetic nephropathy regardless of ethnicity and gender in Malaysian population.

Elhawary, Bogari [24] have investigated the association of ACE gene polymorphisms with nephropathy in the patients with diabetes type 1, in the population of Egypt. The goal of their investigation was among the children/adolescents in Egypt. In their investigation, they were considered 140 patients with diabetes type 1. Among 140 patients only 80 patients were with nephropathy and 60 patients were without nephropathy. They conducted statistical analyses to calculate Odds ratios and chi-square values. Moreover, they also investigated clinical characteristics for patients. According to their result they could not establish any important differences in the ACE gene distribution between the examined groups. They also did not find any association among the ACE gene and PstI polymorphisms with nephropathy in type 1 diabetes.

Marzbanrad, Hambly [25] have studied the relationship between Angiotensinogen Gene Polymorphism and Heart Rate Variability(HRV) in Diabetic control people. They used a small data set of, 231 Australian participants. Among

the participants some were with Type II diabetes (T2D) and some were normal. They divided all participants into eight groups, in the first six groups one of the ACE genotypes combined with either Type II diabetes (T2D) or normal. The remaining two groups are combined ID and II with Type II diabetes (T2D) and normal. A Statistical analysis has been conducted such as Kruskal–Wallis (KW) test, and Mann–Whitney–Wilcoxon (MWW). The results showed there is a significant different in some groups and there is no significant difference in other groups.

Moianu, Blaa [26] have investigated the association between cardiac autonomic neuropathy (CAN) disease and the complications of micro- and, macrovascular with risk factors in diabetes type 2 patients. The investigation included 149 participants, all of them with diabetes type 2, 91 patients were without CAN and 58 patients were with CAN. The assessment of CAN was based on Ewing battery tests. To compare the differences between groups, T- test was performed with continues variable, Chi- Square test was used with categorical variables. Logistic regression also used to assess the independent associations between CAN and predictors. Their study showed there is a significant associated between CAN and the complications of micro- and, macrovascular and risk factors mentioned in their study.

Gupta, Agrawal [27] have studied the role of ACE gene polymorphism in hypertensive among rural population of Haryana in India. In their study, they considered 106 patients and 110 control (healthy) participants. T-test, Chi-square, Hardy–Weinberg Equilibrium tests were conducted to calculate Allele frequencies. However, in their result they did not find any significant differences in the DD, II and I/D genotypes distribution among the two groups.

3 Materials and Methods

3.1 Dataset

For this study we have used data from the diabetes complications screening research initiative in Australia (DiScRi) project [28]. It contains a comprehensive collection of diagnosis tests that are related with CAN. More than 200 tests among 2000 patients are exist in DiScRi. The tests include an Ewing battery tests, ACE genotype, blood biochemistry features, retinal scans, electro cardiogram (ECG) recoding, heart rate variability (HRV) methods, peripheral nerve function plus the demographic data for each patient. The analysis and collection of the data have been done according to the role of the Ethics in Human Research Committee of the university [29]. The participants have been instructed to stop smoking, and drinks such caffeine and alcohol for 24 h before performing the tests. The participants have also requested to fast, starting from the mid-night before the day of tests. Time of performing tests started from 9:00 AM till 12 midday. The tests recorded various clinical data as well as other incident

Table 1. A comprehensive analysis

Study	Purpose of study	Results
Bhaskar, Mahin [22]	Investigated the role of ACE and PPARG genes with diabetic nephropathy in India.	No significant association
Jayapalan, Muniandy [23]	Investigated the relation between diabetic nephropathy disease and ACE genes among Malaysian population.	No significant association
Elhawary, Bogari [24]	Examine the association of ACE gene polymorphisms with nephropathy in the patients with diabetes type 1, in the population of Egypt	No significant association
Marzbanrad, Hambly [25]	Investigated the association between Heart Rate Variability (HRV) and ACE polymorphism.	A significant association
Moianu, Blaa [26]	Investigated the association between CAN disease and the complications of micro- and macrovascular with risk factors in diabetes type 2 patients.	A significant association
Gupta, Agrawal [27]	Investigated the role of ACE gene in hypertensive among rural population of Haryana in India.	No significant association
Our study	The relationship between CAN and ACE genotype, Cholesterol blood biochemistry and the demographic clinical data	No significant association

such as heart attack, palpitations and atrial fibrillation. Based on the ACE gene availability only 299 patients included in this study.

3.2 ACE Genotype (Angiotensin Converting Enzyme)

The variation in ACE gene is called ACE I/D polymorphism [30] as listed in Table 2. The polymorphism is located in the chromosome 17q23 in humans body [24]. It is characterized by insertion (I) allele or deletion (D) allele of a 287-base pair repeat within intron 16 of the gene [31]. Each person has two copies for each gene, either has two I alleles (II), two D alleles (DD), or one allele of each (ID). The DD pattern is the high level of angiotensin-converting enzyme. The ID pattern is the medium level.

Table 2. ACE I/D polymorphism [30]

ACE I/D polymorphism	Description
<i>II</i>	Two insertion (I) allele
<i>ID</i>	One insertion (I) allele and one deletion (D) allele
<i>DD</i>	Two deletion (D) allele

3.3 Cholesterol Blood Biochemistry and Ewing Battery Tests

Number of cholesterol blood biochemistry features were used in this study, as exhibited in Table 3. The Ewing battery tests were listed in Table 4 based on [8,9]. CAN disease was classified into five categories: *Normal*, *Early*, *Definite*, *Severe*, and *Atypical*. The rules to identify CAN categories were presented in Table 5 based on [8,9]. Some demographic clinical data also used, as presented in Table 6.

Table 3. The cholesterol blood biochemistry attributes

Notation	Description
HDL	High-density lipoprotein
LDL	Low-density lipoprotein
TC	Total cholesterol
Triglyceride	Triglyceride level in the blood
HbA1c	The level control of blood sugar over time

Table 4. The tests of Ewing Battery by [8,9]

Notation	Description	Normal	Borderline	Abnormal
LSHR	Lying to standing heart rate change	≥ 1.04	1.01–1.03	≤ 1.00
DBHR	Deep breathing heart rate change	≥ 15	11–14	≤ 10
VAHR	Valsalva manoeuvre heart rate change	≥ 1.21	1.11–1.20	≤ 1.10
HGBP	Hand grip blood pressure change	≥ 16	11–15	≤ 10
LSBP	Lying to standing blood pressure change	≤ 10	11–29	≥ 30

Table 5. The categories of CAN defined by [9]

Category	Test values
<i>Normal</i>	All tests negative or one borderline
<i>Early</i>	One of the three heart rate tests positive or two borderlines
<i>Definite</i>	Two or more of the heart rate tests positive
<i>Severe</i>	Two or more of the heart rate tests positive plus one or both blood pressure tests positive, or both borderline
<i>Atypical</i>	Any other combination of positive tests

Table 6. The demographic clinical data

Notation	Description
Gender	Male / Female
Age	No. of years
Waist circumference (WC)	Excess fat around the waist
BMI	Body Mass Index

3.4 Methods

A number of statistical tests have performed using R language and SPSS software as presented in the Table 7. Hardy Weinberg equilibrium was attained by Chi square test, to evaluate the Allele frequency distribution of ACE genotype. Student t-tests have used to evaluate continuous variables (expressed as mean \pm SD). Chi-square test was done to explore the relationship between two categorical variables (expressed as number and percentage). The Wilcoxon-Mann-Whitney test used with the non-parametric variables (expressed as median, range). Logistic regression was conducted to assess and identify the significant predictors of CAN in our study such as in [23,26]. This test will show the level of change in the predictor variable predicts compare to the outcome variable based on Odds ratio. We have considered the output is statistically significant while $P < 0.05$ with 95% confidence interval (CI). All P values reported for two tailed. The hypotheses were as follow:

- Null hypotheses: ACE genotype, the Cholesterol blood biochemistry attributes, the demographic clinical data have no association with CAN.
- Alternative hypothesis: ACE genotype, the cholesterol blood biochemistry attributes, the demographic clinical data have association with CAN.

Table 7. The statistical tests

Test Type	Variables
T- Test	TC, LDL
Wilcoxon-Mann-Whitney Test	LSHR, HbA1c, Triglyceride, HDL, WC, BMI, DBHR, VAHR, HGBP, LSBP, Age
Chi square Test	Gender
HardyWeinberg equilibrium	ACE genes
Logistic Regression	All variables

4 Experiments

In our experiments, we used data set of 299 patients. The individuals classified into two different categories based on the Ewing battery results such in Table 5. Then we divided the classified individuals into two groups. The first group was (CAN-), included all patients with normal category. The second group was (CAN+), consisted of all patients with early, definite, atypical and serve categories. Among 299 patients 132 individuals were with CAN- and 167 were with CAN+.

5 Results and Discussion

Our experimental results were as follow. Table 8. presented the comparison results between patients with and without CAN. The patients with CAN were significantly older in Age ($P = 0.004$). The HbA1c and Triglyceride were significantly higher ($P = 0.02$) and ($P = 0.03$). There was significant difference in gander. Females were higher while males were lower ($P = 0.008$). Patients with CAN have lower LSHR ($P = 0.000$), DBHR ($P = 0.000$), VAHR ($P = 0.000$), and HGBP rates ($P = 0.000$). However, the LSBP rate ($P = 0.002$) recorded higher. There were no significant differences recorded in WC, BMI, TC, LDL and HDL in the two groups. Table 9. showed the data pertaining to the all genotype distribution for both groups in Hardy Weinberg equilibrium. As we can see, the frequency of ID genotype was higher from the others genotype in both groups. It was observed that both II and DD genotype were very close. However, the differences were not significant statistically ($P > 0.05$).

Table 8. Comparison between the individuals with and without CAN

Variable	CAN-	CAN+	P value
Age	61.5(29–90)	67(32–87)	0.004
WC	96(63–130)	96(165-64)	0.65
BMI	26.5(13.3–44.6)	27.2(15.4–50)	0.2
LSHR	1.15(1.01–1.96)	1.09(0.87–1.66)	0.000
DBHR	17.13(4.5–36)	10.33(1.8–30.7)	0.000
VAHR	1.3(1.1–2.4)	1.1(0.9–1.7)	0.000
HGBP	17.5(0–39)	13(0–36)	0.000
LSBP	5.5(0–29)	8(0–35)	0.002
HbA1c	5.7(5–8)	5.8(4.8–13)	0.024
TC	5.1±0.98940	5.1±1.10195	0.93
Triglyceride	1 (0.2–4.3)	1.2(0.3–5.47)	0.03
LDL	3.1±0.97364	3±0.97803	0.59
HDL	1.4 (0.73–2.9)	1.4(0.7–3.3)	0.90
Gender F/M	67(50.8)/65(49.2)	111(66.5)/56(33.5)	0.008

Table 9. The distribution of ACE genotype

Population(n)	DD	ID	II	Allele Frequencies	
				D	I
CAN-(132)	32(41%)	66(47.8%)	34(41%)	0.492	0.508
CAN+(167)	46(59%)	72(52.2%)	49(59%)	0.491	0.509
Chi-Square(CAN- Vs. CAN+), X-squared = 1.4068, <i>P-value</i> = 0.4949					

5.1 The Result of Logistic Regression

In this experiment, a logistic regression was performed to ascertain the effects of attributes mentioned before. The results were presented in Table 10. As we can see from Table 10, Ewing battery predictors have a significant effect on the CAN output ($P < 0.05$). Increasing LSBP was associated with an increased likelihood of exhibiting CAN disease, but increasing LSHR, DBHR, VAHR, HGBP were associated with a reduction in the likelihood of exhibiting CAN disease. However, the other variables were not an effective predictive factor.

Table 10. Present the significant and odds ratio for each predictor in the model

Features	P value	Odds Ratio	95% C.I. for EXP(B) Lower-Upper
Gender F/M	0.359	1.563/0.640	0.602–0.246 /4.062–1.662
Age	0.658	0.992	0.956–1.029
WC	0.871	1.004	0.957–1.054
BMI	0.602	1.034	0.911–1.174
LSHR	0.011	0.017	0.001–0.400
DBHR	0.000	0.696	0.623–0.779
VAHR	0.000	0.018	0.000–0.006
HGBP	0.000	0.747	0.680–0.821
LSBP	0.003	1.106	1.034–1.183
HbA1c	0.854	1.058	0.577–1.940
TC	0.616	0.492	0.031–7.844
Triglycerid	0.312	2.040	0.511–8.140
LDL	0.585	2.178	0.133–35.581
HDL	0.303	4.766	0.244–92.899
ACEGenotype II	0.416	Ref	
ACEGenotype ID	0.297	1.811	0.593–5.529
ACEGenotype DD	0.958	1.030	0.343–3.095

6 Conclusion

In this study, we investigated the association between CAN and ACE genotype, Cholesterol blood biochemistry and demographic clinical data. The study showed that there was no a significant difference in ACE genotype between people with and without CAN. Furthermore, ACE genotype was not an important risk factor that can be used to predict CAN.

References

1. Vinik, A.I., Maser, R.E., Mitchell, B.D., and Freeman, R.: Diabetic autonomic neuropathy. *Diabetes Care* **26**(5), 1553–1579 (2003)
2. Maser, R.E., Mitchell, B.D., Vinik, A.I., Freeman, R.: The association between cardiovascular autonomic neuropathy and mortality in individuals with diabetes. *Diabetes Care* **26**(6), 1895–1901 (2003)
3. Maser, R.E., Lenhard, J.M., DeCherney, S.G.: Cardiovascular autonomic neuropathy: the clinical significance of its determination. *Endocrinologist* **10**(1), 27–33 (2000)
4. Vinik, A.I., Ziegler, D.: Diabetic cardiovascular autonomic neuropathy. *Circulation* **115**(3), 387–397 (2007)
5. Schumer, M.P., Joyner, S.A., Pfeifer, M.A.: Cardiovascular autonomic neuropathy testing in patients with diabetes. *Diabetes Spectr.* **11**(4), 227 (1998)

6. Hazari, M.A., Khan, R.T., Reddy, B.R., Hassan, M.A.: Cardiovascular autonomic dysfunction in type 2 diabetes mellitus and essential hypertension in a South Indian population. *Neurosciences (Riyadh)* **17**(2), 173–175 (2012)
7. International Diabetes Federation: IDF Diabetes Atlas [Internet]. 6th ed. (2013). Available from: <http://www.idf.org/diabetesatlas>
8. Ewing, D., Campbell, I., Clarke, B.: The natural history of diabetic autonomic neuropathy. *QJM* **49**(1), 95–108 (1980)
9. Ewing, D.J., Martyn, C.N., Young, R.J., Clarke, B.F.: The value of cardiovascular autonomic function tests: 10 years experience in diabetes. *Diabetes Care* **8**(5), 491–498 (1985)
10. Tesfaye, Solomon, Boulton, A.J.M., Dyck, P.J., Freeman, R., Horowitz, M., Kempler, P., Lauria, G., et al.: Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments. *Diabetes Care* **33**(10), 2285–2293 (2010)
11. Stranieri, A., Abawajy, J., Kelarev, A., Huda, S., Chowdhury, M., Jelinek, H.F.: An approach for ewing test selection to support the clinical assessment of cardiac autonomic neuropathy. *Artif. Intell. Med.* **58**(3), 185–193 (2013)
12. Ewing, D., Campbell, I., Murray, A., Neilson, J., Clarke, B.: Immediate heart-rate response to standing: simple test for autonomic neuropathy in diabetes. *Br. Med. J.* **1**(6106), 145–147 (1978)
13. Giebisch, G., Windhager, E.: Integration of Salt and Water Balance. *Medical Physiology*, pp. 861–876. Saunders, New York (2003)
14. Crisan, D., Carr, J.: Angiotensin I-converting enzyme: genotype and disease associations. *J. Mol. Diagn.: JMD* **2**(3), 105 (2000)
15. Dhar, S., Ray, S., Dutta, A., Sengupta, B., Chakrabarti, S.: Polymorphism of ACE gene as the genetic predisposition of coronary artery disease in Eastern India. *Indian Heart J.* **64**(6), 576–581 (2012)
16. Tired, L., Blanc, H., Ruidavets, J.-B., Arveiler, D., Luc, G., Jeunemaitre, X., Tichet, J., Mallet, C., Poirier, O., Plouin, P.-F.: Gene polymorphisms of the renin-angiotensin system in relation to hypertension and parental history of myocardial infarction and stroke: the PEGASE study. *J. Hypertens.* **16**(1), 37–44 (1998)
17. Witzel, I.-I., Jelinek, H.F., Khalaf, K., Lee, S., Khandoker, A.H., Alsafar, H.: *Front. Endocrinol.* **6** (2015)
18. Zhou, Y.-F., Yan, H., Hou, X.-P., Miao, J.-L., Zhang, J., Yin, Q.-X., Li, J.-J., Zhang, X.-Y., Li, Y.-Y., Luo, H.-L.: Association study of angiotensin converting enzyme gene polymorphism with elderly diabetic hypertension and lipids levels. *Lipids Health Dis.* **12**(1), 187 (2013)
19. Cho, B.H., Yu, H., Kim, K.-W., Kim, T.H., Kim, I.Y., Kim, S.I.: Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artif. Intell. Med.* **42**(1), 37–53 (2008)
20. Simon, A.C., Holleman, F., Gude, W.T., Hoekstra, J.B., Peute, L.W., Jaspers, M.W., Peek, N.: Safety and usability evaluation of a web-based insulin self-titration system for patients with type 2 diabetes mellitus. *Artif. Intell. Med.* **59**(1), 23–31 (2013)
21. Wang, F., Fang, Q., Yu, N., Zhao, D., Zhang, Y., Wang, J., Wang, Q., Zhou, X., Cao, X., Fan, X.: Association between genetic polymorphism of the angiotensin-converting enzyme and diabetic nephropathy: a meta-analysis comprising 26,580 subjects. *J. Renin-Angiotensin-Aldosterone Syst.* **13**(1), 161–174 (2012)
22. Bhaskar, L.V., Mahin, S., Ginila, R.T., Soundararajan, P.: Role of the ACE ID and PPARG P12A polymorphisms in genetic susceptibility of diabetic nephropathy in a South Indian population. *Nephro-Urol. Mon.* **5**(3), 813 (2013)

23. Jayapalan, J.J., Muniandy, S., Chan, S.P.: Null association between ACE gene I/D polymorphism and diabetic nephropathy among multiethnic Malaysian subjects. *Indian J. Hum. Genet.* **16**(2), 78 (2010)
24. Elhawary, N.A., Bogari, N., Rashad, M., Tayeb, M.T.: Null genetic risk of ACE gene polymorphisms with nephropathy in type 1 diabetes among Egyptian population. *Egypt. J. Med. Hum. Genet.* **12**(2), 187–192 (2011)
25. Marzbanrad, F., Hambly, B., Ng, E., Tamayo, M., Matthews, S., Karmakar, C., Khandoker, A.H., Palaniswami, M., Jelinek, H.F.: Relationship between Heart Rate Variability and angiotensinogen gene polymorphism in diabetic and control individuals. In: 36th Annual International Conference of the IEEE. Engineering in Medicine and Biology Society (EMBC), IEEE, pp. 6683–6686, (2014)
26. Moianu, A., Blaa, R., Voidzan, S., Bajk, Z.: Cardiovascular autonomic neuropathy in context of other complications of type 2 diabetes mellitus. *BioMed Res. Int.* **2013** (2013)
27. Gupta, S., Agrawal, B.K., Goel, R.K., Sehajpal, P.K.: Angiotensin-converting enzyme gene polymorphism in hypertensive rural population of Haryana, India. *J. Emerg. Trauma Shock* **2**(3), 150 (2009)
28. Jelinek, H.F., Wilding, C., Tinely, P.: An innovative multi-disciplinary diabetes complications screening program in a rural community: a description and preliminary results of the screening. *Aust. J. Prim. Health* **12**(1), 14–20 (2006)
29. Cornforth, D., Jelinek, H.: Automated classification reveals morphological factors associated with dementia. *Appl. Soft Comput.* **8**(1), 182–190 (2008)
30. Purnamasari, D., Widjojo, B.D., Antono, D., Syampurnawati, M.: ACE gene polymorphism and atherosclerotic lesion of carotid artery among offsprings of type 2 diabetes mellitus. *Diabetes* **17**, 18 (2012)
31. Moreira, S., Nbrega, O., Santana, H., Sales, M., Farinatti, P., Simes, H.: Impact of ACE I/D gene polymorphism on blood pressure, heart rate variability and nitric oxide responses to the aerobic exercise in hypertensive elderly. *Revista Andaluza de Medicina del Deporte* (2016)

Modeling of Consumer Interest on E-commerce Products Using Eye Tracking Methods

Juni Nurma Sari^{1,2(✉)}, Lukito Edi Nugroho¹, P. Insap Santosa¹,
and Ridi Ferdiana¹

¹ Department of Electrical Engineering and Information Technology,
Universitas Gadjah Mada, Jogjakarta, Indonesia
juni.s3tel4@mail.ugm.ac.id, {lukito, insap, ridi}@ugm.
ac.id

² Department of Informatics Engineering, Politeknik Caltex Riau, Pekanbaru,
Indonesia

Abstract. In this decade, e-commerce is one of media to conduct sale and purchase transactions. The seller must maintain e-commerce service so that consumer will comfort to shop. One of the services is provide product that consumer interested. There are many ways to get consumer interest in product. Give rating to product or using click-stream data. But both need the interaction of consumer to click product rating or to click product that consumer interested. With the development of sensor technology, consumer interest in e-commerce products can be collected by eye tracking method. Eye tracking method is one way to get consumer interest in product through attention, without requiring consumer interaction with the system. The model of consumer interest in product uses time until first fixation, fixation count, and fixation duration to measure whether the object is attractive and whether the consumer is interested in the product. The measurement variable based on Aga Bojko taxonomy. The value of variable is displayed in graphical form because in graphical form the analysis of consumer interest in e-commerce product more easily to be done. Implementation of modeling consumer interest uses Ogama, eye tracking software analysis by adding a feature of graphic of consumer interest. In the graph, we can see which products are interesting and which products are preferred by consumers. The contribution of this research is the modeling of consumer interest in the product and some procedure to add graphic feature in eye tracking software analysis, ogama.

Keywords: Consumer's interest · Eye tracking · Time to first fixation
Fixation count · Fixation duration

1 Introduction

Shopping through e-commerce is not a new thing because shopping with e-commerce makes it easy for users, and it is also efficient. Various technologies are implemented to provide convenience in online shopping, including personalization to display products

that are often purchased, products by the wishes of consumers and products purchased together with products that have been purchased by consumers. Another technology is cloud computing for storage of product data, transaction data and consumer data. Another thing to consider in shopping with e-commerce is attractive web design, attractive product visual. Those are ways to maintain consumer loyalty and can make consumers more interested in exploring the site.

Consumer interest can be seen from the eye view of something. Eye tracking can capture consumer interest in a particular product. Previously, consumer interest in a product can be detected from the duration when consumer views a page. Duration data can be obtained through server logs, but the data of consumer interest in the product can't be captured yet. The eye tracking technology can capture consumer interest in certain products on the page through consumer attention.

Eye Tracking is a method for knowing the consumer's attention by taking consumer view data, to measure how long the user sees something interesting and to know the eye movements of the user. According to Horsley [1], the eye tracking method is the collection of data by using a visual system that recognizes the attention, cognitive and user behavior. Eye tracking method has been used in various studies in the field of Psychology, Medicine, Learning, Marketing and Web Design. In the field of psychology the eye tracking method is used to identify the user's attention through a virtual agent, which can be in the form of a human or another form [2]. In the field of health, used to evaluate the reading ability of Nystagmus sufferer vision deficiency in the eyes [3]. In learning field, this method has been used to identify students with visual learning techniques and verbal learning techniques [4]. In marketing field, the use of eye tracking method in the most popular marketing field, which is used in digital marketing and online advertising [5].

Consumer interest in product is essential data for marketing department because it can tell which product that consumers like. The data of consumers interest can be collected through eye tracking study. The next process is analyzing consumer interest data with taxonomy analysis methods from Aga Bojko [6]. There are two measurements on the taxonomy, attraction measurement and performance measurement. This study uses attraction measurement. The attraction measurement is divided into three measurements: the measurement of whether the product is attractive, the measurement of consumer interest in the product and the consumer emotional measurement of the product. This study focuses on modeling whether the product is attractive and modeling consumer interest.

The implementation of consumer interest model using software for eye tracking analysis Ogama to collect consumer interest data. Using Ogama, the eye tracking data can be collected directly when consumer shop in e-commerce. Unfortunately, there is no facilitate to display consumer interest data in graphic. This research also develops a function of Ogama to display consumer interest data in graphic. With graphical data of consumer interest make marketing more easily to find the product that consumer like. The experiment of this research using selva-house.com website, the muslimah dress shop with participants of five female students.

Discussion of consumer interest in e-commerce product by using eye tracking data is divided into several sections: Sect. 2 discusses e-commerce and consumer interest; Sect. 3 discusses eye tracking, analytical methods for eye tracking data and modelling of consumer interest in e-commerce products using eye tracking data; Sect. 4 discusses experiments, and experimental results and discussion and Sect. 5 is conclusions.

2 Consumer Interest on E-Commerce Products

2.1 E-Commerce

Technological developments make it possible to perform various activities electronically, such as email or chat; discussion in a forum. Likewise, shopping activities can be done through e-commerce. E-commerce is a combination of technology, applications and business processes related to companies, consumers and communities through electronic transactions and exchange of goods, services and information electronically [7]. E-commerce should pay attention to attractive website appearance, attractive product photos and the reliability of the e-commerce system itself for easy access. The packaging of attractive products and attractive photos of the product will make consumers interested in the product. By uncovering the attention of consumers to specific products, the products favored by consumers can be identified.

2.2 Consumer Interest in the Product

Currently, consumer interest can be seen when consumers provide an assessment of the product by rating [8] on the product. The rating is made when the consumer has made a transaction. Another way to find out consumer interest is by knowing consumers' behavior when browsing, that activity is recorded in the weblog [9]. Both ways have weaknesses. The rating of the product can only be known by the system after the consumer has made a transaction. While the use of weblog only identifies which page is preferred by consumers, by using the data duration of the visit to the page, but which product is preferred can be identified by clicking the mouse.

For marketing purposes, rapid detection of consumer interest in products is helpful. As the development of sensor technology, consumer interest can be known by using the consumer's eyes. The technology is called eye tracking technology.

3 Eye Tracking

Consumer interest in e-commerce product can be detected using eye tracking technology. To find out whether the consumer is interested in a product can be known from the consumer's attention to an object on a web page. The tool used to get the user's attention is the eye tracker. This tool is placed in front of the monitor so that the user's view of the screen can be captured by the eye tracker. Figure 1 is an example of a consumer browsing e-commerce with an eye tracker to record the consumer's eye view.



Fig. 1 Recording user's attention data with eye tracker

Consumer view data on e-commerce will be captured on the screen coordinates (x, y) stored in raw data of eye tracker and then the data is processed and it becomes fixation duration data. The data is visualized as a dot in a particular area that describes the consumer's view of that area. That area is Area of Interest (AOI). AOI is defined when the site already exists. Usually, AOI defined with a square or ellipse form that covers a product, describes in Fig. 2b. The number of dots in AOI is how many times user sees images or text on an AOI. Figure 2a shows the visualization of fixation data.

Each eye tracker has software to analyze the user's attention, such as Eye Tribe using eye proof software and Gaze point has software with several levels: standard level, UX Edition level and Professional Edition level. Eye tracking software analysis is also developed by researcher Adrian Voßküller of Freie Universität Berlin which is Ogama. Ogama is open source software. This study uses ogama software and develops facilities to create graphs to show the amount of consumer attention to an area.

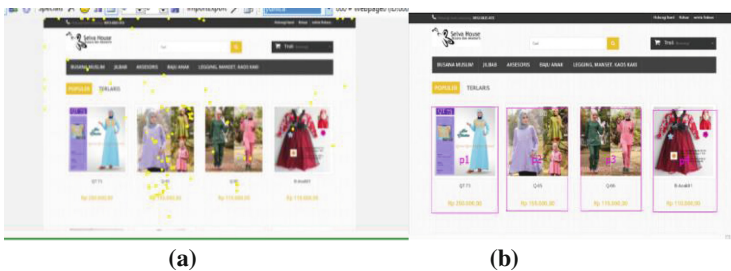


Fig. 2 a Data fixation visualization and b Example of setting AOI

3.1 Eye Tracking Methods for Consumer Interest Analysis of Products

Analysis of consumer interest in e-commerce product in this study uses taxonomy from Aga Bojko that is attraction measurement. The taxonomy of attraction measurement shown in Fig. 3. Attraction measurement is the measurement of the user's interest in the object. These measurements is divided into three measurements: area noticeable measurement, area interest measurement and emotional arousal measurement. Each measurement has a metric.

Area noticeable measurement identify areas that are easily detected by users using three metrics. This research uses only one metric that is the time of the first fixation. Area interest measurement to find out how long the consumer views the area, using three metrics as well. This study uses two metrics, the number of fixation on the AOI (fixation count) and the total dwell time in fixation duration in the AOI (fixation duration). Emotional arousal measurement to know the emotions of the consumer when viewing the area whether happy or casual [10]. The metric is pupil diameter. Pupil diameter will enlarge if the user is happy. This study does not use this measurement.

3.2 Modeling Consumer Interest in E-commerce Product Using Eye Tracking Method

In this research, modeling of consumer interest is applied to e-commerce products, using three metrics to find out products with attractive appearance and products favored by consumers which is time until first fixation, fixation count, and fixation duration. The detail of metric is below. The Consumer Interest model can be seen in Fig. 4.

1. Fixation Count (FC), the number of fixations on AOI:
The higher value means the participant sees a lot in the area, such as looking at the product
2. Fixation Duration (FD), duration of fixation on AOI:
The higher value means participant focus on that product, regarding the deeper study, regarding interested consumers
3. Time to First Fixation (TFF), time calculation when viewing the first AOI:
The smaller value means that the product is quickly found by the consumer, it means that on the product component there is something interesting, whether it is color, body language model and so forth

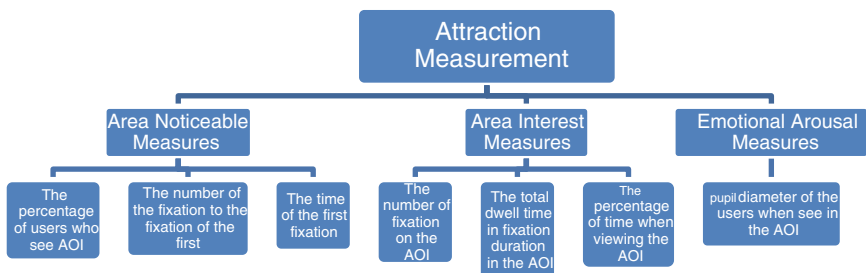


Fig. 3 Taxonomy Aga Bojko to measure interest towards certain objects

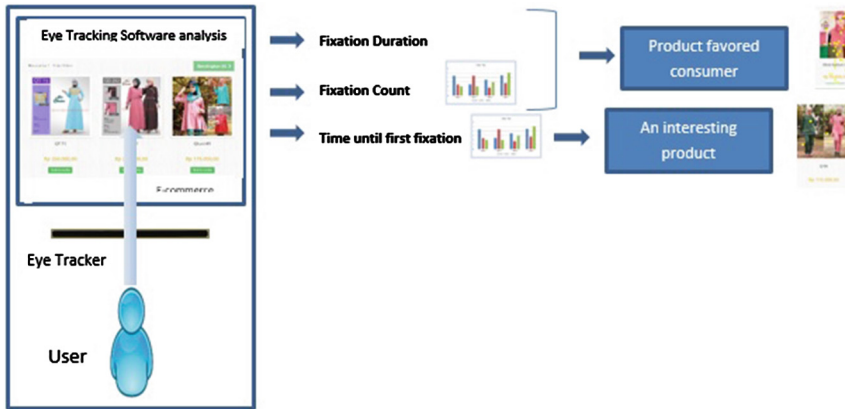


Fig. 4 Consumer interest modeling in e-commerce products

The model implemented in the prototype uses Ogama's eye tracking software analysis by adding features to display graphics on metric FC, FD, and TFF. The graph shows FC, FD, and TFF on each defined AOI, so that the highest AOI can be known, which is the most preferred product of the most interesting product. Graphs can be seen in Figs. 5, 6 and 7.

3.3 Ogama Software Development for Consumer Interest Graphic

Eye tracking software analysis Ogama used for eye tracking study to know how much the value of consumer interest in e-commerce product. Some metrics of consumer interest in product are TFF, FC, and FD. That metric is provided in Ogama in number form. It is difficult for decision makers to see number format of consumer interest in e-commerce product. Even though, consumer interest data is essential to set a strategy to market the product. To ease decision makers reading consumer interest data, so the features that are showing consumer interest data in graphical form is developed.

Procedure to develop feature data charts for consumer interest data are described below:

1. Identified classes in ogama that support feature data chart:
 - a. AOIStatistic, class for saving value of metric calculation in Area of Interest
 - b. VGElementCollection, class for saving the position of Area of Interest
 - c. FormWithTrialSelection, class interface to get data from ogama database
 - d. Statistic, class that contains method to calculate statistic data
 2. Add Graph module and a new Form class named Grafik.cs to implements Ogama. Modules.Common.FormTemplates.FormWithTrialSelection
 3. Add Graphic to ComboBox Form for subject (participant) and to ComboBox Form for consumer interest metric
 4. Get participant data from database and display it in ComboBoxSubject
 5. Get trial data from database one TrialId is a page that visited
 6. Count the consumer interest metric for each TrialId based on subject and metric.
- The procedure below:

- a. Get AOI data from database for each TrialId and save with Document.ActiveDocument.DocDataSet.AOIsAdapter.GetDataBy-TrialID function in VGElementCollection object
 - b. Get GazeFixation from database according to chosen subject and TrialID using method Document.ActiveDocument.DocDataSet.GazeFixationsAdapter.GetDataBySubjectAndTrialID
If all subject is chosen, the method that used is Document.ActiveDocument.DocDataSet.GazeFixationsAdapter.GetDataByTrialID
 - c. Count consumer interest metric for each AOI using method Statistics.Statistic.CalcAOIStatistic, Gaze Fixation data AOI data. The result is saved with AOI name as a keyword and saved in dictionary in AOIStatistic object
7. Get all data according to chosen metric to display in graphical form. The procedure is below:
- a. Fixation Count Graphic
Get data nameAOI and AOIStatistic.FixationCount from dictionary added to datapoint in graphical function. nameAOI as data X and AOIStatistic.FixationCount as Y data
 - b. Fixation Duration Graphic
Get data namaAOI and data AOIStatistic.SumOfTimeOfAllFixations from dictionary added to datapoint in graphical function. nameAOI as data X and AOIStatistic.SumOfTimeOfAllFixations as Y data.
 - c. Time until First Fixation Graphic
Get data namaAOI and data AOIStatistic.HitTimes [1] from dictionary added to datapoint in graphical function. nameAOI as data X and AOIStatistic.HitTimes [1] as Y data.
 - d. If the result is 0, then it is not displayed
8. Add EventHandler on ComboBox that contain the subject and on ComboBox that contains consumer interest metric to repeat the entire process.

4 Experiments of Consumer Interest on Products Using the Eye Tracking Method

This study conducted an experiment to determine consumer interest in e-commerce products. Selva-house.com is an e-commerce that will be attempted by participants to shop online. Selva-house is a website that sells muslimah clothes, accessories, children's clothes, and hijab. Participants were five local private college students. Eye tracker that used the eye tribe with Ogama eye tracking analysis software that is added to display consumer interest in graphical form.

4.1 Experiment

The experimental scenario of eye tracking study is participants are asked to browse on selva-house.com and shop as usual. From viewing then selecting products and filling

out shopping carts. Before filling up the shopping cart, users are prompted to log in firsthand. After completing the shopping cart, user is prompted to logout. And the data will be processed and display in graphical form. Graphic shows data FC, FD and time to first fixation on each AOI. So, the graphs can show which products look attractive and which products are preferred by consumers.

4.2 Experiment Results and Discussion

The metric of consumer interest can be seen in graphical form by click show graph module menu in ogama. The metrics are FC, FD, and TFF. The graphical form is in two type. The first one represent consumer interest data of all participants. And the second one represent consumer interest each participant. The x-axis indicates AOI code. AOI code contains category on front letter, and the number represents product.

Examples such as the first product of hijab category the code is j1, the second hijab product has code j2 and so on. The y-axis indicates the value of metric. Figure 5 describes the value of FC of five participants. Figure 6 describes the value of FD of five participants. And Fig. 7 describes the value of TFF.

The analysis of consumer interest is easier when the data is displayed in graphical form. At Figs. 5, 6 and 7 there are thirteen AOI that seen by consumer. And the categories are popular product and accessories. Based on graphical data the AOI product that consumers like is on Table 1 and the product shown at Fig. 8.

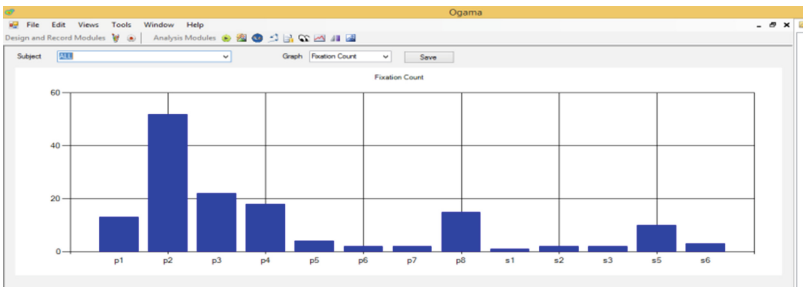


Fig. 5 FC graphic of all participant

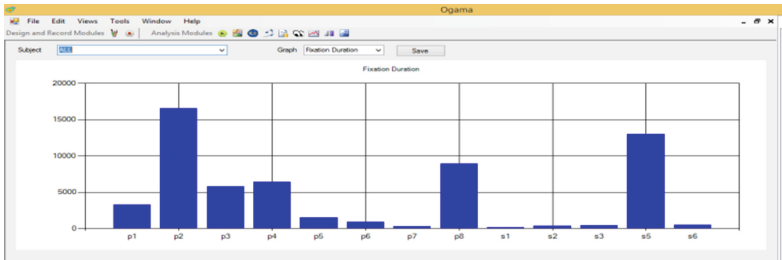


Fig. 6 FD graphic of all participant

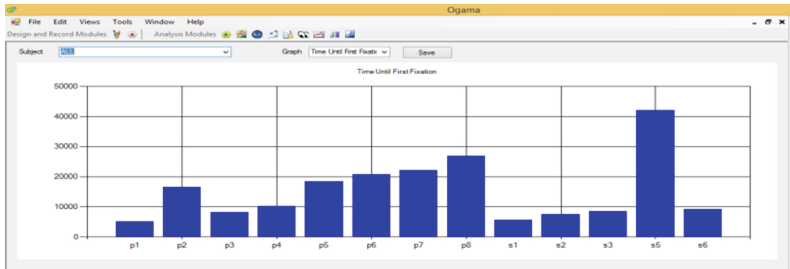


Fig. 7 TFF graph of all participant

Table 1 Value of FD, FC, TFF of interest product

AOI	FD	FC	TFF
p2	16000	50	17000
s5	13000	10	42000
p8	9000	15	26000
p4	6000	19	10000
p3	5000	22	9000

From Table 1, based on the value of FD, FC and TFF can be analyzed:

- AOI p2 is an AOI product with code Q-05. This product is a long-seen, widely seen but not quickly found. It means consumer interested in product and the consumer seeing a lot in part of product but the product could be not attractive.
- AOI s5 is an AOI product with code Acc-Bando-B01. This product is a long-seen, little views but needs a long time to found. It means consumer interested in product and only views the little part of product, but the product could be not attractive.
- AOI p8 is an AOI product with code Jilbab-Rabbani-05. This product is a long-seen, little views but needs a long time to found. It means consumer interested in product and only views the little part of product, but the product could be not attractive.
- AOI p4 is an AOI product with code B-Anak01. This product is a quite long-seen, rather many views and no need long time to found. It means consumer rather interested in product and view some part of product but the product is quite attractive.
- AOI p3 is an AOI product with code Q-06. This product is a quite long-seen, rather many views and no need long time to found. It means consumer rather interested in product and view some part of product but the product is quite attractive.

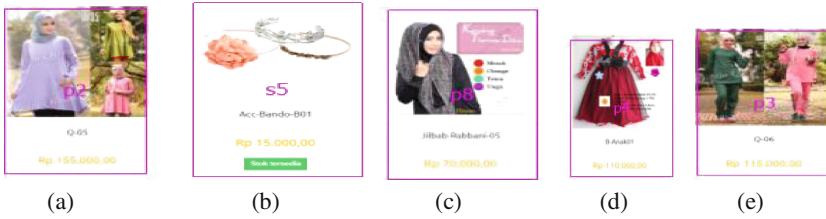


Fig. 8 a Q-05 b Acc-Bando-B01 c Jilbab-Rabbani-05 d B-Anak01 e Q-06

Some researchers discuss decision-making process to select a product to purchase. According to Gidlof [11], the process of product selection is divided into three phases, namely orientation, evaluation, and verification. The orientation process starts from the first fixation on the product rack to the first fixation on product. The evaluation phase starts from first fixation on product to last fixation on product. The verification phase starts from last fixation on product to last fixation on product rack. First fixation can be known through time to first fixation (TFF) metrics. In the experimental, the orientation phase occurs at start time on page until time to first fixation on AOI product. The evaluation phase begins with. TFF on the product, during the fixation duration (FD) with the amount of fixation (FC). The verification phase starts at TFF on product plus FD until end of time on page.

Based on that phase the right metric to show consumer interest in e-commerce product is FD. FC indicates an evaluation process on the product. TFF represented when the product was first viewed. Based on taxonomy Bojko TFF shows interesting product or not. According to Chandon [12], purchase intention is influenced by memory based factor and attention-based factor. Memory-based factor is the impression consumers to the product can be a product experience or purpose to shop for products. Attention-based factor is the consumer's interest in the product. So TFF on product is influenced by memory based factor and attention-based factor. When shopping, there are consumers who already have a purpose to shop or have a liking for a particular brand or something that has a favorite in certain product categories. So if the TFF value is small, it means the product is immediately found by the consumer, means the consumer needs the product or like in that category. Such as AOI products s1, s2, s3, s4. Consumers like the accessories category, so popular product is not seen first. The destination is the category of accessories. In the AOI product s5, TFF value is large, but FD value is large. Indicates that consumers do searching process for some products then interested in AOI product s5.

5 Conclusion

Research on modeling consumer interest to the product on e-commerce using eye tracking method can be concluded that:

1. Attraction measurement from taxonomy of Aga Bojko can be used not only to measure consumer interest in product but also to detect which product is interesting.

2. The addition of graphic facilities makes it easy to analyze the consumer's interest in the product, so it can be known which products are most preferred by consumers also known which products are interesting.
3. From decision making process to select product that consumer wants, TTF is the beginning of evaluation phase, FC is the process of evaluating product and FD is time that used to evaluate process. If FD is big then consumer have cognitive process

Acknowledgements. This work was supported by RisteDikti of the Ministry of Research and Higher Education of the Republic of Indonesia under Doctoral Research Grant.

References

1. Horsley, M., Eliot, M., Knight, B.A., Ronan, R.: Current Trends in Eye Tracking Research (2014)
2. Courgeon, M., Rautureau, G., Martin, J.-C., Grynszpan, O.: Joint attention simulation using eye-tracking and virtual humans. *IEEE Trans. Affect. Comput.* **3045**(3), 1–1 (2014)
3. Shahimin, M., Mohammed, Z., Saliman, N., Mohamad-Fadzil, N., Razali, N., Mutalib, H., Mennie, N.: The use of an infrared eye tracker in evaluating the reading performance in a congenital nystagmus patient fitted with soft contact lens : a case report. In: Horsley, M., Eliot, M., Knight, B.A., Ronan, R. (eds.) *Current Trends in Eye Tracking Research*, pp. 123–128. Springer (2014)
4. Mehigan, T.J., Barry, M., Kehoe, A., Pitt, I.: Using eye tracking technology to identify visual and verbal learners, Department of Computing, Maths & Physics, Waterford Institute of Technology, Ireland. *IEEE* (2011)
5. Borys, M.: Eye Tracking in Marketing Research: a Review of Recent Available Literature, pp. 939–941 (2014)
6. Bojko, A.: Eye Tracking the User Experience. Rosenfeld (2013)
7. Purbo, O., Wahyudi, A.: E-Commerce, Elex Media Komputindo (2001)
8. Zhao, X., Niu, Z., Chen, W.: Interest before liking: two-step recommendation approaches. *Knowl.-Based Syst.* **48**, 46–56 (2013)
9. Su, Q., Chen, L.: A method for discovering clusters of e-commerce interest patterns using click-stream data. *Electron. Commer. Res. Appl.* **14**(1), 1–13 (2015)
10. Sari, J.N., Nugroho, H.A., Nugroho, L.E., Santosa, P.I., Ferdiana, R.: A study on algorithms of pupil diameter measurement. In: *Proceeding of 2016 2nd International Conference on Science and Technology-Computer (ICST)*, pp. 0–5 (2016)
11. Gidlöf, K., Dewhurst, R., Holmqvist, K.: Using eye tracking to trace a cognitive process : gaze behaviour during decision making in a natural environment. **6**(1), 1–14 (1995)
12. Chandon, P., Hutchinson, J.W., Bradlow, E.T., Young, S.H.: Measuring the value of point-of-purchase marketing with commercial eye-tracking data. *INSEAD Bus. Sch. Res. Pap.* (22), 55 (2007)

Part II

Data Mining

A New Concept of Fuzzy TOPSIS and Fuzzy Logic in a Multi-criteria Decision

Ratih Fitria Jumarni^(✉) and Nurnadiah Zamri

Faculty of Informatics and Computing, University Sultan Zainal Abidin, Tembil
Campus, 22200 Besut, Terengganu, Malaysia
ratihfitrial993@gmail.com

Abstract. In reality, humans usually uncertainty or vague in expressing their preference or votes based on crisp number or scale. Much of the information on which decision is based is uncertain, the methods can be used to support the system's decision is to use Fuzzy Multi-Criteria Decision Making (FMCDM). This method was chosen because it can selecting the best alternative from a number of alternatives Criteria. Fuzzy Multi-Criteria Decision Making (FMCDM) is a method of decision-making to determine the best alternative from a number of alternatives based on certain criteria. The criteria usually in the form of action, rules or standards used in decision making. The lack of capability to handle vagueness in the decision making, has been main weakness of Fuzzy TOPSIS. Thus, the purpose of this paper is to introduce Fuzzy TOPSIS and z-number to several criteria fuzzy group decision making (FMCDM). Fuzzy TOPSIS is used to determine the alternative most suitable in relation to different selection criteria and z-number to present experts reability, this method can choose the best alternative from a number of alternatives based on some specific criteria. A numerical example on FMCDM is used to describe the efficiency of the proposed method.

Keywords: Fuzzy TOPSIS · Z-number · Multi-criteria decision-making

1 Introduction

Uncertainties, ambiguities and doubtless, pose a big challenges to any decision-making system. Uncertainties are not just limited to the linguistics evaluation where uncertainties can be referred to as a situation where the current state of knowledge is such that the order or nature of things is unknown or vague [1]. Humans have a capability to make rational decisions based on information which is uncertain, imprecise and/or incomplete [2]. Formalization of this capability, at least to some degree, motivates to the used of fuzzy theory in this paper. Fuzzy theory provides the flexibility to represent the imprecise/vague information resulting from the lack of knowledge/information [3–5].

Fuzzy Multi-Criteria Decision Making (FMCDM) is the process of finding the best option from all of the feasible alternatives [6] and is one of the widely used approaches to deal with a number of uncertainties problems. This approach often requires the experts to provide qualitative and quantitative measurements for determining the

performance of each alternative with respect to the criteria and judgments [7]. One of the related research area in FMCDM is Fuzzy Technique for Order Preference by Similarity to Ideal Solution (FTOPSIS) [8–10].

Based on the previous studies by [11–13], FTOPSIS has been rigorously applied in the fields of medical, environmental and etc. However, FTOPSIS is still lacked in considering the reliability of the information that has been provided. Due to that, this paper introduced the Z-numbers concept. Z-numbers are seen has more ability to describe the knowledge of human. It can describe both restraint and reliability [14]. A Z-number is an ordered pair of fuzzy numbers (\tilde{A}, \tilde{R}) , where \tilde{A} is a value of some variable and \tilde{R} represents an idea of certainty or other closely related concept such as sureness, confidence, reliability, strength of truth, or probability [15]. Therefore, the proposed of this study is to integrate between the concepts of FMCDM with Z-numbers based decision making problems.

The rest of this paper is organized as follows; Sect. 2 presents the required preliminaries and basic concept of Z-numbers. Section 3 proposes a method of FTOPSIS with Z-numbers. Section 4 provides an example to demonstrate the feasibility and applicability of the proposed method. Lastly, a conclusion is presented in Sect. 5.

2 Preliminaries

In the following, we briefly describe the fundamental theories that are involved in this study.

Definition 1 [16, 17] Let F be a Universe of discourse. Where A is a fuzzy subset of F ; and for all $f \in F$, there is a number $[0, 1]$ which is assigned to represent the membership degree of f in A , and is called membership function of A . The definition of FSs based on [zadeh 1965] [5] is:

Definition 1.1 [5] Let X be a space of points (object), with a generic element of X denoted by x . Thus, $X = x$. A fuzzy set (class) A in X is characterized by a membership (characteristic) function $f_A(x)$ which associates with each points in X a real number in the closed interval $[0, 1]$ with the value of $f_A(x)$ at x representing the “grade of membership” of x in A . Specifically, a fuzzy set on a classical set is defined as follows: $A = \{(x, \mu_A(x)) | x \in X\}$

Thus, the nearer the value of $f_A(x)$ to unity, the higher the grade of membership of x in A . When A is a set in the ordinary sense of the term, its membership function can take on only two values 0 and 1, with $f_A(x) = 1$ or 0 according as x does or does not belong to A . (When there is a need to differentiate between such sets and fuzzy sets, the seats with two-valued characteristic function will be referred to as ordinary sets or simply sets).

Definition 1.2 [5] The support of a fuzzy set A , $S(A)$ is the crisp set of all $x \in X$ such that $\mu_A(x) > 0$.

Definition 1.3 [5] The (crisp) set of elements that belongs to the fuzzy set A at least to the degree α is called the α level set: $A_\alpha = \{x \in X | \mu_A(x) \geq \alpha\}$

$A_\alpha = \{x \in X | \mu_A(x) \geq \alpha\}$ is called “strong α —level set” or “strong α cut”.

Definition 1.4 [5] A fuzzy set A is convex if,

$\mu_A = (\lambda x_1 + (1 - \lambda)x_2) \min\{\mu_A(x_1), \mu_A(x_2)\}, x_1, x_2 \in X, \lambda \in [0, 1]$ Alternatively, a fuzzy set is convex if all α -level sets are convex.

Definition 1.5 [5] For a finite fuzzy set A , the cardinality $|A|$ is

$$\text{defined as: } |A| = \sum_{x \in X} \mu_A(x)$$

$\|A\| = \frac{|A|}{|X|}$ is called the relative cardinality of A . The relative cardinality of a fuzzy set has to be in the same universe when making a comparison of fuzzy sets by their relative cardinality.

Definition 2 A Z-number is an ordered pair of fuzzy numbers denote as $Z = (A, B)$, with the first component A , a restriction on the values, is a real-valued uncertain variable. The second component B is a measure of reliability for the first component. The z-numbers concept $Z = (A, B)$ is intended to provide basic calculations with numbers that are not really reliable. Z-numbers can be used to represent information about the variable uncertainty of the type where A represents the value of the variable X , and the second component, B represents the idea of certainty or probability such as the concept of reliability, certainty, self-confidence, strength of trust and possibilities.

Example [5]:

(anticipated budget deficit, about three million USD, likely);

(price of oil in the near future, significantly over 50 dollars/barrel, very likely). In Sect. 2 we present required preliminaries and basic concept of z-numbers. Next Sect. 3 develops the proposed method of multi-criteria decision making using FTOPSIS with z-numbers.

3 The Proposed Method

In this section, the proposed steps involved the whole process simultaneously in order to determine the fuzzy output. This phase includes the FTOPSIS and Z-number. Basically, the concepts of Z-number have capability to represent the reliability of experts to decision making evaluation. A better step has been proposed to guide FMCDM step-by-step starting from the construct a hierarchical diagram, and then scaling the relative of data, until defined the fuzzy output. The whole steps of the proposed FTOPSIS with z-number method is described as follows:

STEP1: Construct a hierarchical diagram of FMCDM problem. The hierarchical diagram of FMCDM problem is constructed. Data for criteria and alternatives must be identified as part of a FMCDM problem. In FMCDM problems, responses from experts are mainly focused on the opinion of the experts regarding rating of the attributes of the problems based on the identified criteria. The experts are asked to specify rating using seven of the linguistic scales varying from Table 1 over the factors associated with FMCDM problems.

STEP2: Scaling the relative of data. In FMCDM problems, responses from experts are mainly focused on the opinion of the experts regarding rating of the attributes of the problems based on the identified criteria. Thus, the proposed method can identify and

aggregate the different opinions of experts with varying influence degrees to suggest the final solution.

Table 1 is used in representing the importance of criteria and the rating of the alternative [18]. In addition to that, Table 2 is proposed to represent the reliability of the experts and to identify the alternative level for the consequent part of the rule.

Table 1. Linguistic variables for importance weight of each criterion and alternative

Linguistic variable	Fuzzy value
Very low (VL)	(0, 0, 0.1)
Low (L)	(0, 0.1, 0.3)
Medium low (ML)	(0.1, 0.3, 0.5)
Medium (M)	(0.3, 0.5, 0.7)
Medium high (MH)	(0.5, 0.7, 0.9)
High (H)	(0.7, 0.9, 1)
Very high (VH)	(0.9, 1, 1)

Table 2. Linguistic variables for importance weight of each criterion and alternative

Linguistic variable	FS with Z-number
Disagree (D)	(0, 0.25, 0.75)
Neutral (N)	(0.25, 0.5, 0.75)
Agree (A)	(0.5, 0.75, 1)
Strongly Agree (SA)	(0.75, 1, 1)

Both Table 1 and 2 are used to construct the decision matrix \tilde{D} and weight matrix \tilde{W} . Let the matrix \tilde{D} be the decision making matrix, where \tilde{x}_{ij} for all $ij = (1, \dots, n)$ and $r\tilde{x}_{ij}$ is the reability of the \tilde{x}_{ij} selection. \tilde{x}_{ij} and $r\tilde{x}_{ij}$ is respectively the constraint and reliability of a Z-number. The knowledge, for example, if an opinion is expressed as “The journey time is critical, very low, and reability is strongly agree.” Then the opinion can be described with Z-number (L, SA).

$$\tilde{D} = (\tilde{x}_{ij}^p, r\tilde{x}_{ij}^p)_{m \times n} = \begin{matrix} & \begin{matrix} X_1 & X_2 & \dots & X_n \end{matrix} \\ \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} & \begin{bmatrix} (\tilde{x}_{11}, r\tilde{x}_{11}) & (\tilde{x}_{12}, r\tilde{x}_{12}) & \dots & (\tilde{x}_{1n}, r\tilde{x}_{1n}) \\ (\tilde{x}_{21}, r\tilde{x}_{21}) & (\tilde{x}_{22}, r\tilde{x}_{22}) & \dots & (\tilde{x}_{2n}^p, r\tilde{x}_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ (\tilde{x}_{m1}, r\tilde{x}_{m1}) & (\tilde{x}_{m2}^p, r\tilde{x}_{m2}) & \dots & (\tilde{x}_{mn}, r\tilde{x}_{mn}) \end{bmatrix} \end{matrix}$$

Then, construct the weighting matrix \tilde{W} of the criteria of the experts

$$\tilde{W} = (\tilde{w}_i, r\tilde{w}_i)_{1 \times m} = [(\tilde{w}_1, r\tilde{w}_1)(\tilde{w}_2, r\tilde{w}_2) \dots (\tilde{w}_n, r\tilde{w}_n)]$$

where $(\tilde{w}_{ij}, r\tilde{w}_{ij})$ is a z-numbers with Fuzzy Set.

STEP3: Construct a normalized weighted fuzzy decision matrix. Next, the normalized weighted decision matrix with respect to aggregated matrix comparison of each criterion and alternatives is constructed and the importance of the experts is considered as linguistic variable. Thus,

$$\overline{D}_w = (\tilde{v}_{ij}, r\tilde{v}_{ij})_{m \times n} = \begin{matrix} & \begin{matrix} X_1 & X_2 & \dots & X_n \end{matrix} \\ \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} & \begin{bmatrix} (\tilde{v}_{11}, r\tilde{v}_{11}) & (\tilde{v}_{12}, r\tilde{v}_{12}) & \dots & (\tilde{v}_{1n}, r\tilde{v}_{1n}) \\ (\tilde{v}_{21}, r\tilde{v}_{21}) & (\tilde{v}_{22}, r\tilde{v}_{22}) & \dots & (\tilde{v}_{2n}, r\tilde{v}_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ (\tilde{v}_{m1}, r\tilde{v}_{m1}) & (\tilde{v}_{m2}, r\tilde{v}_{m2}) & \dots & (\tilde{v}_{mn}, r\tilde{v}_{mn}) \end{bmatrix} \end{matrix} \quad \text{Where}$$

$$(\tilde{v}_{ij}, r\tilde{v}_{ij}) = (\tilde{w}_i, r\tilde{w}_i) \otimes (\tilde{x}_{ij}, r\tilde{x}_{ij})$$

STEP4: Determine the Positive Ideal Solutions (PIS) and Negative Ideal Solutions (NIS), the PIS and NIS is defined as;

$$A^* = (V_1^*, V_2^*, \dots, V_n^*) = \left\{ \left(\max_j v_{ij} | i \in I' \right), \left(\min_j v_{ij} | i \in I'' \right) \right\},$$

$$A^- = (V_1^-, V_2^-, \dots, V_n^-) = \left\{ \left(\min_j v_{ij} | i \in I' \right), \left(\max_j v_{ij} | i \in I'' \right) \right\},$$

STEP5: Calculate the distance of PIS and NIS. The distances alternatives of Z-number, for the PIS is stated as follows:

$$(d_j^*, rd_j^*) = \left(\sqrt{\sum_{i=1}^n (\tilde{v}_{ij} - \tilde{v}_i^*)^2}, \sqrt{\sum_{i=1}^n (r\tilde{v}_{ij} - r\tilde{v}_i^*)^2}, j = 1, \dots, n \right)$$

Similarly, the distances alternatives of Z-number for NIS is given as follows:

$$(d_j^-, rd_j^-) = \left(\sqrt{\sum_{i=1}^n (\tilde{v}_{ij} - \tilde{v}_i^-)^2}, \sqrt{\sum_{i=1}^n (r\tilde{v}_{ij} - r\tilde{v}_i^-)^2}, j = 1, \dots, n \right)$$

STEP6: Calculate the relative closeness. The relative closeness of the alternative x_i with respect to f^* is defined as

$$(C_j^*, rd_j^*) = \left(\frac{\frac{d_j^-}{d_j^* + d_j^-}, \frac{rd_j^-}{rd_j^* + rd_j^-}}{2}, j = 1, \dots, n \right)$$

Alternatives can be ranked based on the large value of closeness coefficient A_j . The best alternative is the one with the greatest relative closeness to the ideal solution.

Next, Sect. 4, contains an example to demonstrate the feasibility and applicability of the proposed method.

4 Numerical Example

This example was taken from B. Kang, et al. (2012) [19] the FMCDM problem in selection of three main vehicle for journey, there are three different choices, namely car, taxi, and train are evaluated by three travelers [19]. The list of criteria is presented as follows:

- C1 = Price: Price is the most important criteria however tariff contents vary in journey.
- C2 = Journey Time: Efficiency of time usage in traveling to destination.
- C3 = Comfort: Some transport offer different comfort in the journey.

The relative importance weights of the three criteria are described using linguistic variables. The proposed method is currently applied to solve this problem and the computational procedure is summarized as follows:

STEP1. Construct the hierarchy structure for vehicle for journey selection. The hierarchical structure of evaluating the vehicle by travelers is given in Fig. 1, where all the criteria and alternatives are drawn horizontally.

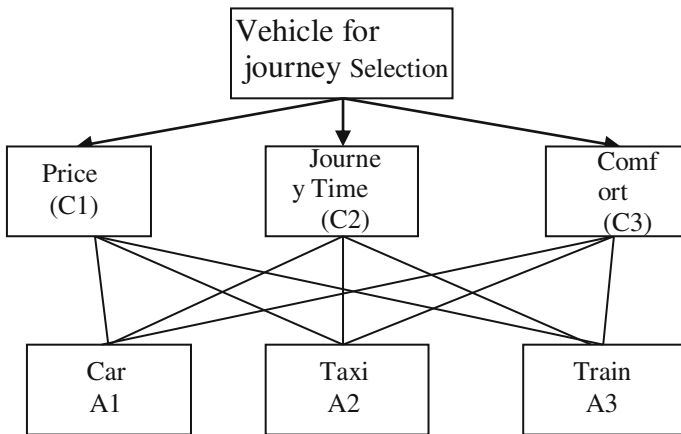


Fig. 1. Hierarchical structure of vehicle for journey selection

STEP2. Scaling the relative of data. The data that considers Z-numbers and its conversion (see Tables 1 and 2) are referred in order to construct the attributes of matrices. The original data has a scale of [1–100], while the fuzzy set has a scale [0–1], we have modified the original data to become the fuzzy set value in order to have an easy calculation

$$x_{11} = (L, SA) = ((0, 0.1, 0.3), (0.9, 1, 1))$$

$$(MH, N) = ((0.5, 0.7, 0.9), (0.3, 0, 5, 0, 7))$$

$$(H, A) = ((0.7, 0.9, 1), (0.7, 0.9, 1))$$

Then, the average for x_{11} is ((0.4, 0.5667, 0.7333), (0.6333, 0.8, 0, 9) as follows (Table 3).

Table 3. Fuzzy decision matrix

	A1	A2	A3
C1	((0.4, 0.5666, 0.7333), (0.6333, 0.8, 0.9))	((0.8333, 0.9666, 1), (0.4, 0.6, 0.7333))	((0.7, 0.8667, 0.9667), (0.7667, 0.9333, 1))
C2	((0.4666, 0.6, 0.7), (0.7666, 0.9333, 1))	((0.2667, 0.4333, 0.6333), (0.7, 0.8333, 0.9))	((0.9, 1, 1), (0.6333, 0.8, 0.9))
C3	((0.3333, 0.4666, 0.6), (0.5333, 0.7333, 0.8333))	((0.3667, 0.5667, 0.7333), (0.4667, 0.7, 0.8333))	((0.5333, 0.6333, 0.7), (0.6333, 0.8, 0.9))

Then, we constructed the weight for fuzzy decision matrix as Table 4. The linguistic variables should be converted into numerical values under the frame of fuzzy set which is described as in Table 2. For example if the value of z-number is stated as (A, SA), thus the numerical values is turn out to be as ((0.5, 0.75, 1), (0.75, 1, 1)

Table 4. The fuzzy normalized decision matrix

	A1	A2	A3
C1	((0.3, 0.5667, 0.7333), (0.475, 0.8, 0.9))	((0.625, 0.9667, 1), (0.3, 0.6, 0.7333))	((0.525,0.8667,0.9667), (0.575,0.9333,1))
C2	((0.2333, 0.45, 0.7), (0.575, 0.9333, 1))	((0.1333, 0.325, 0.6333), (0.525, 0.8333, 0.9))	((0.45, 0.75, 1), (0.475, 0.8, 0.9))
C3	((0.0833,0.2333,0.45), (0.4,0.7333,0.8333))	((0.0916, 0.2833, 0.55), (0.35, 0.7, 0.8333))	((0.1333, 0.3166, 0.525), (0.475, 0.8, 0.9))

STEP3. The weighted normalized fuzzy decision matrix is construction with respect to an aggregated matrix comparison of each criterion and alternatives (Table 5). Therefore, let's take the example on calculating the $\overline{D_w}$

Table 5. Weight

	Weight
C1	((0.75, 1, 1), (0.75, 1, 1))
C2	((0.5, 0.75, 1), (0.75, 1, 1))
C3	((0.25, 0.5, 0.75), (0.75, 1, 1))

$$\begin{aligned}\overline{D_w} &= (\tilde{w}_i, r\tilde{w}_i) \otimes (\tilde{x}_{ij}, r\tilde{x}_{ij}) \\ &= ((0.75, 1, 1), (0.75, 1, 1)) \otimes ((0.4, 0.5667, 0, 7333), (0.6333, 0.8, 0, 9)) \\ &= ((0.0675, 0.1, 0.12), (0.5625, 1, 1))\end{aligned}$$

STEP4. Determine the PIS and NIS. Then, the PIS A^+ and the NIS A^- are successfully determined, as follows:

$$A^+ = (1, 1, 1) \quad A^- = (0, 0, 0)$$

STEP5. The separation measures are calculated using the n -dimensional Euclidean distance, stated as follows (Table 6).

Table 6. The distance of PIS and NIS

	D^+	D^-
A1	((3, 1709), (1, 6782))	((2, 1756), (3, 3831))
A2	((2, 7760), (2, 0974))	((2, 5536), (3, 0377))
A3	((2, 3001), (1, 5714))	((2, 9520), (3, 4541))

STEP6. The results for relative closeness coefficients are successfully stated as in Table 7;

Table 7. Closness coefficients of the proposed method

	Closness coefficients (Cc)	Final result
A1	(0.4069), (0.6684)	0.5376
A2	(0.4791), (0.5915)	0.5353
A3	(0.5620), (0.6873)	0.6246

As a conclusion, the best alternative selection is A3 and the ranking order of the alternative of selecting the best Vehicle for journey is given by $A3 \succ A2 \succ A1$.

5 Comparative Analysis

Since the proposed method introduced a new equilibrium standardized approach in the evaluation process, it is important to compare it with the existing approach. Two different cases are used to employ this comparison. First, the proposed method used FTOPSIS and z-number. Second, the Kang, et al. (2012) method.

Based on Table 8, it can be concluded that the method gives a different rank for the output values. The slightly different results probably due to the combination of FTOPSIS with Z-number. Besides, it is may be due to the different selection of weights from the experts. In addition, due to the different scale opposition.

Table 8. Comparative analysis

Method	Ranking order according to closeness coefficient	Schale (cc)
Proposed method with FTOPSIS and z-number	$A3 \succ A2 \succ A1$	$0.6246 \succ 0.5376 \succ 0.53531$
Kang et al. [2012] method	$A1 \succ A2 \succ A3$	$0.36 \succ 0.29 \succ 0.28$

6 Conclusions

We have proposed a FTOPSIS and Z-number by integrating the ability of fuzzy rule based approach based multi-criteria decision making. The aim this paper is to add the reliability and restriction concept to the previous FTOPSIS method. A numerical example on selection the best vehicle for journey was provided. The ranking based on proposed method is validated comparatively. Based on the results, the proposed method was presented a more flexible and useful way to handle vagueness and imperfect information in decision making problems. Thus, our future studies is to combine fuzzy logic with FTOPSIS and z-number, in order to have more uncertainty in the decision making area.

Acknowledgements. This research is supported by Dana Penyelidikan Universiti UNISZA/2016/DPU/10, Universiti Sultan Zainal Abidin. This support is gratefully acknowledged.

References

1. Syibrah, N., Hani, H.: A general type-2 fuzzy logic based approach for multi-criteria group decision making. In: IEEE International Conference on Fuzzy Systems, pp. 353–358 (2005)
2. Zadeh, L.A.: A note on a Z-number. *Inf. Sci.* **181**, 2923–2932 (2011)
3. Chen, S.J., Hwang, C.L.: *Fuzzy Multiple Attribute Decision Making Methods and Applications*. Springer, Berlin (1992)
4. Chen, M.F., Tzeng, G.H.: Combining grey relation and TOPSIS concepts for selecting an expatriate host country. *Math. Comput. Model.* **40**(13), 1473–1490 (2004)
5. Zadeh, L.A.: The concept of a linguistic variable and its application approximate reasoning, Part 1, 2, and Part 3. *Inf. Sci.* **8**, 199–249 (1975); *Inf. Sci.* **8**, 301–357 (1975); *Inf. Sci.* **9**, 43–58 (1975)
6. Kuo, M.S., Liang, G.S., Huang, W.C.: Extensions of the multicriteria analysis with pairwise comparison under a fuzzy environment. *J. Sci. Direct* **43**, 268–285 (2006)
7. Abdullah, L., Sunadia, J., Imran, T.: A new analytic hierarchy process in multi-attribute group decision making. *Int. J. Soft Comput.* **4**, 208–2014 (2009)
8. Wang, Z.X., Wang, Y.Y.: Evaluation of the provincial competitiveness of the Chinese high-tech industry using an improved TOPSIS method. *Expert Syst. Appl.* **41**, 2824–2831 (2014)
9. Lourenzutti, R., Krohling, R.A.: The Hellinger distance in multicriteria decision making an illustration to the TOPSIS and TODIM methods. *Expert Syst. Appl.* **41**, 4414–4421 (2014)

10. Schneider, E.R.F.A., Krohling, R.A.: A hybrid approach using TOPSIS differential evolution, and Tabu Search to find multiple solutions of constrained non-linear integer optimization problems. *Expert Syst. Appl.* **62**, 47–56 (2014)
11. Joshi, D., Kumar, S.: Intuitionistic fuzzy entropy and distance measure based TOPSIS method for multi-criteria decision making. *Egypt. Inform. J.* **15**, 97–104 (2014)
12. Chen, T.Y.: The inclusion-based TOPSIS method with interval-valued intuitionistic fuzzy sets for multiple criteria group decision making. *Appl. Soft Comput.* **26**, 57–73 (2015)
13. Zhang, X., Xu, Z.: Soft computing based on maximizing consensus and FTOPSIS approach to interval-valued intuitionistic fuzzy group decision making. *Appl. Soft Comput.* **26**, 42–56 (2015)
14. Aliev, R.A., Zeinalova, L.M.: Decision making under Z-Information. In: *Human-Centric Decision-Making Models*, vol. 502, pp. 233–252 (2013)
15. Yager, R.R.: On Z-valuations using Zadeh's Z-numbers. *Int. J. Intell. Syst.* **27**, 259–278 (2012)
16. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)
17. Kauffman, A., Gupta, M.M.: *Introduction to Fuzzy Arithmetic: Theory and Application*. Van Nostrand-Reinhold, NY (1985)
18. Chen, C.T.: Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Sets Syst.* **114**, 1–9 (2000)
19. Kang, B., Wei, D., Li, Y., Deng, Y.: Decision making using Z-numbers under uncertain environment. *J. Comput. Inf. Syst.* **8**, 2807–2814 (2012)

Comparative Studies of Information Retrieval Approaches in User-Centered Health Information System

Ibrahim Umar Kontagora^{1,2} and Isredza Rahmi A. Hamid¹(✉)

¹ Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn, Parit Raja, Malaysia
ibrosoftuk@yahoo.com, rahmi@uthm.edu.my

² Department of Computer Science, Niger State Polytechnic, Zungeru, Niger State, Nigeria

Abstract. In this paper, a comparative studies of different methods deployed in addressing problems of user-centered health information retrieval systems were investigated. The reason for the comparative studies is to identify the approach that best addressed the readability and vocabulary mismatched issues encountered by laymen patients and their relatives in exploring information extracted from medical discharge documents and clinical reports online. We discussed and presented the performance of information retrieval systems in previous research works. We concentrated on classifying and comparing the three approaches used in health information retrieval which are Vector Space Model (VSM), Language Based Approach Model (LM) and Context Based Approach (CBA). The usefulness of incorporating controlled vocabularies such as Metamap, UMLS, external, MeSH, etc. was extensively discussed. The result shows that the Language Based Approach systems achieved better results as compared to the Vector Space Model Approach and Context Based Approach Systems. The Language Based Approach Systems managed to acquire 0.4146, 0.7560 and 0.7445 for Mean Average Precision, Precision @ 10 and Normalized Discounted Cumulative Gains @ 10 respectively. Hence, we conclude based on the outcome of the comparative studies and our experimental results that the language modeling models is best suited to be deployed in addressing the problems of returning relevant information by user centered health information retrievals to users.

Keywords: Language models · Vector space models concept-based approach
External medical resource · Query expansion

1 Introduction

Health information retrieval has become attractive today as a result of the massive rise in medical allied information online [1]. There is no system at present that doubts the information needs of the user context and returns appropriate documents to the best of their knowledge [2]. With the spreading consciousness, probing online health related network mediums and other sources has become a frequent habit [3]. A study

conducted by Pew Research Center [4] revealed that a huge percentage of the population in the United States' search engine operators seek for information on a particular illness online.

The previous campaigns addressed user different information needs. Moreover, it only targeted on a specific group of consumers with skilled health awareness (e.g., health researchers and clinicians) [5]. The ShARe/CLEF eHealth Task 3 launched an information retrieval campaign which focused on the information needs of common people and the queries they created to express their needs. However, the outcome of the (2013) Information retrieval tasks failed to tackle patients' questions which they may come up while reading the clinical report [1]. This same result has showed significant progress over the 2012 results for both the reference line and team submissions using the new query set [3].

The main objective of this study is to carry out comparative studies on the various information retrieval approaches deployed to address the problems of user centered health information retrievals. Hence, we are going to propose a new information retrieval approach for health information system. The remainder of this paper is organized as follows. Section 2 describes related work regarding Information retrieval approaches. Section 3 discusses the comparison of the approaches deployed in addressing information retrieval problems, Sect. 4 contains the result and discussions and Sect. 5 concludes the work and direction for future work.

2 Related Work

Medical Information Retrieval is widely measured as a health related assignment usually executed by a huge range of medical workers and laypeople (patients and their relatives) [6]. Nearly 80% of United States search engine operators' searched for health related facts about a particular illness through online [5]. Various prospective information seekers with different medical understanding, suggests the increased amount of information required. Subsequently, the medical information retrieval systems design requirement must meet up with the health information needs of different categories of users [7].

Patients are always eager of knowing the exact content of their discharge summaries written by medical expert [6]. However, the medical text is highly professional for a layman to follow [1]. Therefore, the medical information retrieval becomes highly accepted as helpful way in answering the patient's questions [7]. Number of evaluation campaigns focusing on health information has increased due to the greater significance of health information retrievals by information seekers. The Text REtrieval Conference (TREC) focused on addressing questions that patients may come up with after reading their medical reports [3]. This task concentrated on a particular topic about a specific disease and relative treatment. All positive contributions made by participants on how to improve the existing retrieval systems were submitted and deliberated [8].

2.1 Information Retrieval Approaches

This section discusses numerous approaches deployed in addressing the problems of user-centered health information retrieval. The study classifies the information retrieval approaches under three fundamental modules; Language Modeling (LM) Models approach, Vector Space Models approach (VSM) and Context Based Approach (CBA).

2.1.1 Language Modeling (LM) Model Approach

Language Modeling Model approach is an arithmetical distribution model that allocate likelihoods to an order of terms, which predicts the possibility of their appearance within the script. All language based built systems operates on likelihoods for each term that come across and these likelihoods are self-determined on the nature of text. Various researchers used Language Modeling model for information retrieval system [2, 7, 9–11].

Shen [2] used language modelling approach and Indri search engine platform as their baseline with integrating Dirichlet smoothing. For the purpose of query expansion, the external resources Unified Medical Language System (UMLS) and Metamap were integrated. Moreover, related information from the user query logs were used to perform query expansion and the connection of shared words from query logs alongside medical vocabulary were used for information retrieval using concept-based approach. This method could not address the readability and vocabulary mismatched issues encountered by laymen in exploring extracted information online. However, this model achieved 0.7560 and 0.4016 for P@10 value and MAP value respectively. Moreso, Oh et al. [10] used Lucene search engine platform as their baseline and proposed the use of multiple-stage re-ranking method. This work is similar with [2] in such a way that they also used Dirichlet smoothing. Medical terms were extracted from discharge summaries to launch query expansion. This method obtained MAP value of 0.3989 and P@10 value of 0.74.

Work by Choi and Choi [7] is similar with [2, 7, 10] where they used indri as search engine with language model and Dirichlet smoothing correspondingly. They combined query expansion using the Metamap vocabulary, with discharge summaries and initial query extracted terms coordinated together. They conducted experiment with the superior feature by integrating learning to rank methods. This feature determines which of documents to be appeared and counted the frequency of terms prior-hand. This model managed to achieve 0.3494 of MAP score and 0.75 for P@10 score. Also Saleh [9] and Pecina [2] presented an interesting variant using Hiemstra language model with terrier search engine. The performance of the system was improved by using pseudo random feedback and Medline resource during query expansion. HTML strip and Boiler pipe resources were incorporated to decrease the data size to 6% of the original by removing insignificant terms. The system overall performance conveyed a MAP value of 0.1677 and a P@10 value of 0.5360.

2.1.2 Vector Space Model (VSM) Approach

Vector space model represents bits of script as vectors of identifiers. It was first tested in the SMART information retrieval system [4]. There are several variants have been offered such as, generalized vector space model and weighted vector space model. The

vector space model based systems are easier to operate due to the fact that they are linear algebra oriented, it calculates the degree of resemblance between huge documents and queries with a limited similarity supports. However, it is deficient when representing long text documents by showing very poor similarity values.

Ksentini et al. [7] used the vector space model for the information retrieval approach. The cosine degree was used to measure the similarity between the query and document. The link between a query term and the total group of documents are computed by increasing the weighted Term Frequency (TF) and Inverse Document Frequency (IDF) measures. The default setup Terrier retrieval system is used for discontinuous term deletion, tokenization and lessening. The overall scheme performance attained a MAP score of 0.167 and a P@10 score of 0.55.

Thesprasith and Jaruskulchai [6] used the Lucene search engine as their baseline system for stemming and tokenization. Pseudo-relevance feedback method was incorporated for the purpose of query expansion. The concepts mined from Medline biomedical dictionary were appended to expand the search query. Rocchio's formula was the determinant factor for extracting terms and was used for SMART system stoppage, with routine assessment of Pseudo-relevance response. The system managed to achieve the MAP value of 0.20 and P@10 score of 0.5540. Also Ozturkmenoglu et al. [5] used terrier search engine as their baseline engine. Unlike work by Thesprasith and Jaruskulchai [6], the query expansion for a specific query was determined by the integrated probabilistic Naïve Bayes. The overall method performance was stated to achieve P@10 value of 0.67 and a MAP score of 0.305.

2.1.3 Context Based Approach (CBA) Approach

The Context Based Approach attempt to extend each query by first extracting all the concepts terms from the search query and the synonyms of these general concepts terms from the incorporated vocabularies e.g. UMLS in order to launch an expanded search query. It uses search engine platforms such as Indri, Lucene etc. as the baseline engines and UMLS, HTML and Medline as incorporated external resources for the purpose of query [12, 13].

Salton [12] used the context based approach with terrier search engine. The performance of the system was improved by using pseudo random feedback and Medline resource during query expansion. MeSH, Metamap resources were incorporated for the purpose of query expansion. The entire system performance experienced a great effect due to the presence of these resources. The system overall performance conveyed a MAP value of 0.1732, P@10 value of 0.5512 and NDCG@10 value of 0.5211.

Work by Suominen et al. [14] used the context based approach with indri as search engine as their base line system. They combined query expansion using the Metamap vocabulary, with discharge summaries and initial query extracted terms coordinated together. They conducted experiment with the superior feature by integrating UMLS and Metamap external resources. This feature determines which of the documents to appear and counted the frequency of terms prior-hand. This model managed to achieve 0.1732 of MAP score, 0.6122 for P@10 score and 0.5523 of NDCG@10 score. Also Koopman et al. [13] used terrier search engine as their baseline engine. Unlike work by

Suominen et al. [14] the query expansion for a specific query was determined by the integrated probabilistic Naïve Bayes. The overall method performance was stated to achieve P@10 value of 0.5324, MAP score of 0.3244 and NDCG@10 score of 0.5788.

3 Information Retrieval Approaches

This paper performs a comparative study on various approaches deployed to address the problem of user centered health information retrievals as shown in Table 1. The problems encountered specifically by laymen patients and the care giver include the readability and vocabulary mismatched issues in exploring information extracted from medical discharge documents and clinical reports online.

3.1 Dataset

The dataset used in this research work were provided by the Unified Medical Language System (UMLS), Medical Subject Heading (MeSH), Metamap and Khresmoi project6 [1]. These datasets covers a wide range of patients' information and medical topics. All documents in the collection are downloaded from numerous online sources, including Health on the Net organization certified websites, Genetics Home Reference, Clinical.gov and Diagnosia7 [12].

3.2 Differences Among the Three Information Retrieval Approaches (LM, VSM and CBA)

The major difference among the three approaches deployed in addressing problems of user-centered health information retrieval systems used in these comparative studies is that, The vector space model based systems are easier to operate due to the fact that they are linear algebra oriented, it calculates the degree of resemblance between huge documents and queries with a limited similarity supports. However, it is deficient when representing long text documents by showing very poor similarity values [3, 4]. The Language Modeling Model approach is an arithmetical distribution model that allocates likelihoods to an order of terms, which predicts the possibility of their appearance within the script. All language based built systems operates on likelihoods for each term that come across and these likelihoods are self-determined on the nature of text. Various researchers used Language Modeling model for information retrieval system [2, 10]. While The Context Based Approach attempt to extend each query by first extracting all the concepts terms from the search query and the synonyms of these general concepts terms from the incorporated vocabularies e.g. UMLS in order to launch an expanded search query. It uses search engine platforms such as Indri, Lucene etc. as the baseline engines and UMLS, HTML and Medline as incorporated external resources for the purpose of query [13, 14].

Table 1. Comparative study of system performances of LM, VSM and CBA approaches

	Authors	Approach	Datasets and query expansion used	Performances of the systems		
				P@10	NDCG@10	MAP
1	Shen et al. [2]	LM	The dataset used were Metamap, UMLS and the Query Expansion used were also UMLS, Mutal Information and Metamap	0.7560	0.7445	0.4016
2	Oh and Jung [10]	LM	The dataset used were None and the Query Expansion used were Abrev. + Pseudo random feedback	0.74	0.73	0.3989
3	Claveau et al. [9]	LM	The dataset used were Ogmios NLP, Metam TreeTagger, UMLS, FASTR, YATEA, and the Query Expansion used were FASTR morpho-syntactic variants, UMLS synonyms and abbreviations	0.6740	0.6793	0.4021
4	Thakkar et al. [11]	LM	The dataset used were Metamap, MeSH and the Query Expansion used were Pseudo relevance feedback and Query-likelihood	0.7060	0.6869	0.4146
5	Choi and Choi [7]	LM	The dataset used were Metamap, UMLS and the Query Expansion used were Discharge summaries and Intersection of terms from query	0.75	0.70	0.3494
6	Yang et al. [8]	LM	The dataset used were Metamap and the Query Expansion used were Pseudo relevance feedback and Markov random field f	0.69	0.6705	0.3589
7	Thesprasith and Jaruskulchai [6]	VSM	The dataset used were Medline and the Query Expansion used were Pseudo relevance feedback	0.5540	0.5471	0.2076
8	Ozturkmenoglu et al. [5]	VSM	The dataset used were Medline and the Query Expansion used were Naïve bayes probabilistic expansion	0.6740	0.6518	0.3049
9	Drame et al. [4]	VSM	The dataset used were MeSH Metamap, UMLS and the Query Expansion used were UMLS Synonyms	0.5460	0.5574	0.2315
10	Ksentini et al. [3]	VSM	The dataset used were None and the Query Expansion used were Weighted vectors for query terms	0.5460	0.5625	0.1677
11	Koopman et al. [13]	CBA	The dataset used were Medline and the Query Expansion used were Naïve bayes probabilistic expansion and Terriers	0.5324	0.5788	0.3244
12	Suominen et al. [14]	CBA	The dataset used were Metamap, UMLS and the Query Expansion used were UMLS Synonyms	0.6122	0.5523	0.2011
13	Salton [12]	CBA	The dataset used were Metamap, MeSH and the Query Expansion used were Pseudo relevance feedback	0.5512	0.5211	0.1732

3.3 Performance Metrics

The Health Information Retrieval approaches will be evaluated using three parameters that are:

(a) *Precision at 10 documents (P@10)*

P@10 calculates the proportion of relevant documents at every 10 documents retrieved from a query. It can be computed as $P@10 = \frac{(A)_{10}}{(A+B)_{10}}$ where P is a Proportion of Relevant Documents Retrieved at every 10 document, A is Retrieved Relevant Documents and B is Retrieved Non relevant Documents [11].

(b) *Normalized Discounted Cumulative Gain at 10 documents (NDCG@10)*

NDCG@10 computes the cumulative gain at each position for a chosen value of p for the entire relevant document in the query. NDCG@10 is computed as $NCDG_p = \frac{DCG_p}{IDCG_p}$ where $IDCG_p = (IDCG_p) = \sum_{i=1}^{REL} \frac{2^{rel_i}-1}{IDCG_p}$. REL represent the list of relevant documents, DCG_p is used to emphasize highly relevant documents appearing early in the result list [11].

(c) *Mean Average Precision (MAP)*

MAP computes the Mean Average Precision of relevant documents retrieved from a query. MAP is computed as $MAP = \frac{1}{N} \sum_{j=1}^N \cdot \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i)$ where, Q_j is number of relevant documents for query j , N is number of queries and $P(doc_i)$ is precision value at i th relevant document [11].

4 Result and Discussions

For the specific task of comparing the various approaches deployed by previous researchers in addressing the problems of user centered health information retrievals and identifying the better approach to be deployed by researchers in addressing the aforementioned problems. The Language Based approach focused on addressing readability issues encountered by laymen patients and their relatives in exploring information extracted from medical discharge documents and clinical reports online; vocabulary mismatched issues between laymen queries and expert vocabulary during query expansion which affects the information retrieval system performance. The outcome of the study is as presented in Table 1 and Fig. 1.

Work by [8, 2] used the Language Model approaches in extracting medical concepts from the search queries and synonym terms from incorporated external resources. Only most specific medical concepts in the layman search queries were extracted and their synonym terms from incorporated external resources were expanded into a new query. The outcome of the search results contain less medical concepts which clearly attempted to better address the readability and vocabulary mismatched issues. There is an increased understanding of retrieved results due to less medical concepts in the returned results. The state-of-art results were reported by [2] using Language Model Based Retrieval System with a P@10 of 0.7560 and a NDCG@10 of 0.7445.

Works by [4–6, 7] used the Vector Space Models approaches and [12–14] used the Context Based Approaches in extracting medical concepts from the laymen search

queries and synonym terms from incorporated external resources. The induced search results returns information with more medical concepts compared to the Language Model approaches. This shows that the Vector Space Model and Context Based Approaches concentrates less on most specific terms during the query search and also fail to incorporate controlled vocabulary during query expansion which affects their results.

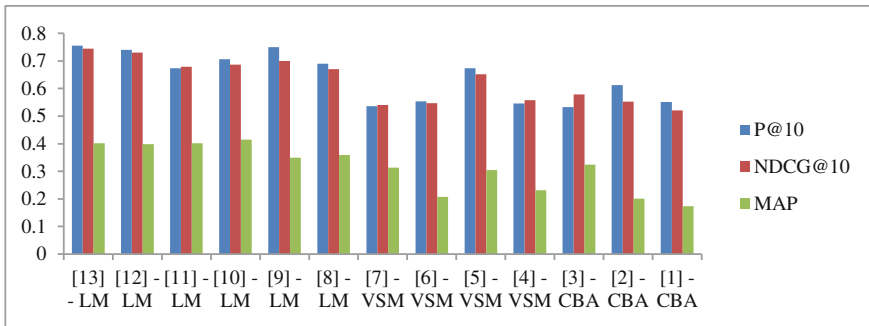


Fig. 1. Comparison of various methods based on P@10, NDCG@10 and MAP values

The scientific reasons for the results obtained by LM, VSM and CBA could be explained in reference to this query “Anoxic Brain Damage” with label number (abd2017001), only the most specific term “Anoxic Brain Damage” and its synonyms “Anoxic Encephalopathy” and “Anoxic Brain Injury”, are expanded into the new query by the language modeling model (LM), thereby disregarding the general terms Brain, Brain Injury, and Injury unlike VSM and CBA that extracts both the general and most specific terms into the new query thereby causing readability and vocabulary mismatched issues in exploring information extracted from medical discharge documents and clinical reports. In the same vain for this second query, “Stroke and Respiratory Failure” with label number (abd2017002), is made up of two labeled most specific terms “Stroke” and “Respiratory Failure”. The extended query will comprise of “Cerebrovascular Accident”, “Vascular Accident, Brain”, “Kidney Failure”, and Renal Failure.

Based on the comparative studies, the best results were achieved by [2] with P@10 value of 0.7560 and NDCG@10 value of 0.7445. Moreover, work by Thakkar et al. [1] presented the highest value of MAP with 0.4146. This work was based on language modeling methods and query expansion performed on specific medical concepts and synonym terms extracted from the search query. They also incorporated two controlled vocabularies that are Metamap and UMLS for the specific purpose of addressing vocabulary mismatched issue.

4.1 Statistical Test for the Validation of Experimental Results

A Language Model (LM) is a statistical distribution model that assigns probabilities to terms in the sequence which predicts the possibilities of such term for appear in the document. The model defines a probability $P(T/D)$, where T refers to terms and D represents Documents. From my experimental results, since the process is repeated 10 times for each performance metrics $P@10$, MAP and $NDCG@10$, picking any relevant term in the document one at a time $T_1, T_2, \dots, T_{n=10}$ in D is given by $P(T_1, T_2, \dots, T_{n=10}) = \sum_{i=1}^{n=10} P(T_i/D)$. The equation will assign zero (0) probability to all irrelevant terms and one (1) to relevant terms in the sequence. The statistical analysis show that out of every 10 terms retrieved from a document (i.e. 100% of the terms), Language Modeling Model (LM) approximately scored 0.76 (76%) as relevant and readable retrieved documents in respect to $P@10$, as against Vector Space Model (VSM) 0.67 (67%) and Context Based Approach (CBA) 0.57 (57%). In respect to $NDCG@10$, LM scored 0.74 (74%), VSM 0.65 (65%) and CBA 0.57 (57%) as relevant and readable retrieved documents and finally in relation to MAP, LM scored 0.41 (41%), VSM scored 0.30 (30%) and CBA scored 0.32 (32%).

5 Conclusion

Based on the outcome of the experimental results obtained from Fig. 1 clearly shows that out of every 10 documents retrieved from a medical search result (i.e. 100% of the documents), Language Modeling Model (LM) approximately scored 0.76 (76%) as relevant and readable retrieved documents in respect to $P@10$, as against Vector Space Model (VSM) 0.67 (67%) and Context Based Approach (CBA) 0.57 (57%). In respect to $NDCG@10$, LM scored 0.74 (74%), VSM 0.65 (65%) and CBA 0.57 (57%) as relevant and readable retrieved documents and finally in relation to MAP, LM scored 0.41 (41%), VSM scored 0.30 (30%) and CBA scored 0.32 (32%).

The experimental results obtained in Fig. 1 shows that the Language Modeling is the best approach to be used in addressing information retrieval system problems. Every 76% of retrieved information by Language Modeling Model are Readable and Vocabulary Mismatched free as against Vector Space Model (VSM) 67% and Context Based Approach (CBA) 57%. The Language Modeling approach overcomes the vector space model and Context Based Approaches between 3 and 8% in relation to Precision@10 and 20% in MAP value. These better results were achieved due to Language Modeling Approaches fully concentrating on most specific terms during query search rather than general terms and also incorporating controlled vocabulary such as Meta-map and UMLS during query expansion. We recommend further work on this study should include the design of algorithm that will address the readability and vocabulary mismatched issues encountered from retrieved images, audios and videos.

Acknowledgements. The authors express appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM). This research is supported by Short Term Grant vot number U653 and Gates IT Solution Sdn. Bhd. under its publication scheme.

References

1. Thakkar, H., Iyer, G., Shah, K., Majumder, P.: Team IRLabDAIICT at ShARe/CLEF eHealth 2014 task 3: user-centered information retrieval system for clinical documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
2. Shen, W., Nie, J.Y., Liu, X., Liui, X.: An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM @ CLEF2014eHealth task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
3. Ksentini, N., Tmar, M., Gargouri, F.: Miracl at CLEF 2014: eHealth information retrieval task. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
4. Drame, K., Mougin, F., Diallo, G.: Query expansion using external resources for improving information retrieval in the biomedical domain. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
5. Ozturkmenoglu, O., Alpkocak, A., Kilinc, D.: Demir at CLEF eHealth: the effects of selective query expansion to information retrieval. In: proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
6. Thesprasith, O., Jaruskulchai, C.: Csku gprf-qe for medical topic web retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
7. Choi, S., Choi, J.: Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at CLEFeHealth2014 task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
8. Yang, C., Bhattacharya, S., Srinivasan, P.: The University of Iowa at CLEF 2014: eHealth task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
9. Claveau, V., Hamon, T., Grabar, N., Maguer, S.L.: RePaLi participation to CLEF eHealth IR challenge 2014: leveraging term variation. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
10. Oh, H.S., Jung, Y.: A multiple-stage approach to re-ranking clinical documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
11. Dybkjaer, L., Hemsén, H., Minker, W.: An overview of evaluation methods. In: Evaluation of Text and Speech Systems in TREC Ad-hoc Information Retrieval and TREC Question Answering. Springer, Dordrecht, the Netherlands (2015)
12. Salton, G.: The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall Inc, Upper Saddle River, NJ, USA (2015)
13. Koopman, B., Zuccon, G., Bruza, P., Sithon, L., Lawley, M.: An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In: Proceedings of CIKM (2012)
14. Suominen, H., et al.: The Proceedings of the CLEFeHealth2012—the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis. NICTA (2015)

A Framework to Cluster Temporal Data Using Personalised Modelling Approach

Muhaini Othman^(✉), Siti Aisyah Mohamed, Mohd Hafizul Afifi Abdullah,
Munirah Mohd Yusof, and Rozlini Mohamed

Faculty of Computer Science and Information Technology, Universiti Tun Hussein
Onn Malaysia, 86400 Batu Pahat, Malaysia
muhaini@uthm.edu.my, {sitiaisyahmohamed, hafizul94}@gmail.com, {munirah,
rozlini}@uthm.edu.my

Abstract. This research paper is focused on the framework design of temporal data by using personalised modelling approach in order to cluster the temporal data. Real world problem on flood occurrences is used as a case study focusing only in Malaysia region. The data are designed according to the criteria needed for temporal data clustering, tested with three clustering techniques including K -means, X -means, and K -medoids. Rapid Miner is used for conducting the clustering processes. Finally, the result from each clustering method is compared to conclude and justify the best clustering approach for clustering temporal data.

Keywords: Personalised modelling · Temporal data · Clustering
Flood case study

1 Introduction

In a world that contain continuously changing systems of interest, a scientific study of temporal data analysis confronts a lot of problems as derived data are dynamic. In most dynamic systems, temporal features are used to describe the changes in its existing variables during the observation period. In this paper, an analysis based on real-world flood case study occurring in Malaysian region is presented. A personalised modelling approach is utilised to study, analyse, and cluster the data, therefore forms a specific model for each data point within a localised problem space [1].

The main problem addressed in this paper is the difficulties in observing and analysing temporal data due to its dynamic behaviour. Hence, a personalised modelling approach is utilised as it focuses on the specific case or data vector; not the model. We have noticed the lack in recent researches which specifically focus on the design of the temporal data clustering technique, although lots of problems occur around us requires temporal data analytic as a solution. In addition, it is difficult to distinguish which of the clustering techniques offers the

best fit to be used for clustering a temporal data since most of the technique had only been tested on a static data type.

Three objectives has been set for this research. The first objective is to design temporal data framework based on the region and features of flood case study. Second objective is to analyse and identify various clustering technique possible for analysing the temporal data. The third objective is to perform the experimentation by using different clustering techniques, justify, and concludes the best technique for clustering temporal data.

Within this research, three clustering techniques is applied for analysing temporal data are K -means, X -means, and K -medoids. These clustering techniques are used to cluster the data from the real-world case study in Malaysian region (provided by Jabatan Meteorologi Malaysia) by grouping it in two types of clusters; flood and non-flood, based on the features values provided by each region. The data contains four features which are rainfall (mm), daily mean wind speed (ms^{-1}), daily mean relative humidity (%), and solar radiation (MJ.m^{-2}). The provided data covers a total of 25 regions in Malaysia including Mersing and Senai (Johor); Alor Setar and Langkawi (Kedah); Kota Bharu and Kuala Krai (Kelantan); Bayan Lepas, Butterworth, and Prai (Pulau Pinang); Tanah Tinggi Cameron, Kuantan, and Muadzam Shah (Pahang); Ipoh and Sitiawan (Perak); Kota Kinabalu and Kudat (Sabah); Kuching, Miri, Sibu, and Sri Aman (Sarawak); Petaling Jaya and Subang (Selangor); and Gong Kedak and Kuala Terengganu (Terengganu).

Here, we introduce a framework to cluster temporal data. Within Methodology in Sect. 2, we introduce several clustering algorithms. Section 3 presents current works by researchers, Sect. 4 presents the approach for the experimentation, Sect. 5 presents the results, and finally Sect. 6 concludes the paper.

2 Clustering Methodology

Clustering [2] attempts to structure data vectors by grouping them together into clusters based on their similar properties and values optimally. Traditional clustering methods, however, are designed to analyse data described with static feature values [2–5].

2.1 K -means Algorithm

K -means [6] clustering algorithm aims to find the position of a centroid by computing for mean position of data vectors at condition which minimises the distance between the data vectors within a cluster and has the furthest separation between different clusters. Separation distance between clusters and data vectors are calculated by solving Euclidean distance, d represented in Eq. 1.

$$d = \frac{y_k - y_i}{x_k - x_i} \quad (1)$$

where distances are measured between two data points, say data point $k(x_k, y_k)$ and data point $i(x_i, y_i)$. Difference measure between position between axis- y and axis- x of two data vectors is the value of distance measured.

The algorithm attempts to minimise the distance between data vectors within the same cluster, thus producing a high accuracy result by using a mathematical formula represented in Eq. 2.

$$\operatorname{argmin}_c \sum_{i=1}^k \sum_{X \in c_i} d(X, \mu_i) = \operatorname{argmin}_c \sum_{i=1}^k \sum_{X \in c_i} \|X - \mu_i\| \quad (2)$$

where c_i is the set of points that belong to cluster i .

The K -means algorithm however requires number of cluster to allow computational of clustering process. Therefore, the value should first pre-defined, or the number of cluster shall be decided using steps as in Table 1.

Table 1. The steps of applying the K -means algorithm

	Steps	Mathematical notation
1	Initialize the centre of the clusters	$\mu_i = \text{some value}; i = 1, \dots, k$
2	Attribute the closest cluster to each data point	$C_i = \{j : d(X_j, \mu_i) \leq d(X_j, \mu_l), l \neq i, j = 1, \dots, n\}$
3	Set the position of each cluster to the mean of all data vectors belonging to that cluster	$\mu_i = \frac{1}{ C_i } \sum_{j \in C_i} X_j, \forall i$
4	Repeat Steps 2–3 until convergence	
*	Notation	$ c = \text{Number of elements in } C$

The algorithm eventually converges several data vectors to a point, and stops when the assignments remain the same from one iteration to the next, defining the position of a centroid for a cluster.

Several drawbacks of K -means algorithm includes difficulties to find k -value; ineffectiveness to be used with global cluster; initial cluster partitioning affects the final result for clustering; and inconsistent density and size of cluster is not handled by the clustering algorithm.

2.2 X -means Algorithm

X -means clustering algorithm is an extension to the K -means which determines the number of centroids based on heuristic methods. X -means sets a minimum set of centroids and improves iteratively, locally decides which subset of the current centroids should split themselves in order to better fit the data. The centroid splitting is decided using the algorithm by Dan Pelleg and Andrew Moore [7], consisting of the following operations and is repeated until completion, as in Algorithm 1.

Algorithm 1 Cluster optimization in X -means

```

while cluster is not optimised do
  Improve parameters
  Improve structure
  if  $K > K_{max}$  then
    cluster is optimised
    stop and report the best scoring model found
  else
    cluster is still not optimised
  end
end

```

Improve parameters consists of conventional K -means to convergence, while improve structure finds out if a new centroid should be created and decides its location on the orthogonal plane. To decide which centroid is to be splitted for the creation of a new centroid, two obvious strategies are described and dismissed, then combine the strengths and avoid their weakness by using X -means.

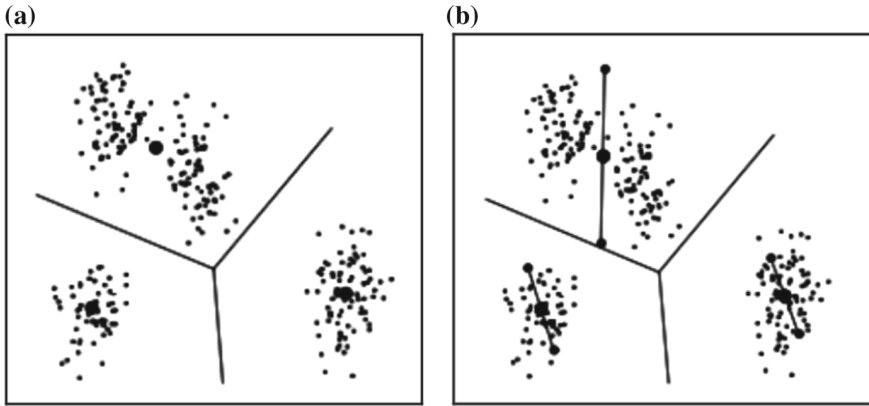


Fig. 1. **a** The result of running K -means with three centroids (left). **b** Each original centroid splits into two children (right)

Figure 1a shows a stable K -means solution with three centroids, where regions covered by each centroid is bounded by the visualised lines. Figure 1b visualises structure improvement operations by splitting each centroid into two children which are moved to a distance that is proportional to the size of the region in opposite directions along a randomly chosen vector. Each parent region then runs a local K -means using $K = 2$ for each pair of children.

2.3 K -medoids Algorithm

The K -medoids algorithm [8], shown in Algorithm 2 is used to find medoids, a centre located point of a cluster. K -medoids is stronger than K -means as it finds

k -value as representative data vector to minimise the sum of dissimilarities of data vectors whereas, K -means utilises sum of squared Euclidean distances for data vectors. This distance metric reduces noise and outliers.

Algorithm 2 K -medoids clustering algorithm

Input: Clusters (K_y); dataset with n data vectors (D_y)

Output: A set of K_y clusters

while *cluster is not optimised* **do**

 Randomly select K_y as the medoids for n data vectors

 Find closest medoids between data vectors and medoids

foreach *medoid* **do**

 Swap m and o to compute the total cost of configuration

 Select the medoids o with the lowest cost of configuration

end

end

3 Current Works on Temporal Data Clustering

Static data consists of time-invariant properties which is different in comparison to temporal data which are capable of capturing dynamic aspects of the system behaviour. Each temporal data has time as a second dimension, in respect to the existing variables. Given a temporal data with K temporal features, each represented as a sequence of L temporal values, a data vector x_i can be represented as a $K \times L$ matrix:

$$x_i = \begin{pmatrix} v_{11}^i & v_{12}^i & \dots & v_{1L}^i \\ v_{21}^i & \ddots & & v_{2L}^i \\ \vdots & & \ddots & \vdots \\ v_{K1}^i & v_{K2}^i & \dots & v_{KL}^i \end{pmatrix}$$

where v_{KL}^i represents the value of temporal feature K at time L . A simple way to represent and interpret the temporal data is by considering the feature values at each time step independently. For a single temporal data vector x_i , create L static data vectors such $x_1^i, x_2^i, \dots, x_L^i$, each represented as a feature value vector of size K :

$$\begin{aligned} x_1^i &= v_{11}^i \ v_{12}^i \ \dots \ v_{1K}^i \\ x_2^i &= v_{21}^i \ v_{22}^i \ \dots \ v_{2K}^i \\ &\dots \\ x_L^i &= v_{L1}^i \ v_{L2}^i \ \dots \ v_{LK}^i \end{aligned}$$

This results in a new data set with N -data vectors and reduces the temporal data set to an equivalent static representation. The clustering algorithms described in the previous section is then applied to group this data. Considering a dataset where data vectors are described by one temporal feature. Table 2

Table 2. Four data vectors described using a single temporal feature

Data vector	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1	10.5	10	9.5	9	9.5	10
2	11	10.5	10	9.5	9	9.5
3	11	10.5	10	9.5	10	10.5
4	10.5	10.5	9.5	9.5	9.5	10.5

shows the feature value description for four such data vectors which length of each temporal value is six, corresponding to $t = 1, t = 2, \dots, t = 6$.

This scheme completely loses information over time because the representation converts time-ordered sequences into a set of independent features, therefore, the temporal aspects of the data are lost. Figure 2 illustrates two cases where, after applying this conversion method, a conventional static data clustering scheme may fail to discover the similar patterns among data vectors and therefore, fail to group them into the same cluster. The first case illustrates a situation where the time varying characteristics such as the trend of two data vectors are very similar, yet the magnitude of the values of the two are different. Figure 2a illustrates two data vectors; data vector 1 and data vector 3, whose feature values differ by 1 unit at each time point. Another case arises when the two data vectors share a similar trend, but the temporal values of the two may have a delay by K . Figure 2b provides such an example, where there is a delay of 1 unit between the feature values of data vector 1 and data vector 2. By comparing values at individual time step separately, conventional static data clustering scheme may fail to capture the similar temporal trends in data.

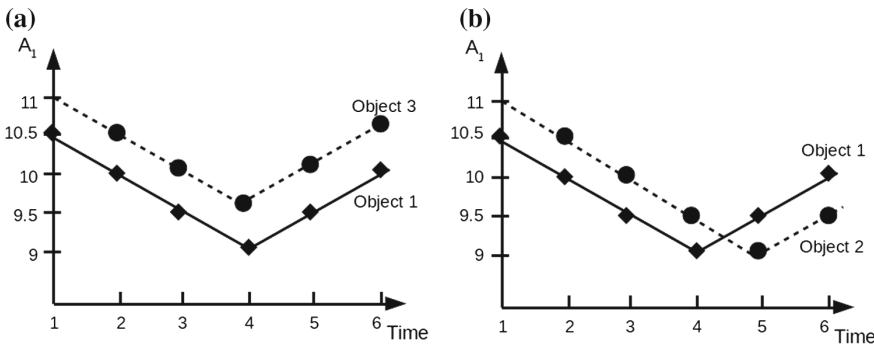


Fig. 2. Two cases where conventional clustering scheme may fail to capture similarities between temporal data vectors

4 Methodology

The research is conducted accordingly to the data preparation phase, experimentation phase, and evaluation phase.

4.1 Data Preparation Phase

During the data preparation phase, the datasets were obtained from Jabatan Meteorologi Malaysia. The datasets consist of 25 samples, each with five features, and covered for 31 days of reading. During data pre-processing, the temporal data is formatted to best fit Rapid Miner data mining tool using a spreadsheet software. Temporal data contain the dynamic spatio-temporal daily record. Hence, each data recorded must be included in the analysis together with the region of the data collected and record of all the features must be at the same time elapse.

4.2 Experimentation Phase

The number of clusters set for each method is $K = 2$, where Class 0 represents flood, and Class 1 represents non-flood. The dataset undergo each clustering algorithm by using cluster model prepared in Rapid Miner and the result obtained is saved.

4.3 Evaluation Phase

The selection of result with the highest accuracy is done by comparing the result obtained in the purity test. Purity test [5] is used to evaluate the cluster quality by computing for percentage of data vectors which is labelled correctly to its corresponding cluster. Purity is computed using Eq. 3:

$$Purity = \frac{1}{N} \sum_{i=1}^k \max_j |c_i \cap t_k| \quad (3)$$

where N represents the total number of data vectors, k represents the number of total clusters available, c_i is a cluster in c , and t_j is the classification with the maximum count for cluster c_i . The better the result of a clustering technique, the closer the result of purity test to value 1. High purity is easily achieved in case of the analysis has a high ratio between data vectors and centroid. In particular, purity value of 1 can be produced by giving data vectors its own cluster, hence it is not suitable to be used as a trade off for the quality of cluster produced.

5 Results and Discussion

The result for each clustering techniques contains a few similarities and difference since each clustering techniques applied a different algorithm. Table 3 presents

the actual class for regions, while Table 4 shows the result gained by each clustering algorithm used. To find the best technique to cluster a temporal data, the result summarized in Table 4 is compared to its actual class of the regions in Table 3, where the most similar and consistent matches is preferred.

Table 3. Actual class for state/region based on data by Jabatan Pengairan dan Saliran Malaysia

State	Class	Region
Johor	Flood	Mersing, Senai
Kedah	Flood	Alor Setar, Langkawi
Kelantan	Flood	Kota Bharu, Kuala Krai
Pulau Pinang	Non-Flood	Bayan Lepas, Butterworth
Pulau Pinang	Flood	Prai
Pahang	Non-Flood	Cameron, Muadzam Shah
Pahang	Flood	Kuantan
Perak	Non-Flood	Sitiawan
Perak	Flood	Ipoh
Sabah	Non-Flood	Kota Kinabalu, Kudat, Sandakan
Sarawak	Non-Flood	Kuching, Sibu, Sri Aman
Sarawak	Flood	Miri
Selangor	Non-Flood	Petaling Jaya, Subang
Terengganu	Flood	Gong Kedak, Kuala Terengganu

5.1 Purity Test

The result of three clustering technique is evaluated using purity test as follows:

1. Create confusion matrix to compute for purity test by iterating each cluster c_i and compute the number of data vector classified as each class t_i .
2. For each cluster c_i , select the maximum value from each row, sum them together and finally divide by the total number of data vectors.

5.2 Discussion

The greater the value obtained from a purity test, the better the clustering technique performs to cluster temporal data. Therefore, by referring to Table 5, the best method that can be used to cluster a temporal data type is using X -means and K -medoids, where both techniques have the same high value of purity test, 0.72, compared to a lower value obtained by K -means, 0.64.

Table 4. Comparison of the result produced by different clustering algorithms

Techniques	Cluster 0 (Non-Flood)	Cluster 1 (Flood)
<i>K</i> -means	Alor Setar, Langkawi, Bayan Lepas, Kuantan, Ipoh, Sitiawan, Kota Kinabalu, Kudat, Sandakan, Kuching, Miri, Sibul, Sri Aman, and Subang	Mersing, Senai, Gong Kedak, Kota Bharu, Kuala, Krai, Butterworth, Prai, Cameron, Muadzam Shah, Petaling Jaya, and Kuala Terengganu
<i>X</i> -means	Alor Setar, Langkawi, Bayan Lepas, Cameron, Kuantan, Muadzam Shah, Ipoh, Sitiawan, Kota Kinabalu, Kudat, Sandakan, Kuching, Miri, Sibul, Sri Aman, and Subang	Mersing, Senai, Gong Kedak, Kota Bharu, Kuala Krai, Butterworth, Prai, Petaling Jaya, and Kuala Terengganu
<i>K</i> -medoids	Alor Setar, Langkawi, Kuala Krai, Bayan Lepas, Butterworth, Prai, Cameron, Kuantan, Muadzam Shah, Ipoh, Sitiawan, Kota Kinabalu, Kudat, Sandakan, Kuching, Miri, Sibul, Sri Aman, Petaling Jaya, and Subang	Mersing, Senai, Gong Kedak, Kota Bharu, and Kuala Terengganu

Table 5. The result of purity test for clustering result using different algorithms

Result	<i>K</i> -means	<i>X</i> -means	<i>K</i> -medoids
Purity test value	0.64	0.72	0.72
Percentage	64%	72%	72%

X-means performs with better accuracy compared to *K*-medoids since the *X*-means cluster result obtained in Table 4 is much more similar than the actual result as compared to the clustering result using *K*-medoids. The different of cluster result in *X*-means and *K*-medoids is only in two regions; Kuala Krai and Prai where *X*-means had successfully cluster both region in Cluster 1 which specifically label as flood region and *K*-medoids cluster both regions in Cluster 0 which is known as a non-flood region. Therefore, we conclude that *X*-means is the most suitable technique for clustering datasets of temporal data type.

6 Conclusion

In the future, the research can be further extended to case study other than this flood case study where a new method or framework for clustering temporal data may be proposed using more suitable techniques such as neural networks and spiking neural networks to achieve a much better result.

Another application for future research includes prediction of the area with risk of flooding by feeding the output gathered from this research frame-

work into a spiking neural network system incorporating personalised modelling approaches, called the NeuCube-M1 [9] which is applicable on various domains including learning brain activity [10], stroke prediction [11, 12], and extracting knowledge from ecological data [13]. The capability of spiking neural networks to process continuous input of data allows us to observe the correlation within spatio and temporal components of the spatio-temporal data, hence allow an early prevention step to be taken in order to minimise losses due to the disaster.

Acknowledgements. This work is supported by the Fundamental Research Grant (Vot 1612) from Ministry of Higher Education Malaysia, Universiti Tun Hussein Onn Malaysia (UTHM), and GATES IT Solution Sdn. Bhd.

References

1. Keim, D.A., Hinneburg, A.: Clustering techniques for large data sets from the past to the future. In: Tutorial Notes of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 141–181. ACM (1999)
2. Biswas, G., Weinberg, J., Li, C.: A conceptual clustering method for knowledge discovery in databases. In: Artificial Intelligence in the Petroleum Industry, pp. 111–139. Editions Technip (1995)
3. Cheeseman, P., Stutz, J.: Bayesian classification (Autoclass): theory and results. *Advances in Knowledge Discovery and Data Mining*, pp. 153–180. MIT Press, Cambridge (1996)
4. Fisher, D.H.: Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.* **2**, 139–172 (1987)
5. Jain, A.K., Dubes, R.C.: *Algorithms for Clustering Data*. Prentice-Hall, Chicago (1988)
6. K-means clustering. <http://www.onmyphd.com/?p=k-means.clustering>. Accessed 16 Nov 2016
7. Pelleg, D., Moore, A.W.: X-means: extending K-means with efficient estimation of the number of clusters. *ICML* **1**, 727–734 (2000)
8. Arora, P., Virmani, D., Varshney, S.: Analysis of K-means and K-medoids algorithm for big data. *Procedia Comput. Sci.* **78**, 507–512 (2016)
9. Kasabov, N., Scott, N.M., Tu, E., Marks, S., Sengupta, N., Capecci, E.: Evolving spatio-temporal data machines based on the NeuCube neuromorphic framework: design methodology and selected applications. *Neural Netw.* **78**, 1–14 (2016)
10. Kasabov, N.: NeuCube: a spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. *Neural Netw.* **52**, 62–76 (2014)
11. Kasabov, N., Feigin, V., Hou, Z.G., Chen, Y., Liang, L., Krishnamurthi, R., Othman, M., Parma, P.: Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke. *Neurocomputing* **134**, 269–279 (2014)
12. Othman, M., Kasabov, N., Tu, E., Feigin, V., Krishnamurthi, R., Hou, Z.G., Chen, Y., Hu, J.: Improved predictive personalized modelling with the use of spiking neural network system and a case study on stroke occurrences data. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 3197–3204 (2014)
13. Tu, E., Kasabov, N., Othman, M., Li, Y., Worner, S., Yang, J., Jia, Z.: NeuCube (ST) for spatio-temporal data predictive modelling with a case study on ecological data. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 638–645 (2014)

Measurement of the Pitch Exploration Amongst Elite Professional Soccer Players: Official Match Analysis

Filipe Manuel Clemente^{1,2(✉)}, Adam Owen³, Aida Mustapha⁴,
Cornelis M. I. (Niels) van der Linden⁵, João Ribeiro⁶,
Bruno Mendes⁷, and Jelle Reichert⁵

¹ School of Sport and Leisure, Viana do Castelo Polytechnic Institute, Complexo Desportivo e Lazer Comendador Rui Solheiro – Monte de Prado, 4960-320 Melgaço, Portugal

flipe.clemente5@gmail.com

² Instituto de Telecomunicações, Delegação da Covilhã, Covilhã, Portugal

³ Centre de Recherche et d'Innovation sur le Sport, Université Claude Bernard Lyon. 1, Lyon, France

⁴ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia

⁵ Department of Sports Sciences, JOHAN Sports, Noordwijk, The Netherlands

⁶ Gabinete de Otimização Desportiva, Sporting Clube de Braga, Braga, Portugal

⁷ BenficaLab, Sport Lisboa e Benfica, Lisbon, Portugal

Abstract. Analysis of the physical and technical aspects of professional soccer is well reported, however the tactical analysis of the elite level is still limited in its conception. The purpose of this exploratory study was to measure the spatial exploration index of elite professional soccer players during official matches inclusive of positional analysis variants. Differences between 1st and 2nd half of match play was also analyzed. The investigation involved the analysis of six-official elite professional soccer matches. Fourteen players participated in the study which included games from the Portuguese National Premier League. Players were tracked with a 10 Hz GPS. Spatial exploration index was computed based on position-data from GPS. Results revealed significant differences between playing positions ($p = 0.001$; $\eta^2 = 0.213$). Wide forwards and centre forwards had the highest values of spatial exploration index with central defenders the lowest. No significant differences were found between 1st and 2nd half of the matches, however repeated measures revealed significant variances between matches ($p = 0.003$; $\eta^2 = 0.995$). In conclusion, it was revealed that that wide forwards and centre forwards are the positional players whom deviate further from their typical middle point when playing this position.

Keywords: GPS · Tactics · External load · Football

1 Introduction

Recent literature has reported how different monitoring devices are used in team sports, especially soccer to instantly track the positional movement of players [1]. One of the most used and well-published devices is Global Positioning System (GPS) that can be used in both training and matches scenarios [2]. GPS is a satellite navigation network device that provides location and real-time data tracking in different contexts [3]. From such a system it is possible to determine the positional point of the players in x- (goal-to-goal) and y-coordinates (lateral-to-lateral) at a frequency between 10 and 15 Hz [3]. The positional data may be then treated as raw data (for different scientific proposals) or using distances at different speed threshold for generic workload monitoring [2]. One of the examples of raw data treatment and application of the positional data is to compute tactical measures that may characterize the individual and collective patterns of players [4, 5].

1.1 Related Work

Position-based tactical measures have significantly increased since research conducted amongst small-sided games (SSGs) tracked the centroid and surface area constituted by soccer players. Furthermore, the integration and the approach proposed by Taki to identify the dominant regions [6], has also played a huge part in the increased attention to research [7]. As a result, collective and individual measures have been proposed to classify patterns of spatio-temporal interactions and the individual behavior in soccer match-related events [8, 9]. In the specific case of player measures, variability of trajectories have been identified by using different non-linear statistics [10]. Kolmogorov, Shannon and approximate entropies are the most used variability measures used to classify the predictability and regularity of player's trajectories during match-related event [11, 12]. Recent literature comparing the effects of different pitch sizes in SSGs proposed that smaller pitches lead to greater unpredictable action zones and less attraction towards the player's spatial positional zones [11]. Another study analyzing the variability of players during a single official match found that midfielders were the most variable and unpredictable players during the match [12].

The spatial region of players have been also analyzed by using dominant regions and Voronoi diagrams [7, 13]. Both measures provide information about the division of labor between teammates and characterize the oscillations of spatial occupation of players in specific moments of the game [8, 13]. In an example provided in a review it was found that the individual area of players varies based on the defensive and attacking status [8].

Variability of trajectories and spatial domain of players provide important information to whom analyze the game. However, the manner that a player explores the pitch can also reveal patterns of behavior that can be used to improve the knowledge about their regularity in different playing scenarios. To analyze that, the spatial exploration index was proposed by Gonçalves et al. [14] to identify the effects of pitch area-restrictions on physical and tactical behavior of players. The main results of the study revealed higher values of spatial exploration index under free-spacing conditions [14].

1.2 Statement of Contribution and Aim of the Study

Despite of the first approach with spatial exploration index, to date according to the authors knowledge no current application examining 11 versus 11 competitive match-play has been performed. The inclusion of this type of analysis may help to characterize the exploration indexes of different playing positions and identify if player's keep such indexes in a set of consecutive matches. The stability of such exploration and the variance between teammates may provide useful information to understand the team's dynamics and the specificity of playing role considering the occupation in the pitch.

Behavior of players can be constrained among others by the context, playing role, match status, teammates, or opponents but mainly positional constraints are generally based on the position of the ball [15]. Due to this the capacity to observe and understand movements, and to explore the pitch is highly related and influenced by the playing position and coaching strategy and demands employed. With this in mind, one of the key aims of this study was to identify the exploration profile of different playing positions during official soccer matches. Moreover, the variance between the 1st versus 2nd half of the game was notably analyzed and linked to identify how different results may influence the figures during the game.

2 Methods

2.1 Sample

Six official competitive elite professional matches including fourteen players (>19 years old) were used across this investigation phase. The players at the time were playing within the Portuguese National Premier League. A minimum playing time accumulating of 45 min during the competitive matches were established a pat of the inclusion criteria. Data collection for the analysis was performed across the mid-section of the in-season phase (2016/2017). Goalkeeper data were excluded from the analysis. Players were informed about the benefits and risks of using GPS trackers during the matches. The study was performed in line with the Declaration of Helsinki.

2.2 Procedures of Data Collection

The six official matches were tracked consecutively, thus including home and away matches. All games lasted the 90 min plus the extra time. Before the commencement of each game, all players performed a standardized pre-game warm-up protocol inclusive of ~25 min consisting of light-to-moderate running, dynamic stretching, mobility, fast running and ball possession and technical drills. All games were played on a natural turf-pitch. Within the game the team tactically assumed the 1-4-3-3 system of play with playing positions defined as 1 goalkeeper (not included in the data acquisition), 2 external defenders, 2 central defenders, 2 midfielders, 1 attacking midfielder, 2 wide forwards and 1 centre forward.

Each starting player within the game wore a GPS tracking pod (10 Hz, including EGNOS correction) which provided key informative data such as accelerometer,

gyroscope and magnetometer (100 Hz, 3 axis, ± 16 g). The GPS sensor used in this study was deemed as highly reliable for the analysis and had been tested with a $2.5 \pm 0.41\%$ (error \pm deviation) reliability for total distance covered [16].

All players were full familiar with the use of the GPS system due to the daily use of the equipment during training and previous competitive matches. Each GPS system for the individual was inserted into the back of tight but comfortable, standard GPS vest to ensure validity throughout the testing procedure remained for the (e.g. body oriented) accelerometer data [16]. Each individual GPS system tracker was uploaded for analysis immediately post-match using the JOHAN Sports¹ online analysis platform.

2.3 Spatial Exploration Index

The spatial exploration index was firstly proposed by Gonçalves et al. [14]. The measure uses the average position of player and all player's positions in the time-series to analyze how far the player can go from their average position. The measure can be calculated as [17]:

$$\text{Spatial exploration index} = \frac{\sum_i^N \sqrt{(x_i - x_m)^2 + (y_i - y_m)^2}}{N}, \quad (1)$$

where N is the number of instants of time for which Spatial Exploration Index is being computed, (x_m, y_m) represents the average position of each player during the period and (x_i, y_i) is the position in the instant i [17].

The player's spatial exploration index was calculated for each half of the match during all matches.

2.4 Statistical Procedures

Analysis of variance compared the spatial exploration index between playing positions (central defenders, external defenders, midfielders, wings and striker) and halves of the match (1st and 2nd half). The two-way ANOVA tested the interaction between factors for the dependent variable of spatial exploration index. The one-way ANOVA tested the variance of exploration index between playing positions, followed by the calculus of partial eta squared (η^2) to estimate the effect size and the Turkey HSD for post hoc analysis. Repeated measures tested the variance of spatial exploration index between matches. The classification of η^2 magnitude was done in the following way [18]: no effect ($\eta^2 < 0.04$); minimum effect ($0.04 < \eta^2 < 0.25$); moderate effect ($0.25 < \eta^2 < 0.64$); and strong effect ($\eta^2 > 0.64$). In the case of analysis between halves, the independent t-test was used, followed by the calculus of Cohen d (d). The classification of magnitude of d was made in the following way [18]: no effect ($d < 0.41$), minimum effect ($0.41 < d < 1.15$), moderate effect ($1.15 < d < 2.70$) and strong effect ($d > 2.70$). All the statistical procedures were done in the SPSS software (version 23.0, IBM, USA) for a $p < 0.05$.

¹ <http://www.johan-sports.com>.

3 Results

The two-way ANOVA revealed no significant interactions between playing position and halves of the match for the spatial exploration index variable ($p = 0.275$; $\eta^2 = 0.893$). No significant variance between halves of the match were found in the spatial exploration index ($t = 0.759$; $d = 0.245$). However, one-way ANOVA revealed significant differences between playing positions ($p = 0.001$; $\eta^2 = 0.213$). Descriptive statistics of spatial exploration index per playing position can be observed in Table 1.

Table 1. Descriptive statistics (mean (M) \pm standard deviation (SD)) of spatial exploration index per playing position

Playing position	Half	Spatial exploration index	
		M	SD
Full back	1st half	22.48	3.22
	2nd half	23.61	3.24
Central defender	1st half	20.06	2.73
	2nd half	22.01	3.89
Midfielder	1st half	22.00	2.71
	2nd half	22.97	3.38
Winger	1st half	24.73	3.68
	2nd half	25.34	3.91
Striker	1st half	26.11	3.96
	2nd half	25.68	2.89

Post hoc test for playing position revealed that central defenders had significant smaller values of spatial exploration index in comparison to wide forwards ($p = 0.001$) and centre forwards ($p = 0.001$). It was also found that both wide forwards and centre forwards had significant greater values of spatial exploration index than midfielders ($p = 0.039$ and $p = 0.032$, respectively).

Repeated measures tested the variance of spatial exploration index between the six matches. Significant differences were found between matches ($p = 0.003$; $\eta^2 = 0.995$). Descriptive statistics can be observed in Fig. 1.

Significant greater values ($p = 0.042$) of spatial exploration index were found in match 4 (25.61 ± 1.54) in comparison with match 1 (20.63 ± 1.31). It was also found significant greater values ($p = 0.027$) of spatial exploration index in match 2 (24.44 ± 1.23) in comparison to match 3 (17.90 ± 1.54).

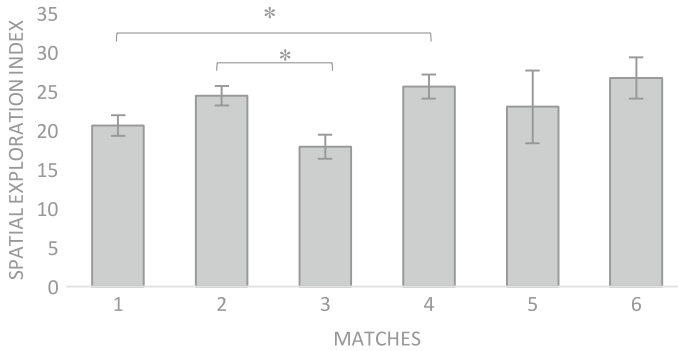


Fig. 1. Descriptive statistics (mean (M) \pm standard deviation (SD)) of spatial exploration index per match

4 Discussion

This exploratory study conducted in six official matches allowed the authors to identify significant differences in spatial exploration index between playing positions. Results revealed that wide-forwards and center forwards were the positional roles that most explored or deviated significantly around the pitch from their central regular position. Using Shannon and approximate entropies, previous research reported how central midfielders are the most unpredictable and variable players into describe trajectories in the pitch [12]. However, in our study wide forwards and centre forwards were who drift apart from their middle position, thus suggesting that the specific role can influence the individual behavior of exploring the pitch. In a study that analyzed the counter-attack profile of Portuguese teams, it was found that wide forwards and centre forwards were the most involved players in terms of receiving the ball highlighting how on transition or winning back possession teams looked to play to an offensive and directly [19]. The most recruited players may be often the players with greater approximation movements to the players from midline and backwards, thus increasing the space to run from their specific location in the pitch.

Another explanation to these results can be associated with movements made during defensive moments. In some cases, centre forwards and wide forwards must retreat a lot in the pitch to be closer to their defensive line, thus also increasing occupied areas in the pitch.

In the other hand, central defenders were the playing positions with smaller values of spatial exploration index. This position can be constrained by the specific role. Some time-motion analysis have been confirming the smaller distances covered by these players [20]. Apart from the smaller distances, these players also are the more stable players in their positions [12], thus confirming their tendency to stretch less than other players from his middle position.

No significant differences of spatial exploration index were found between halves of the match. The values obtained for the players were similar between from the 1st to the 2nd half, thus suggesting the players stabilize their patterns of exploration during

the match, based on his missions and roles. However, the repeated measures found variance of levels between matches, thus suggesting that spatial exploration may vary based on contextual variables as match status, opponents, match location and possession of the ball. Some previous works revealed that these contextual variables may influence the time-motion profile of players and also their actions [21, 22]. In our study, such variables were not taken in account but in future studies must to be considered to better explain the variance from match to match and to identify which factors may contribute to justify such variances in individual behavior.

This study had some limitations. The match location (home vs. away) was not considered for the analysis between matches. Moreover, the final score was not included as contextual variable to explain the variance between matches. These variables may have influenced the final results and for that reason must be considered in future studies. The data used in this study only characterize one team however when performing analysis at the elite level of the game the opportunity to utilize control group and compare squads is minimal as shown in previous literature [23, 24]. In the future both teams must be considered to analyze the opponent's influence on the spatial exploration index. Finally, a bigger data must be considered to generalize the evidences.

Despite of the limitations, this study is the first that have analyzed the spatial exploration of professional soccer players in official matches. It was possible to observe that playing positions lead to different values. As practical implications, we hypothesize that spatial exploration index can be used to identify the individual profile of a players to exploit the pitch during match activities. Moreover, crossing this measures with workload analysis will be possible to identify the influence of such exploration on the external load of players. The use of this measure will be also possible in training sessions to identify if small-sided games are adjusted to the real tactical demands of the game and can provide feedback to coaches to change some patterns of individual behavior.

5 Conclusions

It can be concluded from this specific, novel and exploratory investigation conducted at the elite level of professional soccer that wide forwards and centre forwards have shown to be the positional roles that significantly variate positionally from their typical centre playing or middle point. It may be associated with the positional demand and interchangeable roles they play when trying to become creative and individually focused in upsetting the opposition tactical strategy with their behaviors. Additionally, the potential defensive roles adhered to and required by the technical staff within these positions may also require larger deviations from their normal attacking roles which subsequently may influence the distance from the central point. Further examination of the data revealed that central defenders are the most stable positional roles or players within the team. Contextual variables associated with the match can contribute to justify the variances of spatial exploration index between matches. Secondary analysis of the data formed within the study revealed no significant difference in terms of spatial exploration index were found when comparing 1st versus 2nd halves of games, but

significant variances were found between matches. It may therefore be proposed that the exploration profile of players may vary based on playing position but importantly the contextual variables of the match may influence the capacity to explore space more or less within the dimensions of the pitch. As a result, future studies must consider the use of in-game real-time and post-game analysis of contextual variables to better justify the variances between matches with the profile of opponents being used to identify the association with the team's results.

Acknowledgements. We would like to thank to JOHAN sports for allowing the use of their GPS trackers.

References

1. Buchheit, M., Allen, A., Poon, T.K., Modonutti, M., Gregson, W., Di Salvo, V.: Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and GPS technologies. *J. Sports Sci.* **32**(20), 1844–1857 (2014)
2. Cummins, C., Orr, R., O'Connor, H., West, C.: Global positioning systems (GPS) and microtechnology sensors in team sports: a systematic review. *Sport. Med.* **43**(10), 1025–1042 (2013)
3. Malone, J.J., Lovell, R., Varley, M.C., Coutts, A.J.: Unpacking the black box: applications and considerations for using GPS devices in sport. *Int. J. Sports Physiol. Perform.* **12**(Suppl 2), S2-18–S2-26 (2017)
4. Memmert, D., Lemmink, K.A.P.M., Sampaio, J.: Current approaches to tactical performance analyses in soccer using position data. *Sport. Med.* **47**(1), 1–10 (2017)
5. Clemente, F.M., Couceiro, M.S., Martins, F.M.L., Mendes, R.S., Figueiredo, A.J.: Practical implementation of computational tactical metrics for the football game: towards an augmenting perception of coaches and sport analysts. In: Murgante, Misra, Rocha, Torre, Falcão, Taniar, Apduhan, Gervasi (eds.) *Computational Science and Its Applications*, pp. 712–727. Springer (2014)
6. Frencken, W., Lemmink, K.: Team kinematics of small-sided football games: a systematic approach. In: Reilly, T., Korkusuz, F. (eds.) *Science and Football VI*, pp. 161–166. Routledge Taylor & Francis Group, Oxon (2008)
7. Taki, T., Hasegawa, J.: Visualization of dominant region in team games and its application to teamwork analysis. *Proc. Comput. Graph. Int.* **2000**, 227–235 (2000)
8. Duarte, R., Araújo, D., Correia, V., Davids, K.: Sports teams as super organisms: implications of sociobiological models of behaviour for research and practice in team sports performance analysis. *Sport. Med.* **42**(8), 633–642 (2012)
9. Travassos, B., Davids, K., Araújo, D., Esteves, P.T.: Performance analysis in team sports: advances from an ecological dynamics approach. *Int. J. Perform. Anal. Sport* **13**(1), 83–95 (2013)
10. Fonseca, S., Milho, J., Passos, P., Araújo, D., Davids, K.: Approximate entropy normalized measures for analyzing social neurobiological systems. *J. Mot. Behav.* **44**(3), 179–183 (2012)
11. Silva, P., Aguiar, P., Duarte, R., Davids, K., Araújo, D., Garganta, J.: Effects of pitch size and skill level on tactical behaviours of association football players during small-sided and conditioned games. *Int. J. Sport. Sci. Coach.* **9**(5), 993–1006 (2014)
12. Couceiro, M.S., Clemente, F.M., Martins, F.M.L., Machado, J.A.T.: Dynamical stability and predictability of football players: the study of one match. *Entropy* **16**(2), 645–674 (2014)

13. Fonseca, S., Milho, J., Travassos, B., Araújo, D.: Spatial dynamics of team sports exposed by Voronoi diagrams. *Hum. Mov. Sci.* **31**(6), 1652–1659 (2012)
14. Gonçalves, B., Esteves, P., Folgado, H., Ric, A., Torrents, C., Sampaio, J.: Effects of pitch area-restrictions on tactical behavior, physical and physiological performances in soccer large-sided games. *J. Strength Cond. Res.* (2017) (vol. ahead-of-p)
15. McGarry, T.: Soccer as a dynamical system: some theoretical considerations. In: Reilly, T., Cabri, J., Araújo, D. (eds.) *Science and Football V*, pp. 570–579. Routledge, Taylor & Francis Group, London and New York (2005)
16. Clemente, F.M., Nikolaidis, P.T., Van Der Linden, C.M.I.N., Silva, B.: Effects of small-sided soccer games on internal and external load and lower limb power: a pilot study in collegiate players. *Hum. Mov.* **18**(1), 50–57 (2017)
17. Clemente, F.M., Sequeiros, J.B., Correia, A.F.P.P., Silva, F., Martins, F.M.L.: *Computational Metrics and Its Applications on the Analysis of Soccer: Connecting the dots*. Springer Singapore, Singapore (2017)
18. Ferguson, C.J.: An effect size primer: a guide for clinicians and researchers. *Prof. Psychol. Res. Pract.* **40**(5), 532–538 (2009)
19. Malta, P., Travassos, B.: Characterization of the defense-attack transition of a soccer team. *Motricidade* **10**(1), 27–37 (2014)
20. Carling, C.: Interpreting physical performance in professional soccer match-play: should we be more pragmatic in our approach? *Sports Med.* **43**(8), 655–663 (2013)
21. Lago-Peñas, C., Dellal, A.: Ball possession strategies in elite soccer according to the evolution of the match-score: the influence of situational variables. *J. Hum. Kinet.* **25**, 93–100 (2010)
22. Lago-Peñas, C., Lago-Ballesteros, J.: Game location and team quality effects on performance profiles in professional soccer. *J. Sport. Sci. Med.* **10**, 465–471 (2011)
23. Owen, A.L., Lagos-Penas, C., Gómez, M.A., Mendes, B., Dellal, A.: Analysis of a training mesocycle and positional quantification in elite European soccer players. *Int. J. Sports Sci. Coach.* **12**(5), 1–8 (2017)
24. Owen, A.L., Dunlop, G., Rouissi, M., Haddad, M., Mendes, B., Chamari, K.: Analysis of positional training loads (ratings of perceived exertion) during various-sided games in European professional soccer players. *Int. J. Sports Sci. Coach.* **11**(3), 1–8 (2016). <https://doi.org/10.1177/1747954116644064>

RMIL/AG: A New Class of Nonlinear Conjugate Gradient for Training Back Propagation Algorithm

Sri Mazura Muhammad Basri¹, Nazri Mohd Nawi^{2(✉)},
Mustafa Mamat³, and Norhamreeza Abdul Hamid²

¹ Faculty of Science, Technology and Human Development, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

² Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
nazri@uthm.edu.my

³ Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Kampus Tembila, 22200 Besut, Terengganu Darul Iman, Malaysia

Abstract. The conventional back propagation (BP) algorithm is generally known for some disadvantages, such as slow training, easy to getting trapped into local minima and being sensitive to the initial weights and bias. This paper introduced a new class of efficient second order conjugate gradient (CG) for training BP called Rivaie, Mustafa, Ismail and Leong (RMIL)/AG. The RMIL uses the value of adaptive gain parameter in the activation function to modify the gradient based search direction. The efficiency of the proposed method is verified by means of simulation on four classification problems. The results show that the computational efficiency of the proposed method was better than the conventional BP algorithm.

Keywords: Second order method · Back-propagation · Conjugate gradient Search direction · Classification

1 Introduction

During the past decade, Artificial Neural Networks (ANN) had been widely applied in the different fields. The ANN have been designed to model the relationships between independent and dependent variables and are capable of modeling complex, non-linear relationships directly from the raw data. Unlike classical statistical techniques, ANN requires less formal statistical training where it able to implicitly detect complex nonlinear relationships between dependent and independent variables, able to detect all possible interactions between predictor variables and the availability of multiple training algorithms. Several ANN algorithms have been developed and proposed. The most popular algorithm, called the back propagation (BP) can be used to update weights by the method of steepest descent [1]. It is capable of approximating most problems with high accuracy and generalization ability. Despite its popularity, this algorithm has some drawbacks, for example, converging to local minima instead of the

global minimum, temporal instability and poor scaling properties. However, BP's main problem is slow convergence when applied in different fields.

Therefore to overcome the problem of slow convergence in BP, many researchers have tried to improve the learning efficiency of BP algorithm through various enhancements [2–9]. Ooyen and Nienhuis [10] presented modifications in the error function used to measure the global net performance. Haykin [11] discussed several data-driven optimization training algorithms, such as Levenberg-Marquardt algorithm and scaled conjugate gradient algorithm, which may overcome these problems. While, [8] Kumar et al. [12] employed the Bayesian regularization for neural network training in order to improve the performance in pulse radar detection.

Other approaches in improving the speed of convergence of BP algorithm have been adopted from the numerical optimization theory. Bishop [13] in his book suggested the use of many optimization techniques in order to improve the efficiency of error minimization. Among these are second order methods such as Fletcher and Powell [14] and the Fletcher and Reeves [15] which improved the CG method of Hestenes and Stiefel [16]. It has been shown that these methods provide stable learning, robustness to oscillations and improved convergence rates.

A few researchers [17, 18] studied the influence of gain parameter to control the steepness of the activation function and found that a larger gain value has an equivalent effect of increasing the learning performance. Nazri et al. [19] discovered the adaptive gain (AG) in the activation function can accelerate the training time. The value of gain is modified adaptively for each node. Yet, a further improvement on [19] had been proposed and tested on the classification problem in [20, 21]. The results show that the proposed method in [20, 21] significantly improved the performance of BP.

Later, Nazri et al. [22, 23] introduced a new fast learning algorithm for neural networks based on second order method. In their research, they identified that the proposed algorithm improved the training efficiency of BP algorithms by adaptively modifying the gradient search direction using gain value. Also, the proposed algorithm is generic and easy to implement in all commonly used gradient based optimization processes [24–27]. The simulation results showed that the proposed algorithm is robust and has a potential to significantly enhance the computational efficiency of the training process.

Motivated by the previous works [19–21, 28, 29], this paper proposed an improved approach by using a new class of CG proposed by Rivaie et al. [30] to train BP algorithm. The experimental results on the proposed Rivaie, Mustafa, Ismail and Leong with Adaptive Gain (RMIL/AG) indicate that the proposed CG methods are efficient and have a potential to significantly enhance the computational efficiency and robustness of the training process.

The remainder of this paper is organized as follows: In Sect. 2, we present the CG algorithm and Sect. 3 presents the proposed algorithms. The implementation of the proposed algorithm with CG algorithm is discussed in Sect. 4. Section 5 presents our experiments and simulation results. The final section contains concluding remarks and short discussion for future research.

2 Conjugate Gradient

The Conjugate gradient (CG) methods are an alternative second order optimization method to find the minimum value of function for unconstrained optimization problems. This method was driven by Hestenes and Stiefel in solving an equation of symmetric positive definite linear [16] and improved by Fletcher and Reeves in solving unconstrained minimization problems [14]. The CG methods are probably the most famous iterative methods for efficiently training neural networks due to their simplicity, numerical efficiency and their very low memory requirements [31, 32].

The CG methods perform the search along the conjugate directions in order to determine the step size which defines the function of performance along that line. Although the function decreases most rapidly, when applied to non-linear function, the performance methods diverse. Thus, this research proposed gain function in the search direction in order to improve learning performance of the CG methods.

These methods are given by an iterative method of the form

$$x_{n+1} = x_n + \eta_n d_n, \quad n = 0, 1, 2, \dots \quad (1)$$

where n is the current iteration usually called epoch, x_n is the current iterate point, $\eta_n > 0$ is the learning rate and d_n is the descent search direction defined by

$$d_n = \begin{cases} -g_n & \text{if } n = 0, \\ -g_n + \beta_n d_{n-1} & \text{if } n \geq 1, \end{cases} \quad (2)$$

where g_n denotes the gradient of $f(x)$ at the point x_n . $\beta_n \in \mathbb{R}$ is a scalar known as the CG coefficient. Some well-known methods for β_n are Fletcher-Reeves (FR), Polak-Ribiere (PR) and Hestenes-Steifel (HS) respectively specified as follows:

$$\beta_n^{FR} = \frac{g_n^T g_n}{\|g_{n-1}\|^2} \quad (3)$$

$$\beta_n^{PR} = \frac{g_n^T (g_n - g_{n-1})}{\|g_{n-1}\|^2} \quad (4)$$

$$\beta_n^{HS} = \frac{g_n^T (g_n - g_{n-1})}{(g_n - g_{n-1})^T d_{n-1}} \quad (5)$$

where, $\| \cdot \|$ denotes the Euclidean norm of vectors.

In recent years, much effort has been placed on designing and constructing a new formula for CG methods with good numerical performance and that has global convergent properties. A new class of CG proposed by Rivaie et al. [30] is known as β_n^{RMIL} where RMIL denotes Rivaie, Mustafa, Ismail and Leong as below:

$$\beta_n^{RMIL} = \frac{g_n^T(g_n - g_{n-1})}{\|d_{n-1}\|^2} \quad (6)$$

Here, RMIL is introduced by modifying the denominator while retaining the original numerator as in the PR and HS. This new conjugate gradient is tested and shows that this method is very efficient when compared to the early CG coefficients for a given optimization test problems. Thus, this research proposed further improvement on [30] by adopting the used of adaptive gain.

3 The Proposed Method

In this section, a novel approach for improving the training efficiency of gradient descent method (back propagation algorithm) is presented. The proposed method modifies the gradient based search direction by changing the gain value adaptively for each node.

With an optimization perspective, the objective of a learning process in neural networks is to find a weight vector w that minimizes the difference between the actual output and the desired output on both training and testing data sets. Namely,

$$\min_{w \in \mathbb{R}^n} E(w) \quad (7)$$

Consider a multilayer feed forward neural network (MFNN) [1] with one output layer, one input layer and one or more hidden layers. Each layer has a set of units, nodes, or neurons. It is usually assumed that each layer is fully connected with a previous layer without direct connections between layers which are not consecutive. Each connection has a weight. For a particular input pattern, define an error function on that pattern as,

$$E = \frac{1}{2} \sum_k (t_k - o_k^L)^2 \quad (8)$$

where o_k^L be the activation of the k th node of layer L .

Let w_{ij}^L be the weight on the connection from the i th node in layer $L-1$ to the j th node in layer L . The overall error on the training set is simply the sum, across patterns, of the pattern error E .

The net input to the j th node of layer L is defined as $net_j^L = (w_j^L, o^{L-1}) = \sum_k w_{j,k}^L o_k^{L-1}$, The activation of a node o_j^L is given by a function of its net input,

$$o_j^L = f(c_j^L net_j^L) \quad (9)$$

where f is any function with bounded derivative, and c_j^L is a real value called the gain of the node. Note that at $c_j^L = 1$ this activation function becomes the usual logistic activation function.

The weight update expression in Eq. (6) with a non-unit gain value is derived by differentiating the error term as given in Eq. (2) with respect to w_{ij}^L as follows:

$$\begin{aligned} \frac{\partial E}{\partial w_{ij}^L} &= \frac{\partial E}{\partial net^{L+1}} \cdot \frac{\partial net^{L+1}}{\partial o_j^L} \cdot \frac{\partial o_j^L}{\partial net_j^L} \cdot \frac{\partial net_j^L}{\partial w_{ij}^L} \\ &= [-\delta_1^{L+1} \dots - \delta_n^{L+1}] \cdot \begin{bmatrix} w_{ij}^{L+1} \\ \vdots \\ w_{nj}^{L+1} \end{bmatrix} \cdot f'(c_j^L net_j^L) c_j^L \cdot o_j^{L-1} \end{aligned} \quad (7)$$

In particular, the first three factors of (4) indicate that

$$\delta_1^L = (\sum_k \delta_k^{L+1} w_{kj}^{L+1}) f'(c_j^L net_j^L) \quad (8)$$

As we noted that the iterative in Eq. (5) for δ_1^L is the same as standard back propagation [1] except for the appearance of the value gain. By combining (4) and (5) yields the learning rule for weights:

$$\begin{aligned} \Delta w_{ij}^L &= \eta \delta_j^L c_j^L o_j^{L-1} \\ &= \eta \frac{\partial E}{\partial w_{ij}^L} \end{aligned} \quad (9)$$

$$w_{new} = w_{old} + \Delta w_{ij}^L \quad (10)$$

where η is a “learning rate” and the search direction or gradient vector at point w_{ij}^L is $d = \frac{\partial E}{\partial w_{ij}^L} = g$.

In the proposed method the calculation of the gradient of error $g^{(n)}$ at step n is a function of gain $c_j^{L(n)}$.

$$d^{(n)} = -\frac{\partial E}{\partial w_{ij}^{L(n)}} (c_j^{L(n)}) = g^{(n)} (c_j^{L(n)}) \quad (11)$$

The gain value at step n is calculated using gradient of error *w.r.t.* to gain,

$$\frac{\partial E}{\partial c_j^L} = (\sum_k \delta_k^{L+1} w_{k,j}^{L+1}) f'(c_j^L net_j^L) net_j^L \quad (12)$$

Then the gradient descent rule for the gain value becomes,

$$\Delta c_j^L = \eta \delta_j^L \frac{net_j^L}{c_j^L} \quad (13)$$

At the end, each iterations of the new gain value is updated using a simple gradient based method as given by the formula,

$$c_j^{new} = c_j^{old} + \Delta c_j^L \quad (14)$$

The following section will discussed further the implementation of the proposed algorithm with new second order method known as RMIL.

4 The Implementation of the Proposed Algorithm with Conjugate Gradient Rivaie, Mustafa, Ismail and Leong (RMIL)

One of the remarkable properties of the conjugate gradient (CG) method is its ability to generate, in a very economical fashion, a set of vectors with a property known as conjugacy [13]. Most widely used CG algorithms are given by Fletcher and Powell [14] and Fletcher and Reeves [15]. Both these procedures generate conjugate search directions and therefore aim to minimize a positive definite quadratic function of n variables in n steps. The proposed algorithm referred to RMIL/AG begins the minimization process with an initial estimate w_0 and an initial search direction as:

$$d_0 = -\nabla E(w_0) = -g_0 \quad (15)$$

The search direction at $(n+1)$ th iteration is calculated as:

$$d_{(n-1)} = -\frac{\partial E}{\partial w_{(n+1)}}(c_{i,n+1}) + \beta_{(n+1)}d_n(c_{i,n}) \quad (16)$$

where the scalar $\beta_{(n+1)}$ is to be determined by the requirement that d_n and d_{n+1} must fulfil the conjugacy property [13]. There are many formulae for the parameter $\beta_{(n+1)}$ and the choice of the formulae for selection of $\beta_{(n+1)}$ is problem dependent [30]. In this paper, common formula as referred by Rivaie et al. [30] (RMIL) is used which has been stated in Eq. (6). Like $\beta_{(n)}$, the computation of learning rate η also requires knowledge as that of $\beta_{(n)}$. The learning rate η can be optimally chosen as to minimize the error $E(\eta)$ along the chosen search direction d_n .

$$E(\eta) = E(w_{(n)}(\eta)) = E(w_{n-1} + \eta_{n-1}d_{n-1}) \quad (17)$$

The given us an automatic procedure for the setting the learning rate, once the search direction is chosen. This procedure is also referred to as ‘line search’ method. In this paper we used golden section search technique to obtain optimized learning rate. The golden search technique starts by restricting η in $[\eta_1, \eta_h]$. In this paper wet set $\eta_1 > 0$ and $\eta_h < 1$, then the following steps are performed.

- Compute $E(\eta_1)$, $E(\eta_h)$
- If $E(\eta_1) < E(\eta_h)$, then set $\eta_h = \eta_1 - 0.618(\eta_1 - \eta_h)$

- If $E(\eta_1) > E(\eta_h)$, then set $\eta_h = \eta_1 + 0.618(\eta_1 - \eta_h)$
- The process is repeated until $(\eta_1 - \eta_h) < \varepsilon$ and then set $\eta = \frac{\eta_1 + \eta_h}{2}$.

The complete RMIL/AG algorithm works as follows:

- Step 1** Initialize the weight vector randomly, the gradient vector g_0 to zero and gain vector to unit values. Let the first search direction d_0 be g_0 . Set $\beta_0 = 0$, $epoch = 1$ and $n = 1$. Let Nt be the total number of weight values. Select a convergence tolerance CT .
- Step 2** At step n , evaluate gradient vector $g_n(c_n)$.
- Step 3** Evaluate $E(w_n)$. If $E(w_n) < CT$ then STOP training ELSE go to **step 4**.
- Step 4** Calculate a new gradient based search direction which is a function of gain parameter: $d_n = -g_n(c_n) + \beta_n d_{n-1}$.
- Step 5** IF $n > 1$ THEN,
 update $\beta_n = \frac{g_n^T(c_n)(g_n(c_n) - g_{n-1}(c_n))}{\|d_{n-1}(c_n)\|^2}$
 ELSE go to **step 6**.
- Step 6** IF $[(epoch + 1)/Nt] = 0$ THEN ‘restart’ the gradient vector with $d_n = -g_{n-1}(c_{n-1})$ ELSE go to **step 7**.
- Step 7** Calculate the optimal value for learning rate η_n^* by using line search technique such as described in Eq. (14).
- Step 8** Update $w_n : w_{n+1} : w_n - \eta_n^* d_n$
- Step 9** Evaluate new gradient vector $g_{n+1}(c_{n+1})$ with respect to gain value c_{n+1} .
- Step 10** Calculate new search direction:
 $d_{n+1} = -g_{n+1}(c_{n+1}) + \beta_{n+1}(c_n) d_n$
- Step 11** Set $n = n + 1$ and go to **step 2**.

5 Experiments and Simulation Results

In order to illustrate the advantage of the proposed algorithm, four benchmark problems have been used and compared among the Back Propagation (BP), Rivaie, the original Mustafa, Ismail and Leong (RMIL) and the proposed Rivaie, Mustafa, Ismail and Leong with Adaptive Gain (RMIL/AG). The performance criteria used in this research focuses on the speed of convergence, measured in number of iterations, CPU time and accuracy. Four benchmark problem dataset used for verification process are taken from the open literature [33] including Iris [34], Glass [35], Pima Indian diabetes [36] and breast cancer Wisconsin [37]. The simulations have been carried out on a Intel Core i3 with 2.40 GHz, 2 GB RAM and using MATLAB version 7.10.0 (R2010a).

To compare the performance of the proposed algorithm with BP and RMIL, network parameters such as network size and architecture (number of nodes, hidden layer and etc), values for the initial weight and gain parameters were kept the same. For all problems, three layers in multilayer perceptron are used for testing the models, the hidden layer is kept fixed to 5 nodes, while input and output layer nodes are vary according to the dataset given. Sigmoid activation function was used for all nodes.

Prior to the training, the weights are initialised to small random values range $[0, 1]$ and received the input patterns for training in the same sequence. The initial value for the gain parameter was 1. Learning rate value and momentum value was set to 0.3 and 0.4 respectively. All parameters were finalised based on an initial studies by Nazri et al. [38].

For each algorithm, 100 trials are tested. In order to compare the convergence rate, for each run, the number of epoch is required. For an experiment of 100 runs, the means of epoch, CPU times, accuracy and the number of failures are collected. A failures occurs when the network exceeds the maximum iteration limit; each experiment is run to a 1000 epoch to reach the target error; otherwise, it is halted and the run is reported as failure e. Convergence is achieved when the outputs of the network confirm to the error criterion as compared to the desired outputs.

- A. **Iris Classification Problem.** This dataset was a classical classification dataset made famous by Fisher, who used it to illustrate principles of discriminant analysis [34]. There were 75 instances, 4 inputs, and 3 outputs in this dataset. The classification of Iris dataset involves classifying the data of petal width, petal length, sepal width, and sepal length into three classes of species, which are Iris Sentosa, Iris Versicolor, and Iris Verginica. The selected network topology for Iris classification problem is 4-5-3, which is 4 input nodes, 5 hidden nodes and 3 output nodes. The target error was set as 0.02.

Table 1. Performance comparison for IRIS classification problem

	BP	RMIL	RMIL/AG
Mean of epoch	357	49	35
CPU time	2.57	0.54	0.38
Accuracy (%)	84.596	76.4655	75.1197
Failures	0	0	0

Table 1 proved that the proposed algorithm (RMIL/AG) outperforms other algorithms in terms of CPU time and number of epochs. The proposed algorithm (RMIL/AG) needs only 35 epochs to converge as opposed to the conventional BP at about 357 epochs while RMIL needs 49 epochs to converge. Apart from speed of convergence, the time required for training the classification problem is another important factor when analysing the performance. The results in Table 1 clearly show that the proposed algorithm (RMIL/AG) outperforms conventional BPGD with an improvement ratio, 10.2 s while RMIL, the proposed algorithm outperformed 1.4 s for the total time of converge. However, the accuracy of RMIL is much better than RMIL/AG and BP algorithm. This is mainly because the proposed method improved the search direction which reduced the convergence time and the total number of epochs.

- B. **Glass Classification Problem.** This dataset was collected by B. German on fragments of glass encountered in forensic work. The glass dataset is used for separating glass splinters into six classes, namely float processed building windows,

non-float processed building windows, vehicle windows, containers, tableware, or head lamps [35]. The selected architecture of the network is 9-5-6 with target error was set to 0.01.

Table 2. Performance comparison for glass classification problem

	BP	RMIL	RMIL/AG
Mean of epoch	1000	66	41
CPU time	12.93	1.73	1.07
Accuracy (%)	75.5778	79.4752	76.7274
Failures	100	0	0

Table 2 reveals that BP needs more than 12 s with 1000 epochs to converge, whereas RMIL needs 1.7318 s with 66 epochs to converge. Conversely, RMIL/AG performed significantly better with only 1.0666 s and took only 41 epochs to converge. We find that, the RMIL/AG requires essentially the same generalization results as RMIL and BP. The table also demonstrates that the performance of the RMIL/AG not only faster than two conventional method, but also much more stable which almost 96% faster than BP and almost 38% faster than RMIL. This is due to the improved approaches in finding the optimal search direction in the each iteration. Besides, the BP did not perform well in this dataset since 100% of simulation results failed in classified the patterns. It can also be seen that the conventional BP failed to converge to reach the target error.

C. *Pima Indians Diabetes*. Pima Indians Diabetes dataset is taken from the UCI Machine learning repository. Consisting of 768 instances, 8 inputs and 2 outputs, this dataset contains all the information of the chemical changes in a female body whose imbalance can cause diabetes [36]. The feed forward network architecture for this classification problem is set to 8-5-2. The target error is again set to 0.01.

Table 3. Performance comparison for Pima Indians diabetes classification problem

	BP	RMIL	RMIL/AG
Mean of epoch	869	93	64
CPU time	11.36	2.51	1.69
Accuracy (%)	74.2782	70.6469	67.183
Failures	72	0	0

Table 3 shows that the proposed RMIL/AG exhibit an excellent average performance to reach the target error with only required 64 epochs in 1.6946 s CPU times by 67.183% accurate. Whereas RMIL required 93 epochs in 2.513 s CPU times to reach target error with 70.6469% accurate. At the same time, BP needs 869 epochs in 11.358 s CPU times and 74.2782% accurate. As we can see in Table 3, the average number of learning iterations for the RMIL/AG was reduced up to 13.58 and 1.453

faster as compared to BP and RMIL respectively. BP not performed very well in this dataset since 72% failed to learn the patterns.

- D. **Breast Cancer Wisconsin.** This dataset was generated from University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [37]. The input attributes are for instance the clump thickness, the uniformity of cell size, the uniformity of cell shape, the amount of marginal adhesion, and the single epithelial cell size, frequency of bare nuclei, bland chromatin, normal nucleoli and mitoses. This problem tries to diagnosis of Wisconsin breast cancer by trying to classify a tumor as either benign or malignant based on cell description gathered by microscopic examination. The selected architecture of the network is 9-5-2 with target error 0.02.

Table 4. Performance comparison for breast cancer Wisconsin classification problem

	BP	RMIL	RMIL/AG
Mean of epoch	72	27	26
CPU time	0.86	0.67	0.69
Accuracy (%)	91.2433	93.6507	88.0926
Failures	18	0	0

From Fig. 4, it is worth noticing that the performance of the proposed RMIL/AG substantially improved than the RMIL and BP algorithm. Table 4 reveals that the RMIL/AG approximately took 2.65×10^{-2} CPU time per epoch to reach target error as well as 88.0926% accurate. While RMIL took 2.5×10^{-2} CPU time per epoch to reach target error as well as 93.6507% accurate and BP took 1.189×10^{-2} CPU time per epoch to reach target error as well as 91.2433% accurate. Yet, BP has 18% failure to learn the patterns which indicate that it is more unstable as compare to the RMIL/AG. Still, the second order surpasses the BP algorithm in terms of total time of converge and accuracy to learn the pattern. The results show that the new classes of conjugate gradient perform better as compared to the conventional BP algorithm. Moreover, when comparing those algorithms, it has been empirically demonstrated that the RMIL performed highest accuracy than BP algorithm. This conclusion enforces the usage of the proposed method as an alternative training algorithm of BP algorithm.

6 Conclusions

This research proposed a new approach of conjugate gradient algorithm namely Rivaie, Mustafa, Ismail and Leong with adaptive gain (RMIL/AG). The proposed method improved the training efficiency of the conventional back propagation (BP) neural network algorithms by adaptively modifying the gradient search direction. The gradient search direction is modified by introducing the gain value. The effectiveness of the proposed method has been compared with the conventional BP algorithm and RMIL algorithm. The three algorithms were been verified by means of simulation on four

classification problems including iris, glass, Pima Indian diabetes and breast cancer Wisconsin. The results show that the proposed method (RMIL/AG) has a better convergence rate and learning efficiency as compared to conventional BP method.

Acknowledgements. The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) Ministry of Higher Education (MOHE) Malaysia for financially supporting this Research under Trans-disciplinary Research Grant Scheme (TRGS) vote no. T003. This research also supported by GATES IT Solution Sdn. Bhd under its publication scheme.

References

1. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: David, E.R., James, L.M., C.P.R. Group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, pp. 318–362. MIT Press (1986)
2. Zhang, S.L., Chang, T.C.: A study of image classification of remote sensing based on back-propagation neural network with extended delta bar delta. *Math. Problems Eng.* **2015**, 10 (2016)
3. Nawi, N.M., Rehman, M., Khan, A.: WS-BP: an efficient wolf search based back-propagation algorithm. In: *International Conference on Mathematics, Engineering and Industrial Applications 2014 (ICoMEIA 2014)*. AIP Publishing (2015)
4. Rehman, M.Z., Nawi, N.M.: The effect of adaptive momentum in improving the accuracy of gradient descent back propagation algorithm on classification problems. In: Mohamad Zain, J., Wan Mohd, W.M.B., El-Qawasmeh, E. (eds.) *Second International Conference on Software Engineering and Computer Systems, ICSECS 2011, Kuantan, Pahang, Malaysia, 27–29 June 2011, Proceedings, Part I*, pp. 380–390. Springer, Berlin, Heidelberg (2011)
5. Nawi, N.M., Khan, A., Rehman, M.Z.: A new back-propagation neural network optimized with cuckoo search algorithm. In: *International Conference on Computational Science and Its Applications*. Springer, Berlin, Heidelberg (2013)
6. Liu, Y., Jing, W., Xu, L.: Parallelizing backpropagation neural network using MapReduce and cascading model. *Comput. Intell. Neurosci.* **2016**, 11 (2016)
7. Chen, Y., et al.: Three-dimensional short-term prediction model of dissolved oxygen content based on PSO-BPANN algorithm coupled with Kriging interpolation. *Math. Problems Eng.* **2016**, 10 (2016)
8. Cao, J., Chen, J., Li, H.: An adaboost-backpropagation neural network for automated image sentiment classification. *Sci. World J.* **2014**, 9 (2014)
9. Abdul Hamid, N., et al.: A review on improvement of back propagation algorithm. *Glob. J. Technol.* **1** (2012)
10. Van Ooyen, A., Nienhuis, B.: Improving the convergence of the back-propagation algorithm. *Neural Netw.* **5**(3), 465–471 (1992)
11. Haykin, S.: *Neural Networks: A Comprehensive Foundation*, 842 pp. Prentice Hall PTR (1998)
12. Kumar, P., Merchant, S.N., Desai, U.B.: Improving performance in pulse radar detection using Bayesian regularization for neural network training. *Digit. Signal Proc.* **14**(5), 438–448 (2004)
13. Bishop, C.M.: *Neural Networks for Pattern Recognition*, 482 pp. Oxford University Press, Inc. (1995)

14. Fletcher, R., Powell, M.J.D.: A rapidly convergent descent method for minimization. *Comput. J.* **6**(2), 163–168 (1963)
15. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. *Comput. J.* **7**(2), 149–154 (1964)
16. Hestenes, M., Stiefel, E.: Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bureau Stand.* **49**(6), 409–436 (1952)
17. Maier, H.R., Dandy, G.C.: The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environ. Model Softw.* **13**(2), 193–209 (1998)
18. Thimm, G., Moerland, P., Fiesler, E.: The interchangeability of learning rate and gain in backpropagation neural networks. *Neural Comput.* **8**(2), 451–460 (1996)
19. Nawi, N.M., Ransing, R., Hamid, N.A.: BPGD-AG: a new improvement of back-propagation neural network learning algorithms with adaptive gain. *J. Sci. Technol.* **2**(2) (2010)
20. Nawi, N.M., et al.: An improved back propagation neural network algorithm on classification problems. In: Zhang, Y., et al. (eds.) *Database Theory and Application, Bio-Science and Bio-Technology: International Conferences, DTA and BSBT 2010, Held as Part of the Future Generation Information Technology Conference, FGIT 2010, Jeju Island, Korea, 13–15 December 2010, Proceedings*, pp. 177–188. Springer, Berlin, Heidelberg (2010)
21. Nawi, N.M., et al.: Enhancing back propagation neural network algorithm with adaptive gain on classification problems. *Networks* **4**(2) (2011)
22. Nawi, N.M., et al.: Predicting patients with heart disease by using an improved back-propagation algorithm. *J. Comput.* **3**(2) (2011)
23. Nawi, N.M., Wahid, N., Idris, M.M.: A new gradient based search direction for conjugate gradient algorithm. *Int. J. Adv. Data Inf. Eng.* **1**(1), 1–7 (2016)
24. Abdul Hamid, N., et al.: Learning efficiency improvement of back propagation algorithm by adaptively changing gain parameter together with momentum and learning rate. In: Zain, J. M., Wan Mohd, W.M.B., El-Qawasmeh, E. (eds.) *Second International Conference on Software Engineering and Computer Systems, ICSECS 2011, Kuantan, Pahang, Malaysia, 27–29 June 2011, Proceedings, Part III*, pp. 812–824. Springer, Berlin, Heidelberg (2011)
25. Abdul Hamid, N., Nawi, N.M., Ghazali, R.: The effect of adaptive gain and adaptive momentum in improving training time of gradient descent back propagation algorithm on classification problems. *Int. J. Adv. Sci. Eng. Inf. Technol.* **1**(2), 178–184 (2011)
26. Abdul Hamid, N., et al.: Improvements of back propagation algorithm performance by adaptively changing gain, momentum and learning rate. *Int. J. New Comput. Archit. Appl. (IJNCAA)* **1**(4), 866–878 (2011)
27. Abdul Hamid, N., et al.: Solving local minima problem in back propagation algorithm using adaptive gain, adaptive momentum and adaptive learning rate on classification problems, In: *International Journal of Modern Physics: Conference Series*, pp. 448–455. World Scientific Publishing Company (2012)
28. Mohd Nawi, N., et al.: Predicting patients with heart disease by using an improved back-propagation algorithm. *J. Comput.* **3**(2) (2011)
29. Nawi, N.M., Wahid, N., Idris, M.M.: A new gradient based search direction for conjugate gradient algorithm. *Int. J. Adv. Data Inf. Eng.* **1**(1), 1–7 (2013)
30. Rivaie, M., et al.: A new class of nonlinear conjugate gradient coefficients with global convergence properties. *Appl. Math. Comput.* **218**(22), 11323–11332 (2012)
31. Angiulli, G., et al.: Accurate modelling of lossy SIW resonators using a neural network residual Kriging approach. *IEICE Electron. Express* **14**(6), 20170073 (2017). 20170073

32. Rodríguez-Quiriones, J.C., et al.: Improve a 3D distance measurement accuracy in stereo vision systems using optimization methodsâ€™ approach. *Opto-Electron. Rev.* **25**(1), 24–32 (2017)
33. Prechelt, L.: PROBEN1-A Set of Benchmarks and Benchmarking Rules for Neural Network Training Algorithms. Technical Report 21/94, Fakultät für Informatik, Universität Karlsruhe, /Anonymous FTP: /pub/papers/techreports/1994/1994-21 (1994)
34. Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**(2), 179–188 (1936)
35. Evett, I.W., Spiehler, E.J.: Rule induction in forensic science. In: *Knowledge Based Systems*, pp. 152–160. Halsted Press (1988)
36. Smith, J.W., et al.: Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261–265 (1988)
37. Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.* **43**(4), 570–577 (1995)
38. Nawi, N.M., et al.: Second order back propagation neural network (SOBPNN) algorithm for medical data classification. In: Phon-Amnuaisuk, S., Au, T.W. (eds.) *Computational Intelligence in Information Systems: Proceedings of the Fourth INNS Symposia Series on Computational Intelligence in Information Systems (INNS-CIIS 2014)*, pp. 73–83. Springer International Publishing, Cham (2015)

A Regulative Norms Mining Algorithm for Complex Adaptive System

Moamin A. Mahmoud^{1,2(✉)}, Mohd Sharifuddin Ahmad^{1,2},
Mohd Zaliman M. Yusoff^{1,2}, and Salama A. Mostafa^{2,3}

¹ College of Computer Science and Information Technology, Universiti Tenaga Nasional, Kajang, Malaysia

{moamin, sharif}@uniten.edu.my, zaliman.yusoff@tnb.com.my

² Business Development Unit, TNB Integrated Learning Solution Sdn. Bhd. (ILSAS), Kajang, Malaysia
salama@uthm.edu.my

³ Faculty of Computer Science and Information Technology, UTHM, Parit Raja, Malaysia

Abstract. In complex adaptive system, a visitor agent is not usually and explicitly given the norms of its host agent. Thus, when it is not able to adapt the host agent's norms, it is totally deprived of accessing resources and services from the host. Such circumstance severely affects its performance resulting in failure to achieve its goal. Consequently, this paper attempts to resolve the problem by enabling the agent to identify the host's regulative norms via an algorithm called the Regulative Norms Mining Algorithm (RNMA). Regulative norms constitute the recommendation, obligation, and prohibition norms, which the RNMA identifies by analyzing exceptional events that trigger rewards or penalties. In this paper, we argue that existing norms identifications algorithms are inadequate to detect different regulative norm types. Consequently, we propose the RNMA algorithm, which could alleviate the problem. We demonstrate the merit of the algorithm by apply it on a typical scenario.

Keywords: Social norms · Normative systems · Regulative norms
Norm mining

1 Introduction

The literature on normative multi-agent systems suggests three types of essential norms [1]. The first type is regulative norms, which specify the ideal and varying degrees of sub-ideal behavior of a system by means of recommendation, obligations, prohibitions and permissions. The second type is constitutive norms, which normalize the creation of institutional norms, in addition to the revision of the normative system itself [2]. The third type is the procedural norms, which are instrumental norms [3] addressed to agents acting on roles in the normative system intending to perform the social order, particularly in terms of substantive norms [4].

Regulative norms are intended for regulating activities by imposing recommendation, obligation or prohibition in performing an action [5, 6]. As Peczenik [7] commented, a regulative norm qualifies an action or a state of affairs as prescribed, permitted or prohibited. Because a regulative norm qualifies an action, it can be treated as a norm of conduct, for example, the responsibility to lodge a police report without a reasonable delay upon finding lost or stolen things. A norm of conduct can prescribe punishment or a sanction for a person who violates a norm. One can thus make a distinction between a sanctioned and sanctioning norm.

In complex adaptive system [8], the agent is not conferred with the community's norms automatically or in offline mode. Instead the agent must be able to identify the norms by using some mining algorithm. This paper presents a new algorithm called the Regulative Norms Mining Algorithm (RNMA) to identify the regulative norms, which constitute the recommendation, obligation and prohibition norms [6]. The norms are identified by analyzing exceptional events, which are defined as those events that trigger rewards or penalties. When an agent gets a reward or penalty or detects such occurrence on other agents, it exploits the resources of the host system and uses the RNMA algorithm to implement data formatting, filtering, and extracting the exceptional events (constituting the regulative norms) to identify the norms and subsequently, the normative protocol. Savarimuthu et al. [9] defined the normative protocol as "the order of occurrence of events or protocols that are related to a set of norms". For example, arrive, order, eat, pay, tip and depart in that order is the normative protocol for dining in a restaurant. The idea of this study is motivated by the need to resolve potential norms conflict between two people of different culture [10].

2 Related Work

Several studies have addressed the issues of norms identification [11–15]. Epstein [11] proposed an imitation model that relies on embracing the behavior of a majority and the model is based on the local environment state and the agent's amount of thinking regarding its behavior.

Andrighetto et al. [16] developed norm innovation theory in coping with specific types of complex entities such as a social system. The theory is classified into a two-way dynamics that entails emergent processes from interactions among individual agents and emergent effects involving norms at the aggregate level into the agents' minds. The project, known as EMIL, consists of three main components which are EMIL-M, EMIL-S, and EMIL-T that are designed to collect real scenario data and to simulate findings that deal with the simulation executions. EMIL-M achieves a general model of norm-innovation, a complex social dynamics between intelligent social agents to be verified by tools of agent-based simulation in a particular context. EMIL-S is a simulation platform for experiments and has been built to test EMIL-M by running simulations and compare results with available data, specifically in the area of Open Source. Finally, EMIL-T assesses the achievement of the model by comparing the results of the simulations with the experiential data documented in the open source.

Campos et al. [17] proposed a norm adaptation mechanism based on social power, which providing support to the coordination of agents. They used a generic

level-assisted architecture to develop systems that self-adapt their organization depending on their evolution. They modeled peer-to-peer sharing network as a multi-agent system with two levels of organization. Both levels share the same goal, that all participant agents obtain data by using minimum time. They add an assistance layer to the first level, model the set of computers that share some data as agents within a domain-level. In supporting the coordination of agents, it adapts domain level's organization to changing circumstances. They used Case-based Reasoning (CBR), a learning technique to decide how to adapt domain-level norms depending on current system status. CBR learning is based on a heuristic that aligns the amount of serving/receiving capacity, and this heuristic is used by the CBR to suggest a solution when no similar cases are found.

Savarimuthu et al. [18] developed two algorithms, Obligation Norm Identification (ONI), and Candidate Norm Inference (CNI), to identify obligation norms and prohibition norms, respectively.

3 Development of RNMA

The Regulative Norm Mining Algorithm (RNMA) is an algorithm that mines the Record Base to identify the domain's regulative norms and the subsequent normative protocol.

RNMA consists of five steps. The first step is grouping the data of the Record Base into three classes based on whether the data entails reward or penalty. The classes are rewarded data (Class 1), no rewarded and no penalized data (Class 2), and penalized data (Class 3). Having grouping the data, the second step is filtering the events of Class 2 (no rewarded and no penalized data) by applying some filtering processes which are sequencing, cleaning, ordering, and final setting.

The third step is subset filtering, which is applied on Class 1 and Class 3 by removing all the subset episodes that have the same sequence of event of superset episodes. The fourth step is extracting the recommendation norms, obligation norms, and prohibition norms. The final step is identifying the normative protocol that comprises the optimal and neutral normative protocol. We define six terms before we detail out the RNMA steps.

Definition 1 A Convention, Cnv, is a type of norm that is expected to be used by every agent in the domain.

Definition 2 A Recommendation Norm, RcmN a type of norm that if an agent exercises it, it gets a reward, but if it does not, it is not penalized.

Definition 3 An Obligation Norm, OblgN is a type of norm that if an agent exercises it, it avoids a penalty, but if it does not do so, it is penalized.

Definition 4 A Prohibition Norm, ProhbN is a type of norm that if an agent exercises it, it is penalized, but if it does not do so, it avoids the penalty.

Definition 5 All episodes, A-episodes, refer to the entire episodes of a specific history.

Definition 6 Subset episodes (S-episodes) refer to the episodes which have the same sequence of events of superset episodes in a specific history.

To illustrate the novelty of the RNMA algorithm, we provide an example of the elevator scenario in identifying and recognizing the obligation, prohibition, and recommendation norms. In this scenario, we assume a set of norms that are commonly practiced in a typical elevator domain, which are: (*arrive*; *wait*; *enter*; *greet*; *litter*; *excuse*; *depart*).

We also assume that some host agents, $A_\lambda = \{\alpha_{\lambda 1}, \alpha_{\lambda 2}, \dots, \alpha_{\lambda n}\}$ are penalized, Π , by the authorized agent, α_σ , because of failing to excuse or littering, while other agents are rewarded, Ω , for performing a commendable action such as greeting.

In this scenario, a visitor agent, α_v , observes exceptional events and starts detecting the regulative norms. We assume the agent moved from its local domain, D_X , to the new domain, D_Y . In the new domain, there is an elevator and the agent has knowledge of the normative protocol from the previous domain which is *arrive*, *wait*, *enter*, *depart*. The agent observes and discovers exceptional events in the new domain and consequently collects the following episodes and records them in its Record Base. The episodes are collected from the host agents $\alpha_{\lambda 1}$ to $\alpha_{\lambda 13}$. We represent these episodes as follows:

$(\alpha_{\lambda 1})$	= (enter, excuse, depart)
$(\alpha_{\lambda 2})$	= (arrive, wait, enter, excuse)
$(\alpha_{\lambda 3}, \alpha_\sigma)$	= (wait, enter, litter, excuse, Π)
$(\alpha_{\lambda 4})$	= (arrive, wait, enter, excuse, depart)
$(\alpha_{\lambda 5}, \alpha_\sigma)$	= (arrive, wait, enter, depart, Π)
$(\alpha_{\lambda 6}, \alpha_\sigma)$	= (enter, depart, Π)
$(\alpha_{\lambda 7})$	= (wait, enter, excuse)
$(\alpha_{\lambda 8}, \alpha_\sigma)$	= (arrive, wait, enter, litter, excuse, depart, Π)
$(\alpha_{\lambda 9}, \alpha_\sigma)$	= (wait, enter, litter, Π)
$(\alpha_{\lambda 10}, \alpha_\sigma)$	= (arrive, wait, enter, greet, excuse, depart, Ω)
$(\alpha_{\lambda 11})$	= (arrive, wait, enter)
$(\alpha_{\lambda 12}, \alpha_\sigma)$	= (wait, enter, depart, Π)
$(\alpha_{\lambda 13}, \alpha_\sigma)$	= (greet, excuse, depart, Ω)

The RNMA steps are as follow:

Step 1: Grouping the data, GrDt, is the process of grouping the Record Base data (RB) into three classes. Class 1, C_1 , involves the rewarded events. Class 2, C_2 , groups the neutral events (no rewards or penalties). Class 3, C_3 , involves the penalized events. If V_Ω are rewarded events; V_Π are penalized events; and V_N are neutral events, then,

$$\begin{aligned}
 RB &= \{V_\Omega, V_N, V_\Pi\} \\
 C_1 &= RB / \{V_N, V_\Pi\} \\
 C_2 &= RB / \{V_\Omega, V_\Pi\} \\
 C_3 &= RB / \{V_N, V_\Pi\}
 \end{aligned}$$

$$\text{GrDt} = C_1(V_\Omega), C_2(V_N), C_3(V_\Pi) \quad (1)$$

The aim of Step 1 is to group similar data to facilitate the mining process in the next steps. Applying Eq. 1 on the episodes of the elevator example:

Classes	Episodes
$C_1(V_\Omega)$ (Class 1)	<ul style="list-style-type: none"> • (arrive, wait, enter, greet, excuse, depart) • (greet, excuse, depart)
$C_2(V_N)$ (Class 2)	<ul style="list-style-type: none"> • (enter, excuse, depart) • (arrive, wait, enter, excuse) • (arrive, wait, enter, excuse, depart) • (wait, enter, excuse) • (arrive, wait, enter)
$C_3(V_\Pi)$ (Class 3)	<ul style="list-style-type: none"> • (wait, enter, litter, excuse) • (arrive, wait, enter, depart) • (enter, depart) • (arrive, wait, enter, litter, excuse, depart) • (wait, enter, litter) • (wait, enter, depart)

Step 2: Events filtering (E-episode) is applied on the C_2 (Class 2), which is the neutral class by applying the following procedure:

$$E - \text{episode} : Q(C_2) \rightarrow L(C_2) \rightarrow D(C_2) \rightarrow C_{F2} \quad (2)$$

- Sequencing, Q: Numbers each episode's event as a sequence from 1 to k (1, 2, 3, 4, ..., k). For example, the episode (a, b, c, d) becomes ($a_{[1]}$, $b_{[2]}$, $c_{[3]}$, $d_{[4]}$).

Applying Sequencing $Q(C_2)$ on Class 2 of the elevator scenario:

$$\begin{aligned}
 Q(C_2) = & (\text{enter}_{[1]}, \text{excuse}_{[2]}, \text{depart}_{[3]}) \\
 & (\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{excuse}_{[4]}) \\
 & (\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{excuse}_{[4]}, \text{depart}_{[5]}) \\
 & (\text{wait}_{[1]}, \text{enter}_{[2]}, \text{excuse}_{[3]}) \\
 & (\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]})
 \end{aligned}$$

- Cleaning, L: Select the event, which has the highest sequence number from the repeated events and removing the rest. For example, given these two episodes: ($a_{[1]}$, $b_{[2]}$, $c_{[3]}$, $d_{[4]}$), and ($b_{[1]}$, $d_{[2]}$) we remove $b_{[1]}$ and $d_{[2]}$ retain $b_{[2]}$ and $d_{[4]}$, since they hold the highest sequence number for repeated events.

Applying Filtering $L(C_2)$ on $Q(C_2)$ of the elevator scenario:

$$L(C_2) = (\text{enter}_{[3]}, \text{excuse}_{[4]}, \text{depart}_{[5]}, \text{arrive}_{[1]}, \text{wait}_{[2]})$$

- Ordering, D: After getting the highest number of sequence for non-repeated set, the set is ordered from 1 to n (counting up). For example, the episode $(b_{[2]}, a_{[1]}, d_{[4]}, c_{[3]})$ becomes $(a_{[1]}, b_{[2]}, c_{[3]}, d_{[4]})$ upon ordering the sequence.

Applying Ordering $D(C_2)$ on $L(C_2)$ of the elevator scenario:

$$D(C_2) = (\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{excuse}_{[4]}, \text{depart}_{[5]})$$

- Final setting, C_{F2} : Get the final set of Class 2 ordering to use it in the next steps. For example, the final set of ordering episode $(a_{[1]}, b_{[2]}, c_{[3]}, d_{[4]})$ is (a, b, c, d) ,

Getting C_{F2} from $D(C_2)$ of the elevator scenario:

$$C_{F2} = (\text{arrive}, \text{wait}, \text{enter}, \text{excuse}, \text{depart})$$

This step defines the filtered and ordered set of events that do not trigger any reward or penalty. C_{F2} represents the identity class that can be exploited to extract all the types of regulative norms and identify the normative protocol of the domain.

Step 3. Subset filtering (F-episodes) is applied on A-episodes of C_1 and C_3 by removing the S-episodes if these subset episodes have the same events sequence of bigger or superset episodes.

$$F - \text{episodes } (C_1) = A - \text{episodes } (C_1) \setminus S - \text{episodes } (C_1) \Rightarrow C_{F1} \quad (3)$$

$$F - \text{episodes } (C_3) = A - \text{episodes } (C_3) \setminus S - \text{episodes } (C_3) \Rightarrow C_{F3} \quad (4)$$

The subset filtering removes all the subsets from C_1 and C_3 to get the final set of Class 1 (C_{F1}) and Class 3 (C_{F3}).

This step filters the data of Class 1 and Class 3 by removing the unnecessary episodes to determine the candidate episodes and to reduce the cost of extracting the regulative norms because the candidate episodes are used in the extraction process which is costly and this cost increases as the volume of data increases. Moreover, based on the Venn diagram, the removed episodes do not have any effect on the results because they are represented by the bigger episodes.

Getting C_{F1} from Class 1 in the elevator scenario (Eq. 3),

A-episodes (C_1) (arrive, wait, enter, greet, excuse, depart) (greet, excuse, depart)	\setminus	S-episodes (C_1) (greet, excuse, depart)	$=$	C_{F1} (arrive, wait, enter, greet, excuse, depart)
--	-------------	--	-----	---

We obtain, $C_{F1} = (\text{arrive}, \text{wait}, \text{enter}, \text{greet}, \text{excuse}, \text{depart})$

Getting C_{F3} from Class 3 in the elevator scenario (Eq. 4):

A-episodes (C_3)	\setminus	S-episodes (C_3)	$=$	C_{F3}
(wait, enter, litter, excuse)		(wait, enter, litter,		(arrive, wait, enter, depart)
(arrive, wait, enter, depart)		excuse)		(arrive, wait, enter, litter,
(enter, depart)		(enter, depart)		excuse, depart)
(arrive, wait, enter, litter,		(wait, enter, litter)		
excuse, depart)		(wait, enter,		
(wait, enter, litter)		depart)		
(wait, enter, depart)				

We obtain, $C_{F3} = (\text{arrive, wait, enter, depart}); (\text{arrive, wait, enter, litter, excuse, depart})$

Step 4. Extracting the regulative norms.

Based on Definition 2, Recommendation norms [6] can be extracted from C_{F1} (the storage of rewarded events). A Recommendation type norm must appear in C_{F1} only, i.e.

$$\text{norm}_k \in \text{RcmN} \Leftrightarrow \text{norm}_k \in (C_{F1}) \wedge \text{norm}_k \notin (C_{F2}, C_{F3}) \quad (5)$$

Based on Definition 3, Obligation norms [15, 19] can be extracted from C_{F2} (the storage of neutral events, i.e., no rewarded and or penalized events). An Obligation type norm must appear in C_{F2} only, i.e.,

$$\text{norm}_k \in \text{OblgN} \Leftrightarrow \text{norm}_k \in (C_{F2}) \wedge \text{norm}_k \notin (C_{F1}, C_{F3}) \quad (6)$$

Based on Definition 4, Prohibition norms [15] can be extracted from C_{F3} (the storage of penalized events). A Prohibition type norm must appear in C_{F3} only, i.e.,

$$\text{norm}_k \in \text{ProhbN} \Leftrightarrow \text{norm}_k \in (C_{F3}) \wedge \text{norm}_k \notin (C_{F2}, C_{F1}) \quad (7)$$

The extraction procedure is as follow,

Recommendation Norm: An agent detects a Recommendation norm by finding the set-theoretic complement between the episodes of C_{F1} and the segments from C_{F2} class that have the same first and last event of C_{F1} , ($V_\Omega C_{F2}$).

Extracting the Recommendation Norm,

$$\begin{aligned} &\text{If } ((C_{F1} \setminus V_\Omega C_{F2}) = \emptyset) \Rightarrow (\text{RcmN} = \emptyset), \\ &\text{else } ((C_{F1} \setminus V_\Omega C_{F2}) = \text{Norm}) \Rightarrow (\text{Norm} \in \text{RcmN}) \end{aligned} \quad (8)$$

which means that if C_{F1} set without $V_\Omega C_{F2}$ set equals \emptyset , then no reward events appear in the C_{F1} , otherwise, the results represent the Recommendation norm.

Identifying the Recommendation norms in the elevator scenario:

$C_{F1} = (\text{arrive, wait, enter, greet, excuse, depart})$

$C_{F2} = (\text{arrive, wait, enter, excuse, depart})$

$V_\Omega C_{F2} = (\text{arrive, wait, enter, excuse, depart})$

Therefore, from Eq. 8,

$$\begin{aligned} &(\text{arrive, wait, enter, greet, excuse, depart} \setminus \text{arrive, wait, enter, excuse, depart}) \\ &= \text{greet}) \Rightarrow (\text{greet} \in \text{RcmN}) \end{aligned}$$

- **Obligation Norm:** An agent detects an Obligation norm by finding the set-theoretic complement between the episodes $V_{\Omega} C_{F2}$ and C_{F3} that have the same first and last event of $V_{\Omega} C_{F2}$.

Extracting the obligation norm,

$$\begin{aligned} &\text{If } ((V_{\Omega} C_{F2} \setminus C_{F3}) = \emptyset) \Rightarrow (\text{ObligN} = \emptyset), \\ &\text{else } ((V_{\Omega} C_{F2} \setminus C_{F3}) = \text{Norm}) \Rightarrow (\text{Norm} \in \text{ObligN}) \end{aligned} \quad (9)$$

which means that if $V_{\Omega} C_{F2}$ set without C_{F3} set equal \emptyset , then no obligation norms appear in the $V_{\Omega} C_{F2}$, otherwise, the results represent the obligation norms.

Identifying the Obligation norms in the elevator example:

$C_{F2} =$	(arrive, wait, enter, excuse, depart)
$C_{F3} =$	(arrive, wait, enter, depart)
	(arrive, wait, enter, litter, excuse, depart)

C_{F3} has two episodes

- Episode1 $C_{F3} = (\text{arrive, wait, enter, depart})$
 $C_{F2} = (\text{arrive, wait, enter, excuse, depart})$
 $V_{\Omega} C_{F2} = (\text{arrive, wait, enter, excuse, depart})$

Therefore, from Eq. 9,

$$\begin{aligned} &(((\text{arrive, wait, enter, excuse, depart}) \setminus \text{arrive, wait, enter, depart}) \\ &= \text{excuse})) \Rightarrow (\text{excuse} \in \Pi) \end{aligned}$$

- Episode2 $C_{F3} = (\text{arrive, wait, enter, litter, excuse, depart})$
 $C_{F2} = (\text{arrive, wait, enter, excuse, depart})$
 $V_{\Omega} C_{F2} = (\text{arrive, wait, enter, excuse, depart})$

Therefore, from Eq. 9,

$$\begin{aligned} &((\text{arrive, wait, enter, excuse, depart} \setminus \text{arrive, wait, enter, litter, excuse, depart}) \\ &= \emptyset)) \Rightarrow (\text{ObligN} = \emptyset) \end{aligned}$$

- **Prohibition Norm:** An agent detects a prohibition norm by finding the set-theoretic complement between each episode in C_{F3} and episode of $V_{\Omega} C_{F2}$.

Extracting the prohibition norm,

$$\begin{aligned} &\text{If } ((C_{F3} \setminus V_{\Omega} C_{F2}) = \emptyset) \Rightarrow (\text{ProhbN} = \emptyset), \\ &\text{else } ((C_{F3} \setminus V_{\Omega} C_{F2}) = \text{Norm}) \Rightarrow (\text{Norm} \in \text{ProhbN}). \end{aligned} \quad (10)$$

which means that if C_{F3} set without $V_{\Omega} C_{F2}$ set equal \emptyset , then no prohibition norm appears in the C_{F3} , otherwise, the results represent the prohibition norm of penalized events.

Identifying the Prohibition norms in the elevator example:

$C_{F2} =$	(arrive, wait, enter, excuse, depart)
$C_{F3} =$	(arrive, wait, enter, depart)
	(arrive, wait, enter, litter, excuse, depart)

C_{F3} has two episodes

- Episode1 $C_{F3} = (\text{arrive, wait, enter, depart})$
 $C_{F2} = (\text{arrive, wait, enter, excuse, depart})$
 $V_{\Omega} C_{F2} = (\text{arrive, wait, enter, excuse, depart})$

Therefore, from Eq. 10,

$$\begin{aligned} &((\text{arrive, wait, enter, depart} \setminus \text{arrive, wait, enter, excuse, depart}) \\ &= \emptyset) \Rightarrow (\text{ProhbN} = \emptyset) \end{aligned}$$

- Episode2 $C_{F3} = (\text{arrive, wait, enter, litter, excuse, depart})$
 $C_{F2} = (\text{arrive, wait, enter, excuse, depart})$
 $V_{\Omega} C_{F2} = (\text{arrive, wait, enter, excuse, depart})$

Therefore, from Eq. 10,

$$\begin{aligned} &((\text{arrive, wait, enter, litter, excuse, depart} \setminus \text{arrive, wait, enter, excuse, depart}) \\ &= \text{litter}) \Rightarrow (\text{litter} \in \text{ProhbN}) \end{aligned}$$

The RNDT detects *greet* as a recommendation norm; *excuse* as an obligation norms and *litter* as a prohibition norm. Given the identified regulative norms, the agent then infers the normative protocol for the scenario.

Step 5. Identify the normative protocol, NorProt. In this last step, the normative protocols are identified. As a consequence of the RNMA algorithm, two types of normative protocols are identified:

- Neutral Normative Protocol, NuNorProt, is a normative protocol that does not contain recommendation and prohibition norms. The agent identifies the NuNorProt from C_{F2} (neutral episode) because C_{F2} contains the set of norms that if exercised by the agent, is not rewarded or penalized.

$$\text{NuNorProt} = C_{F2} \quad (11)$$

Identifying the Neutral Normative Protocol in the elevator scenario:
From Eq. 11,

$$\text{NuNorProt} = C_{F2} = (\text{arrive}, \text{wait}, \text{enter}, \text{excuse}, \text{depart})$$

- Optimal Normative Protocol, OptNorProt, is a normative protocol that contains recommendation norms but without prohibition norms. The agent can identify the NuNorProt by filtering events, which is mentioned in Step 2. Following the complement of C_{F1} and C_{F2} events filtering, sequencing, filtering, and ordering are performed. Then the final set represents the set of norms that are rewarded.

$$Q(C_{F1}, C_{F2}) \rightarrow L(C_{F1}, C_{F2}) \rightarrow D(C_{F1}, C_{F2}) \rightarrow \text{OptNorProt} \quad (12)$$

Identifying the Optimal Normative Protocol in the elevator scenario:
From Eq. 12,

$$C_{F1} = (\text{arrive}, \text{wait}, \text{enter}, \text{greet}, \text{excuse}, \text{depart})$$

$$C_{F2} = (\text{arrive}, \text{wait}, \text{enter}, \text{excuse}, \text{depart})$$

- Sequencing, Q
 $(\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{greet}_{[4]}, \text{excuse}_{[5]}, \text{depart}_{[6]})$
 $(\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{excuse}_{[4]}, \text{depart}_{[5]})$
- Filtering, L
 $(\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{greet}_{[4]}, \text{excuse}_{[5]}, \text{depart}_{[6]})$
- Ordering, D
 $(\text{arrive}_{[1]}, \text{wait}_{[2]}, \text{enter}_{[3]}, \text{greet}_{[4]}, \text{excuse}_{[5]}, \text{depart}_{[6]})$
- Final setting
 $\text{OptNorProt} = (\text{arrive}, \text{wait}, \text{enter}, \text{greet}, \text{excuse}, \text{depart})$

The advantages of this algorithm are:

- It does not rely on the use of association rule mining, which detects high frequency candidate norms but omit the low frequency norms, i.e., the RNMA algorithm identifies norms that are independent of the frequency count.
- It identifies any number of regulative norms in the domain. In the elevator scenario, if we assume that there are two or more regulative norms (litter, excuse, and so on) by one or two different agents, for example agent $\alpha_{\lambda 1}$ is penalized because it did not excuse itself and agent $\alpha_{\lambda 2}$ is penalized because it littered the elevator, both regulative norms can be detected by the algorithm.
- The RNMA algorithm detects all regulative norm types, which are obligation, prohibition and recommendation with each norm under its corresponding type. For example in the elevator scenario, the RNMA algorithm detects *litter* as under prohibition norms, *excuse* as under obligation norms and *greet* as under recommendation norms

- The RNMA algorithm exploits both reward and penalty actions to identify the regulative norms which gives more flexibility and accuracy in assigning the type of each detected norm.

4 Conclusion and Future Work

This paper presents the regulative norm mining algorithm as a new tool to identify the regulative norms from a domain. It could provide a suitable solution for software agents and robot community to adapt to the various environments and learn new behaviors which lead to improve capabilities.

Regulative norm mining algorithm entails the processes of data formatting, filtering and extracting the different types of norms and normative protocols. The preliminary results show that the RNMT succeeded in identifying the regulative norms in general.

For our future work, we shall study the issue of norm's awareness. In this work and other work in the literature, agents are not aware of the context of the enacted norms. Consequently, it would be interesting to look into semantic agents that can deal with ontology-based contexts. When agents could understand the meaning of the norms, they would have greater reasoning ability about the norms' effects on their performance.

References

1. Caire, P.: A normative multi-agent systems approach to the use of conviviality for digital cities. In: *Proceedings of the International Conference on Coordination, Organizations, Institutions, and Norms in Agent Systems III, COIN'07*, pp. 245–260 (2007)
2. Boella, G., Torre, L.V.D.: Regulative and constitutive norms in normative multi-agent systems. In: *Proceedings of the 9th International Conference on the Principles of Knowledge Representation and Reasoning, Whistler (CA)*, pp. 255–26 (2004)
3. Boella, G., Torre, L.V.D.: Substantive and procedural norms in normative multiagent systems. *J. Appl. Log.* **6**(2), 152–171 (2008)
4. López, F., Luck, M., d'Inverno, M.: A normative framework for agent-based systems. *Comput. Math. Organ. Theory* **12**(2), 227–250 (2006)
5. Rubino, R., Omicini, A., Denti, E.: Computational institutions for modelling norm-regulated MAS: an approach based on coordination artifacts. In: *Proceedings of AAMAS Workshops*, pp. 127–141 (2005)
6. Mahmoud, M.A., Ahmad, M.S., Mohd Yusoff, M.Z., Mustapha, A.: A review of norms and normative multiagent systems. *Sci. World J.* **2014** (2014)
7. Peczenik, A.: *On Law and Reason*. Springer (2009)
8. Chan, S.: Complex adaptive systems. In: *ESD, 83 Research Seminar in Engineering Systems*, vol. 31 (2001)
9. Savarimuthu, B.T.R., Cranefield, S., Purvis, M., Purvis, M.: Obligation norm identification in agent societies. *J. Artif. Soc. Soc. Simul.* **13**(4) (2010)
10. Mahmoud, M.A., Ahmad, M.S., Ahmad, A., Yusoff, M.Z.M., Mustapha, A.: Norms detection and assimilation in multi-agent systems: a conceptual approach. In: *Knowledge Technology*, pp. 226–233. Springer, Berlin, Heidelberg (2012)

11. Epstein, J.: Learning to be thoughtless: social norms and individual computation. *Comput. Econ.* **18**(1), 9–24 (2001)
12. López, López: Social Powers and Norms: Impact on Agent Behaviour. Ph.D. thesis, Department of Electronics and Computer Science, University of Southampton, United Kingdom (2003)
13. Andrighetto, G., Campenni, M., Cecconi, F., Conte, R.: The complex loop of norm emergence: a simulation model. In: Chen, S.-H., Cioffi-Revilla, C., Gilbert, N., Kita, H., Terano, T., Takadama, K., Cioffi-Revilla, C., Deffuant, G. (eds.) *Simulating Interacting Agents and Social Phenomena*, volume 7 of *Agent-Based Social Systems*, pp. 19–35. Springer (2010)
14. Mahmoud, M.A., Ahmad, M.S., Zaliman, M.Y.M.: Development and implementation of a technique for norms-adaptable agents in open multi-agent communities. *J. Syst. Sci. Complex.* **29**(6), 1519–1537 (2016)
15. Mahmoud, M.A., Ahmad, M.S., Yusoff, M.Z.M.: A norm assimilation approach for multi-agent systems in heterogeneous communities. In: *Asian Conference on Intelligent Information and Database Systems*, pp. 354–363. Springer, Berlin, Heidelberg (2016)
16. Andrighetto, G., Conte, R., Turrini, P., Paolucci, M.: Emergence in the loop: simulating the two way dynamics of norm innovation. In: *Normative Multi-agent Systems*, in *Dagstuhl Seminar Proceedings. Internationales Begegnungs-und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany* (2007)
17. Campos, J., López-Sánchez, M., Esteva, M.: A case-based reasoning approach for norm adaptation. In: *5th International Conference on Hybrid Artificial Intelligence Systems (HAIS'10)*, pp. 168–176. Springer, Spain (2010)
18. Savarimuthu, B.T.R., Craneffeld, S., Purvis, M., Purvis, M.: Norm Identification in Multi-agent Societies. Discussion Paper, Department of Information Science, University of Otago (2010)

Violence Video Classification Performance Using Deep Neural Networks

Ashikin Ali and Norhalina Senan^(✉)

Faculty of Computer Science and Information Technology, Universiti Tun
Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Batu Pahat, Johor, Malaysia
gil50038@siswa.uthm.edu.my, halina@uthm.edu.my

Abstract. Violence is autonomous, the contents that one would not let children to see in movies or web videos. This is a challenging problem due to strong content variations among the positive instances. To solve this problem, implementation of deep neural network to classify the violence content in videos is proposed. Currently, deep neural network has shown its efficiency in natural language processing, fraud detection, social media, text classification, image classification. Regardless of the conventional methods applied to overcome this issue, but these techniques seem insufficiently accurate and does not adopt well to certain webs or user needs. Therefore, the purpose of this study is to assess the classification performances on violence video using Deep Neural Network (DNN). Hence, in this paper different architectures of hidden layers and hidden nodes in DNN have been implemented using the try-error method and equation based method, to examine the effect of the number of hidden layers and hidden nodes to the classification performance. From the results, it indicates 53% accuracy rate for try and error approach, meanwhile for equation based approach it indicates 51% accuracy rate.

Keywords: Violence video · Artificial neural network · Deep neural network
Classification

1 Introduction

In recent years, multimedia content is growing vast. Almost every user will be able to provide content, which is accessible by huge portions of population with limited central control. The violence characterization is subjective and this creates difficulty in defining violent content. Violence may outline any situation or action that may cause harm mentally or physically to users. Violence scenes in video documents regard the content that includes such actions. These scenes normally displayed depending on characteristic audio signals as well, for instance like screams or gunshots. Provided many web-filtering systems are commercially available and possibilities for users to download from the internet is high. However, these techniques seem insufficiently accurate for classification tasks to meet the user needs. To solve this drawback, artificial neural network has been proposed to classify the violence content more efficiently. At present, Deep neural network (DNN) is widely in use for classification tasks. Despite the fact, DNN still face issues regarding its complex architecture, especially on multiple

hidden layers and hidden nodes. DNN are also computationally intensive. Also requires high amount of data to avoid overfitting and to perform reliably [1]. In the literature, these focus is very limited. Researchers focuses on audio features [2, 3], audio-visual features [4–6], inter feature and inter-classes [7], multimodal features [7] and exploiting feature and class relationships [8]. Thus, in this paper, the features as in [4–6] have been implemented and all the attributes of audio-visual features have been used to fed into the DNN classifier. To examine the accuracy performance, the try-error method and equation based method was implemented. The experiments were conducted using the generalized set of VSD2014 consist of 86 YouTube videos consist of violence and non-violence video, segregated uniformly.

2 Deep Neural Network

The backpropagation algorithm [9] is used in layered feed-forward MLP. The backpropagation algorithm uses supervised learning, where the algorithm is provided with the inputs and outputs which the network should compute and then the error is calculated [10]. The idea of the backpropagation algorithm is to reduce this error, until the MLP learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. The weighted sum of a neuron is written as:

$$A_j(x, w) = \sum_{i=0}^n X_i W_{ji}, \quad (2.1)$$

where the sum of input X_i is multiplied by their respective weights, W_{ji} . The activation depends only on the inputs and the weights. If the output function would be the identity, then the neuron would be called linear. The most used output function is sigmoid function:

$$O_j(x, w) = \frac{1}{1 + e^{-A(x, w)}} \quad (2.2)$$

The sigmoid function is very close to one for large positive numbers and very close to zero for large negative numbers. This allows a smooth transition between the low and high output of the neuron. The output depends only in the activation, which in turn depends on the values of the inputs and their respective weights. The goal of the training process is to obtain a desired output when certain inputs are given. Since the error is the difference between the actual and desired output, the error depends on the weights and preferred to be adjusted to minimize the error. The error function for the output of each neuron can be defined as:

$$E_j(x, w, d) = (O_j(x, w) - d_j)^2 \quad (2.3)$$

The output will be positive and the desired target will be greater if the difference is big and lesser if the difference is small. The error of the network will simply be the sum of the errors of all the neurons in the output layer:

$$E(x, w, d) = \sum_j (O_j(x, w) - d_j)^2 \quad (2.4)$$

where O_j is the target output and d_j is the target output of the experiment. After finding this, the weights can be adjusted using the method of gradient descent:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}} \quad (2.5)$$

This equation can be inferred in the following way: the adjustment of each weight (Δw_{ji}) will be the negative of a constant eta (η), where η is the learning rate. Multiplied by the dependence of the previous weight on the error of the network, which is derivative of E in respect to w_{ji} . The size of the adjustment will depend on η , and on the contribution of the weight to the error of the function. This is, if the weight contributes a lot to the error, the adjustment will be greater than if it contributes in a smaller amount. Equation (2.5) is used until appropriate weights with minimal error founded.

Henceforth, derivative of E in respect to w_{ji} discovered. This is the goal of the backpropagation algorithm, since the backwards need to be achieved. First, calculate the error depends on the output, which is the derivate of E in respect to O_j from Eq. (2.3).

$$\frac{\partial E}{\partial O_j} = 2(O_j - d_j) \quad (2.6)$$

The reliance of the output on the activation depends on the weights from Eqs. (2.1) and (2.2). Can be seen that from Eqs. (2.6) and (2.7):

$$\frac{\partial O_j}{\partial w_{ji}} = \frac{\partial O_j}{\partial A_j} \frac{\partial A_j}{\partial w_{ji}} = O_j(1 - O_j)x_i \quad (2.7)$$

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial O_j} \frac{\partial O_j}{\partial w_{ji}} = 2(O_j - d_j)O_j(1 - O_j)x_i \quad (2.8)$$

The adjustment to each weight will begin from Eqs. (2.5) and (2.8):

$$\Delta w_{ji} = -2\eta(O_j - d_j)O_j(1 - O_j)x_i \quad (2.9)$$

Equation (2.9) can be used as it is for training ANN with two layers. For training the network with one more layer, some considerations are needed particularly on training time which can be affected by the architecture of the network. For practical reasons, ANNs implementing the backpropagation algorithm do not have too many layers, since the time for training the networks grows exponentially. These processes clearly explained in the following algorithm of DNN:

Algorithm 1 General training

Input: Training-set, model

Output: model trained

```

1: for i = 1 to (numEpoch) do
2:    $\Delta^{(i)} \leftarrow 0$ 
3:   for m = 1 to (sizeDataset) do
4:     do forward Propagation
5:     get loss
6:      $\Delta_{\text{backProp}} \leftarrow$  do backward Propagation
7:      $\Delta^{(i)} \leftarrow \Delta^{(i)} + \Delta_{\text{backProp}}$ 
8:   end for
9:    $D^{(i)} \leftarrow \frac{1}{\text{sizeDataset}} (\Delta^{(i)} + \lambda \Theta^{(i)})$ ,  $D^{(i)} \leftarrow \frac{1}{\text{sizeDataset}} \Delta^{(i)}$ 
10: end for
  
```

Algorithm 2 Forward Forward

Input: Input X

Output: $h_{\theta}(x)$, A , Z for each layer

```

1: for l = 1 to (numLayers) do
2:    $Z^{(l+1)} \leftarrow \Theta^{(l)} . A^{(l)}$ 
3:    $A^{(l+1)} \leftarrow g(Z^{(l+1)})$ 
4: end for
  
```

Algorithm 3 Backward Forward

Input: Loss, Model activations and pre-activations

Output: Δ_{backprop}

```

1:  $\delta^{(\text{numLayers})} = y - A^{(\text{numLayers})}$ 
2: for l=(numLayers-1) downto 1 do
3:    $\delta^{(l)} \leftarrow ((\Theta^{(l)})^T \delta^{(l+1)}) . * g'(Z^{(l)})$ 
4: end for
5: for l=(numLayers-1) downto 1 do
6:    $\Delta_{\text{backprop}}^{(l)} \leftarrow \delta^{(l+1)} (a^{(l)})^T$ 
7: end for
  
```

Fig. 1. Algorithm of deep neural network

2.1 Architecture of Hidden Layers and Hidden Nodes

Despite input and output layer, hidden layers are the essential part in the architecture of artificial neural network. Therefore, focus of this paper is to adjust the number of

hidden layers and hidden nodes using two approaches, namely, try-error method and equation method. From the algorithm above, hidden layer adjustments were conducted on Algorithm 2 based on Fig. 1.

Try-error method results was obtained by running simulations until the network learns the optimum architecture. Meanwhile, equation method was calculated based on $N_s = (N_i + N_o) * N_h \sqrt{N_{hiddenlayers}}$, where N_s is number of samples in training data set, N_i is number of input neurons, N_o is number of output neurons and N_h is number of hidden nodes.

3 Experimental Design

The experimental design in this paper consists of feature extraction process and the classification of deep neural network. The program is developed in Eclipse using Java programming language. The simulations have been conducted in GPU type device Macbook Pro, the processor is 2.5 GHz Intel Core i5 with Intel HD Graphics 4000.

3.1 Dataset—VSD2014

Benchmark VSD2014 dataset obtained from Technicolor Group [11, 12] has been used as a domain in this study. 86 videos have been used with 434 attributes and 68, 830 instances. This data consists of two classes, namely, violence represented as 1 and non-violence represented as 0.

3.2 Feature Extraction

This section explains the audio and visual features that has been used in the experiment. Basically, total of 29 audio features and 405 visual features.

3.2.1 Audio Features

For the audio features, consist of 8 features. Each feature is provided by per-video-frame-basis. The 8 features are accordingly, amplitude envelope (AE), root-mean-square energy (RMS), zero-crossing rate (ZCR), band energy ratio (BER), spectral centroid (SC), frequency bandwidth (BW), spectral flux (SF) and mel-frequency cepstral coefficients (MFCC). For each window, 22 MFCC is computed while all other features are computed in 1-dimensional, by total of 29 features. The equations are as shown in Table 1.

Table 1. Audio features

Feature names	Feature equations
Amplitude envelope (AE)	$F(x, t) = \sin\left[2\pi\left(\frac{x}{\lambda - \Delta\lambda} - (f + \Delta f)t\right)\right] + \sin\left[2\pi\left(\frac{x}{\lambda + \Delta\lambda} - (f + \Delta f)t\right)\right]$
Root-mean-square energy (RMS)	$RMS = \sqrt{\frac{a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2}{n}}$
Zero-crossing rate (ZCR)	$ZCR = \frac{1}{2N} \sum_{n=1}^N sign(x[n]) - sign(x[n-1]) $
Band energy ratio (BER)	$B(t) = \frac{\sum_{i=0}^{J-1} \sum_{m=0}^{M-1} s_i^2(Mt+m) \cdot h(M-1-m)}{\sum_{i=J}^{N-1} \sum_{m=0}^{M-1} s_i^2(Mt+m) \cdot h(M-1-m)}$
Spectral centroid (SC)	$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$
Frequency bandwidth (BW)	$BW = f_2 - f_1$
Spectral flux (SF)	$F_r = \sum_{k=1}^{N/2} (X_r[k] - X_{r-1}[k])^2$
Mel-frequency cepstral coefficients (MFCC)	$F_{mel} = \frac{1000}{\log(2)} \cdot \left[1 + \frac{F_{Hz}}{1000}\right]$

3.2.2 Visual Features

As for the visual features, it includes 4 features, namely color naming histogram (CNH) with 99-dimensional, local binary patterns (LBP) with 144-dimensional, both color moments (CM) and histogram of oriented gradients (HOG) with 81-dimensional, by total of 405 features. The equations are as shown in Table 2.

Table 2. Visual features

Feature names	Feature equations
Color naming histogram (CNH)	$H = \cos^{-1} \frac{\frac{1}{3}[(R-G) + (R-B)]}{\sqrt{(R-G)^2 + (R-B)(G-B)}}$ $S = 1 - \frac{3[(\min(R, G, B))]}{R + G + B}$ $V = \left(\frac{R + G + B}{3}\right)$
Color moments (CM)	$E_i = \sum_{j=1}^{j=1} \frac{1}{N} P_{ij}$
Local binary patterns (LBP)	$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p$
Histogram of oriented gradients (HOG)	$L2-norm: f = \frac{v}{\sqrt{ v _2^2 + e^2}}$ $L1-norm: f = \frac{v}{(v _1 + e)}$ $L1-sqrt: f = \sqrt{\frac{v}{(v _1 + e)}}$

4 Results

Thus, to quantify the performance of classification method, the performance metrics: classifier accuracy, precision, and recall will be used to access the performance of the classifier. There are four terms that will be used for this study, namely TP, TN, FP and FN. Evaluation measures of this study is as following as shown in Table 3.

Table 3. Evaluation measures

Measure	Formula
Accuracy (A)	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall (R)	$\frac{TP}{TP+FN}$
Precision (P)	$\frac{TP}{TP+FP}$

Table 4 shows the classification accuracy of different hidden layers and fixed 434 hidden nodes.

Table 4. Classification accuracy for 434 hidden nodes, try and error approach

DNN architecture	Accuracy (%)
Hidden layer 3, hidden node 434	52
Hidden layer 10, hidden node 434	53
Hidden layer 11, hidden node 434	48
Hidden layer 21, hidden node 434	54
Hidden layer 30, hidden node 434	45

Through simulations, different architecture is used to get the optimum architecture and promising accuracy. However, amongst the highest accuracy rate with 54% was by architecture of 21 hidden layer and 434 hidden nodes. Meanwhile, Table 5 shows the classification accuracy of different hidden layers and fixed 21 hidden nodes.

Table 5. Classification accuracy for 21 hidden nodes, equation based approach

DNN architecture	Accuracy (%)
Hidden layer 3, hidden node 21	49
Hidden layer 10, hidden node 21	46
Hidden layer 21, hidden node 21	51
Hidden layer 22, hidden node 21	50
Hidden layer 30, hidden node 21	45

Afterwards, intended to examine the architecture by using equation based approach. Amongst, the highest accuracy rate with 51% was by architecture of 21 hidden layer and 21 hidden nodes. It can be summarized that at both approaches, the architecture will be optimum and provide the desired accuracy at 21 hidden layers and hidden nodes.

5 Conclusion and Future Works

As a conclusion from the experiments, there are 2 possibilities in determining the number of hidden layers and hidden nodes. However, the number of hidden layers and hidden nodes can be either proportional or non-proportional to each other. Meanwhile, according to the provision of architecture portions this paper studies the classification performance using DNN. The experimental results show that all the examined architecture of hidden layers and hidden nodes able to achieve less than 60% accuracy only. Testing different architecture of neural network provides different desired performances as the learning process may differ for each architecture. Based on the practices, DNN still have a lot of room of improvements.

Thus, this study intended to improve the DNN by optimizing its hidden layers and hidden nodes in future. At present in the process of implementing Butterfly Optimization algorithm to examine the compatibility with DNN to improve the accuracy and complexity of the architecture, focusing to optimize the hidden layers and hidden nodes based on supervised learning. As a summary, the importance to find optimum number of hidden layers and hidden nodes is crucial for future experiments.

Acknowledgements. This research funded by Ministry of Higher Education (MOHE) under the Fundamental Research Grant Scheme (FRGS)—Vot. No. 1608. Besides, partially supported by Office for Research, Innovation, Commercialization and Consultancy Management (ORICC), UTHM.

References

1. EnterpriceTech. <https://www.enterprisetech.com/2017/07/10/deep-neural-networks-not-use/>
2. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Hellenic Conference on Artificial Intelligence, pp. 502–507. Springer, Berlin, Heidelberg (2006)
3. Mu, G., Cao, H., Jin, Q.: Violent scene detection using convolutional neural networks and deep audio features. In: Chinese Conference on Pattern Recognition, pp. 451–463. Springer, Singapore (2016)
4. Mironică, I., Duță, I.C., Ionescu, B., Sebe, N.: A modified vector of locally aggregated descriptors approach for fast video classification. *Multimed. Tools Appl.* **75**(15), 9045–9072 (2016)
5. Ali, A., Senan, N.: A review on violence video classification using convolutional neural networks. In: International Conference on Soft Computing and Data Mining, pp. 130–140. Springer, Cham (2016)

6. Dai, Q., Wu, Z., Jiang, Y.G., Xue, X., Tang, J.: Fudan-NJUST at MediaEval 2014: violent scenes detection using deep neural networks. In: MediaEval (2014)
7. Wu, Z., Jiang, Y.G., Wang, J., Pu, J., Xue, X.: Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 167–176. ACM (2014)
8. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks (2015). [arXiv: 1502.07209](https://arxiv.org/abs/1502.07209)
9. Eyben, F., Wenginger, F., Lehment, N., Schuller, B., Rigoll, G.: Affective video retrieval: violence detection in hollywood movies by large-scale segmental feature extraction. PLoS ONE 8(12), e78506 (2013)
10. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gonçalves, M.: Detecting spammers and content promoters in online video social networks. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 620–627. ACM (2009)
11. Zhang, B., Yi, Y., Wang, H., Yu, J.: MIC-TJU at MediaEval violent scenes detection (VSD) 2014. In: MediaEval (2014)
12. Schedi, M., Sjöberg, M., Mironică, I., Ionescu, B., Quang, V.L., Jiang, Y.G., Demarty, C.H.: Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos. In: 2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6. IEEE (2015)

Fibonacci Polynomials Based Functional Link Neural Network for Classification Tasks

Umer Iqbal^{1(✉)}, Rozaida Ghazali¹, and Habib Shah²

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
umeriqball5@gmail.com, rozaida@uthm.edu.my

² Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia
hurrahman@kku.edu.sa

Abstract. The artificial neural network has been proved among the best tools in data mining for classification tasks. The concept of obtaining more accurate classifier with less computational complexity has been gaining importance, because of day by day increase in the data. Several numbers of models have been developed for classification problems. This paper is the depiction of higher order neural networks especially Fibonacci Functional Link Neural Network (FFLNN) for data classification. The coefficients of individual terms in Fibonacci polynomials are smaller than those of individual terms in the classical orthogonal polynomials. Additionally, less number of terms make it a preferable classifier regarding functional expansion. These properties lead this FFLNN to produce more accurate higher order neural network with less computational complexity to tackle the classification problems. Four datasets were collected from KEEL and LIBSVM dataset repositories. Computational results were compared with three benchmarked models including Chebyshev Functional Link Neural Network (CFLNN), Chebyshev Multilayer Perceptron (CMLP) and Multilayer Perceptron Neural Network (MLP). A t-test was applied to check the significance of the proposed classifier based on classification performance. The findings showed that the proposed classifier outperformed all benchmarked models in all evaluation measures.

Keywords: Data mining · Classification · Fibonacci polynomials · Functional link neural network

1 Introduction

Classification has become more active and commonly encountered decision-making activity in the field of Artificial Neural Networks (ANNs) [1–6]. It appears as a revolutionary task in data mining. This problem occurs when an object needs to be assigned to a specific class or group on the base of their attributes that related to objects. Classification task is based on two steps. The first step is to construct the model, which represents the group of precogitated classes. The second step is the model usage, which is used for classifying the unknown objects. In order to solve the classification problem, the first step is the construction of model where the set of

example records known as training set is needed, which is presented to ANNs so that network can “learn” the pattern. During the training of network, each record set in the training set consists of numerous features. In features contained training set, one attribute known as classifying attribute is mainly used for the indication of the class to which each record related. After that, based on the functional relationship between classifying attribute and other attributes of training set record, ANNs creates classifier (classification model). In the second step, this new build classifier is used to classify the unseen record (out of sample record). Numbers of real-world application examples on neural classification tasks include data classification [7, 8], temperature time-series prediction [9], financial time-series prediction [10] and microarray classification [11].

Various techniques have been developed for classification, such as statistical and neural network techniques, which are prominent. With the passage of time, ANNs have gained much popularity as a useful alternative to statistical techniques and due to having the variety of applications in real life [12]. Multilayer perceptron known as MLP is a feed forward neural network that consists of input, hidden and output nodes. Every node is interconnected with other node in next layer which makes a connection between them. MLP has been used to find the missing values from data [13]. Due to its ability of fast learning, MLP has been tested for stock trading problems for better prediction [14]. MLP is also successfully implemented for early fault detection in gearboxes [15].

However, difficulties in fixing appropriate number of neurons and determining the appropriate number of hidden layers have to make the MLP architecture not rather easy to train. To overcome this, Higher Order Neural Networks (HONNs) were used. These HONNs are the type of artificial neural networks that are found to be very active for data classification as given in [16]. Functional Link Neural Network (FLNN) [17], Chebyshev Functional Link Neural Network (CFLNN) [18] and Chebyshev Multilayer Perceptron (CMLP) [7] are some known HONNs.

In this paper, we propose Fibonacci polynomials as functional expansion with FLNN. These polynomials are used to make standard FLNN more accurate for classification tasks. The contributions made by this study are as follow:

- This paper proposed Fibonacci polynomials based Functional Link Neural Network for classification.
- The two properties of Fibonacci polynomials such as; firstly, less number of terms as compared to the shifted Chebyshev, Legendre and Chebyshev polynomials, which means that with increasing degree of polynomials, the number of terms also increases. Secondly, the coefficients of individual terms in Fibonacci polynomials are smaller than the coefficients of individual terms in the classical orthogonal polynomials. These properties encourage us to implement these polynomials as functional expansion.
- A comparative analysis of proposed model with MLP, CMLP and CFLNN was conducted using four datasets. The performance of all techniques was verified in terms of four evaluation measures.

The rest of the paper is organized as follows: In Sect. 2, the proposed model is presented. Some experimental setups are presented in Sect. 3. Section 4 is devoted to results and discussions. Finally, Sect. 5 outlines the conclusion.

2 Proposed Model: Fibonacci Functional Link Neural Network

According to approximation theory, the non-linear approximation capacity of the Fibonacci orthogonal polynomial is very dominant [19]. The proposed method is a combination of the characteristics of Fibonacci polynomial and FLNN, which is, named as FFLNN. This model utilizes the FLNN input-output pattern with non-linear capabilities of the Fibonacci orthogonal polynomial for classification. The Fibonacci Functional Link Neural Network is single layer neural network. The structure consists of two parts, first one is the transformation part and learning is the second part. Transformation means that from a lower feature space to higher feature space. This transformation is also known as functional expansion, where Fibonacci polynomial basis can be seen as a new input layer. Table 1, describes the recurrence relation to find the Fibonacci polynomials of degree n .

Table 1. Recursive formula for Fibonacci polynomials

$F_0(x) = 0,$	(2.1)
$F_1(x) = 1,$	(2.2)
$F_n(x) = xF_{n-1}(x) + F_{n-2}(x).$	(2.3)

Where $F_0(x), F_1(x)$ is the Fibonacci polynomials when $n = 0, 1$ and $F_n(x)$ is the equation for the n th polynomial. The reason for which we are using Fibonacci polynomials is that, firstly, Fibonacci polynomials have less number of terms than the shifted Chebyshev, Legendre and Chebyshev polynomials which means that with increasing degree of polynomials, the number of terms also increases. Secondly, the coefficients of individual terms in Fibonacci polynomials are smaller than the coefficients of individual terms in the classical orthogonal polynomials. Since the computational errors in the product are related to the coefficient of individual terms, the computational errors are less by using Fibonacci polynomials. Motivated from above characteristics of Fibonacci polynomials, a new model has been proposed. The proposed method can be seen in Fig. 1. Where, $F_0(t), F_1(t), F_2(t) \dots F_n(t)$, are the enhanced inputs pattern and $w_0(t), w_1(t), w_2(t), \dots, w_n(t)$ are the neural network weights.

Where, $X_1, X_2 \dots X_m$ is the inputs and $Y_1, Y_2, \dots Y_n$ indicate the output of the neural network. The output Y of the GFLNN can be computed with the following Eq. (2.4):

$$Y = \sigma \left(\sum_{j=1}^n W_j * F(X_k) \right) \quad (2.4)$$

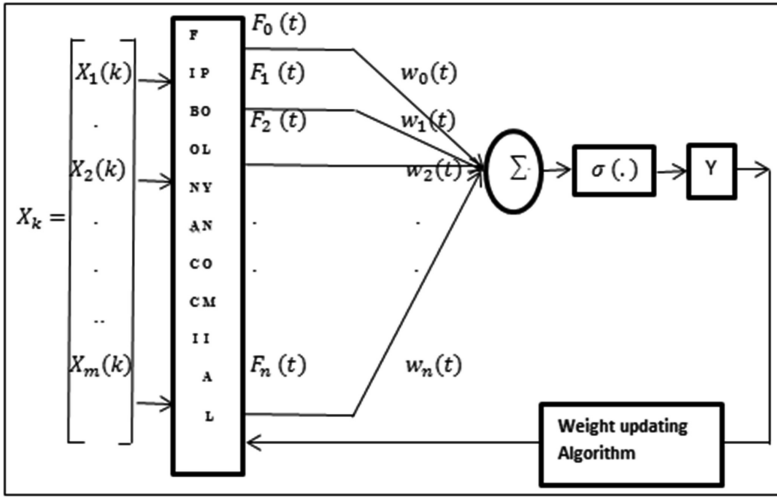


Fig. 1. Fibonacci functional link neural network

where, σ , $W(t) = [w_1(t), w_2(t) \dots w_n(t)]$ and $X(k) = [X_1(k), X_2(k) \dots X_m(k)]$ are activation function, weight vector and network inputs respectively. The dimension of input pattern expanded from 1 to n by Fibonacci polynomials (basis function) as shown in Eq. (2.5).

$$F(X_k) = [F_1(X_k), F_2(X_k), \dots F_n(X_k)] \quad (2.5)$$

3 Experimental Setup

This section illustrates the data description, evaluation metrics and network topology of proposed and comparison models.

3.1 Data Collection

The data that have been used for classification task is taken from KEEL and LIBSVM dataset repositories. The detail of four different datasets is described in Table 2.

Table 2. Description of datasets

Datasets	No. of samples	No. of attributes	No. of classes
Svmguide4	300	10	6
Banana	5300	2	2
Titanic	2210	3	2
Ringnorm	7400	20	2

3.2 Evaluation Measures

The performance evaluation of FFLNN against CFLNN, CMLP and MLP was done based on following evaluation measures as given in Table 3.

Table 3. Description of evaluation measures

Accuracy	$= \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$
Sensitivity	$= \frac{Tp}{Tp + Fn} \times 100$
Specificity	$= \frac{Tn}{Tn + Fp} \times 100$
Area under the curve (ROC)	$= \int_{-\infty}^{\infty} A_0(t) a_1(t) dt,$ $AUC = \int_{-\infty}^{\infty} A_0(A'_1(x)) dx,$ Here, $a_1(t) = \frac{dA_1(t)}{dt}$

Where Tp, Tn, Fp and Fn are the true positive, true negative, false positive and false negative values respectively.

3.3 Training and Network Topology

The proposed model topology of FFLNN is shown in Table 4. Settings were selected empirically. Levenberg Marquardt back propagation (LMBP) was used as learning algorithm with all techniques.

Table 4. Network topology

Setting	Value
Activation function	Sigmoid function
Fibonacci polynomial degree	3
Stopping criteria	Maximum no. of epochs = 1000
Learning rate	[0.1–0.3]
Momentum	[0.3–0.9]
Learning algorithm	LM back propagation

4 Results and Discussion

In this section, simulation results for classification of the four datasets are discussed and presented. To verify the experimental results, trial and error method was performed 10 times per experiments with an average of each experiment. This help to avoid any influence due to the initial internal state like initialization of random weights. The FFLNN performance was compared with CFLNN, CMLP and MLP. Performance

of all techniques was based on pre-defined evaluation measures. All simulations were carried on Intel Core i7-3770XPU@3.40 GHz, and 4 GB of RAM machine.

4.1 Best Average Simulation Results

Simulation results were presented in Fig. 2. It can be seen that in Ringnorm dataset, FFLNN performs 95.73% accurate classification, whereas, CFLNN, CMLP and MLP shown 86.73%, 77.00% and 60.85.00% respectively. The reason behind these significant results is that the proposed solution helps FFLNN to raise and find more appropriate settings during the training that enhances the classification performance for the network. After that, in Svmguide4 and Banana datasets, FFLNN has shown approximately same results but at the same time, it provides better results as compared to rest of the models. The same parameters setting and the number of epochs are the reason behind the same results. The performance of CFLNN was also well after FFLNN on all datasets as compared to CMLP and MLP. Finally, results reveal that the proposed model was found to be competitive on all data sets in terms of classification accuracy, which means that Fibonacci polynomials provide the better classification accuracy.

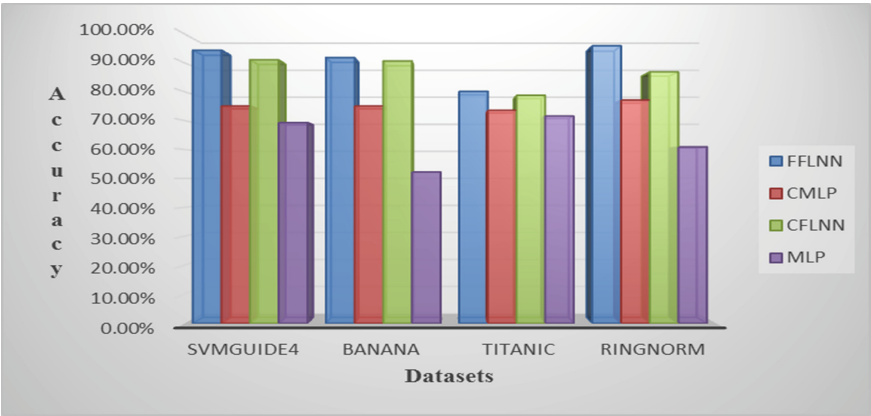


Fig. 2. Comparison in terms of classification accuracy

Table 5, describes the simulation results of FFLNN, CFLNN, CMLP and MLP in terms of sensitivity and specificity. Results clearly indicate that performance of FFLNN is competitive with CFLNN on all datasets in terms of sensitivity, specificity. In Ringnorm classification problem, FFLNN shows a clear boundary. After that, FFLNN has gained better results in Svmguide4 and Banana dataset. Similarly, in the comparative study of CMLP and MLP, CMLP was competitive with MLP in all datasets. Simply, results have shown the clear edge of FFLNN over CFLNN, CMLP and MLP in all four classification problems in terms of sensitivity and specificity. Silver colour was used to highlight the performance of the proposed model.

Table 5. Simulation results of sensitivity and specificity of all models on all datasets (%)

Datasets	Models	Sensitivity	Specificity
Svmguide4	FFLNN	93.11	94.73
	CFLNN	90.91	94.51
	CMLP	75.00	69.00
	MLP	69.23	65.00
Ringnorm	FFLNN	95.00	90.93
	CFLNN	86.73	87.03
	CMLP	77.00	70.10
	MLP	60.85	61.03
Titanic	FFLNN	80.74	75.23
	CFLNN	78.64	69.30
	CMLP	73.50	65.00
	MLP	71.55	63.87
Banana	FFLNN	92.38	91.22
	CFLNN	90.46	90.13
	CMLP	75.00	73.45
	MLP	52.25	52.30

To measure the classification accuracy of each class, the area under roc curve (AUC) was used to measure the exact percentage. Figure 3, illustrates the performance of all models in terms of AUC. It can be judged that FFLNN is coping well with CFLNN, CMLP and MLP. It is good to note that higher will be the percentage of AUC, higher will be the accuracy. In Svmguide4, FFLNN shows its best results, whereas, CFLNN performance was found to be good on Titanic dataset as compared to CMLP and MLP. In Svmguide4 and Banana datasets, CMLP and MLP have shown same performance. In Short, FFLNN was remarkable on all datasets.

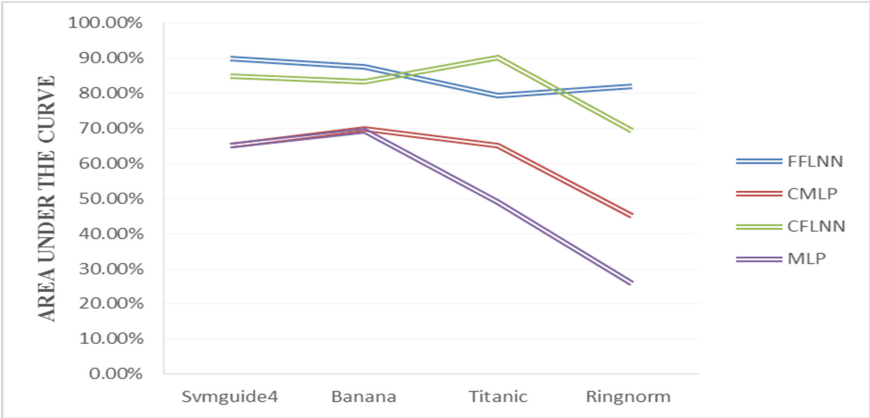


Fig. 3. Comparison in terms of area under the curve (ROC)

Significance testing using t-Test

To check that how much significant is our proposed model, we have applied paired two sample for means t-test in a relation to accuracy [11]. We found that in all datasets our proposed method has shown significance except in Banana dataset when compared with CFLNN. Insignificance has caused due to the approximately similar high accuracy of both proposed and comparison method. In rest of the datasets, all comparison techniques remained insignificant in all datasets. The hypothetic value was taken 0.05. If we achieved 'value < 0.05' then called as significant and 'value > 0.05' then called as insignificant.

5 Conclusion

This research highlights an important contribution to the proposed Fibonacci Functional Link Neural Network (FFLNN) model to achieve multi-class classification task. FFLNN was proposed based on FLNN. FFLNN was successfully applied to solve nonlinear higher dimension problems. In this model, Fibonacci polynomials were used as a functional expansion for better accuracy and to reduce computational work. In comparison with the other techniques like CFLNN, CMLP and MLP; this model was very fast and efficient. To provide the fair comparison among all the models, four-evaluation measures were adopted. Irrespective of used models and datasets, it is clear that Fibonacci polynomials were found to be effective to improve the classification accuracy. Moreover, when four models namely FFLNN, CFLNN, CMLP and MLP were compared, proposed model outperforms other three models. Many classification models such as statistical and neural network techniques were developed to solve the different types of classification problems. With the passage of time, more models are under consideration to make classification task more accurate than before. Although this may seem like a never-ending cycle, it is not possible to obtain such model which can classify all kinds of classification problems. In this article, different types of neural networks and higher order neural networks such as FFLNN, CFLNN, CMLP and MLP were discussed and compared. The advantages and drawbacks of MLP, CMLP and CFLNN were explained as well as a solution of these techniques was deliberated to make them more accurate for classification. The performance of all comparison techniques was evaluated on four bench-mark datasets in terms of accuracy, sensitivity, specificity and area under the curve. Simulation results have shown that FFLNN performance was better than the rest of the techniques due to its less number of coefficients and degree properties. T-test has confirmed the significance of FFLNN over all techniques.

Acknowledgements. The authors would like to thank King Khalid University to provide the International Research Grant with Grant number A134 for supporting this research.

References

1. Zhang, G.P.: Neural networks for classification: a survey. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **30**(4), 451–462 (2000)
2. Misra, B.B., Dehuri, S.: Functional Link Artificial Neural Network for Classification Task in Data Mining, vol. 1 (2007)
3. Chen, C., Duan, S., Cai, T., Liu, B.: Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Sol. Energy* **85**(11), 2856–2870 (2011)
4. Al-Jarrah, O., Arafat, A.: Network intrusion detection system using neural network classification of attack behavior. *J. Adv. Inf. Technol.* **6**(1) (2015)
5. Mason, M.: Classification of Handwritten Digits Using an Artificial Neural Network (2015)
6. Babaei, T., Lim, C.P., Abdi, H., Nahavandi, S.: A Modified functional link neural network for data classification. In: *Emerging Trends in Neuro Engineering and Neural Computation*, pp. 229–244. Springer, Singapore (2017)
7. Iqbal, U., Ghazali, R.: Chebyshev multilayer perceptron neural network with Levenberg Marquardt-back propagation learning for classification tasks. In: *International Conference on Soft Computing and Data Mining*, pp. 162–170. Springer, Cham (2016)
8. Mohmad Hassim, Y.M., Ghazali, R.: Using artificial bee colony to improve functional link neural network training. In: *Applied Mechanics and Materials*, pp. 2102–2108. Trans Tech Publications (2013)
9. Ghazali, R., Husaini, N.A., Ismail, L.H., Herawan, T., Hassim, Y.Y.M.: The performance of a Recurrent HONN for temperature time series prediction. In: *2014 International Joint Conference on Neural Network*, pp. 518–524 (2014)
10. Ghazali, R., Hussain, A.J., Merabti, M.: Higher order neural networks and their applications to financial time series prediction. In: *Artificial Intelligence and Soft Computing*, pp. 120–125 (2006)
11. Kumar, M., Singh, S., Rath, S.K.: Classification of microarray data using functional link neural network. *Proc. Comput. Sci.* **57**, 727–737 (2015)
12. Paliwal, M., Kumar, U.A.: Neural networks and statistical techniques: a review of applications. *Expert Syst. App.* **36**(1), 2–17 (2009)
13. Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M.: Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Appl. Soft Comput.* **29**, 65–74 (2015)
14. Mabu, S., Obayashi, M., Kuremoto, T.: Ensemble learning of rule-based evolutionary algorithm using multi-layer perceptron for supporting decisions in stock trading problems. *Appl. Soft Comput.* **36**, 357–367 (2015)
15. Jedliński, Ł., Jonak, J.: Early fault detection in gearboxes based on support vector machines and multilayer perceptron with a continuous wavelet transform. *Appl. Soft Comput.* **30**, 636–641 (2015)
16. Bebarta, D.K., Rout, A.K., Biswal, B., Dash, P.K.: Forecasting and classification of Indian stocks using different polynomial functional artificial neural networks. In: *2012 Annual IEEE India Conference (INDICON)*, pp. 178–182 (2012)
17. Pao, Y.H., Takefuji, Y.: Functional-link net computing: theory, system architecture, and functionalities. *Computer* **25**(5), 76–79 (1992)
18. Li, M., Liu, J., Jiang, Y., Feng, W.: Complex-Chebyshev functional link neural network behavioral model for broadband wireless power amplifiers. *IEEE Trans. Microw. Theory Tech.* **60**(6), 1979–1989 (2012)
19. Ozdemir, G., Simsek, Y.: Generating functions for two-variable polynomials related to a family of Fibonacci type polynomials and numbers. *Filomat* **30**(4), 969–975 (2016)

Decision Support Model in Determining Factors and Its Dominant Criteria Affecting Cholesterol Level Based on Rough-Regression

Riswan Efendi^{1,2(✉)} and Mustafa Mat Deris¹

¹ Faculty of Computer Science, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

{riswan, mmustafa}@uthm.edu.my

² Mathematics Department, Faculty of Science and Technology, State Islamic University of Sultan Syarif Kasim Riau, 28294 Panam, Pekanbaru, Indonesia

Abstract. The statistical regression models have been frequently used to explain the causal relationship between exogenous factors and the cholesterol level of patients. While, the dominant criteria for each exogenous factor which give impact to the cholesterol level are not yet investigated by previous studies. In this paper, we are interested to introduce a decision making model based on rough-regression in handling the significant contribution between the dominant criteria, exogenous and endogenous factors, respectively. The result showed the proposed model is able to investigate the dominant criteria and factors affecting cholesterol level patients. This model may help the counterparts in the decision making.

Keywords: Rough-regression · Decision making · Dominant criteria
Cholesterol level

1 Introduction

Monitoring of cholesterol level is very essential activity for patients to continue their life. Besides that, this monitoring is also help the counterparts in providing information for decision making and health budget planning. Many models have been presented to investigate the factors (variables), such as, blood pressure [1], sleeping hour [2, 3], weight or obesity [4] and calorie level [5] which affect the patient cholesterol level. However, the dominant criteria (category) for each factor which give the significant impact to the cholesterol level is still issue and not easy to handle.

In the previous work, rough set model has been implemented to medical diseases, especially in prediction and management the diabetes mellitus [6]. In this paper, we introduce a decision making model to solve the issue and problem using rough-regression model (RRM). This model is very appropriate to handle the relationship among the qualitative variables with categorical data. The detail of theories, proposed model, and implementation are discussed in Sects. 2, 3 and 4, respectively.

2 The Basic Theories

2.1 Rough Set Theory

The rough set theory has been proposed by Pawlak in 1982 [7], this theory has been well divided by researchers into information systems, indiscernibility relation, set approximations, rough clustering, and others. An information system $S = (U, \Omega, V_q, f_q)$ consists of [8–12]:

U : a nonempty, finite set called the universe;

Ω : a nonempty, finite set of attributes;

$\Omega = C \cup D$, in which C is a finite set of condition attributes and D is a finite set of decision attributes;

for each $q \in \Omega$, V_q is called the domain of q ;

f_q : an information function $f_q : U \rightarrow V_q$.

Objects can be interpreted as cases, states, processes, patients and observations. Attributes can be assumed as features, variables, and characteristic information. A special case of information systems called decision table or attribute-value table is applied in the following analysis. In a decision table, the row and column correspond to objects and attributes, respectively. The starting point of rough set theory is the indiscernibility relation, generated by information about objects of interest. Let $S = (U, \Omega, V_q, f_q)$ be an information system, then any subset B of A determines a binary (equivalence) relation $IND(B)$ on U , which will be called B -indiscernibility relation, and is defined as follows:

$$IND(B) = \{(x, y) \in U^2 : \forall a \in B, a(x) = a(y)\}, \quad (1)$$

where $a(x)$ denotes the value of attribute a for element x in U . The collection of all equivalence classes determined by $IND(B)$, denoted by U/B . An equivalence class of U/B , containing x , is denoted by $[x]_B$. In rough set theory, an equivalence class is the basic concepts of our knowledge. The indiscernibility relation will be used next to define approximations, basic concepts of rough set theory.

Let $S = (U, \Omega, V_q, f_q)$ be an information system and let $B \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained in B by constructing the B -lower and B -upper approximations of X . Both approximations are denoted as:

$$\underline{B}(X) = \{ \{x \in U | [x]_B \subseteq X\} \}, \quad (2)$$

and

$$\overline{B}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\}, \quad (3)$$

where $[x]_B$ is an equivalence class containing x . While, the difference between both approximations and its accuracy can be written:

$$\text{BND}(X) = \underline{B}(X) - \overline{B}(X), \quad (4)$$

$$\alpha(X) = \frac{\underline{B}(X)}{\overline{B}(X)}, \quad (5)$$

Moreover, the dependency attributes is formulated as [13]:

$$k = \frac{\sum_{x \in U/D} |\underline{C}(X)|}{|U|}; C, D \subseteq A \wedge C \cap D = \emptyset. \quad (6)$$

The maximum value of k can be interpreted as a dominant attribute or criteria.

2.2 Regression Model

A simple linear regression can be written as [14]:

$$Y = \beta_0 + \beta_1 X + e, \quad (7)$$

where Y is a dependent (endogenous) variable, X is an independent (exogenous) variable, β_0 and β_1 are intercept and slope, while e is error of model. The algorithm in building Eq. (7) can be explained by following steps:

- Step 1: Check correlation between Y and X .
- Step 2: Estimate intercept and slope, respectively using ordinary least square (OLS) method.
- Step 3: Verify the significance X and Y using ANOVA and F -test.
- Step 4: Verify the significance intercept and slope using t -test.
- Step 5: Verify the normality error.

3 Proposed Decision Making Model

Let we have a multiple regression equation as follows [14]:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + e, \quad (8)$$

Based on Eq. (8), X_1, \dots, X_n are independent variables which affect to dependent variable, Y . While, a_0, \dots, a_n are estimated parameters (coefficients). In this equation, the main objective to determine the significant independent variables and its coefficients by following steps given in Sect. 2.2. We assume all independent variables affect to the dependent variable. In the previous studies, there is no approach can be implemented to investigate the dominant criteria (decisive condition) from each independent variable which give the maximum dependency to dependent variable. In this paper, we are interested to implement rough set approximations into regression model for investigating the dominant criteria from each independent variable by following steps:

Step 1: Transform real values of independent and dependent variables into categorical values (data) as presented in Table 1.

Table 1. Categorical data of exogenous and endogenous factors

ID	X_1	X_2	\dots	X_n	Y
P_1	Low	Low	\dots	Very rarely	Low
P_2	Low	Moderate	\dots	Very rarely	High
P_3	Very low	Low	\dots	Rarely	Medium
P_4	Many	Very low	\dots	Sometimes	High
\dots	\dots	\dots	\dots	\dots	\dots
P_n	Many many	Sometimes	\dots	Very rarely	High

Step 2: Based on Eqs. (2), (3), (5) and Table 1, determine the accuracy approximation as follows:

$$\begin{aligned} \text{Let } U/Y &= U/\{\text{Low cholesterol, Medium cholesterol, High cholesterol}\} \\ &= [\{R_{1,\dots,R_{n-3}}\}, \dots, \{R_{3,\dots,R_{n-5}}\}]. \end{aligned}$$

$$\underline{B}(X_1, \dots, X_n) = \{R_2, R_7, \dots, R_{n-3}, R_{n-1}\},$$

$$\overline{B}(X_1, \dots, X_n) = \{R_7, R_{n-7}, \dots, R_{n-5}, R_{n-1}\},$$

Then, accuracy approximation can be written as:

$$\alpha(X_1, \dots, X_n) = \frac{\underline{B}(X_1, \dots, X_n)}{\overline{B}(X_1, \dots, X_n)}. \quad (9)$$

Step 3: Based on Step 2, determine universe of decision as follows:

$$U/D = U/Y = \{\{R_{1,\dots,R_{n-3}}\}, \dots, \{R_{1,\dots,R_{n-5}}\}\}, \quad (10)$$

Step 4: Determine universe of each criteria (category) and variable (factors) as follows:

For X_1 ,

$$U/X_1(\text{Criteria} - 1) = \{\{R_1, \dots, R_{n-3}\}, \dots, \{R_4, \dots, R_{n-5}\}\},$$

$$U/X_1(\text{Criteria} - p) = \{\{R_3, \dots, R_{n-2}\}, \dots, \{R_7, \dots, R_{n-4}\}\},$$

For X_2 ,

$$U/X_2(\text{Criteria} - 1) = \{\{R_4, \dots, R_{n-1}\}, \dots, \{R_7, \dots, R_{n-8}\}\},$$

$$U/X_2(\text{Criteria} - p) = \{\{R_1, \dots, R_{n-2}\}, \dots, \{R_4, \dots, R_{n-4}\}\},$$

For X_n ,

$$\begin{aligned} U/X_n(\text{Criteria} - 1) &= \{\{R_1, \dots, R_{n-3}\}, \dots, \{R_4, \dots, R_{n-5}\}\}, \\ &\dots \\ U/X_n(\text{Criteria} - p) &= \{\{R_1, \dots, R_{n-3}\}, \dots, \{R_4, \dots, R_{n-5}\}\}, \end{aligned}$$

Step 5: Based on Eq. (6), determine dominant criteria for each variable or factor as follows:

For X_1 ,

$$\begin{aligned} \text{Criteria} - 1 &\rightarrow k_1^Y = \frac{U/X_1(\text{Criteria}-1)}{U/Y}, \\ &\dots \\ \text{Criteria} - p &\rightarrow k_p^Y = \frac{U/X_1(\text{Criteria}-p)}{U/Y}. \end{aligned}$$

For X_2 ,

$$\begin{aligned} \text{Criteria} - 1 &\rightarrow k_1^Y = \frac{U/X_2(\text{Criteria}-1)}{U/Y}, \\ &\dots \\ \text{Criteria} - p &\rightarrow k_p^Y = \frac{U/X_2(\text{Criteria}-p)}{U/Y}. \end{aligned}$$

For X_n ,

$$\begin{aligned} \text{Criteria} - 1 &\rightarrow k_1^Y = \frac{U/X_n(\text{Criteria}-1)}{U/Y}, \\ &\dots \\ \text{Criteria} - p &\rightarrow k_p^Y = \frac{U/X_n(\text{Criteria}-p)}{U/Y}. \end{aligned}$$

Step 6: Based on Step 5, determine the max-value (dominant) for each category (criteria) and variable as follows:

For X_1 ,

$$\text{Dominant criteria}(X_1) = \max \left\{ (k_1^Y), \dots, k_p^Y \right\}, \quad (11)$$

For X_2 ,

$$\text{Dominant criteria}(X_2) = \max \left\{ (k_1^Y), \dots, k_p^Y \right\}, \quad (12)$$

For X_n ,

$$\text{Dominant criteria}(X_n) = \max \left\{ (k_1^Y), \dots, k_p^Y \right\}, \quad (13)$$

4 Implementation

Based on steps given in Sect. 2.2 and using row data from 20 patients, a linear relationship (regression model) between sleeping hour (X_1), weight (X_2), calorie level (X_3), blood pressure (X_4) and cholesterol level (Y) can be written mathematically:

$$Y = 71.791 + 3.012X_1 - 0.460X_2 + 0.002X_3 + 0.119X_4. \quad (14)$$

Based on Eq. (14), the coefficient (slope) of sleeping hour is highest among them, so that this variable give the biggest impact to cholesterol level if compared with weight, calorie level and blood pressure. On the other hand, we do not know which criteria (condition) from all exogenous factors or variables really affect to the patient cholesterol level. By using proposed model in Sect. 3, we can investigate the dominant condition or criteria from each exogenous variable by following steps:

Step 1: Transform numerical values of exogenous and endogenous variables into categorical values (data) as presented in Table 2.

Table 2. Transformation numerical into categorical values of variables

Patient ID	X_1	X_2	X_3	X_4	Y
P_1	Not normal	Small	L_2	Normal	High level
P_2	Not normal	Small	L_1	Normal	High level
P_3	Normal	Medium	L_3	Pre-H	High level
...
P_{20}	Normal	Big	L_3	Pre-H	High level

Step 2: Based on Table 2, determine universe of decision as follows:

$$\begin{aligned} U/Y &= \{\text{High level cholesterol}\}, \\ U/Y &= \{P_1, \dots, P_{20}\}. \end{aligned} \quad (15)$$

Step 3: Determine universe of each criteria (category) and variable (factors) as follows:

For X_1 ,

$$\begin{aligned} U/(\text{Not normal}) &= \{P_1, P_2, P_8, P_{11}, P_{13}, P_{18}\}, \\ U/(\text{Normal}) &= \{P_3, P_5, P_6, P_9, P_{12}, P_{14}, P_{15}, P_{16}, P_{17}, P_{20}\}, \\ U/(\text{Over sleep}) &= \{P_4, P_7, P_{10}, P_{19}\}. \end{aligned}$$

For X_2 ,

$$\begin{aligned}
 U/(\text{Small}) &= \{P_1, P_2, P_6, P_{11}\}, \\
 U/(\text{Medium}) &= \{P_3, P_5, P_8, P_9, P_{12}, P_{13}, P_{18}\}, \\
 U/(\text{Big}) &= \{P_4, P_7, P_{10}, P_{14}, P_{15}, P_{16}, P_{17}, P_{19}, P_{20}\}.
 \end{aligned}$$

For X_3 ,

$$\begin{aligned}
 U/(\text{L1}) &= \{P_2, P_8, P_{13}, P_{17}, P_{18}\}, \\
 U/(\text{L2}) &= \{P_1, P_9, P_{11}, P_{12}\}, \\
 U/(\text{L3}) &= \{P_3, P_{15}, P_{16}, P_{19}, P_{20}\}, \\
 U/(\text{L4}) &= \{P_2, P_8, P_{13}, P_{17}, P_{18}\}.
 \end{aligned}$$

For X_4 ,

$$\begin{aligned}
 U/(\text{Low}) &= \{P_3, P_5, P_8, P_9, P_{12}, P_{13}, P_{18}\}, \\
 U/(\text{Normal}) &= \{P_{11}, P_{13}\}, \\
 U/(\text{Pre} - \text{H}) &= \{P_3, P_4, P_5, P_9, P_{10}, P_{18}, P_{19}, P_{20}\}, \\
 U/(\text{H} - 1) &= \{P_7\}.
 \end{aligned}$$

Step 4: Based on Eq. (6), determine dominant criteria for each variable or factor as follows:

For X_1 ,

$$\text{Not normal} \rightarrow k_1^Y = \frac{6}{20}, \text{Normal} \rightarrow k_2^Y = \frac{10}{20}, \text{Over sleep} \rightarrow k_3^Y = \frac{4}{20}.$$

For X_2 ,

$$\text{Small} \rightarrow k_1^Y = \frac{4}{20}, \text{Medium} \rightarrow k_2^Y = \frac{7}{20}, \text{Big} \rightarrow k_3^Y = \frac{9}{20}.$$

For X_3 ,

$$\text{L1} \rightarrow k_1^Y = \frac{5}{20}, \text{L2} \rightarrow k_2^Y = \frac{5}{20}, \text{L3} \rightarrow k_3^Y = \frac{5}{20}, \text{L4} \rightarrow k_3^Y = \frac{5}{20}.$$

For X_4 ,

$$\begin{aligned}
 \text{Low} \rightarrow k_1^Y &= \frac{2}{20}, \text{Normal} \rightarrow k_2^Y = \frac{9}{20}, \text{Pre} - \text{H} \rightarrow k_3^Y = \frac{8}{20}, \text{H1} \rightarrow k_3^Y \\
 &= \frac{1}{20}.
 \end{aligned}$$

Step 5: Based on Step 4, determine the max-value (dominant) for each category (criteria) and variable as follows:

For X_1 ,

$$\text{Dominant criteria of } X_1 = \max \left\{ \frac{6}{20}, \frac{10}{20}, \frac{4}{20} \right\} = \frac{10}{20}. \quad (16)$$

For X_2 ,

$$\text{Dominant criteria of } X_1 = \max \left\{ \frac{4}{20}, \frac{7}{20}, \frac{9}{20} \right\} = \frac{9}{20}.$$

For X_3 ,

$$\text{Dominant criteria of } X_3 = \max \left\{ \frac{5}{20}, \frac{5}{20}, \frac{5}{20}, \frac{5}{20} \right\} = \text{all}. \quad (17)$$

For X_4 ,

$$\text{Dominant criteria of } X_4 = \max \left\{ \frac{2}{20}, \frac{9}{20}, \frac{8}{20}, \frac{1}{20} \right\} = \frac{9}{20}. \quad (18)$$

Step 6: Based on Step 5, provide a decision making table for the dominant criteria (category) and variables affecting cholesterol level as presented in Table 3.

Table 3. Decision making for dominant criteria and factors

Variable/Impact	Criteria	Y (High level cholesterol)
X_1 (+)	Normal	10/20
X_2 (-)	Big	9/20
X_3 (+)	All criteria	5/20
X_4 (+)	Normal	9/20

Based on Table 3, the dominant criteria and factors which affect to cholesterol level (Y) are the sleeping hour (X_1) with positive impact and normal criteria (7–9 h) per day, the weight (X_2) with negative impact and big criteria (63–69 kg), the calorie level (X_3) with positive impact and all criteria (2235–3455 Cal) and blood pressure (X_4) with positive impact and normal criteria (90–110 mm/Hg). While, for other criteria are not dominant. In this case, most of condition cholesterol patients fall in the normal criteria, but we assume that the life style and the food consumption are also very significant in influencing their cholesterol level.

5 Conclusion

In this paper, we propose a procedure to investigate the dominant criteria (category) of each independent variable which influence the dependent variable. This proposed procedure has been examined to evaluate the dominant criteria and factors affecting

student achievement. While this procedure is no yet discussed in the previous studies. Interestingly, the proposed procedure is also help the decision makers in determining the dependency variables to the dominant criteria precisely. Based on our perspective, rough-regression model (RRM) is suggested to handle the qualitative variables (data) in the social-economics research domains.

Acknowledgements. This study is supported by Research, Innovation, Commercialization, and Consultancy Management Office (ORICC) at Universiti Tun Hussein Onn Malaysia (UTHM) and in part by Contract Research Grant Vot. U689.

References

1. Sakurai, M., Stamler, J., Miura, K., Brown, I.J., Nakagawa, H., Elliot, P., Ueshima, H., Chan, Q., Tzoulaki, I., Dyer, A.R., Okayama, A., Zhao, L.: Relationship of dietary cholesterol to blood pressure: the intermap study. *J. Hypertens.* **29**, 222–228 (2011)
2. Wolk, R., Somers, V.K.: Sleep and the metabolic syndrome. *Exp. Physiol.* **92**, 67–78 (2007)
3. Kaneita, Y., Uchiyama, M., Yoshiike, N., Ohida, T.: Associations of usual sleep duration with serum lipid and lipoprotein levels. *Sleep* **31**, 645–652 (2008)
4. Miettinen, T.A.: Cholesterol production in obesity. *Circulation* **64**, 842–850 (1971)
5. Ueshima, H., Iida, M., Shimamoto, T., Konishi, M., Tanigaki, M., Doi, M., Nakashini, N., Takayama, Y., Ozawa, H., Komachi, Y.: Dietary intake and serum total cholesterol level: their relationship to different lifestyles in several Japanese populations. *Circulation* **66**, 519–526 (1982)
6. Ali, R., Hussain, J., Siddiqi, M.H., Hussain, M., Lee, S.: A hybrid rough set reasoning model for prediction and management of Diabetes Mellitus. *Sensors* **15**, 15921–15951 (2015)
7. Pawlak, Z.: Rough sets. *Int. J. Compt. Inf. Sci.* **11**, 341–356 (1982)
8. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publisher Dordrecht (1991)
9. Hampton, J.: Rough set theory: the basics (part 1). *J. Compt. Intel. Finance* **5**, 25–29 (1997)
10. Hampton, J.: Rough set theory: the basics (part 1). *J. Compt. Intel. Finance* **6**, 40–42 (1998)
11. Hampton, J.: Rough set theory: the basics (part 1). *J. Compt. Intel. Finance* **6**, 35–37 (1998)
12. Tay, F.E.H., Shen, L.: Economic and financial using rough sets model. *Eur. J. Oper. Res.* **141**, 641–659 (2002)
13. Herawan, T., Deris, M.M., Abawajy, H.: A rough set approach for selecting clustering attribute. *Knowl. Based Syst.* **23**, 220–231 (2010)
14. Wooldridge, M.: *Introductory Econometrics a Modern Approach*, 3rd edn. Thomson, South Western, USA (2006)

A Numerical Classification Technique Based on Fuzzy Soft Set Using Hamming Distance

Iwan Tri Riyadi Yanto^{1(✉)}, Rd Rohmat Saedudin²,
Saima Anwar Lashari³, and Havaluddin⁴

¹ Department of Information Systems, University of Ahmad Dahlan, Kampus III
UAD, Jalan Prof. Dr. Soepomo, Yogyakarta, Indonesia
yanto.itr@is.uad.ac.id

² School of Industrial Engineering, Telkom University, 40257 Bandung, West
Java, Indonesia
rdrohmat@telkomuniversity.ac.id

³ Faculty of Computer Science and Information Technology, Universiti Tun
Hussein Onn, Parit Raja, Johor, Malaysia

⁴ Faculty of Computer Science and Information Technology, Mulawarman
University, Samarinda, Indonesia
havaluddin@unmul.ac.id

Abstract. In recent decades, fuzzy soft set techniques and approaches have received a great deal of attention from practitioners and soft computing researchers. This article attempts to introduce a classifier for numerical data using similarity measure fuzzy soft set (FSS) based on Hamming distance, named HDFSSC. Dataset have been taken from UCI Machine Learning Repository and MIAS (Mammographic Image Analysis Society). The proposed modeling consists of four phases: data acquisition, feature fuzzification, training phase and testing phase. Later, head to head comparison between state of the art fuzzy soft set classifiers is provided. Experiment results showed that the proposed classifier provides better accuracy when compared to the baseline fuzzy soft set classifiers.

Keywords: Fuzzy soft set (FSS) · Similarity measure · Hamming distance, classification

1 Introduction

In recent years, computers and their peripherals have been made cheaper and more readily available and in line with the development of information technology, various kinds of advanced data mining techniques have hit the market. These new age data mining techniques embrace traditional and more recent sophisticated classification algorithms. Both classification techniques are for handling complex datasets such as multidimensionality, user inference and prior knowledge, web data, spurious data points that cause overfitting of models, improvement in human ability, noisy datasets cleaning, mining multimedia datasets and incremental datasets. Interdisciplinary data mining techniques and approaches can be used for all the above mentioned databases

for forecasting the impact and discovering meaningful relationships in the data with the purpose of extracting useful information for knowledge generation [1].

Thus, a variety of models have been fitted in order to determine hidden patterns in the data [2, 3]. The approach that is able to produce the most accurate output and relationships pattern in the observed datasets is considered to be the most efficient in the particular model. Such approach fulfills the objective of data mining. Current data mining practices utilizes a range of model functions including classification, regression, clustering, discovering association rules and sequence analysis [4]. Hence, solutions are needed in order to manage and analyze such as complex, diverse, and huge datasets in a reasonable time complexity and storage capacity for enhanced insight and decision-making.

Molodtsov [5] investigated soft set theory that classifies the objects with help of binary information. In principle, the initial description of any object has an approximate nature and one do not need to introduce the notion of exact solution. Therefore, the problem of membership function setting does not arise in this theory. Currently, soft set theory is being rapidly progressing in several fields of sciences, engineering, economics and medicals sciences. Maji et al. [6] did further exploration and analysis on soft set theory by providing some operations and viability of soft set into decision making problems. Thus, the scope application of fuzzy soft set theory is still available to be expand. The numerical data classification is one of the potential applications of it. A novel classification method using notions on soft set theory has been proposed by Mushrif et al. [7] on natural texture where the type data consist of a numerical value between $[0,1]$.

The measurement of the similarity has an important role in classification using FSS. Currently, some use on measuring similarity have been carried out [8, 9]. Handaga et al. [10] proposed an algorithm namely FSSC which provides high accuracy and the proposed FSSC used general similarity measure of FSS. However, there are various distance measures in mathematics. A new similarity measures of FSS based on different distance measures has been proposed by Feng and Zheng [11]. The similarity measure based on Hamming distance in this paper is more reasonable. Thus, we propose an alternative technique for classification based on FSS similarity measurement using hamming distance which is has good performance in term of accuracy and time responses as compared existing FSS classifiers. Eight dataset have been used.

The rest of the paper is organized as follows: soft set and FSS theory are introduced in Sect. 2. Section 3 discusses similarity measure and distance measure in details. Section 4 demonstrates fuzzy soft set classification using hamming distance. Section 5 exhibits the experimental results and summary of paper are given in Sect. 6.

2 Soft Sets and FSS (Fuzzy Soft Sets)

2.1 Soft Sets

Let U be an initial universe set and E be a set of parameters. Parameters are properties of objects, it can be known as attribute, factor or characteristics of objects. Let the power set of U as $P(U)$ and $A \subset E$ [3]. A family of subsets parameters of the universe U is

called as soft set and can be defined as A pair (F, A) , where F is a mapping given by, $F: A \rightarrow P(U)$.

2.2 Fuzzy Soft Sets

Let U be an initial universe set and E be a set of parameters (which are fuzzy words or sentences involving fuzzy words). Let $P(U)$ denotes the set of all fuzzy sets of U . Let $A \subset E$. A pair (F, A) is called a fuzzy soft set (FSS) over U , where F is a mapping given by $F: A \rightarrow P(U)$.

Example 1 Suppose a FSS (F, E) describes desirability of the gowns with respect to the given parameters, which Mr X going to wear $U = \{g_1, g_2, g_3, g_4, g_5\}$ which is the set of gown under consideration and E is a set of decision parameters $E = \{e_1, e_2, e_3, e_4, e_5\}$. Let $P(U)$ be the collection of all fuzzy subsets of U . Also let $E = e_1 = \text{"exclusive"}$, $e_2 = \text{"striking"}$, $e_3 = \text{"vibrant"}$, $e_4 = \text{"inexpensive"}$, $e_5 = \text{"warm"}$.

Meanwhile, mapping models $F: E \rightarrow P(U)$ is given by where $(.)$ is to be filled in by one of parameters $e \in E$. Suppose that;

$$\begin{aligned} F(e_1) &= \{g_2, g_4\} \\ F(e_2) &= \{g_1, g_3\} \\ F(e_3) &= \{g_3, g_4, g_5\} \\ F(e_4) &= \{g_1, g_3, g_5\} \\ F(e_5) &= \{g_1\} \end{aligned}$$

Hence, $F(e_1)$ means expensive shirt whose functional value is the set $\{g_2, g_4\}$. Therefore, soft set (F, E) as a collection of approximates as follows:

$$(F, E) = \left\{ \begin{array}{l} \text{exclusive gown} = \{g_2, g_4\}, \\ \text{striking gown} = \{g_1, g_3\}, \\ \text{vibrant gown} = \{g_3, g_4, g_5\}, \\ \text{inexpensive gown} = \{g_1, g_3, g_5\}, \\ \text{warm gown} = \{g_1\}. \end{array} \right\}$$

Now suppose that
Let

$$\begin{aligned} F(e_1) &= \{y_1/0.2, y_2/0.3, y_3/0.5, y_4/0.5, y_5/0.0\}, \\ F(e_2) &= \{y_1/1.0, y_2/0.6, y_3/0.8, y_4/0.6, y_5/0.0\}, \\ F(e_3) &= \{y_1/0.2, y_2/0.4, y_3/0.4, y_4/0.7, y_5/1.0\}, \\ F(e_4) &= \{y_1/0.3, y_2/1.0, y_3/0.1, y_4/0.3, y_5/0.2\}. \end{aligned}$$

Then a FSS (F, E) represents the family $\{F(e_i); i = 1, 2, 3, 4\}$ of $P(U)$ and the FSS (F, E) can be represented as shown in Table 1.

Table 1. Representation of FSS (F, E)

U/E	e_1	e_2	e_3	e_4
y_1	0.2	1.0	0.2	0.3
y_2	0.3	0.6	0.4	1.0
y_3	0.2	0.8	0.4	0.1
y_4	0.5	0.6	0.7	0.3
y_5	0	0	1.0	0.2

3 Similarity Measure and Distance Measure

A similarity between two entities is one of the measurement models in data grouping and clustering. In this study, the fuzzy soft set were measured based on the normalized Hamming distance [10]. Where, assume that the fuzzy soft set (F, A) and (G, B) have the same parameter set, namely, $A = B$. The normalized Hamming distance and normalize distance in FSS using Eqs. (1) and (2).

$$d_1((F, A), (G, B)) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n |F(e_i)(x_j) - G(e_i)(x_j)| \quad (1)$$

and

$$d_2((F, A), (G, B)) = \frac{1}{mn} \left(\sum_{i=1}^m \sum_{j=1}^n |F(e_i)(x_j) - G(e_i)(x_j)|^2 \right)^{\frac{1}{2}} \quad (2)$$

Example 2 As in [10] let $U = \{u_1, u_2, u_3\}$ be as set with parameters $= \{a_1, a_2, a_3\}$. Given two FSS (G, A) and (H, A) are represented by two tables, Tables 2 and 3.

Table 2. Fuzzy set (G, A)

(G, A)	a_1	a_2	a_3
u_1	0.7	0.8	0.6
u_2	0.6	0.7	0.5
u_3	0.5	0.8	0.8

Table 3. Fuzzy set (H, A)

(H, A)	a_1	a_2	a_3
u_1	0.5	0.6	0.9
u_2	0.7	0.8	0.6
u_3	0.4	0.8	1

From Eqs. (1) and (2), respectively, the distance between (G, A) and (H, A) can be calculated as follows

$$\begin{aligned} d_1((G, A), (H, A)) \\ &= \frac{1}{3 \times 3} \sum_{i=1}^3 \sum_{j=1}^3 (0.2 + 0.1 + 0.1 + 0.2 + 0.1 + 0 + 0.3 + 0.1 + 0.2) \\ &\approx 0.144 \end{aligned}$$

and

$$\begin{aligned} d_2((F, E), (G, E)) \\ &= \frac{1}{3 \times 3} \sum_{i=1}^3 \sum_{j=1}^3 (0.2^2 + 0.1^2 + 0.1^2 + 0.2^2 + 0.1^2 + 0^2 + 0.3^2 + 0.1^2 + 0.2^2)^{\frac{1}{2}} \\ &\approx 0.056 \end{aligned}$$

Feng and Zheng [11] extend Eq. (3) into a generalized normalized distance in FSS by using Eq. (3).

$$d((F, A), (G, B)) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (|F(e_i)(x_j) - G(e_i)(x_j)|^p)^{\frac{1}{p}}, (p \in N_+) \quad (3)$$

Clearly, if $p = 1$, then Eq. (3) is reduced to Eq. (4).

From Eq. (4), it can be know that

$$d' = \frac{1}{n} \sum_{j=1}^n |F(e_i)(x_j) - G(e_i)(x_j)| \quad (4)$$

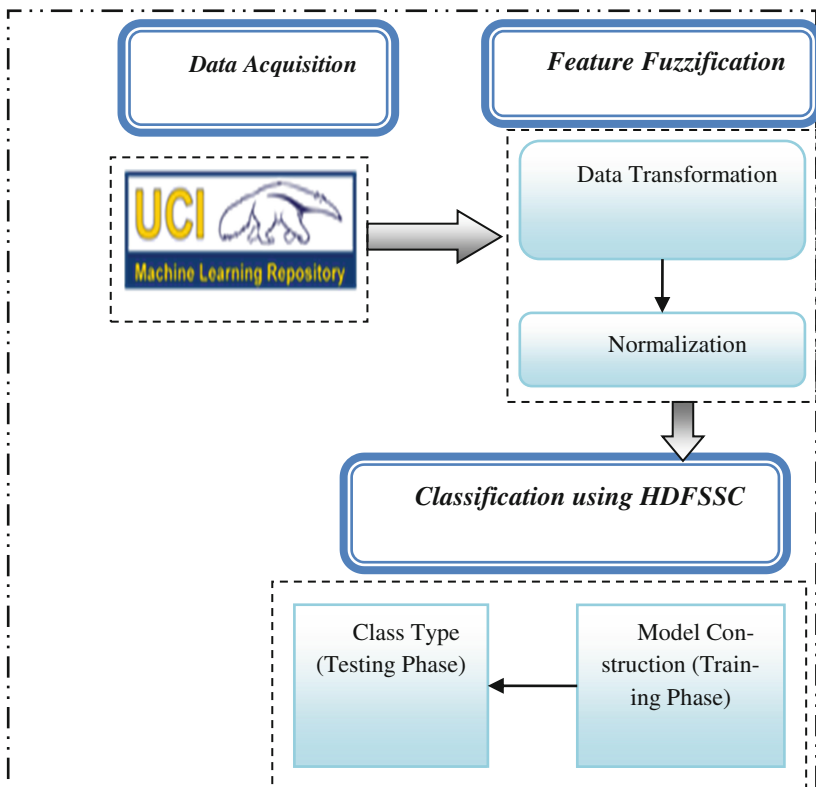
Indicate the distance between the i th parameter of (F, A) and (G, B) , and $d_1((F, A), (G, B))$ indicates that the distance among all parameters of (F, A) and (G, B) .

4 Fuzzy Soft Set Classification Using Hamming Distance (HDFSSC)

As shown in Fig. 1, the proposed modelling comprises of three phases that feature *fuzzyfication*, training phase and testing phase. As, data acquisition is one of the crucial elements to design and develop a successful classifier. Data collections were gathered from University of California at Irvine (UCI) machine learning repository and Mam-mographic Image Analysis Society (MIAS) datasets. Therefore, Table 4 provides description of these dataset.

Table 4. Dataset description

NO	Dataset	Description
1.	BCWO	Breast cancer Wisconsin (original)
2.	SYM8HARD	Sym8 (Hard threshold) Level 1
3.	DB3ROISOFT	Daub3 (Soft threshold) Level 1
4.	DB3SOFT LEVEL4S1	Daub3 (Soft threshold) Level 4
5.	SYM8HARD LEVEL4	Sym8 (Hard threshold) Level 4
6.	DB3SOFT LEVEL4S2	Daub3 (Soft threshold) Level 4
7.	SYM8HARD LEVEL8	Sym8 (Hard threshold) Level 4
8.	DB3HARD LEVEL8S2	Daub3 (Hard threshold) Level 8

**Fig. 1.** Proposed modelling for HDFSSC

The HDFSSC algorithm is divided into three phases. The first is feature *fuzzification*. It is to obtain a feature vector for all data including training and testing dataset. The second is training, which to obtain a fuzzy soft set model for each class. The last is classification, which is to label the unknown data to the target class. The algorithm is shown in the Fig. 2.

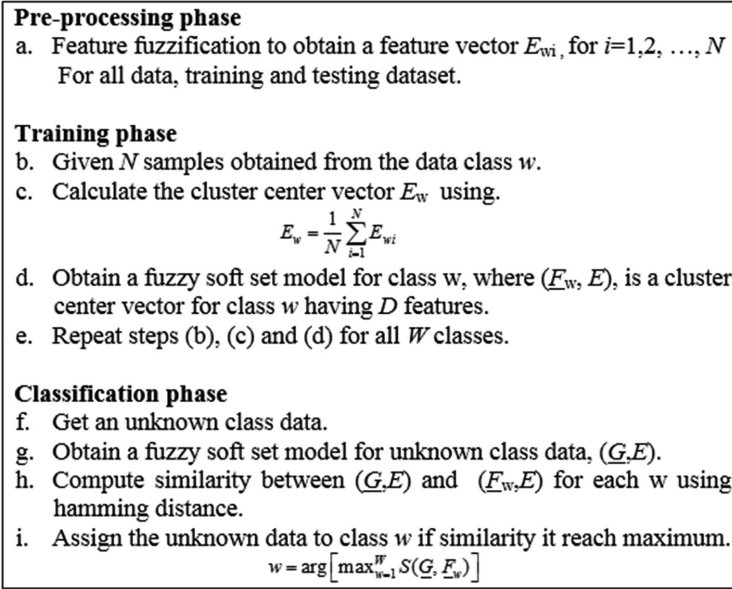


Fig. 2. Classification using HDFSSC

5 Experiment Results

The proposed method is compared to baseline algorithms which have been implemented in MATLAB version 8.6.0.267246 (R2015b). The Algorithms were executed on a processor Intel @1.5 GHz (4CPUs) with 2G total memory using Windows 7 Professional 32 bit operating system sequentially. The 80% sample of data is reserved randomly for training and 20% for testing purpose. Table 5 presents the accuracy of proposed method for classification on UCI benchmark datasets. The experimental results strongly suggest that HFSSC has high accuracy even had lower complexity in the testing phase. Figure 3 illustrate the time response of BCWO data set, for the

Table 5. The comparisons in term of accuracy

Dataset	FSSC	FussCyier	HDFSSC
BCWO	0.9258	0.9439	0.9578
SYM8HARD	0.5918	0.6563	0.6603
DB3ROISOFT	0.6485	0.7002	0.7695
DB3SOFT LEVEL4S1	0.7458	0.7976	0.8181
SYM8HARD LEVEL4	0.6859	0.6292	0.7035
DB3SOFT LEVEL4S2	0.6413	0.6345	0.6765
SYM8HARD LEVEL8	0.7295	0.6750	0.7504
DB3HARD LEVEL8S2	0.7283	0.6679	0.7598

HDFSSC can reduce the time response up to 13.09% and 72.53% comparing to the *FussCyier* and FSSC, respectively.

Figure 4 illustrate the data set number 2–8. Based on the Fig. 4, the HDFSSC can reduce the time response up to 4.91% and 61.72% in average comparing to the *FussCyier* and FSSC, respectively

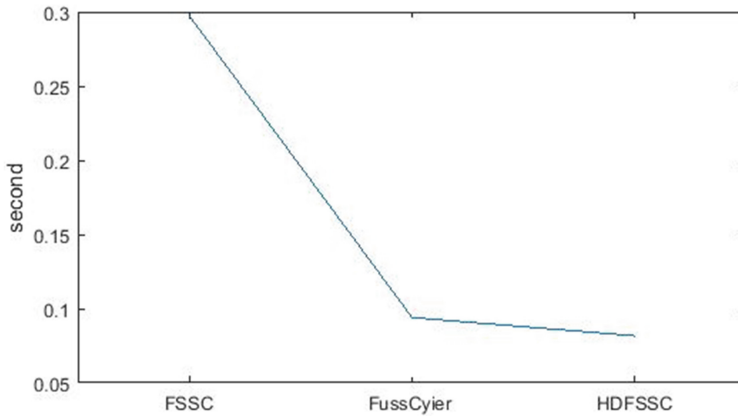


Fig. 3. Time response of BCWO

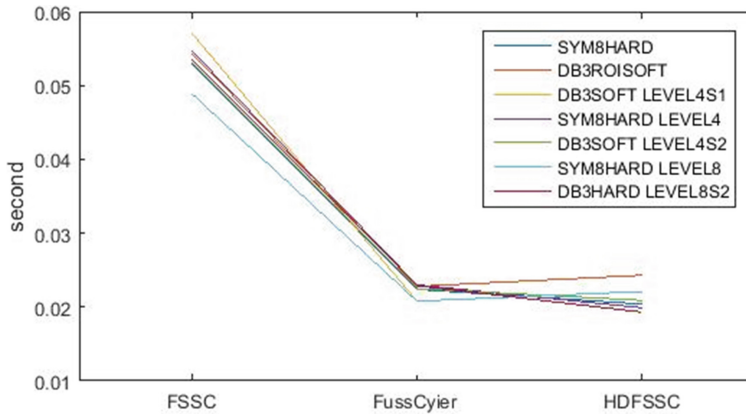


Fig. 4. Time response of dataset number 2–8

6 Conclusion

In this paper, a frameworks of hamming distance approach have been proposed, in order to obtain a balanced solution of a FSS based decision making problem. Moreover, we have justified and compared with the existing FSS classification methods i.e.,


FussCyier and FSSC by using medical datasets. In other words, we confidence that these theories have a lot of future and may serve to solve many decision-making problems.

In the current implementation of this research, the experimental setup involved UCI benchmark datasets, however, for the future works mammogram images classification will be taken into consideration which can be improved by incorporating feature selection phase before classification and the classification results can be more compact and precise.

References

1. Lashari, S.A., Ibrahim, R., Senan, N., Yanto, I.T.R., Herawan, T.: Application of wavelet de-noising filters in mammogram images classification using fuzzy soft set. In: 2016 International Conference on Soft Computing and Data Mining, pp. 529–537 (2016)
2. Yanto, I.T.R., Ismail, M.A., Herawan, T.: A modified fuzzy k-partition based on indiscernibility relation for categorical data clustering. *Eng. Appl. Artif. Intell.* **53**, 41–52 (2016)
3. Purnawansyah, Haviluddin: K-Means clustering implementation in network traffic activities. In: 2016 International Conference on Computational Intelligence and Cybernetics, Makassar, Indonesia, pp. 51–54 (2016)
4. Beniwal, S., Arora, J.: Classification and feature selection techniques in data mining. *Int. J. Eng. Res. Technol.* **1**(6) (2012)
5. Molodtsov, D.: Soft set theory—first results. *Comput. Math. Appl.* **37**(4–5), 19–31 (1999)
6. Maji, P.K., Roy, A.R., Biswas, R.: An application of soft sets in a decision making problem. *Comput. Math. Appl.* **44**(8–9), 1077–1083 (2002)
7. Mushrif, M., Sengupta, S., Ray, A.: Texture classification using a novel, soft-set theory based classification algorithm. In: Computer Vision—ACCV 2006, pp. 246–254 (2006)
8. Roy, A.R., Maji, P.K.: A fuzzy soft set theoretic approach to decision making problems. *J. Comput. Appl. Math.* **203**(2), 412–418 (2007)
9. Kharal, A.: Distance and similarity measures for soft sets. *New Math. Nat. Comput.* **6**(3), 321–334 (2010)
10. Handaga, B., Herawan, T., Deris, M.M.: FSSC: an algorithm for classifying numerical data using fuzzy soft set theory. *Int. J. Fuzzy Syst. Appl.* **2**(4), 29–46 (2012)
11. Feng, Q., Zheng, W.: New similarity measures of fuzzy soft sets based on distance measures. *Ann. Fuzzy Math. Inf.* **7**(4), 669–686 (2014)

Is SVM+FS Better to Satisfy Decision by Majority?

Yao Lin() , Kohei Yamaguchi, Tsunenori Mine, and Sachio Hirokawa

Department of Advanced Information Technology, Kyushu University, 744 Motooka,
Nishi-ku, Fukuoka 819-0395, Japan

lynn901230@ma.ait.kyushu-u.ac.jp, mine@ait.kyushu-u.ac.jp

Abstract. Government 2.0 activities have become very attractive and popular. Using the platforms to support the activities, anyone can any-time report issues in a city on the Web and share the reports with other people. Since a variety of reports are posted, officials in the city management section have to give priorities to the reports. However, it is not easy task to judge the importance of the reports since importance judgments vary depending on the officials and consequently the agreement rate becomes low. To remedy the low agreement rate problem of human judgment, it is necessary to create an automatic method to find reports with high priorities. Hirokawa et al. employed the Support Vector Machine (SVM) with word feature selection method (SVM+FS) to detect signs of danger from posted reports because signs of danger is one of high priority issues to be dealt with. However they did not compare the SVM+FS method with other conventional machine learning methods and it is not clear whether or not the SVM+FS method has better performance than the other methods. This paper compared the results of the SVM+FS method with conventional machine learning methods: SVM, Random Forest, and Naïve Bayse with conventional word vectors, an LDA-based document vector, and word embedding by Word2Vec. Experimental results illustrate the validity and effectiveness of the SVM+FS method.

Keywords: Government 2.0 · Machine learning · Support vector machine

1 Introduction

Government 2.0 is the concept proposed by Tim O'Reilly, which means Government as a platform, where citizens are encouraged to participate government activities¹. Even in Japan, some platforms such as the Chiba citizen

¹ <http://techcrunch.com/2009/09/04/gov-20-its-all-about-the-platform/> Gov 2.0: It's All About The Platform. (2016/7/14).

coordination report (ChibaRepo for short)² have been established and enable the citizen to publish complaints that the citizen holds about the region with the location data and image data related to the complaint on the Web, check the correspondence situation of the administrative side and argue with others about improvement. However, there are lots of problems hard to be solved. For example, delay of taking action by the local government is one of the problems, result from the overload on the government side. Therefore, it is indispensable to reduce the burden on the government side by introducing automatic classification method for complaint report processing supporting features such as automatic or semi-automatic judgment of the urgency of dealing with complaint reports. In this paper, we try to find out a better method of detecting signs of danger in ChibaRepo. The danger is one of urgent issues in the city to be dealt with by the officials with the highest-priority. Actually, lots of researchers have studied about detection of emergency events such as disaster or criminal offense events from micro blogs or information network such as twitter³ or social networking services such as facebook (e.g. [4, 6]). Dealing with the signs of danger with high-priority can prevent citizens from facing accidents or emergency events forecasted by the signs. Hirokawa et al. [5] employed Support Vector Machine (SVM) [7] with word feature selection (SVM+FS for short) [1, 10] to detect the signs of danger and achieved higher performance in detecting the signs of danger compared to four human subjects. However, since they only used SVM+FS in their paper, we can not judge whether or not the problem of detecting signs of danger is a difficult one to be dealt with by machine learning methods in the first place, in other words, whether or not SVM+FS is the best method. In fact, effective machine learning methods change depending on the analysis target, so it is essential to compare with other methods. Therefore, this paper aims to confirm whether or not SVM+FS is truly the best method or if there are other better methods than SVM+FS. To this end, we conducted compare SVM+FS with several machine learning methods such as SVM, Random Forest (RF) [3], and Naïve Bayse (NB) [9] using other feature selections than word feature selection such as part of speech, sentiment polarity words, Latent Dirichlet allocation (LDA) [2] as a topic model, and Word2Vec (W2V) [11] as a word embedding method. Experimental results show the SVM+FS method is the best compared to the conventional machine learning methods employed in this paper. In what follows, Sect. 2 explains data set we used in this paper. Section 3 explains the method we used in the experiments. Section 4 explains the way of experiments conducted, and illustrates the experimental results obtained by the method. Section 5 discusses the experimental results and finally we conclude the paper in Sect. 6.

² <http://chibarepo.force.com/>.

³ <https://www.cnet.com/news/twitters-not-a-social-network/>.

2 Data Set

ChibaRepo is a platform which has three good functions to help the citizens.⁴ On the ChibaRepo Web site, citizens have issued 4069 reports so far; 3903 have already been dealt with, 60 is now being tackled, and 106 are in a waiting list.⁵ Among them, 1874 reports being in CSV format are open to the public.⁶ Table 1 shows the categories in each report data. We have tackled this research since 2015, which was before the operational secretariat of ChibaRepo gave the data to the public; we obtained our data by crawling ChibaRepo in 2015. Hirokawa et al. [5] used the data that did not include some latest reports in the ChibaRepo open data mentioned above. The number of reports in our data is 656. Hirokawa et al. [5] asked four subjects to read the report in the data and to put a mark on a report if signs of danger were included in the report. 36% reports in the whole were judged as reports including signs of danger ('danger report' for short), by at least one subject. Table 2 shows the numbers of danger reports and their percentages in the whole judged by N subjects. They defined a report as a positive example if the report was judged as danger one by N or more subjects. In this paper, we also used this data to conduct experiments. We apply several machine learning methods to the reports, build models to judge if a report is an

Table 1. Data categories of ChibaRepo open data

Id	Identification
Name	Name label denoted from the issued date of a report
Address_c	Address
CBC_M_Sections_c	Section related to address
CBC_M_WebUser_c	User ID
Category_c	Category {Road, Park, Garbage, Others}
Comment_c	Text message in a report
CompleteDate_c	Date of completing the issue
CopeImageId_c	Photo image
ImageId_c	Photo image
LatitudeWG84_c	latitude
LongitudeWGS84_c	longitude
ReportDateTime_c	Date when a report was issued
Status_c	State of correspondence
Subject_c	Report subject, Title
VideoURL_c	URL of video

⁴ https://chibarepo.secure.force.com/CBC_VF_WebBasicPhilosophy.

⁵ <https://chibarepo.secure.force.com/> confirmed on the 20th July 2017.

⁶ The last date of the reports is the 27th of February, 2016. confirmed on the 20th July 2017.

Table 2. Percentage of danger reports judged by N subjects [5]

N	Count	Percentage	Comments
1	235	0.36	At least one subject judged as danger report
2	111	0.17	Two subjects judged as danger report
3	66	0.10	Three subjects judged as danger report
4	22	0.03	All the subjects judged as danger report

danger report and compare the discrimination performance of the models with the SVM+FS method. Especially, focusing on the decision by majority, we only show the experimental results on the cases of $N \geq 3$ and $N \geq 4$.

3 Experimental Methods

3.1 Settings

Hirokawa et al. [5] only selected five categories: Comment.c, Subject.c, Status.c, Category.c, and CBC_M_Sections.c. However, they did not evaluate the effect of the categories. Then, we evaluate the effects of the categories compared to only using Comment.c. We set three cases so as to evaluate the effects of the categories. In case 1, we only use the Comment.c data to create input vector. In case 2, we use the five category data as well as Hirokawa et al. [5]. In case 3, we use sentiment polarity word tags from the Japanese Sentiment Polarity Dictionary⁷ in addition to the five category data used in case 2. In case 1, instead of word feature selection, we use one-hot bag-of-words (BoW for short), Term-Frequency Inverse-Document-Frequency term weighting (TFIDF for short), Word2Vec⁸ (W2V), and LDA when creating report vectors. Then we apply RF, SVM and NB to the vectors. Only for W2V, to increase the text volume, we use other text data: the four year records of complaint calls from citizens about city parks in Kashiwa City (Kashiwa complaint data for short). We examined three data sets: (1) ChibaRepo data consisting of 656 reports [5], (2) ChibaRepo open data consisting of 1873 reports, and (3) Kashiwa complaint data consisting of 5665 reports. We combined them and created four data sets: (1), (1) + (2), (1) + (3), (1) + (2) + (3). For the combination, we applied three machine learning methods: SVM, RF, and NB. Through experiments, we confirmed that as the volume of the data set increases, the average F-measure score becomes higher. Due to the limitation of spaces, we only show the best score of each method using the data set (1) + (2) + (3) in $N \geq 3$ and $N \geq 4$. We used Scikit-learn library to determine parameter values of each machine learning classifier. Since the numbers of positive and negative examples are dis-balanced,

⁷ <http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FJapanese%20Sentiment%20Polarity%20Dictionary>.

⁸ <https://en.wikipedia.org/wiki/Word2vec>.

we took `class_weight` = “balanced” option for RF and SVM and default values for other parameters. For NB, we took GaussianNB and used default parameters for the rest. Although Hirokawa et al. [5] did not use parts of speech information of words in the data set and not remove stop words either, we used parts of speech features, which automatically removes stop words. In case 2, we use BoW and *SVM^{light}*⁹[8] with the default parameter settings, which is the same as Hirokawa et al. [5]. Furthermore, as well as Hirokawa et al. [5], we denoted the five categories by “cOntent,” “Title,” corresponDing state,” “Genre,” and “Region,” and used one character in each category as a tag to distinguish the same word, say ‘danger,’ appearing in different categories, such as o:danger in “content” and t:danger in “title.” In case 3, tags of “Negative” and “Positive” were added to the words in case 2. For example, if the danger, which is a negative word by the Japanese Sentiment Polarity Dictionary, is appeared in “Title” category, we add tag “NT”¹⁰ to “danger” and describe “NT:danger”. We also use BoW and *SVM^{light}* with the default parameter settings. The reason why we do not use *SVM^{light}* in case 1 is because when we conducted experiments by *SVM^{light}*, all the test data were predicted as negative. Consequently, the variations of the vector’s dimensional feature values made by the methods are too large.

3.2 Creation of Input Vector

We describe the procedure of our experiments below.

Morphological Analysis: Since words in a Japanese sentence are not separated with each other, first, we perform morphological analysis using a Japanese morphological analyzer MeCab¹¹ and extract words and their parts of speech from sentences in the data set. Second, we normalize sentences by several 2-byte characters such as alphanumeric characters, signs and spaces into 1-byte characters and 1-byte Japanese katakana characters into 2-byte characters, and deleting spaces and line breaks before and after the sentence. Then we extract words with some specific parts of speech: adjectives, auxiliary verbs, verbs, and nouns, and reconstruct the sentences using the extracted words. In **case 1**, we use the original form word in constructing input vector by BoW, TFIDF, W2V and LDA for each. In **case 2**, we add the category tag in front of the word and make input vectors by BoW. In **case 3**, we use the Japanese Sentiment Polarity Dictionary to judge a word’s sentiment polarity: Positive or Negative, and add the tag of sentiment polarity in front of the word, and make input vectors by BoW.

Input Vectors by BoW: We vectorize a report d_i as a vector $vec(d_i) = (s_1, \dots, s_M)$, where M is the total number of words that appear in the reports and $s_i = 1$ if w_i appears in d_i , 0 otherwise.

Input Vectors by TFIDF: We calculate the TFIDF value of word w_i in a report d_i as s_i and vectorize the report as a vector $vec(d_i) = (s_1, \dots, s_M)$.

⁹ <http://svmlight.joachims.org>.

¹⁰ N means Negative and T means Title.

¹¹ <http://taku910.github.io/mecab/> (in Japanese).

Input Vectors by W2V: In this research, we used the W2V library offered by gensim¹² to make a model from the sentences consisting of separated words, using the default parameter of 100 dimensionality. From the model, the vector of a word w_i can be represented as $w_i = (v_{i1}, \dots, v_{i100})$. v_i is the feature value calculated by the W2V model. The vector of a report d_i can be represented as the average of the word vectors in d_i as $d_i = \sum_{i=0}^n (w_i/n)$. d_i means a report vector, while n means the number of words which appeared in the report.

Input Vectors by LDA: We used LDA to perform the dimension compression; the number of dimensions before compression was 2483 and 200 after compression. The vector of a report d_i can be represented as $vec(d_i) = (s_1, \dots, s_{200})$. s_i means the feature value.

4 Results of Experiments

To compare the results the SVM+FS method with other machine learning methods, we conducted experiments in 3 cases. The results of the SVM+FS method [5] are shown in Table 3. In the experiment, we performed 10-fold cross validation for ten times and took the average.

4.1 Results in Case 1

In case 1, when we only used the Comment_c data, the average F-measure value of SVM with TFIDF is the best in $N \geq 3$, but that of SVM with BoW is the best in $N \geq 4$. The differences between SVM with TFIDF and SVM with Bow in $N \geq 3$ and in $N \geq 4$ are less than 0.02 and greater than 0.04, respectively. Then we decided that SVM with BoW is the best among 12 methods when considering both cases of $N \geq 3$ and $N \geq 4$. However, the results of the SVM+FS method outperformed those of SVM with BoW in both $N \geq 3$ and $N \geq 4$. For W2V, although we added another data: the four year records of complaint calls from

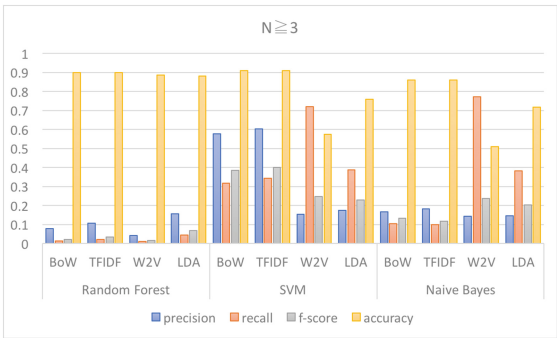


Fig. 1. Result in case 1: $N \geq 3$

¹² <https://radimrehurek.com/gensim/models/word2vec.html>.

Table 3. Prediction performance by SVM and feature selection compared with by human [5]

N	1	2	3	4
Precision	0.8829	0.6429	0.6075	0.4847
(Human)	1.0000	0.7916	0.5993	0.2471
Recall	0.8913	0.7833	0.8434	0.8052
(Human)	0.4617	0.6982	0.8333	1.0000
F-measure	0.8712	0.7008	0.7001	0.5887
(Human)	0.6002	0.7054	0.6671	0.3875
Accuracy	0.8094	0.8590	0.8950	0.8845
(Human)	0.8072	0.9017	0.9017	0.8681

Table 4. Result in case1

N	Classifier	Vectorizer	Precision	Recall	f-score	Accuracy
$N \geq 3$	RF	BoW	0.0783	0.0133	0.0219	0.898
		TFIDF	0.1075	0.0222	0.0356	0.899
		W2V	0.1567	0.0322	0.0515	0.895
		LDA	0.1557	0.0455	0.0669	0.882
	SVM	BoW	0.578	0.318	0.385	0.909
		TFIDF	0.6025	0.3431	0.4013	0.91
		W2V	0	0	0	0
		LDA	0.174	0.388	0.23	0.759
	NB	BoW	0.1659	0.1033	0.113	0.859
		TFIDF	0.1832	0.1003	0.117	0.859
		W2V	0.1603	0.868	0.2664	0.5296
		LDA	0.1469	0.383	0.204	0.716
$N \geq 4$	RF	BoW	0	0	0	0.966
		TFIDF	0	0	0	0.966
		W2V	0	0	0	0.966
		LDA	0	0	0	0.966
	SVM	BoW	0.11	0.0619	0.0751	0.965
		TFIDF	0.05	0.027	0.0334	0.965
		W2V	0	0	0	0
		LDA	0.0417	0.1569	0.0605	0.851
	NB	BoW	0.045	0.0258	0.0302	0.958
		TFIDF	0.0533	0.0324	0.038	0.957
		W2V	0.0539	0.744	0.099	0.525
		LDA	0.0696	0.175	0.0898	0.861

citizens about city parks in Kashiwa city, to increase the data volume, the results of the three methods with W2V were worse than those of SVM with BoW. Experimental results in $N \geq 3$ and $N \geq 4$ are shown in Table 4 and Figs. 1 and 2. Although we conducted grid search on SVM and RF, the experimental results were almost the same as those with default parameters. Figure 3 shows the best parameters by grid search.

4.2 Results in Case 2

In case2, we used the five categories: ChibaRepo’s Comment_c, Subject_c, Status_c, Category_c, and CBC_M_Sections_c’s. We only show the results obtained by SVM^{light} with BoW in Table 5. because the method was the best among all the methods employed in case 1. We used default parameter settings for SVM^{light} as Hirokawa et al. [5] did. As shown in Table 5, the average F-measure values of SVM^{light} with BoW in $N \geq 3$ and $N \geq 4$ are 0.508 and 0.268.

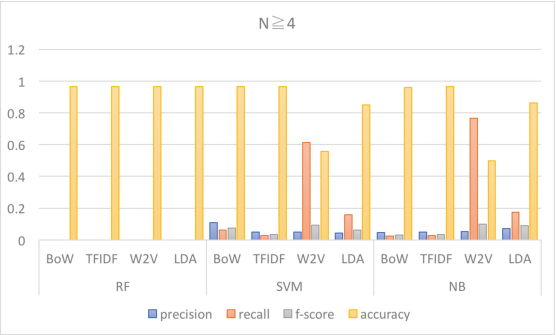


Fig. 2. Result in case1: $N \geq 4$

Table 5. Result in case 2 and case 3 by BoW and SVM^{light}

Case 2				
N	Precision	Recall	f-score	Accuracy
$N \geq 3$	0.359	0.937	0.508	0.823
$N \geq 4$	0.2	0.5	0.268	0.908
Case 3				
N	Precision	Recall	f-score	Accuracy
$N \geq 3$	0.341	0.891	0.485	0.814
$N \geq 4$	0.23	0.55	0.298	0.911

Table 6. Comparison of prediction F-measure results among SVM+FS, SVM+BoW, and human subjects

N	3	4
SVM+FS	0.700	0.589
SVM+BoW	0.508	0.298
Human	0.667	0.388

```

RandomForestClassifier
(bootstrap=True, class_weight='balanced',
criterion='gini', max_depth=None, max_features='auto',
max_leaf_nodes=None, min_impurity_split=1e-07,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1,
oob_score=False, random_state=None, verbose=0,
warm_start=False)

SVC
(C=10, cache_size=200, class_weight='balanced', coef0=0.0,
decision_function_shape=None, degree=3, gamma=0.001, kernel='rbf',
max_iter=-1, probability=False, random_state=None, shrinking=True,
tol=0.001, verbose=False)

```

Fig. 3. Parameters of SVM and RF

4.3 Results in Case 3

In case 3, we used sentiment polarity word tags: “Negative” and “Positive” from the Japanese Sentiment Polarity Dictionary in addition to the five category data in case 2. We used SVM^{light} with default parameter settings and BoW because this method was the best among 12 methods employed in case 1. Compared to the results in case 2, the average F-measure value in $N \geq 3$ became 0.023 worse, but that in $N \geq 4$ became 0.03 better. Results are shown in Table 5.

5 Discussion

In case 1, using vectors made by BoW, TFIDF, W2V and LDA, we built models by RF, SVM, and NB and conducted experiments. When creating vectors with W2V, we combined three kinds of data: ChibaRepo data used in [5], ChibaRepo open data, and Kashiwa complaint data. The best F-measure values 0.4013 in $N \geq 3$ and 0.0985 in $N \geq 4$ were taken by SVM+TFIDF and by NB with W2V, respectively. However, both are lower than the results by the SVM+FS method which were 0.7001 and 0.5887 in $N \geq 3$ and $N \geq 4$, respectively [5]. We found that the F-measure value became higher as the volume of data used for W2V increased. Therefore, we may be able to improve the F-measure results of machine learning methods with W2V by using big volume data. This is our future work. In case 2, we used the five category data in [5] and compared the results with those in case 1. The F-measure values in $N \geq 3$ and $N \geq 4$ are 0.508 and 0.268, obtained by SVM+BoW. However, the both results were lower than

those of SVM+FS: 0.700 and 0.589. In case 3, we used the data tagged with five categories in case 2 and sentiment polarity from the Japanese Sentiment Polarity Dictionary. The F-measure values in $N \geq 3$ and $N \geq 4$ are 0.485 and 0.298, respectively. Compared to the results in case 2, the F-measure value in $N \geq 3$ became 0.02 lower, but that in $N \geq 4$ became 0.03 higher. With the results in three cases, we found that the average F-measure values improved by using tagged data. As our future work, we will find other tagged data to improve the F-measure, especially for W2V. However, the results of SVM+FS were the best in $N \geq 3$ and $N \geq 4$, and the results by human were also better than the conventional methods used in this paper as shown in Table 6. Therefore the problem of detecting signs of danger to satisfy decision by majority is not easy.

6 Conclusion

In this paper, we conducted comprehensive experiments to compare the SVM+FS method with several machine learning methods such as SVM, RF, and NB using other feature selections than word feature selection such as parts of speech, sentiment polarity words, LDA as a topic model, Word2Vec as a word embedding method. We conducted experiments considering three cases to evaluate the effects of five categories and of the sentiment polarity words. In case 1, we only used the Comment_c data to create input vector by BoW, TFIDF, Word2Vec and LDA. Then we applied RF, SVM and NB to the vectors. We confirmed the effects of tagged words of five categories and sentiment polarity information only in $N \geq 4$. We also found that increasing data volume for Word2Vec effects to improve the F-measure value. However, all the results obtained by the conventional methods employed in the paper were worse than those by human. Therefore we confirmed that the problem of detecting signs of danger to satisfy decision by majority is not easy task. In addition, all the experimental results obtained in three cases illustrate the superiority of the SVM+FS method.

Acknowledgement. This work was partially supported by JSPS KAKENHI Grant No. JP15H05708, JP16H02926, and JP17H01843.

References

1. Adachi, Y., Onimura, N., Yamashita, T., Hirokawa, S.: Standard measure and SVM measure for feature selection and their performance effect for text classification. In: Proceedings of the 18th iiWAS2016, pp. 262–266. ACM (2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1002 (2003)
3. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
4. Cresci, S., Cimino, A., Dell’Orletta, F., Tesconi, M.: Crisis mapping during natural disasters via text analysis of social media messages. In: International Conference on Web Information Systems Engineering, pp. 250–258 (2015)
5. Hirokawa, S., Suzuki, T., Mine, T.: Machine learning is better than human to satisfy decision by majority. In: WI’17. IEEE/ACM (2017)

6. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: a survey. *ACM Comput. Surv. (CSUR)* **47**(4), 67 (2015)
7. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: *ECML 1998: Machine Learning: ECML-98*, pp. 137–142 (1998)
8. Joachims, T.: Learning to classify text using support vector machines. Dissertation, Kluwer (2002)
9. McCallum, A., Nigam, K., others.: A comparison of event models for naive bayes text classification. In: *AAAI-98 Workshop on Learning for Text Categorization*, vol. 752, pp. 41–48 (1998)
10. Sakai, T., Hirokawa, S.: Feature words that classify problem sentence in scientific article. In: *Proceedings of the 14th iiWAS2012*. pp. 360–367. ACM (2012)
11. Tomáš, M.: Statistical language models based on neural networks. Ph.D. thesis, Brno University of Technology (2012)

Exploiting LabVIEW FPGA in Implementation of Real-Time Sensor Data Acquisition for Rowing Monitoring System

Zarina Tukiran^{1,2(✉)} and Afandi Ahmad^{1,2(✉)}

¹ Department of Computer Engineering, Faculty of Electrical & Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

{zarin, afandia}@gmail.com

² Reconfigurable Computing for Analytics Acceleration (ReCAA) Research Laboratory, Microelectronics and Nanotechnology—Shamsuddin Research Centre (MiNT-SRC), Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

Abstract. Field Programmable Gate Arrays (FPGAs) platform has been increasingly used in sensor-based applications because of reconfigurable and parallelisms features offered in the FPGA. However, most of the application designers are unfamiliar with hardware programming and design concepts of the FPGA. This paper presents an implementation of real-time sensor data acquisition (ReSDAq) for rowing monitoring system using LabVIEW FPGA which utilising the high-level synthesis (HLS) technique. The HLS allows application designers to use high-level language for configuring the FPGA. The ReSDAq application comprises of a tri-axis accelerometer sensor, an LCD monitor, and a National Instrument (NI) sbRIO-9632 board. The sbRIO-9632 board was targeted programmed on the Xilinx FPGA core to acquire sensor data and compute acceleration of the arm movement of the rower. From this study, it was found that the compilation time to convert G-code into hardware description language (HDL) code depends on the size of the code. Apart from having an interesting experience in graphical programming approach, the LabVIEW FPGA module could be used by application designers to facilitate and accelerate the development of FPGA-based systems.

Keywords: LabVIEW · FPGA · Rowing · Data acquisition · Sensor

1 Introduction

Rowing is a strenuous on-water sport that requires excellent physical performance and proficient rowing technique of a rower to increase the boat velocity in order to reach the targeted destination in the shortest time [1]. Training is a compulsory aspect to adapt the human body with loads in obtaining high performance result in sports. Indoor rowing employing rowing ergometer is one of the methods that enable physical development among the rowers. Rowing ergometer allows rowers to imitate the movement of on-water rowing before the actual rowing on the water. Even though the

rowing ergometer has a small system to record the performance of a rower, however, the monitoring of rowing technique of a rower can only be done with human intervention via the observation of coaches or another athletes themselves. There is no system site available to assist the coaches or athletes to review any progress that the rowers have made throughout subsequent training sessions. Consequently, the physical condition and performance of the athletes are difficult to be monitored systematically. According to Bernstein et al. [2], rowers are exposed to 10-fold higher risk of injury per hour during land-based training than the on-water based training. The usage of rowing ergometer during the indoor rowing training is one of the leading causes. Therefore, if the training outcomes of an athlete on rowing ergometer can be captured, recorded and revised from time to time by their coaches, biomechanists and the athletes themselves, the injury of the athlete in training could be reduced and both physical condition and performance of the athlete could be enhanced.

Motion capture (MoCap) is a system that captures the process of actual motion events to provide valuable and precise movement of data that magnifies the human visual sense through MoCap sensing technology [3, 4]. MoCap has attracted various research areas including the athletic training [5–7]. As shown in Fig. 1, MoCap comprises of three (3) main parts; sensing the motion, processing the sensor data, and storing the processed data [8]. The first two parts are referred as motion tracking. The tracking data can be used in two ways, either for real-time interactive applications or stored for future reference.

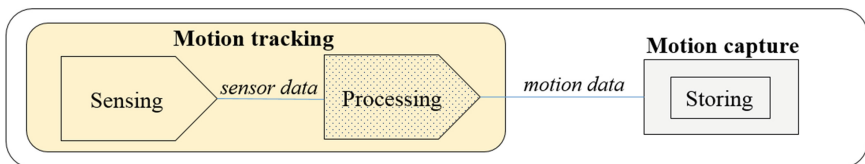


Fig. 1. Motion capture basics [8]

By referring to Fig. 1, processing unit is frequently performed by Digital Signal Processor (DSP) or microcontroller [9, 10]. In the market, around 50% of the sensors are analog-type sensors [11] which explicitly explains, for instance, the need for converting an analog signal into a digital signal as one of the common approaches for data acquisition from the sensors. In order to complete the task, the microcontroller executes the task using a sequential processing technique. However, this technique can result in an increased propagation delay which will affect the overall performance of the system.

A field-programmable gate array (FPGA) is another approach that has been increasingly used as processing unit in sensor-based applications [12–14]. The reason is that the FPGA provides reconfigurable sensor system [15] and can be configured to perform parallel processing [14]. However, most of application domain experts are not acquainted with hardware programming and design. In order for them to understand this foreign concept, it requires steep learning curves. Therefore, Laboratory Virtual Instrument Engineering Workbench Field Programmable Gate Arrays (LabVIEW

FPGA) module was suggested in [16–18] to be used for configuring the FPGA with less burden without sacrificing the benefits that could be gained from the hardware implementations in reconfigurable devices. Therefore, the purpose of this paper is to describe an implementation of data acquisition from on-body sensors in real-time using LabVIEW FPGA on Xilinx FPGA platform for rowing monitoring system.

The rest of the paper is organised as follows. Section 2 describes LabVIEW with the focus of its FPGA module. Section 3 presents the implementation of real-time sensor data acquisition (ReSDAq) for rowing monitoring system using LabVIEW FPGA. The results are presented in Sect. 4. Finally, conclusion and future work are provided in Sect. 5.

2 LabVIEW FPGA

An FPGA is a programmable silicon chip that comprises of configurable logic blocks (CLBs), I/O resources, and programmable routing circuitry. When an FPGA is configured, programmable routing circuitry is wired together the I/O blocks, CLBs, and memory resources to create a hardware implementation of the software application [16] (Fig. 2).

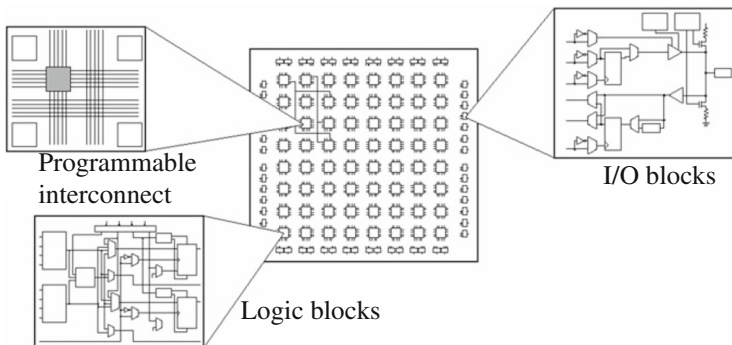


Fig. 2. Inside an FPGA chip [16]

Xilinx is one of the leading FPGAs manufacturer besides Altera, Lattice Semiconductor, Microsemi, and QuickLogic. Xilinx FPGA devices have a number of families such as Spartan, Virtex, Kintex, and Artix.

Typically, FPGAs are configured using a hardware description language (HDL). Currently, there are two HDLs dominated their use in industry; Very High Speed Integrated Circuit Hardware Description Language (VHDL), and Verilog. High-level synthesis (HLS) is another technique that allows domain experts to use a high-level language for configuring the FPGA [19]. One of the tools that provide the HLS technique is LabVIEW software [20].

National Instruments Corporation (NI) developed the LabVIEW software in the mid-1980s [18]. LabVIEW software uses a graphical programming language, namely

G which is similar to the data flow model of computation [18]. Front panel (FP) and block diagram (BD) are two (2) main sections of LabVIEW programming. The FP provides an area to create graphical user interface (GUI) using a control palette. The BD is an area to program in G-code using function palette. By establishing the relations between FP and BD, an application is developed and termed as Virtual Instrument (VI). An example of FP and BD to flash onboard LED is shown in Fig. 3.

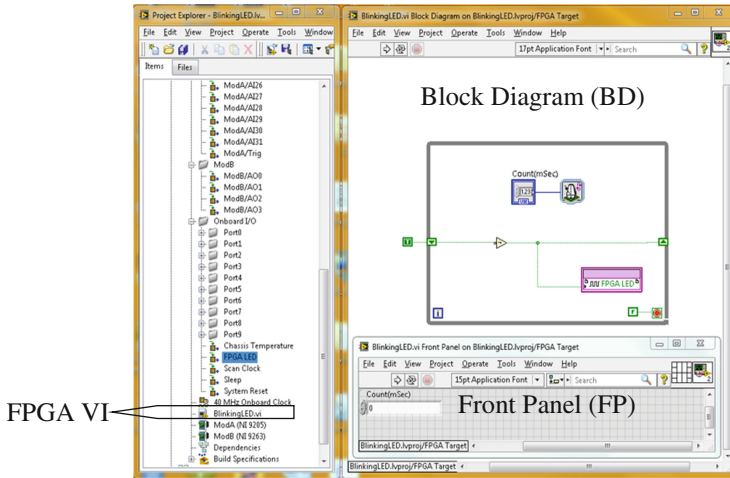


Fig. 3. The FP, and BD for flashing the on-board LED application

In LabVIEW, HLS technique can be implemented by using LabVIEW FPGA module in order to configure the FPGA devices. The LabVIEW FPGA module is divided into two (2) VIs applications; the host VI and the FPGA VI. The host VI can be executed either on a computer desktop or a real-time system. FPGA VI, on the other hand, runs on targeted FPGA. Typically, the host VI is suitable for processing, logging, and analysing data on the host. Besides that, by using a host VI, the process of testing and simulation of FPGA code can be executed before implementing it on the targeted FPGA. The FPGA VI is a VI targeted to configure the behaviour of the FPGA to match the specific requirements of a system. When the created FPGA VI is downloaded into the FPGA, the functionality of the FPGA device is programmed with custom timing, triggering, and I/O solutions. It is important to note that the host VI and its FPGA VI do not support the same VIs. Likewise, the FPGA can only support fixed-point operations.

3 Proposed Design

The proposed design of real-time sensor data acquisition (ReSDAq) comprises of an ADXL 335, an NI sbRIO-9632 board, and an LCD monitor (Fig. 4). The ADXL 335 is a tri-axis accelerometer sensor [21] that is used to obtain raw data of arm movement of

a rower. The NI sbRIO-9632 provides powerful features such as 2 M gate Xilinx Spartan FPGA, 110 3.3 V (5 V tolerant/TTL compatible) digital I/O lines, 400 MHz industrial processors, 32 single-ended/16 differential 16-bit analog input channels at 250 kS/s, and four 16-bit analog output channels at 100 kS/s [22]. The LCD is an output device to display the raw sensor data on the screen. The raw data can be stored for further analysis.

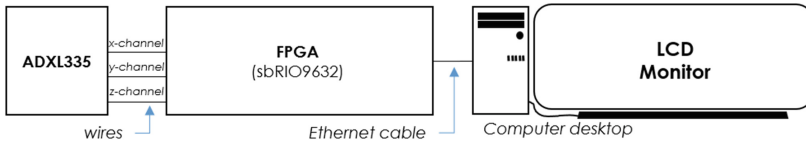


Fig. 4. Block diagram of ReSDAq application

The proposed design of ReSDAq was designed and implemented in two phases; hardware and software. The hardware phase involved the configuring physical connection between ADXL335, FPGA sbRIO-9632, and computer desktop. This physical connection is needed for streaming sensor data of ADXL335 to computer desktop via FPGA sbRIO-9632 board. The software phase involved programming of two main tasks; (i) acquiring sensor data from ADXL335 via FPGA sbRIO-9632, and (ii) providing a graphical user interface.

3.1 Hardware Configuration

The hardware configuration of ReSDAq application of this work involved two (2) parts; (i) ADXL335 and FPGA sbRIO-9632, and (ii) FPGA sbRIO-9632 and computer desktop. The latter configuration was performed by using Measurement & Automation (MAX) software [22] which is not included as a research scope of this paper.

The configuration of ADXL335 and FPGA sbRIO-9632 is crucial in order to obtain sensor data of the arm movement of the rower. In this configuration, the ADXL335 was supplied with 5.0 V and grounded to AI GND via sbRIO-9632 connectors. The tri-axis of sensor channels were connected to analog input channels of sbRIO-9632.

3.2 Software Programming

In the design and implementation of ReSDAq application of this paper, LabVIEW2011 Service Pack 1 (LV2011 SP1) was used to program the FPGA device and the application. As shown in Fig. 5, this application is designed using two (2) VIs; an FPGA VI, and a host VI. The FPGA VI contains G-code language for acquiring sensor data via FPGA sbRIO-9632, and the host VI contains G-code for displaying the acquired data on the screen. Programmatic front panel communication is a data transfer method used for sending commands between the FPGA VI and host VI.

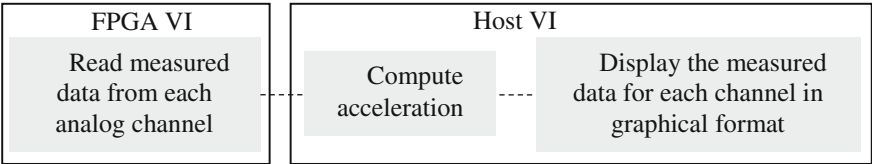


Fig. 5. Flow block diagram for ReSDAq application

Both FPGA VI and host VI are saved and run. The G language programming in FPGA VI is transformed into VHDL code. After the VHDL code is generated, the Xilinx tools will convert this VHDL code in “Bitstream” file [16]. Standard Category-5 (CAT-5) Ethernet cable must be installed properly through RJ-45 port on the sbRIO-9632 board and computer desktop before downloading the generated hardware design into the FPGA board.

The compilation process time varies depending on the size of the code. Hence, to speed up the development time, it is worth to firstly simulate the code by choosing *Simulated of Execution VI* option before performing any compilation and downloading of hardware design into the FPGA board. In this design, LabVIEW took approximately 12 min to convert the G-code of FPGA VI into VHDL code. For every successful compilation, there were eight (8) types of reports generated by LabVIEW to determine whether the design of FPGA VI fit the FPGA and met the timing constraints [23]. Figure 6 shows a report of final device utilisation (map) which indicates the number of FPGA elements that has been compiled by FPGA VI for ReSDAq application.

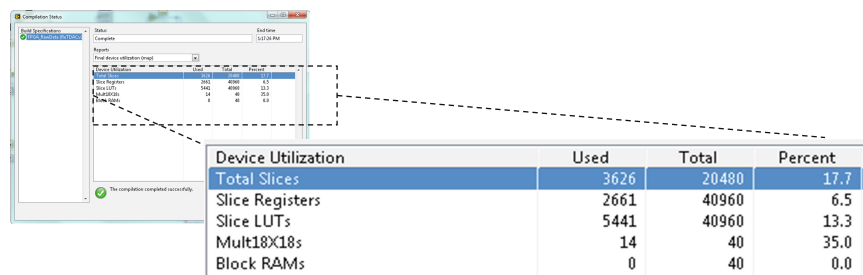


Fig. 6. Report of final device utilisation (map) for ReSDAq application

4 Results

The sensor is mounted on a breadboard, and the use of a goniometer as a reference for angle position to perform sensor measurement are shown in Fig. 7a. The results in Fig. 7b and c show a visual of the raw, and calibrated data of a sensor in a stationary position, respectively. According to [21], when the sensor is in the position as shown in Fig. 7a, the acceleration on Y-axis of the sensor is approximately 0 g (Fig. 7d).

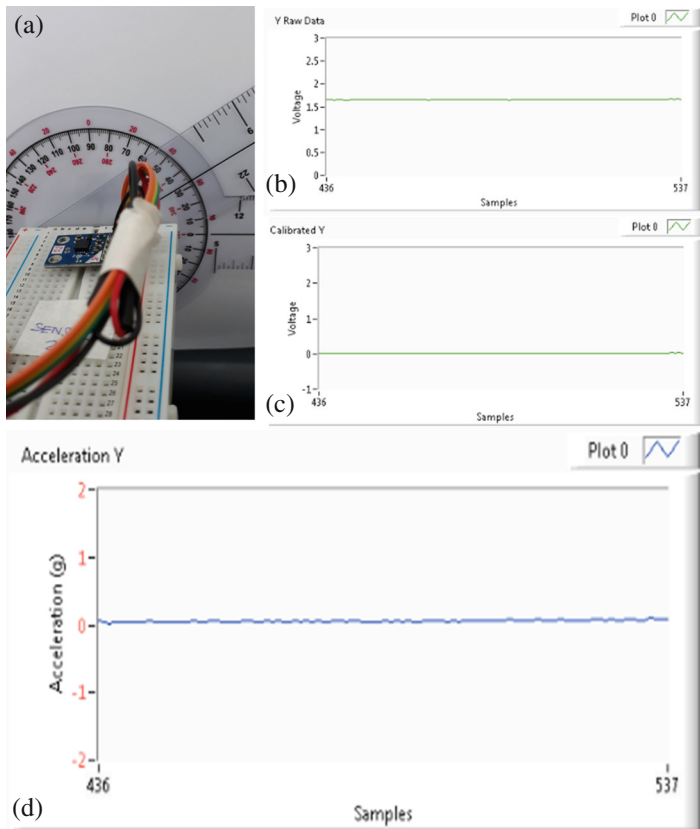


Fig. 7. **a** Sensor in a stationary position, sensor measurement for **b** raw data **c** calibrated data on Y-axis, and **d** acceleration measurement on Y-axis

Figure 8 shows the experimental setup to collect sensor data measurement for rowing movement. The rower is instrumented with the sensor by using a Velcro strap as shown in Fig. 8a. Then, the rower is asked to set his posture at the catch position for calibration and initial measurement as shown in Fig. 8b before performing a motion of sculling style rowing (Fig. 8b and c) at their self-selected race pace on an ergometer for 5 min. Figure 9 shows the results of the forearm motion in four (4) cycles of rowing.

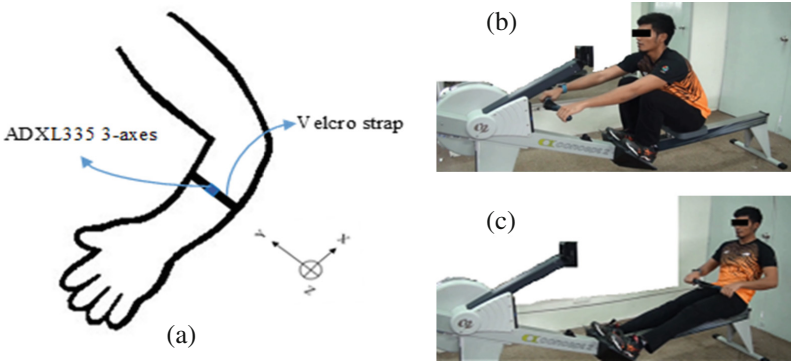


Fig. 8. Experimental setup—**a** sensor placement on arm, **b** arm position at catch position, and **c** arm position at finish position

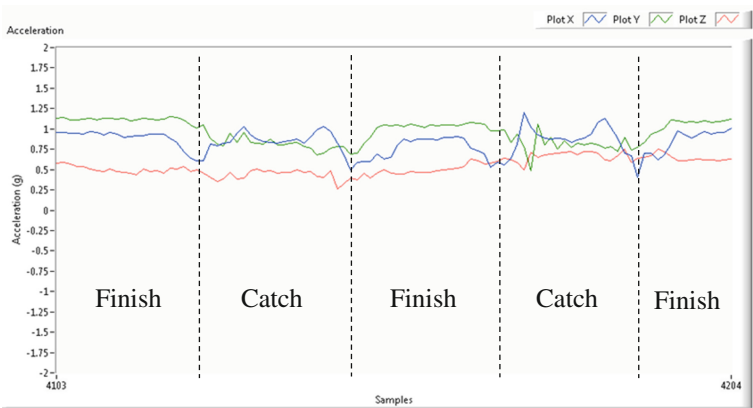


Fig. 9. Acceleration of forearm movement from catch to finish positions at x-axis, y-axis, and z-axis of the sensor

5 Conclusion

In conclusion, this paper discusses the use of LabVIEW FPGA in design and implementation of the proposed real-time sensor data acquisition for rowing monitoring system. This proposed system was developed with the National Instrument devices targeted on the Xilinx FPGA core through LabVIEW FPGA. From this study, it shows that the compilation time to convert G-code into HDL code depends on the size of the code. Hence, this study suggests that the G programming language in the LabVIEW FPGA module could be used to facilitate and accelerate the development of FPGA-based systems.

Further research involving the extraction of motion sensor data from multiple sensor placements on the body of an athlete can be carried out to measure whole body kinematics of rowing motion by using FPGA devices as processing unit.

Acknowledgements. The authors would like to thank the Ministry of Higher Education, Malaysia and Universiti Tun Hussein Onn Malaysia (UTHM) for funding this study.

References

1. Sforza, C., Casiraghi, E., Lovecchio, N., Galante, D., Ferrario, V.F.: A three-dimensional study of body motion during Ergometer rowing. *Open Sports Med. J.* **1**(6), 22–28 (2012)
2. Bernstein, I.A., Webber, O., Woledge, R.: An ergonomic comparison of rowing machine designs: possible implications for safety. *Br. J. Sports Med.* **36**, 108–112 (2002)
3. Shi, G., He, Y., Ye, F., Yang, J., Wang, P., Jin, Y.: Towards an ubiquitous motion capture system using inertial MEMS sensors and ZigBee network. In: *International Conference on Cyber Tech. in Automation, Control, and Intelligent System*, pp. 230–234. IEEE, Kunming (2011)
4. Borghetti, M., Sardini, E., Serpelloni, M.: Evaluation of bend sensors for limb motion monitoring. In: *International Symposium on Medical Measurements and Applications*, pp. 1–5. IEEE, Lisboa (2014)
5. Byrd, G.: 21st Century Pong. *Computer* **48**(10), 80–84 (2015)
6. Valeria, R., Stefan, L., Vesa, L., Yves, V., Walter, R., Laura, G.: Trunk kinematics during cross country Sit-skiing Ergometry: Skiing strategies associated to neuromusculoskeletal impairment. In: *International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6. IEEE, Benevento (2016)
7. Taha, Z., Hassan, M.S.S., Yap, H.J., Yeo, W.K.: Preliminary investigation of an innovative digital motion analysis device for badminton athlete performance evaluation. In: *11th Conference of the International Sports English Association*. *Procedia Engineering* **147**, 461–465 (2016)
8. Chapter 3 Motion Capture. <http://www.uio.no/imv/literature/knap3-4>
9. Zhu, R., Zhaoying, Z.: A real-time articulated human motion tracking using tri-axis inertial/magnetic sensors package. *IEEE Trans. Neural Syst. Rehabil. Eng.* **12**(2), 295–302 (2004)
10. King, R.C., McIlwraith, D.G., Lo, B., Pansiot, J., McGregor, A.H., Yang, G.Z.: Body sensor networks for monitoring rowing technique. In: *2009 Proceedings on 6th International Workshop on Wearable and Implantable Body Sensor Networks*, pp. 251–255. IEEE, Berkeley (2009)
11. Yurish, S.Y.: High Performance Digital Sensors Design: How to Make It Smarter, http://www.iaria.org/conferences2014/filesSENSORCOMM14/Yurish_Tutorial_2014.pdf
12. De La Piedra, A., Braeken, A., Touhafi, A.: Sensor systems based on FPGAs and their applications: a survey. *Sensors* **12**(9), 12235–12264 (2012)
13. Minouni, E.H.E., Karim, M., Kouache, M.E., Amarouch, M.Y.: An FPGA-based system for real-time electrocardiographic detection of STEMI. In: *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 830–835. IEEE, Monastir (2016)
14. Oballe-Peinado, Ó., Vidal-Verdú, F., Sánchez-Durán, J.A., Castellanos-Ramos, J., Hidalgo-López, J.A.: Smart capture modules for direct sensor-to-FPGA interfaces. *Sensors* **15**(12), 31762–31780 (Dec 16, 2015)
15. García, G.J., Jara, C.A., Pomares, J., Alabdo, A., Poggi, L.M., Torres, F.: A Survey on FPGA-based sensor systems: towards intelligent and reconfigurable low-power sensors for computer vision. *Control Signal Process. Sensors* **14**, 6247–6278 (2014)

16. Ponce-Cruz, P., Molina, A., MacCleery, B.: LabVIEWTM FPGA. In: Fuzzy Logic Type 1 and Type 2 Based on LabVIEWTM FPGA. Series Studies in Fuzziness and Soft Computing. vol. 334, pp 71–138. Springer International Publishing, Switzerland (2016)
17. Andrade, H.A., Ahrends, S., Hogg, S.: Making FPGAs Accessible with LabVIEW. In: Koch, D., Hanning, F., Ziener, D. (eds.) FPGAs for Software Programmers, pp. 63–79. Springer International Publishing, Switzerland (2016)
18. Wang, G., Tran, T.N., Andrade, H.A.: A graphical programming and design environment for FPGA-based hardware. In: 2010 International Conference on Field Programmable Technology, pp. 337–340. IEEE, Beijing (2010)
19. Nane, R., Sima, V.M., Pilato, C., Choi, J., Fort, B., Canis, A., Chen, Y.T., Hsiao, H., Brown, S., Ferrandi, F., Anderson, J., Bertels, K.: A Survey and evaluation of FPGA high-level synthesis tools. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **35**(10), 1591–1604 (2016)
20. Fuller, D.: The Future of FPGA Design Software, Jan. 24, 2013, <http://www.ni.com/newsletter/51624/en/>
21. Accelerometer ADXL335 Datasheet, 2009–2010 Analog Devices. <https://www.sparkfun.com/datasheets/Components/SMD/adxl335.pdf>
22. NI sbRIO-961x/963x/964x and NI sbRIO-9612XT/9632XT/9642XT National Instruments, <http://www.ni.com/pdf/manuals/375052c.pdf>
23. FPGA Module Help National Instruments, http://zone.ni.com/reference/en-XX/help/371599K-01/lvfpgsahelp/fpga_compile_window_reports/

A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses

Abdullahi O. Adeleke^(✉), Noor Azah Samsudin, Aida Mustapha,
and Nazri Mohd Nawi

Faculty of Computer Science and Information Technology, Universiti Tun
Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
hi150007@siswa.uthm.edu.my, {azah, aidam, nazri}@uthm.
edu.my

Abstract. Most existing feature selection approach is limited to determine features from a single source of data. In this paper, a feature selection approach is proposed to consider multiple sources of textual data. The proposed GBFS approach is then applied to label Quranic verses based on two major references, the English translation and tafsir (Commentary). The verses were selected from two chapters, Surah Al-Baqarah and Surah Al-Anaam. The verses are classified into three categories: Faith, Worship, and Etiquette. The textual data from the translation and commentary were preprocessed using StringToWord Vector with weighted TF-IDF. Feature selection algorithms: information gain, chi square, Pearson correlation coefficient, relief, and correlation-based were experimented on four classifiers: naïve Bayes, libSVM, k -NN, and decision trees (J48). The proposed group-based feature selection approach has shown promising results in terms of Accuracy and Area under Receiver Operating Characteristics (ROC) curve (AUC) by achieving Accuracy of 94.5% and AUC of 0.944.

Keywords: Holy Quran · Text classification · Feature selection techniques
K nearest neighbor · Support vector machine · Naïve Bayes · Decision trees

1 Introduction

The massive technological growth over the years has made the field of machine learning one of the mainstays of information technology [1]. Machine learning as defined by Arthur Samuel is a field of study that gives computers the ability to learn without being explicitly programmed [2]. Learning is considered as a parameter for intelligent machines to be able to take decisions in a more optimized form as well as work smoothly [3]. The concept of ML is based on training machines to be able to detect patterns and adapt to new circumstances.

Important task in machine learning is classification [4]. A problem of identifying to which set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known [5]. Most common to classification is text categorization, a task of automatically sorting a set of documents into categories from a predefined set [6].

The Holy Quran is the religious text for Muslims. A divine, highly comprehensive and detailed book from Almighty God, considered as an essential reference for over 1.6 billion Muslims in the world [7, 8]. Analyzing a chapter (or chapters) of the Holy Quran for better understanding of mentioned issues leads to knowledge discovery. This requires looking into multiple related sources of data. For example, the translation and tafsir (commentary) as two sources of Holy Quran data are not individually sufficient for the analysis purpose. In view of this, automation of the analysis may apply techniques in classification problem.

However, most existing feature selection approach to classification task is about determining features from individual source of data [9]. Therefore, this paper proposes an extension of the existing feature selection approach. The new approach determines features from multi-related sources of data referred to as Group-based feature selection. The proposed feature selection approach will be used for classifying verses in chapter two (Surah al-Baqarah) and chapter six (Surah al-Anaam) of the Holy Quran into three distinct predefined categories namely: Faith (Iman), Worship (Ibadah), and Etiquettes (Akhlaq).

The remainder of this paper is organized as follows. Section 2 presents the works related to Quranic text classification, Sect. 3 presents the methodology with detailed account of the dataset, classification experiment as well as the evaluation metrics, Sect. 4 reports the experiment results, and finally Sect. 5 concludes with some directions for future work.

2 Related Work

Compared to other textual data, there are very few research studies in the classification of Quranic verses based on the English translation [10, 11]. Most research focuses on classification of Quranic verses in Arabic [7, 12, 13]. Furthermore, in these existing works, analyses of features are dependable on individual sources as summarized below.

Jamil et al. [14] proposed a subject identification method based on term frequency to categorize groups of text into specified subjects. The dataset for the experimental work comprises of 224 verses of the Holy Quran with the verses containing 16 keywords on female chosen. The keywords include: daughter, female, woman, damsel, niece, mother, aunt, consort, divorcee, girl, lady, maiden, sister, widow, wife, and queen. The predefined labels for classifying the selected verses were: ‘inheritance’; ‘marriage’; and ‘divorce’. Results were evaluated in terms of ranking score and the term ‘dower’ in verse 237 of surah al-Baqarah identified as the subject.

Goudjil et al. [15] proposed in their research work an active learning method for Arabic text classification using multi-class SVM. To conduct the experiment, the dataset used consist of 363 documents divided into three categories namely: ‘Quran’; ‘Economic’; and ‘Sport’. The authors experimented both the Quran dataset and the standard Reuter’s corpus (R8) dataset. The results obtained showed the method had good accuracies by significantly reducing the training data of R8 and Quran datasets by 1.5 and 2.5% respectively.

In [13], the work focused on the review of Quranic web portals and their predictions using data mining tools such as Oracle Data Miner (ODM), Weka, SPSS. The

dataset was obtained from Alexa's web information company, a part of Amazon.com company that provides website analytics for all websites counting wise. The specific objective of the research was on studying the access pattern of some websites region wise using classification based data mining tools. The results from the four selected websites (quran.com, islamicity.com, quranexplore.com, tanzil.net) showed the web portal (islamicity.com) had the highest AUC value of 0.69 and Accuracy of 0.87 while the portal (tanzil.net) had the lowest AUC of 0.37 and Accuracy of 0.81.

Hassan et al. [16] implemented a k-Nearest Neighbor (kNN) algorithm to classify the Holy Quran Tafseer verses into predefined categories. To achieve this task, a database of 1000 verses of the Quran was divided into two document sets with the training set consist of 800 verses and the test set had 200 verses. Seven (7) predefined categories of the Tafseer texts were chosen (Marriage, Inheritance, Pray, Zakat, Respecting Parents, Halal, and Jihad) after transforming from Arabic to Malay language. The results were evaluated based on Precision and Recall metrics with 'marriage' category has the highest recall value of 0.9 and 'inheritance' has the lowest recall value of 0.74.

3 Method

Feature selection is a process commonly used in machine learning to select subset of features available in a data for application of a learning algorithm. Feature selection (FS) is essentially a task of removing irrelevant and/or redundant features from a dataset [5]. In this paper, standard feature selection algorithms namely: Information gain (IG); Chi square (CH); Pearson correlation coefficient (PCC); ReliefF; and Correlation based (CFS) are experimented on four classifiers: Naïve bayes (NB); Support vector machines (LibSVM); Nearest Neighbor (kNN); and Decision trees (J48). Figure 1 illustrates the framework of the proposed GBFS approach comprising of four phases: Data Acquisition; Preprocessing; Implementation; and Prediction (Results). Data on Quranic verses for analysis are gathered from multiple sources of Holy Quran comprising of Quranic Translation (source A) and Tafsir (source B). The resulting data is a fused data of both the translation and tafsir (otherwise called group-based data). The group-based data is preprocessed for feature generation using the standard StringToWordVector tool along with TF-IDF weighting method. Feature selection algorithms are applied to select the most relevant feature subsets (group-based features). The group-based features are further used in the implementation phase by the classifiers. The results obtained are compared in order to evaluate the significant influence of feature selection on the classification model.

3.1 Dataset

The dataset consists of 451 verses (instances); 286 verses from chapter two (Surah al-Baqarah) and 165 verses from chapter six (Surah al-Anaam) of the Holy Quran. The datasets were from the classical Holy Quran (English) translation by Abdullah Yusuf Ali obtained from (www.qurandatabase.org) and Holy Quran (English) tafsir of Imam

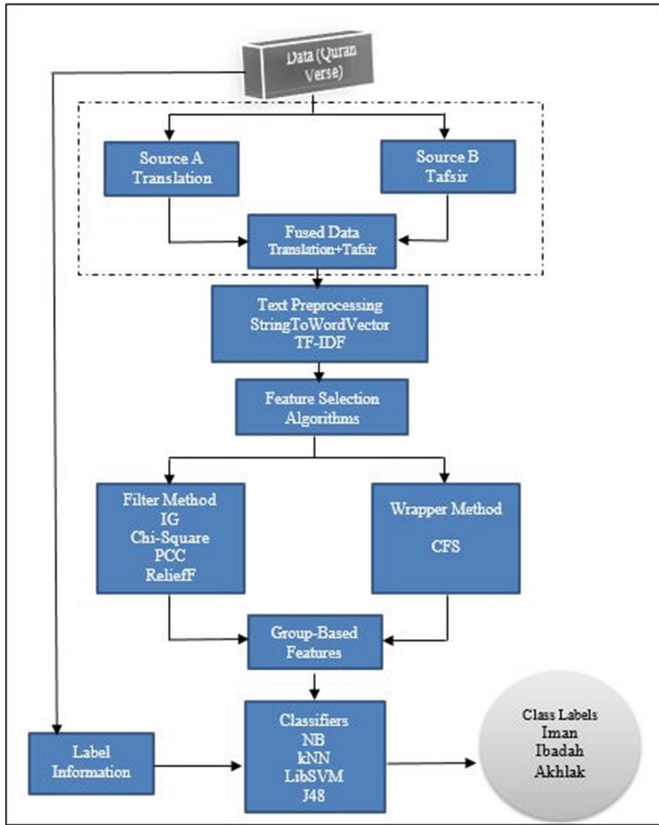


Fig. 1 The proposed group-based feature selection approach

Ismail ibn Kathir obtained from (www.allahsword.com). At present, there is no standard English Quran dataset for machine learning.

3.2 Data Preprocessing

Preprocessing is an essential and necessary task in text classification [5]. It is the process of generating features, transformation, and cleaning the text data in order to remove noise. It also reduces high level of dimensionality present in the text data before the training phase where the classification algorithms are trained to predict [5].

In the experimental work, the text data is first converted to Attribute-Relation File Format (ARFF) which is the standard file format for machine learning using WEKA (Waikato Environment for Knowledge Analysis). The following are the steps taken to preprocess the text:

StringToWordVector

StringToWordVector is an unsupervised filter standard tool in Weka used for converting string attributes into a set of attributes representing word occurrence

information from the text contained in the strings. This process can also be termed as Tokenizing. StringToWordVector directory in Weka is:

```
Java.lang.Object
    weka.filters.Filter
        weka.filters.unsupervised.attribute.StringToWordVector
```

Its command function syntax by default in Weka is:

```
public StringToWordVector ()
```

Where the empty argument is set to output 1000 words.

However, a desired number of words to output could be specified using the syntax:

```
public StringToWordVector (int wordsToKeep)
```

wordsToKeep is the number of words in the output vector (per class if assigned).

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a standard term weighting scheme used in accessing and measuring the significance of a word to a document in a collection [17]. Its value increases proportionally to the number of times a word appears in a document. In text classification, TF-IDF is one of the method used for stop-words filtering. It is an important step in dimensionality reduction process involving the removal of redundant, irrelevant words from text. TF-IDF is a product of two statistical weighting methods, Term Frequency (TF) and Inverse Document Frequency (IDF).

Term Frequency

Term frequency $Tf(t, d)$ is defined as the number of times a given term t (word/token) appears in a document d [18].

Mathematically, Term Frequency ($Tf(t, d)$) is defined as:

$$Tf(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\text{Maximum Occurrences of words}} \quad (1)$$

where *Maximum Occurrences of words* is denoted with: $\text{Max}\{f_t^d, d : t \in d\}$.

Inverse-Document Frequency (IDF)

The Inverse-Document Frequency (IDF) is a measure of how much information a word provides. In other words, IDF is a method of evaluating if a term is common or rare across all documents in a collection [19].

Mathematically, IDF is given as:

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (2)$$

where N is the total number of documents in D . $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears.

TF-IDF could then be given as:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (3)$$

3.3 Feature Selection Techniques

Due to the high level curse of dimensionality present in data (of most interest text data), which often leads to problems such as overfitting and performance degeneration of the classifiers, the feature selection techniques are employed. There are two possible ways to feature selection: the ranking features approach and subset selection approach [5].

The ranking features approach ranks features according to a certain criterion of the feature selection algorithms and the top k features are selected while on the other hand, the subset selection approach selects a minimum subset of features without learning performance deterioration [5]. For this paper, we experimented using the ranking features approach standard algorithms namely: information gain (IG), chi square (CH), Pearson correlation coefficient (PCC), and reliefF. For the subset selection approach, we experimented the correlation-based feature selection algorithm.

Information Gain (IG) is one of the filter feature selection method used in measuring the dependence between features and labels. The information gain between X and Y is calculated as:

$$I(X : Y) = H(X) - H(X|Y) \quad (4)$$

where $H(X)$ is the entropy of discrete random variable X defined as:

$$H(X) = - \sum_{x_i \in X} P(x_i) \log(P(x_i)) \quad (5)$$

x_i denotes a specific value of the variable X , $P(x_i)$ denotes the probability of x_i over all possible values of X

$H(X|Y)$ is the conditional entropy of variable X given another discrete random variable Y and its defined as:

$$H(X|Y) = \sum_{y_j \in Y} P(y_j) \sum_{x_i \in X} P(x_i|y_j) \log(P(x_i|y_j)) \quad (6)$$

$P(y_j)$ is the prior probability of y_j , $P(x_i|y_j)$ is the conditional probability of x_i given y_j .

Generally, Information Gain (IG) is calculated as:

$$I(X : Y) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \quad (7)$$

Chi-square filter algorithm is used as a test of independence to access the independence of the class label of a particular feature. Given a feature with r different values and c classes, Chi-square feature score can be defined as:

$$x^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (8)$$

where n_{ij} is the number of samples with the i th feature value.

$$\mu_{ij} = \frac{n * j n_{i*}}{n} \quad (9)$$

n_{i*} is the number of samples with the i th value for a particular feature, $n * j$ is the number of samples in class j , and n is the number of samples.

Pearson correlation coefficient (also called linear correlation coefficient) is a method used to evaluate the strength of relationship between two vectors. Given a pair of variables x and y , the Pearson correlation coefficient P is given as:

$$P = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2 \sum_i (y_i - \bar{y}_i)^2}} \quad (10)$$

where \bar{x}_i and \bar{y}_i are the mean of variables x and y respectively.

The value of p lies between $(-1,1)$ if x and y are linearly dependent (i.e., correlated), and $p = 0$ if x and y are independent (i.e., uncorrelated) [20]. Thus, to detect if redundancy exist among features, a feature must be strongly correlated to some other features [20].

ReliefF is an extension of the Relief binary classification task filter algorithm. It is a multi-class supervised filter algorithm that select features to separate instances from different classes. ReliefF (F_i) is defined as:

$$F_i = \frac{1}{c} \sum_{j=1}^l \left(-\frac{1}{m_j} \sum_{x_r \in NH_{(j)}} d(X(j, i) - X(r, i)) + \sum_{y \neq y_i} \frac{1}{h_{jy}} \frac{p(y)}{1 - p(y)} \sum_{x_r \in NM_{(j,y)}} d \right) (X(j, i) - X(r, i)) \quad (11)$$

where $NH_{(j)}$ and $NM_{(j,y)}$ indicate the nearest data instances to x_r in the same class and a different class y , respectively m_j and h_{jy} . $p(y)$ is the ratio of instances with class label y . d is a distance metric usually by default Euclidean distance. F_i represent feature score, l data instances and c classes.

Correlation-based Feature selection (CFS) search feature subsets according to the degree of redundancy among the features. The aim is to find the subsets of features that are individually highly correlated with the class but have low inter-correlation [5]. CFS usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search, and genetic search. CFS is given by:

$$r_{zc} = \frac{k \bar{r}_{zi}}{\sqrt{k + k(k-1) \bar{r}_{ii}}} \quad (12)$$

where r_{zc} is the correlation between the summed feature subsets and the class variables. k is the number of subset features. $\overline{r_{zi}}$ is the average of the correlations between the subset features and the class variables. $\overline{r_{ii}}$ is the average inter-correlation between subset features.

3.4 Classification Model

A growing number of data mining techniques have been applied to text classification problem, including the Bayes probabilistic approach [21], decision trees [22], neural networks [23], support vector machines (SVM) [24], and k-nearest neighbor [25]. In this study, four classification algorithms: nearest neighbor (k -NN), support vector machines (LibSVM), naïve bayes (NB), and decision trees (J48) classifiers are implemented for the labeling task.

The k -NN classifier is an instance-based learning algorithm that has shown to be very simple but effective for text classification problem [26]. It is a non-parametric method used in classification and works by calculating the Euclidean distance between points [27]. In classifying a new document x , the algorithm ranks the document's neighbors in the training set, and then uses the class of k most similar neighbors to predict the class of a new document (also known as majority vote). The Euclidean distance is given as:

$$d(x, x_i) = \sqrt{\sum_{i=1}^n (x_j - x_{ij})^2} \quad (13)$$

where x is the new point, x_i is the existing point across all input attributes j .

Naïve bayes classifier greatly simplify learning by assuming that features are independent given class and has proven effective in many practical applications, including text classification [28]. The classifier is a simple probabilistic model based on the Bayes rule [29]. Given a class C , the probability of a particular document d to belong to C is given as:

$$P(C_i|d) = \frac{P(d|C_i) * P(C_i)}{P(d)} \quad (14)$$

SVM is one of the most widely used and applied classification methods. It has been successfully applied to many application domains. SVMs are typically used for learning classification, regression, or ranking function. The algorithm works by searching a separating hyperplane to separate between samples with a maximal margin [30]. The equation for hyperplane is:

$$w^T x + b = 0 \quad (15)$$

To classify an unseen document d , the sign of $w^T x + b$ must be known [30]. This is further shown as:

$$w^T x_i + b \geq 1 \text{ or } w^T x_i + b \leq -1 \quad (16)$$

Decision tree is a way of representing a sequence of rules that leads to a class or value. It consists of three fundamentals: root node, internal node, and leaf node [31]. The algorithm is a tree like structure which classifies an input sample into one of its possible classes [32]. In decision tree classification algorithm, each node specifies a test to be performed on a single attribute [33]. The goal is to create a model that predicts the value of a target variable based on several input variables. The data generally takes the form:

$$(x, Y) = (x_1, x_2, x_3, \dots, x_k, Y) \quad (17)$$

where Y is a dependent variable to be classified or predict, x is a vector with input variables $x_1, x_2, x_3, \dots, x_k$ to be used for the classification of Y .

The classification experiment used the standard 10 fold cross-validation method for training and testing phase. The input to the classifier is a verse represented by a vector of group-based term count. Meanwhile, the output of the classifier is the class; ‘iman’, ‘ibadah’, and ‘akhlak’.

3.5 Evaluation Metrics

The objective of this study is to compare the significant influence of the feature selection algorithms on the classification process. The classification experiments were set to measure the accuracy and area under the receiver operating characteristics curves (AUC) across the selected feature selection algorithms on the classifiers.

Classification Accuracy

Classification Accuracy is the percentage or proportion of the total number of predictions that are correctly classified [34]. Accuracy can be calculated as:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (18)$$

- TP is True Positive (instances correctly classified as Positive)
- TN is True Negative (instances correctly classified as Negative)
- FP is False Positive (instances incorrectly classified as Positive)
- FN is False Negative (instances incorrectly classified as Negative)

Area under (ROC) curve (AUC)

AUC is one of the most popularly used performance metrics in classification problems [35]. AUC values reflect the overall ranking performance of a classifier. The performance metric over the years have been proven to be better than classification accuracy performance metric for evaluating the classifier performance [36].

Its value ranges from 0 to 1, where $AUC = 1$ corresponds to perfectly correct classification, $AUC = 0.5$ corresponds to classification by chance, and $AUC = 0$

corresponds to an inverted classification. As AUC value tends to 1 it means most perfect classification algorithm [37].

4 Results and Discussions

This section detailed the experimental results of the study following carefully the proposed research framework. Implementation was carried out using five selected standard feature selection algorithms across four conventional classification algorithms based on three scenarios. The experimental work is based on the following scenarios:

- i. Case 1: Using the Holy Quran (English) translation only.
- ii. Case 2: Using the Holy Quran (English) tafsir only.
- iii. Case 3: Using the combination of translation and tafsir.

The experimental results obtained using the proposed GBFS approach were compared with the existing traditional approach to feature selection in terms of classification accuracy and AUC. Tables 1, 2, 3 and 4 shows the results comparison in terms of accuracy using naïve bayes, libSVM, k-NN, and J48 classifiers respectively.

Table 1 Results comparison in term of accuracy using NB classifier

Dataset	All features (%)	IG (%)	CH (%)	PCC (%)	ReliefF (%)	CFS (%)
Translation	83.4	91	91	83.7	81.4	91
Tafsir	87.5	88.3	88.3	87.7	86.4	88.3
GBFS	90.4	92.6	92.6	90.3	87.1	92.5

Table 2 Results comparison in term of accuracy using LibSVM classifier

Dataset	All features (%)	IG (%)	CH (%)	PCC (%)	ReliefF (%)	CFS (%)
Translation	84	88.6	88.6	84	84.8	90
Tafsir	84.3	86.6	86.6	84.3	84.3	89.2
GBFS	84.8	93.1	93.1	84.8	86.7	94.5

Table 3 Results comparison in term of accuracy using Avg kNN classifier (k = 1, 3, 5)

Dataset	All features (%)	IG (%)	CH (%)	PCC (%)	ReliefF (%)	CFS (%)
Translation	83.2	86.9	86.9	83.2	84.8	87
Tafsir	84.2	87.5	87.5	84.2	85	90.9
GBFS	84.5	86	86	84.5	85.5	89.4

Table 4 Results comparison in term of accuracy using J48 classifier

Dataset	All features (%)	IG (%)	CH (%)	PCC (%)	ReliefF (%)	CFS (%)
Translation	82.3	85.7	85.7	81.8	83.2	85.2
Tafsir	85.7	86.6	86.7	86	86.6	86.7
GBFS	84.1	87	86.8	83.3	84.2	87.3

From the above results, the proposed GBFS had consistently above 90% accuracy results across several of the feature selection algorithms using naïve bayes and libSVM classifiers. However, with the kNN and J48 classifiers, the GBFS approach had relatively low accuracy results. These show that the feature selection algorithms as well as the classifiers have their strengths and weaknesses. The nature of data and features selected may influence the performance of a classification algorithm. Furthermore, the experimental results were evaluated based on the AUC values across several classifiers as visualized in Figs. 2, 3, 4 and 5.

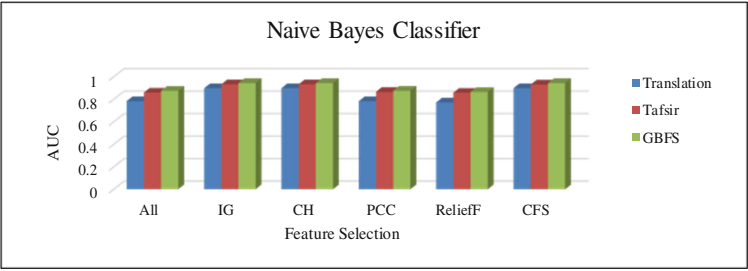


Fig. 2 Results comparison in term of AUC using NB classifier

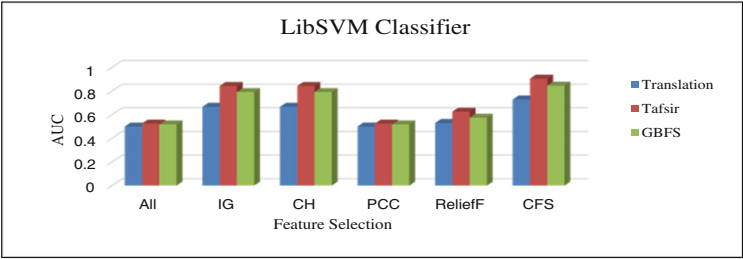


Fig. 3 Results comparison in term of AUC using LibSVM classifier

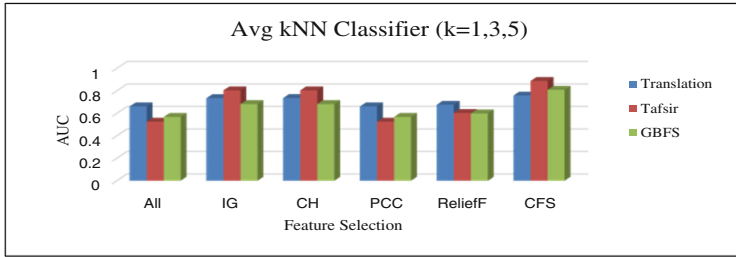


Fig. 4 Results Comparison in term of AUC using k -NN classifier

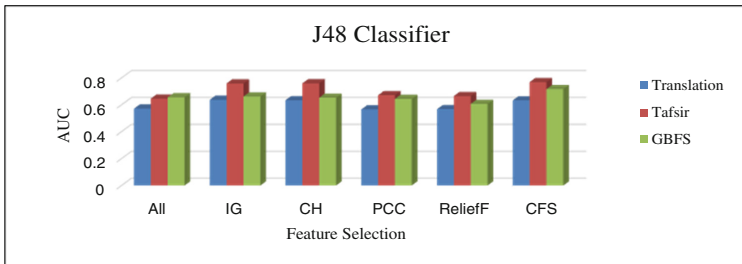


Fig. 5 Results comparison in term of AUC using J48 classifier

Assessing the classification performances of the classifiers, the proposed GBFS approach improved over the existing traditional approach to feature selection with the overall highest classification accuracy of **94.5%** for correlation-based feature selection (CFS) algorithm using the libSVM classifier as shown in Table 2. In addition, the closer an AUC value is to 1 the more efficient is a classifier. The proposed approach had the overall highest Area Under (ROC) Curve (AUC) value of **0.944** for information Gain, chi-square, and correlation-based feature selection algorithms using naïve bayes classifier as visualized in Fig. 2. Thus, the results show that selecting the most relevant features to reduce the curse of dimensionality is an essential step in text classification problem.

Furthermore, in comparison with the traditional feature selection approach, the proposed group-based feature selection approach has shown to be applicable and more efficient in some real world classification problems such as the Quranic verses labeling. The implementation generates several ROC curves, but due to the limitation of space, selected ROC curves of the proposed GBFS approach are plotted in Fig. 6 and 7.

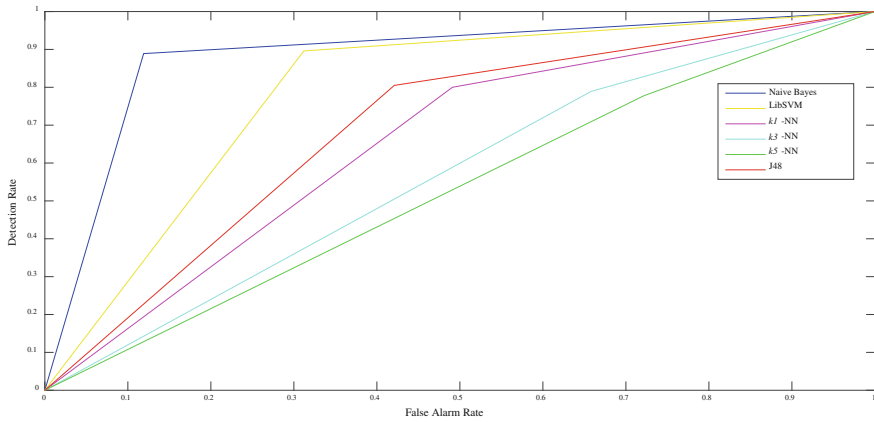


Fig. 6 ROC curve of the proposed GBFS approach using information gain (IG) feature selection algorithm

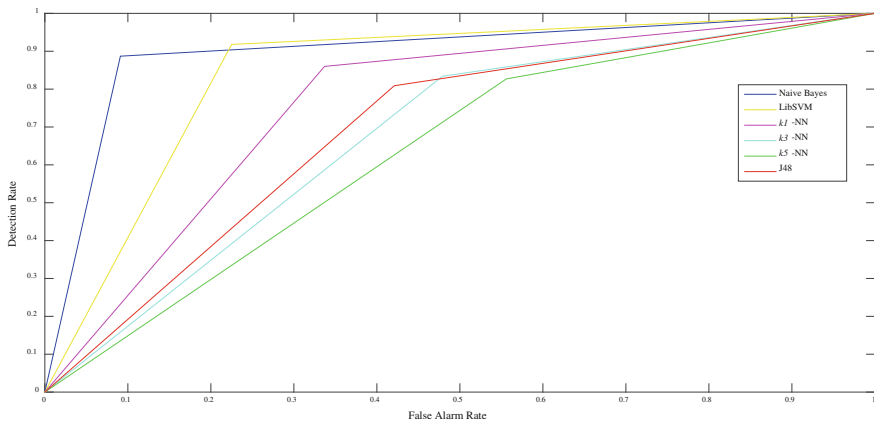


Fig. 7 ROC curve of the proposed GBFS approach using correlation-based feature selection (CFS) algorithm

5 Conclusion

Classifying the Quranic verses into pre-defined categories is an essential task in Quranic studies. In this paper, we presented a feature selection approach to automatically label Quranic verses in order to relate the verses with the three most fundamental aspects of Islam; ‘Iman’ (profession of faith), ‘Ibadah’ (worship), and ‘Akhlaq’ (etiquettes).

The dataset comprising of 451 instances (verses) from chapter two and six of the Holy Quran were normalized using StringToWordVector tool in WEKA with the TF-IDF method. Five of the most common feature selection algorithms from the ranking features and subset selection approaches were adopted and experimented in the

classification task. Finally, the NB, LibSVM, k -NN, and J48 classifiers were implemented independently on the feature selection algorithms to determine class membership for each verse and measure the results in terms of accuracy and area under the receiver operating characteristics curve (AUC). The experimental results have shown that the feature selection algorithms employed in the proposed approach had significant impacts on the classifiers implemented in the verse classification task. The proposed GBFS had the highest accuracy of 94.5% and AUC value of 0.944.

In conclusion, we hope to develop a complete standard English Quran dataset and further extend the proposed GBFS approach in the classification of the entire Holy Quran verses into labels as defined by the Quranic scholars. In addition, this approach could also be implemented on the Prophetic sayings (*Hadith*). Further enquiries on the ROC curves, kindly contact the authors.

Acknowledgements. This study was supported in part by a grant from the Ministry of Education of Malaysia, Research Acculturation Grant Scheme (RAGS) Vot R045, a grant from Universiti Tun Hussein Onn Malaysia Vot U611, and in part by a grant from Research Gates IT Solution Sdn. Bhd.

References

1. Ivanovic, M., Radovanovic, M.: Modern machine learning techniques and their applications. In: International Conference on Electronics, Communications and Networks (2015)
2. Das, S., Dey, A., Pal, A., Roy, N.: Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *J. of Comput. Appl.* **115**, 31–41 (2015)
3. Talwar, A., Kumar, Y.: Machine Learning: An Artificial Intelligence Methodology. *J. Eng. Comput. Sci.* **2**, 3400–3404 (2013)
4. Pundir, P., Gomanse, V., Krishnamacharya, N.: Classification and prediction techniques using machine learning for anomaly detection. *J. Eng. Res. Appl.* **1**, 1716–1722 (2013)
5. Tang, J., Alelyani, S., Lin, H.: Feature selection for classification: a review. In: *Data Classification: Algorithms and Applications*. CRC Press (2014)
6. Faraz, A.: An elaboration of text categorization and automatic text classification through mathematical and graphical modelling. *Comput. Sci. Eng. Int. J.* **5**, 1–11 (2015)
7. Hilal, A., Srinivas, N.: Analytical of the initial holy Quran letters based on data mining study. *Am. Int. J. Res. Formal Appl. Nat. Sci.* **10**, 1–8 (2015)
8. Alhawarat, M.: Extracting Topics from the Holy Quran using generative models. *J. Advanc. Comput. Sci. Appl.* **6**, 288–294 (2015)
9. Prusa, J.D., Khoshgoftaar, T.M., Dittman, D.J.: Impact of feature selection techniques for tweet sentiment classification. In: *Proceedings of the Twenty-Eight International Florida Artificial Intelligence Research Society Conference*. pp. 299–304 (2015)
10. Hamed, S.K., Ab Aziz, M.J.: A question answering system on holy Quran translation based on question expansion technique and neural network classification. *J. Comput. Sci.* **12**, 169–177 (2016)
11. Hamoud, B., Atwell, E.: Quran question and answer corpus for data mining with WEKA, pp. 211–216. *IEEE Conference of Basic Sciences and Engineering Studies*, Leeds (2016)
12. Akour, M., Alsmadi, I., Alazzam, I.: MQVC: measuring Quranic verses similarity and Surah classification using N-Gram. *WSEAS Trans. Comput.* **13**, 485–491 (2014)

13. Siddiqui, M.K., Naahid, S., Khan, M.N.I.: A review of Quranic web portals through data mining. *VAWKUM Trans. Comput. Sci.* **5**, 1–7 (2014)
14. Jamil, N.S., Ku-mahamud, K.R., Din, A.M., Ahmad, F., Chepa, N., Ishak, W.H.W., Din, R., Ahmad, F.K.: A subject identification method based on term frequency technique. *J. Advanc. Comput. Res.* **7**, 103–110 (2017)
15. Goudjil, M., Bedda, M., Koudil, M., Ghoggali, N.: Using active learning in text classification of Quranic sciences. In: *International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, pp. 209–213 (2015)
16. Hassan, G.S., Mohammad, S.K., Alwan, F.M.: Categorization of ‘Holy Quran Tafseer’ using k-Nearest neighbour algorithm. *Int. J. Comput. Appl.* **129**, 1–6 (2015)
17. Leskovec, J., Rajaraman, A., Ullman, J.D.: *Mining of Massive Datasets*, 2nd edn. Cambridge University Press, England (2014)
18. Menaka, S., Radha, N.: Text classification using keyword extraction technique. *J. Advanc. Res. Comput. Sci. Software Eng.* **3**, 734–740 (2013)
19. Chen, J., Chen, C., Liang, Y.: Optimized TF-IDF algorithm with the adaptive weight of position of word. *Advanc. Intelligen. Syst. Res.* **133**, 114–117 (2016)
20. Eid, H.F., Hassanien, A.E., Kim, T.H., Banerjee, S.: Linear correlation-based feature selection for network intrusion detection model. *Advanc. Security Informat. Commun. Netw.* **381**, 240–248 (2013)
21. Tang, B., He, H., Baggenstoss, P.M., Kay, S.: A Bayesian classification approach using class-specific features for text categorization. *IEEE Trans. Knowl. Data Eng.* **28**, 1602–1606 (2016)
22. Zharmagambetov, A.S., Pak, A.A.: Sentiment analysis of document using deep learning and decision trees. In: *Twelve IEEE International Conference on Electronics Computer and Computation*, pp. 1–4 (2015)
23. Wang, J.H., Wang, H.Y.: Incremental Neural Network Construction for Text Classification. In: *IEEE International Symposium on Computer Consumer and Control*, pp. 970–973 (2014)
24. Sabbah, T., Selamat, A.: Support vector machine based approach for Quranic words detection in online textual content. In: *8th IEEE Malaysian Software Engineering Conference, Malaysia*, pp. 325–330 (2014)
25. Townsend, K.R., Sun, S., Johson, T., Attia, O.G., Jones, P.H., Zambreno, J.: k-NN text classification using an FPGA-based sparse matrix vector multiplication accelerator. In: *IEEE International Conference on Electro/Information Technology*, pp. 257–263 (2015)
26. Gharehchopogh, F.S., Khaze, S.R., Maleki, I.: A new approach in bloggers classification with hybrid of k-nearest neighbor and artificial neural network algorithms. *Indian J. Sci. Technol.* **8**, 237–246 (2015)
27. Dey, L., Chakraborty, S., Biswas, A., Bose, B., Tiwari, S.: Sentiment analysis of review datasets using Naïve Bayes’ and k- NN classifiers. *J. Informat. Eng. Electron. Business.* **4**, 54–62 (2016)
28. Narayanan, V., Arora, I., Bhatia, A.: Fast and accurate sentiment classification using an enhanced Naïve Bayes model. In: *Intelligent Data Engineering and Automated Learning. 14th International Conference Proceedings*, Springer, Berlin Heidelberg, vol. 8206, pp. 194–201 (2013)
29. Nikam, S.S.: A comparative study of classification techniques in data mining algorithms. *Comput. Sci. Technol.* **8**, 13–19 (2015)
30. Amarappa, S., Sathyanarayana, S.V.: Data classification using support vector machine (SVM), a simplified approach. *J. Electron. Comput. Sci. Engineering.* **3**, 435–445 (2014)
31. Sewaiwar, P., Verma, K.K.: Comparative study of various decision tree classification algorithm using WEKA. *J. Emerging Res. Manag. Technol.* **4**, 87–91 (2015)

32. Teli, S., Kanikar, P.: A survey on decision tree based approaches in data mining. *J. Advanc. Res. Comput. Sci. Soft. Eng.* **5**, 613–617 (2015)
33. Adamatti, D.F., Silveira, J.A., Carvalho, F.A.H.: Analyzing brain signals using decision trees: an approach based on neuroscience. *Revista Eletronica Argentina-Brasil de Technologies da informacao e da Comunicacao.* **1**, 5 (2016)
34. Santra, A.K., Christy, C.J.: Genetic algorithm and confusion matrix for document clustering. *Int. J. Comput. Sci. Iss.* **9**, 322–328 (2012)
35. Yang, J., Qu, Z., Liu, Z.: Improved Feature-Selection Method Considering the Imbalance Problem in Text Categorization. *Scientific World J.* 1–17 (2014)
36. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. *Int. J. Data Mining Knowledge Manag. Process.* **5**, 1–11 (2015)
37. Adeleke, A.O., Samsudin, N.A., Mustapha, A., Nawi, N.M.: Comparative analysis of text classification algorithms for automated labelling of Quranic verses. *Int. J. Advanc. Sci. Eng. Info. Tech.* **7**, 1419–1427 (2017)

A Review: Image Analysis Techniques to Improve Labeling Accuracy of Medical Image Classification

Mazniha Berahim¹(✉), Noor Azah Samsudin²,
and Shelena Soosay Nathan¹

¹ Department of Information Technology, Center for Diploma Studies,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor,
Malaysia

{mazniha, shelena}@uthm.edu.my

² Faculty of Computer Science and Information Technology, Universiti Tun
Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
azah@uthm.edu.my

Abstract. Medical images contain the Region of Interest (ROI) from the affected area in human body and provide useful information to support clinical decision-making for diagnostics as well as the treatment planning. Unfortunately, medical image data may contain noise, missing values, inhomogeneous ROI that may give inaccurate diagnostic. Therefore, image analysis techniques are needed to improve the quality of an image. Then, features extraction task will be performed to produce best feature of images which leads to better classification result for accurate diagnostic. Many techniques have been used for image analysis. However, limited review have been done in categorize the list of related techniques for each image analysis task in medical imaging application. Thus, the aims of this paper is to gather and present general overview of image analysis task and their techniques in order to inspire researcher, pathologist or radiologist to adapt it when analyzing different types of medical image. The current study of image analysis task was summarized and discussed in this paper.

Keywords: Image analysis technique · Medical image · Image classification

1 Introduction

Images of body, organs, and cells are reflecting most medical issues. In disease diagnostics field, imaging technologies and automated analysis is growing interest recently [1]. There are many types of medical images which are Magnetic Resonance (MR), Computerized Tomography (CT), Ultrasound, Microscopic, Mammogram, and X-Ray. Even though the analysis and diagnosis of medical issues from the images are done by experienced radiologists [2], there are human error possibility [3]. Thus, various image analysis techniques have been used to develop an automated analysis to overcome problem from different types of medical images. However, limited review has been conducted in categorizing the list of related techniques for each image

processing task in medical imaging application. The aim of this paper is to gather and present an overview of image analysis task and their techniques. The current study of image analysis tasks in handling medical images are reviewed and summarized. The contribution of this study is overview of previous related studies and to inspire researcher or radiologist to adapt the technique during analyzing different type of medical images.

2 Imaging Task Analysis

Image processing refers to the operations that transforming [4], improving [5] or manipulating images. It is considered to be one of the most rapidly growing issue in medical as well as other fields [6]. The input and output for this process are images [5]. The output will be a set of features or character. A desired image for different purposes obtained using computer algorithms based on related techniques [7]. This process is crucial and used as input for classification task to get the best accuracy result.

This section provides some typical imaging task for recognition, registration, enhancement, fusion, segmentation, feature extraction, feature selection, fusion and analysis. The task basically performed during pre-processing phase. The implementation of imaging task is very useful to enhance the clinical applicability of medical images for diagnosis and assessment of medical issues [1]. The most crucial is imaging task for classification phase which used in most medical studies of feature extraction. All imaging task were described briefly and current studies for automated imaging task are listed.

2.1 Enhancement

To obtain quality image, this task depends in the image part to be enhanced from input image [8]. The image improvement is achieved if a certain feature is easily interpreted. Without making any loss in image information, scope of the enhancement are such as brightness preservation, reduction or removal of noise, altering the contrast, intensities and colors [9]. For instance, the boundaries of region of interest (ROI) in an image can be sharpen using image enhancement techniques [2]. Then, the clear boundaries increased the contrast between abnormal and the surrounding normal tissue. Selected studies in image enhancement technique applied for medical are Histogram Equalization (HE) [9, 10], Principal Components Transformation [11] and Wavelet Transform [12, 13]. HE are well-known technique for contrast enhancement [8, 10] due to simplicity and moderately better performance on images [14]. HE increase image visibility [15] and avoid the dark edges [16]. A study by [13] prove that the disease diagnosis accuracy after using image enhancement improved from 71.58–76.28% to high performance 87.30–94.01% which effected in various of features using discrete wavelet transform (DWT) technique.

2.2 Segmentation

Segmentation is a process of subdividing an image into its component parts such as objects instance, contiguous regions and similar region pixels [17]. This is crucial tasks which identify ROI in the images [1]. For instance, abnormal regions such as reflective of tumors [1] or level of a cancer malignancy [18]. Study by [19] found this task can aid in better disease diagnosis with 0.03% improved in accuracy level which also considered significant. There were many types of segmentation techniques implemented in literature such as, morphological smoothing on X-ray image of cervical vertebrate [9] and mammogram image [20], Level set (LS) on mammogram image [18, 20], Gray level co-occurrence matrix on mammogram image [21], Fuzzy C-Means [18]. Rouhi and Jafari [20] show after applied Region Growing(RG)-LS segmentation method produce more accurate primary boundary of tumors and subsequently reach 96.89% accuracy level and give 100% sensitivity.

2.3 Registration

Image registration is a process of referencing or transforming an image to another image based on mapping points [17] or spatial transformation that gives correct match [22] either with same device (mono-modal) or different devices of same scene (multi-modal) [23]. The concept of multi-modal image registration first introduced by [24]. The objective of this task is to recover the correspondences between the images [25]. This task implemented by [26] to reconstruct any “missing” analogous image and de-noising the target image to transform desired image. The image registration works in four steps: feature detection, feature matching, transform model estimation and image resampling and transformation [27]. The scope of this analysis includes volumes, landmarks, surfaces, contours [23], visible and infrared [25]. Mutual information (MI) is well-known technique for registration task [28, 29] due to efficient similarity measure for multimodal image [29].

2.4 Fusion

Fusion imaging task is a process of merging the relevant information from multiple images to be a single image [1, 8, 30, 31]. This task will integrate redundant information from multimodality images to produce an accurate description and get obtain complete overview of image [30]. The information also can be derived from single-view multi-modality or multi-view of images from mono-modality of image which referred to the same object. This task also enhance the imaging quality, reduce redundancy and randomness [1]. Basic scope for image fusion is pixels intensities [8]. Without affecting the image quality, image fusion is similar with image compression [7]. Table 1 shows fusion techniques used in multi-view medical images either from mono-modal or multi-modal.

Table 1. Automated techniques in image fusion task

Images	Techniques
PET and MR images of brain	Framelet transform [30]
Multi-view MR images of Alzheimer Disease	Propagation graph fusion (PGF) [32]
Multi-modality image	Low level (LL) and high level (HL) [33]
Positron emission tomography (PET), single photon emission computed tomography (SPECT) of brain	Generalized intensity-hue-saturation [34], Union Laplacian pyramid [31]

2.5 Features Extraction

Feature extraction is a process of transforming the input data into the set of features [35] to obtain information in a lower dimensionality space [35]. Feature extraction is done after the preprocessing phase [35]. Common feature extracted from this task includes color, texture and shape textures [36]. Most of studies from Table 2 used Gray-Level Co-Occurrence Matrix (GLCM) technique to extract image features. The GLCM is a Second Order Histogram (SOH) techniques. Usually, mathematical descriptions are expressed from textures of ROIs [37] as mean, standard deviation, entropy and skewness. Besides that, Principal Component Analysis (PCA) is a powerful technique to extract features from high dimensional data [38].

Table 2. Automated techniques in image feature extraction task

Image	Techniques
MR images of brain	Wavelet Entropy [39] First Order Histogram (FOH) [40], GLCM [38, 40], Ripplet transform Type-I (RT) [41] PCA [38]
Mammograms of breast cancer	Higher-order spectra(HOS) [2], Local Binary Pattern (LBP) [2, 42], Laws' Texture Energy (LTE), Discrete Wavelet Transform (DWT) [2, 37] and GLCM [37, 42]
Thermography images Breast cancer	LBP [43], Markov Random Field (MRF) [43]
3D confocal microscopy of Corneal	FOH [44], GLCM [44], Averages SOH [44], Grey Run Length Matrix [44]
Alzheimer disease	Atlas-based analysis [45]
MR image of lung	Morphology [46]

2.6 Feature Selection

Feature selection task is a process of selecting an optimum and potentially useful subset of features [47]. This task will ignore irrelevant or redundant features in order to produces more intelligible results [48] and minimize computation time [3]. The collected data contain many attributes (features) for each entity, some of which are irrelevant or redundant. These features does not have any role in the process of knowledge discovery, and increase the complexity as well as incomprehensibility of the results [1, 48]. The accuracy of the inferred function strongly depends on how

features are being chosen. According to Table 3, the typical technique used for feature selection is PCA.

Table 3. Automated techniques in image feature selection task

Image	Techniques
Computed tomography images of the chest for lung diseases	Generalized matrix learning vector quantization [49], MI ranking [49]
MR images of brain	PCA [41, 50], linear discriminant analysis (LDA) [50]
Mammograms of breast cancer	Sequential forward, backward (SFS) [2], plus-l-takeaway-r, and branch-and-bound selections [2]
Colonoscopy image for colon	SFS [11], Sequential floating forward selection (SFFS) [11],
MR image of Alzheimer’s disease (AD)	Support vector machine recursive feature elimination (SVM-RFE) [45]

3 Discussion

Generally, image classification consists of four main stage: image acquisition, image preprocessing, features extraction or selection and classification. Segmentation, feature extraction, feature selection are important in view of dimension reduction for improving learning with generalization ability and identifying significant features [45, 51]. The features captured the subtle variation in the pixel intensities and contours in the medical images and serve as significant indicators for classification [2].

For diagnostic, some disease use radiological image processing and analysis [52] such as ultrasonic X-Ray Image for Ovarian Disease [53], MRI for Brain Tumor [54]. However, some other disease like breast cancer, it need a microscopic image to confirm the existing of cancer tissue [52]. The mammogram image only show the abnormality of the breast only [55]. In literature, analysis task of radiological image is easier compared to histopathology or microscopic image [52]. Therefore, the use of image analysis task are differ based on the image modality and the disease diagnostic procedure.

In addition to the best image analysis techniques used, the analysis time need to be considered to measure the features can be reduced in order to improve the classification performance [47]. For example, the feature selection using the proposed technique by [50] is more beneficial as it analyses the data according to grouping class variable and gives reduced feature set with high classification accuracy. While, [56] found that morphology is a popular method in image processing field due to its rigorous mathematical description and its proven applicability in imaging problems including noise elimination, feature extraction and image compression.

The textural appearances of objects features from medical images may contain composed of repetitive and non-repetitive patterns [44]. For example, medical x-ray images are grayscale with almost the same texture characteristic of image regions [57].

The patterns of features may extracted using different methods and will be used as an input to the classifier which assign them into the class that they represent for automated diagnosis [50, 58]. Thus, the image analysis task is very crucial in order to produce best features of significant disease indicators [2, 20] and will leading to enhance the classification performance [47].

As conclusion, besides the technique used in the classification phase, the performance of automated medical image diagnostic can also be influenced by image quality [13, 20] besides considering features selection task [59, 60] issues. Rather than that, literature noticed that classification performance also influenced by the issue of the images used, either view from mono-modality versus multi-modality of images [22, 32, 61] or single view versus multi-view images [62, 63]. The above issues need to be considered when developing a classification framework.

4 Conclusion

Review showed that image analysis task is very crucial and important in classifying medical issues. When the pre-processing phase get the desired image from enhancement, registration, fusion or segmentation task, the features extraction and features selection phase will be applied to finds the appropriate features to represent the input images which can be used in classification task. Classifier phase trained to classify normal or abnormal ROI based on quantitative features measured. From the current studies listed in Tables 1 and 2, image fusion and registration task are found to be typically used for multi-modal or multi-view image. However, limited studies has been conducted on multimodal or multi-view medical image analysis. Thus, as future work, multi-view images can be considered to be applied for suitable analysis task.

Acknowledgement. This work is supported by UTHM under Short Term Grant Vot U660.

References

1. James, A., Dasarathy, B.: Medical image fusion: a survey of the state of the art. *Inf. Fusion* (2014)
2. Ganesan, K., Acharya, R.U., Chua, C.K., Min, L.C., Mathew, B., Thomas, A.K.: Decision support system for breast cancer detection using mammograms. *Proc. Inst. Mech. Eng. H*. **227**(7), 721–732 (2013)
3. Ghasemian, F., Mirroshandel, S.A., Monji-Azad, S., Azarnia, M., Zahiri, Z.: An efficient method for automatic morphological abnormality detection from human sperm images. *Comput. Methods Prog. Biomed.* **122**(3), 409–420 (2015)
4. Fu, K.-S., Rosenfeld, A.: Pattern recognition. *IEEE Trans. Comput.* **C-25**, 1336–1346 (1976)
5. Lee, L., Liew, S.-C.: A Survey of medical image processing tools. In: *IEEE 4th International Conference Software Engineering Computer System (ICSECS)* (2015)
6. Chiuchisan, I.: A New FPGA-based real-time configurable system for medical image processing. In: *4th IEEE International Conference E-Health Bioengineering-EHB 2013*, pp. 0–3 (2013)

7. Asaduzzaman, A., Martinez, A., Sepehri, A.: Time-efficient image processing algorithm for multicore/ manycore parallel computing. In: *Proceedings of IEEE Southeast Conference 2015* (2015)
8. Ahirwar, V., Yadav, H., Jain, A.: Hybrid model for preserving brightness over the digital image processing. *IEEE 4th International Conference Computerized Communication Technology* **1**, 48–53 (2013)
9. Abdullah, S., Asy, M., Mimi, W., Wan, D., Ibrahim, F.: X-Ray image enhancement for anterior osteophyte diagnosis. *IEEE Int. Electron. Symp.* pp. 47–52 (2015)
10. Senthilkumaran, N., Thimmiraja, J.: Histogram equalization for image enhancement using MRI brain images. In: *2014 World Congress Computing Communicable Technologies*, pp. 80–83 (2014)
11. Fu, J.J.C., Yu, Y.W., Lin, H.M., Chai, J.W., Chen, C.C.C.: Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging. *Comput. Med. Imaging Graph.* **38**(4), 267–275 (2014)
12. Li, X., Kang, Y.: A novel medical image enhancement method based on wavelet multi-resolution analysis. In: *IEEE I8th International Conference Biomedical Engineering Informatics*, pp. 727–731 (2015)
13. Beheshti, S.M.A., Ahmadi Noubari, H., Fatemizadeh, E., Khalili, M.: Classification of abnormalities in mammograms by new asymmetric fractal features. *Biocybern. Biomed. Eng.* **36**(1), 56–65 (2014)
14. Oak, P.V., Kamathe, P.R.S.: Contrast enhancement of brain MRI images using histogram based techniques. *Int. J. Innov. Res. Electr. Electron. Instrument. Control Eng.* **1**(3), 90–94 (2013)
15. Albadarneh, A., Albadarneh, I., Alqatawna, J.: Iris Recognition System for Secure Authentication Based on Texture and Shape Features. *IEEE Jordan Conference Applied Electronic Engineering Computerized Technology, Iris* (2015)
16. Sivasundari, S., Kumar, R.S., Karnan, M.: Review of MRI Image Classification Techniques. *Int. J. Res. Stud. Comput. Sci. Eng.* **1**(1), 21–28 (2014)
17. Rayudu, M., Jain, V., Kunda, M.R., Review of image processing techniques. In: *IEEE Sixth International Conference Sensor Technology Review*, pp. 320–325 (2012)
18. Krawczyk, B., Galar, M., Jelen, L., Herrera, F.: Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Appl. Soft Comput. J.* **38**, 714–726 (2016)
19. GeethaRamani, R., Balasubramanian, L.: Retinal blood vessel segmentation employing image processing and data mining techniques for computerized retinal image analysis. *Biocybern. Biomed. Eng.* **36**(1), 102–118 (2015)
20. Rouhi, R., Jafari, M.: Classification of benign and malignant breast tumors based on hybrid level set segmentation. *Expert Syst. Appl.* **46**, 45–59 (2016)
21. Angayarkanni, A.S.P., Kamal, B.N.B.: Automatic classification of mammogram MRI using dendograms. *Asian. J. Comput. Sci. Inf. Technol. J.* **4**, 78–81 (2012)
22. Legg, P.A., Rosin, P.L., Marshall, D., Morgan, J.E.: Feature neighbourhood mutual information for multi-modal image registration: an application to eye fundus imaging. *Pattern Recogn.* **48**(6), 1937–1946 (2015)
23. Chakraborty, S., Ray, R., Ghosh, S., Chatterjee, S., Chowdhuri, S., Dey, N.: Rigid image registration using parallel processing. In: *Proceedings of International Conference Circuits, Communicable Control Comput.* (I4C 2014) Rigid, no. November, pp. 21–22 (2014)
24. Woods, R.P., Mazziotta, J.C., Cherry, S.R.: MRI-PET registration with automated algorithm. *J. Comput. Assist. Tomogr.* **17**(4), 536–546 (1993)
25. Han, J., Pauwels, E.J., De Zeeuw, P.: Visible and infrared image registration in man-made environments employing hybrid visual features. *Pattern Recognit. Lett.* **34**(1), 42–51 (2013)

26. Cao, T., Zach, C., Modla, S., Powell, D., Czymmek, K., Niethammer, M.: Multi-modal Registration for Correlative Microscopy using image analogies. *Med. Imag. Anal.* **18**(6), 914–926 (2014)
27. Bedi, S.S., Agarwal, J., Agarwal, P.: Image fusion techniques and quality assessment parameters for clinical diagnosis: a review. *Int. J. Adv. Res. Comput. Commun. Eng.* **2**(2), 1153–1157 (2013)
28. Sahoo, P.K., Pati, U.C.: Image Registration using Mutual Information with Correlation for Medical Image. *IEEE* (2015)
29. Patra, D., Pradhan, S.: Enhanced mutual information based medical image registration. *IET Image Process.* **10**(5), 418–427 (2016)
30. Bhatnagar, G., Wu, Q.M.J., Liu, Z.: Human Visual system inspired multi-modal medical image fusion framework. *Expert Syst. Appl.* **40**(5), pp. 1708–1720 (2013)
31. Du, J., Li, W., Xiao, B., Nawaz, Q.: Union Laplacian Pyramid with Multiple Features for Medical Image Fusion. *Neurocomputing* **194**, 326–339 (2016)
32. Liu, S., Cai, W., Liu, S. Pujol, S., Kikinis, R., Feng, D.: Subject-centered multi-view feature fusion for neuroimaging retrieval and classification. *IEEE Int. Conf. Image Process.* 2505–2509 (2015)
33. Dimitrovski, I., Koccev, D., Kitanovski, I., Loskovska, S., Džeroski, S.: Improved Medical Image Modality Classification Using a Combination of Visual and Textual features. *Comput. Med. Imag. Graph* **39**, 14–26 (2015)
34. Wang, Q., Li, S., Qin, H., Hao, A.: Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis. *Inf. Fusion* **26**, 103–121 (2015)
35. Kumar, G., Bhatia, P.K.: A detailed review of feature extraction in image processing systems. In: 2014 Fourth International Conference Advanced Computerized Communication Technology. February 2014, pp. 5–12 (2014)
36. Tian, D.P.: A review on image feature extraction and representation techniques. *Int. J. Multimed. Ubiquitous Eng.* **8**(4), 385–395 (2013)
37. Beura, S., Majhi, B., Dash, R.: Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. *Neurocomputing* **154**, 1–14 (2015)
38. Khan, A., Syed, N.A., Reyaz, M.: Image processing techniques for brain tumor extraction from MRI images using SVM classifier. *Int. J. Recent Innov. Trends Comput. Commun.* **3** (May), 2707–2711 (2015)
39. Saritha, M., Joseph, K.P., Mathew, A.T.: Classification of MRI brain images using combined wavelet entropy based spider web plots and probabilistic neural network. *Pattern Recogn. Lett.* **34**(16), 2151–2156 (2013)
40. Hemanth, J.D., Vijila, C.K.S., Selvakumar, A.I., Anitha, J.: Performance improved iteration-free artificial neural networks for abnormal magnetic resonance brain image classification. *Neurocomputing* **130**, 98–107 (2014)
41. Sudeb, D., Manish, C., Kundu, M.K.: Brain MR image classification using multi- scale geometric analysis of ripplelet. *Prog. Electromagn. Res.* **137**(February), 1–17 (2013)
42. Liu, X., Zeng, Z.: A new automatic mass detection method for breast cancer with false positive reduction. *Neurocomputing* **152**, 388–402 (2015)
43. Rastghalam, R., Pourghassem, H.: Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images. *Pattern Recognit.* **51**, 176–186 (2014)
44. Sharif, M.S., Qahwaji, R., Ipson, S., Brahma, A.: Medical image classification based on artificial intelligence approaches: a practical study on normal and abnormal confocal corneal images. *Appl. Soft Comput. J.* **36**, 269–282 (2015)

45. Ota, K., Oishi, N., Ito, K., Fukuyama, H.: Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease. *J. Neurosci. Methods* **256**, 168–183 (2015)
46. Thamilselvan, P., Sathiaselvan, J.G.R.: An enhanced k nearest neighbor method to detecting and classifying MRI lung cancer images for large amount data. *Int. J. Appl. Eng. Res.* ISSN 0973-4562, **11**(6), 4223–4229 (2016)
47. Mbaga, A.H., ZhiJun, P.: Pap Smear images classification for early detection of cervical cancer. *Int. J. Comput. Appl.* **118**(7), 8887 (0975–8887) (2015)
48. Feizi-Derakhshi, M.-R., Ghaemi, M.: Classifying different feature selection algorithms based on the search strategies. *Int. Conf. Mach. Learn. Electr. Mech. Eng. (ICMLEME' 2014)*, 17–21 (2014)
49. Huber, M.B., Bunte, K., Nagarajan, M.B., Biehl, M., Ray, L.A., Wismüller, A.: Texture feature ranking with relevance learning to classify interstitial lung disease patterns. *Artif. Intell. Med.* **56**(2), 91–97 (2012)
50. Rathi, V.P.G.P., Palani, S.: Brain tumor MRI image classification with feature selection and extraction using linear discriminant analysis. *Int. J. Comput. Inf. Sci. Eng.* **2**(4), 131–146 (2012)
51. Mlambo, N., Cheruiyot, W.K., Kimwele, M.W.: A survey and comparative study of filter and wrapper feature selection techniques. *Int. J. Eng. Sci.* **5**(8), 57–67 (2016)
52. Aswathy, M.A., Jagannath, M.: Detection of breast cancer on digital histopathology images: present status and future possibilities. *Informat. Med. Unlocked*, November, pp. 0–1 (2016)
53. Khazendar, S., Al-Assam, H., Du, H., Jassim, S., Sayasneh, A., Bourne, T., Kaijser, J., Timmerman, D.: Automated classification of static ultrasound images of ovarian tumours based on decision level fusion. In: 6th Computerized Science Electron Engineering Conference Proceedings, pp. 148–153 (2014)
54. Sanjeev Kumar, P.M., Chatterjee, S.: Computer aided diagnostic for cancer detection using MRI images of brain (brain tumor detection and classification system). *IEEE Annual Indian Conference*, (2016)
55. Chen, Y., Ling, L., Huang, Q.: Classification of breast tumors in ultrasound using biclustering mining and neural network. In: 9th Int. Congr. Image Signal Process. Biomed. Eng. Informatics (CISP-BMEI 2016), pp. 1787–1791 (2016)
56. Zeng, N., Wang, Z., Zineddin, B., Li, Y., Du, M., Xiao, L., Liu, X., Young, T.: Image-based quantitative analysis of gold immunochromatographic strip via cellular neural network approach. *IEEE Trans. Med. Imag.* **33**(5), 1129–1136 (2014)
57. Mohammadi, S.M., Helfroush, M.S., Kazemi, K.: Novel shape texture feature extraction for medical x-ray image classification. *Int. J. Innov. Comput. Inf. Control* **8**(1) (2012)
58. Sudarshan, V., Acharya, U.R., Ng, E. Y.-K.S., Chou, M., Tan, R.S.: Automated identification of infarcted myocardium tissue characterisation using ultrasound images: a review. *IEEE Rev. Biomed. Eng.* **PP**(99), 1 (2014)
59. Aalaei, S., Shahraki, H., Rowhanimanesh, A., Eslami, S.: Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iran. J. Basic Med. Sci.* **6**, 476–482 (2016)
60. Al-Kadi, O.S.: A multiresolution clinical decision support system based on fractal model design for classification of histological brain tumours. *Comput. Med. Imaging Graph.* **41**, 67–79 (2015)
61. Liberman, G., Louzoun, Y., Aizenstein, O., Blumenthal, D.T., Bokstein, F., Palmon, M., Corn, B.W., Ben Bashat, D.: Automatic multi-modal mr tissue classification for the assessment of response to Bevacizumab in patients with glioblastoma. *Eur. J. Radiol.* **82**, 2, e87–e94 (2013)

62. Battula, B.P., Prasad, R.S.: An overview of recent machine learning strategies in data mining. *Int. J. Adv. Comput. Sci. Appl.* **4**(3), 50–54 (2013)
63. Xiang, Z., Lv, X., Zhang, K.: An Image Classification Method Based on Multi-feature Fusion and Multi-kernel SVM, 2014 Seventh Int. Symp. Comput. Intell. Des. **2**(1), 49–52 (2014)

A New Adaptive Energy-Aware Job Scheduling in Cloud Computing

Ali Aghababaeipour^{1,2(✉)} and Shamsollah Ghanbari^{1,2(✉)}

¹ Department of Computer Science, Islamic Azad University, Ashtian Branch, Iran
aghababaeipoor@gmail.com

² Iranian Non-profit Association of Distributed Computing and Systems, Qom, Iran
myrshg@gmail.com

Abstract. In the last decade, with the significant growth of the calculation and data concerns over energy use and carbon dioxide emissions caused by the servers have increased. Various scheduling algorithms have been created all of which attempt to reduce the execution time of tasks and have not paid enough attention to reduce energy consumption. Other scheduling algorithms try to reduce the makespan and the energy consumption simultaneously that are known as the energy-aware scheduling algorithms. The algorithm presented in this article schedules the tasks with a focus on reducing makespan and energy consumption. The proposed method provides a new scheduling algorithm using four factors of communication between tasks, the distance between nodes, virtual machines' status and energy consumption forecasts to reduce makespan and energy consumption. The purpose of this scheduling algorithm is to reduce the displacement between the nodes and optimize VMs execution that using the analytical hierarchy process (AHP) the best decision is made for task implementation.

Keywords: AHP · Task scheduling · Energy-aware scheduling · Cloud computing

1 Introduction

Cloud computing has become an extremely important tool for technology-related industries and IT organizations and has solved many of the problems in providing fast and high-quality services to customers [1]. Service provider organizations tend to use methods such as pay-as-you-go or elastic service in cloud environments to provide faster services to their customers [2]. With the rapid growth of cloud computing and increase in energy consumption by data centers researchers have devised different methods in terms of reducing energy consumption. In these methods efforts have been made to optimize energy consumption along with reducing the makespan similar to the dynamic voltage and frequency scaling (DVFS) hardware method or software methods with the change in the

scheduling algorithm [3–5]. With respect to [6], carbon emissions from information and communications technology (ICT) activities has had 6% growth per year that the highest energy consumption is associated with performing calculation in data centers and about 40% of total energy is consumed by the data centers themselves. Other parts of the data center such as the cooling system, the Energy Supply System, communication tools, etc., are other sectors that consume energy in the data center. As noted in [7] one of the important factors effective in the growth of carbon emission and the risks associated with it is the increased energy consumption by data centers. One of the effective methods in reducing energy consumption in cloud environments is to use energy-aware scheduling algorithms such that the energy reduction factor does not have an adverse effect on the makespan [8]. In the proposed method, the number of transmissions for each task to reach the selected node is considered as one of the required parameters to improve the makespan and energy consumption. Another parameter is the communication between the tasks which is obtained by the directed acyclic graph (DAG) [9]. The DAG contains the input tasks and the communication between them. Also the VMs' status is considered as another parameter to prevent unnecessary switching. In the end, the energy required for each task is calculated based on the method presented in [10] and the best node for executing the task is selected by the four parameters and the AHP decision-making method.

2 Related Work

Most scheduling algorithms in cloud infrastructure attempt to reduce execution time without energy consumption concerns. Energy consumed in cloud can be reduced by methods such as DVFS, maximizing cooling efficiency or scheduling algorithm. The other important factor for reducing energy consumption is VMs migration management.

A dynamic energy-aware scheduling model for task-based application in cloud has been proposed in [10]. The input task is converted to DAG graph and the relationship between tasks is determined. First, the necessary parameters for energy consumption in cloud environment are calculated. The mentioned scheduling algorithm tries to use energy usage parameter and tasks' execution time to set a bi-objective method for task allocation based on a ranking model. In the proposed method, energy consumption and execution time are taken into consideration in same way and a framework with multi-dimensional scheduler is created that allocates tasks based on transfer rate and more related task distance. The scheduler tries to keep related tasks in nearest distance and prevents many specific task transfers. In addition, VMs is managed by estimating the task behavior to prevent unnecessary VMs shut down and start up. The suggested framework doesn't use ranking model, it is attempted to keep locality and reduce transfer between nodes inside of VMs behavior manager for decrease the energy consumption and execution time.

Other work proposed by Aupy et al. [11] try to aim at minimizing the energy consumption while enforcing two constraints: a prescribed bound on the execution time (or makespan), and a reliability threshold. Authors for reducing energy consumption let some task to re-execution based on 3 parameters, makespan, reliability and energy. In [12], authors proposed an algorithm to executing by a linear combination of processor's maximum and minimum frequency. They calculate energy consumption by task based on constrained optimization problem. Jeffrey Chase et al. [13] proposed an economic method to managing server resources to reduce energy consumption by servers. The system continuously monitors load and plans resource allotments by estimating the value of their effects on service performance. In [14], authors proposed a queuing model for detect minimum number of servers and estimate when use these resources. Algorithm try to reduce energy consumption by switching idle server to power saving modes.

Table 1. Definitions and notations

Notation	Description
t_i	Current task
n	Represent node of cloud
DAG graph	Contains tasks and relation between tasks
VM_{on_m}	Represent online VM of current node
VM_{off_m}	Represent off-line VM of current node
<i>Coordinator</i>	VM for running scheduler
Power profile	Contains energy parameter from [10]
Rel_{t_i}	Relation between t_i and other tasks
Trn_{st_i}	Transfer rate of t_i
<i>MonitoredVM</i>	State of VMs in related nodes
e_{t_i}	Energy usage by t_i

3 Proposed Method

3.1 Notations and Definitions

Table 1 indicates the basic notations used in this paper.

3.2 Description of the Model

The proposed algorithm uses four parameters including the relationship between the tasks, the number of required transmissions to transfer the task to the corresponding node, VMs' condition of each node and the amount of energy used to execute the task to schedule the input tasks of the DAG. In the first steps

the scheduling algorithm reduces the number of transfers between nodes so that related tasks are carried out in neighboring nodes so that fewer transfers are applied so that the related tasks meet each other. In the second part of the scheduling algorithm the VM status is monitored to execute tasks by online VMs so that VMs' switching is avoided as much as possible. As the final monitoring the algorithm calculates the energy consumption to implement the input task using the given energy profile in [10] so that it has a balanced choice between makespan and energy consumption to execute the task. In other words, this algorithm looks for the nearest node that contains the largest number of tasks related to the input task and the VMs of this node do not require running and the least amount of energy to execute the input task is needed in that state. For example the algorithm may choose a node is more distant from the current node but there is no need to run a new VM on that node. This selection is done using the AHP method [15–18] and based on the defined priorities. In this way the algorithm chooses the most suitable place to run the input task with the lowest energy consumption and makespan based on the above mentioned priority. Since there is a need for separate processing resources and storage space to cache some information to calculate the mentioned parameters, a VM is considered as the coordinator to run the scheduling algorithm. The general schema of the proposed method is shown in Fig. 1.

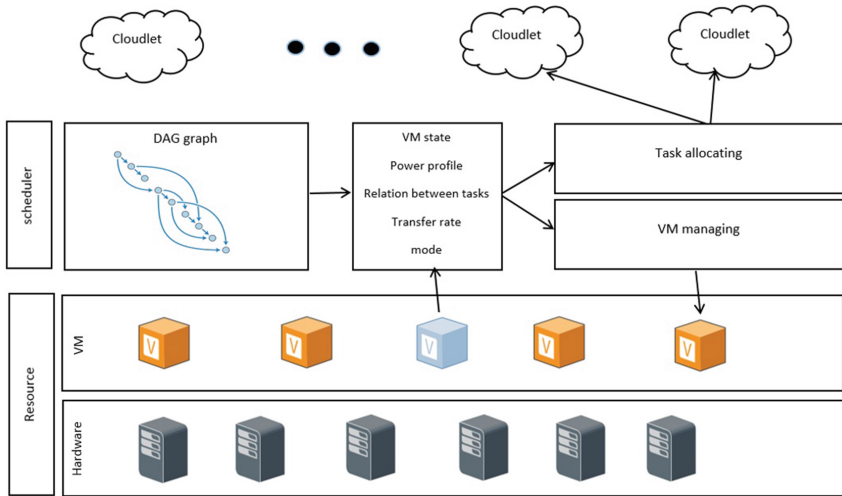


Fig. 1. Schema of proposed energy-aware framework

3.3 Model Formulation

The input data is initially converted into DAG to determine the relationship between the Tasks. The DAG represented as $G = (T, D)$ consists of vertices and

edges where the vertices represent the tasks and the edges display the relationship between the tasks. For example $i \rightarrow j$ indicates that the task j needs task i to run. Therefore i and j are vortices and $i \rightarrow j$ is equivalent to an edge in the DAG. This graph is sent to the cloud environment as the set of input tasks. The cloud environment includes a set of processing nodes in different point with different distances that each node itself consists of a number of on or off VMs. The list of nodes in the cloud is represented as $N = \{n_1, n_2, \dots, n_m\}$, where m represents the number of nodes in the cloud. The list of online VMs in the node m is $V_{on_m} = \{V_{on_{m_1}}, V_{on_{m_2}}, \dots, V_{on_{m_k}}\}$ and the list of off-line VMs in the node m is $V_{off_m} = \{V_{off_{m_1}}, V_{off_{m_2}}, \dots, V_{off_{m_l}}\}$ that k represents the online VMs in the node m and l denotes the off-line VMs in the node m . Also in the presented cloud environment VM coordinator represents the coordinator VM that is responsible for the execution of the scheduling algorithm. The scheduling algorithm S used the energy profile provided in [10] to reduce energy consumption during task assignment which is shown in Table 2.

Table 2. Power profile framework

Parameter	Description
e_{task}	Energy consumed by task
$e_{transfer}$	Energy consumed for transfer between nodes
e_v	Energy consumed per VMs
e_n	Energy consumed by node

3.4 Parameters

As mentioned, the algorithm used to schedule the tasks uses four parameters described as follows.

- **Related task** (Rel_{t_i}) This value is obtained by the DAG formation. In this graph the connection between each t_i input task with other tasks that have been performed so far. The scheduling algorithm prioritizes each node based on the number of related tasks with t_i so that the t_i runs in the node with the highest number of related tasks. This will prevent large displacements in the next steps of task processing and reduces time span and energy consumption. The Eq. 1 indicates how to calculate Rel_{t_i} .

$$Rel_{t_i} = \{Rel_{t_{i_1}}, Rel_{t_{i_2}}, \dots, Rel_{t_{i_m}}\} \quad (1)$$

where m is count of nodes and relation array of every node calculate as Eq. 2

$$Rel_{t_{i_k}} = \sum_{n=0}^{i-1} relation(t_n, t_i) \quad (2)$$

- **Transfer of task** ($Trans_{t_i}$) After the scheduling algorithm specified the nodes related to t_i , it calculates the number of transfer required to reach each of these nodes and gives a priority to each node based on the obtained distances. In fact the scheduling algorithm tries to select the node that has the lowest distance to run t_i by calculating the needed transfers to run the task. Equation 3 shows that $Trans_{t_i}$ is calculated based on Rel_{t_i} .

$$Trans_{t_i} = \{Trans_{t_{i_1}}, Trans_{t_{i_2}}, \dots, Trans_{t_{i_m}}\} \quad (3)$$

where m is count of related nodes and transfer rate to every related node calculate as Eq. 4

$$Trans_{t_{i_k}} = distance(n_i, n_k) \quad (4)$$

- **VM state** ($VMState$) The priority to execute the tasks is with the online VMs and the scheduling algorithm tries to prevent launching the off-line VMs as much as possible. The $VMState$ is a momentary variable, for example when nodes are checked a node may contain two free VMs but as long as the scheduler sends t_i for that node, these two VMs are used by other cloud sections. The scheduling algorithm adopts three different policies to solve this problem. First, in the $VMState$ parameter nodes that have the freest online VMs have a higher priority to increase the chances of implementing t_i . In addition, the scheduling algorithm chooses two nodes to run t_i , the main node is the best possible choice, and the second node is the closest choice to the main node, so that in the event that t_i is not executed in the main node, the second node would execute it. The third policy used by the algorithm to increase the $VMState$'s reliability coefficient is to build a cache in the coordinator called "trust". The scheduler analyzed task execution status after each task assignment. As the task runs on each node, the scheduler increases its rating in the trust variable so that the scheduler would increase the priority of the nodes that have a higher trust score to form $VMState$ in the next rounds, the $VMState$ formation is shown in Eq. 5

$$VMState = \{evaluate(V_{onm}, V_{offm}, Trust_m)\} \quad (5)$$

- **Energy usage.** The scheduling algorithm uses the energy profile presented in [10] to estimate the consumed energy to run t_i on any of the nodes in Rel_{t_i} . This profile calculates the consumed energy for each component of the cloud member separately which includes the energy used to transfer between nodes, the energy needed to start the VM and task processing by it and the energy consumed by the task and the node. Equation 6 calculates the energy to run t_i .

$$e_{total_t} = Min\{e_{total_{t_1}}, e_{total_{t_2}}, \dots, e_{total_{t_m}}\} \quad (6)$$

where m is count of related nodes and total energy used by every task calculate as Eq. 7

$$e_{total_{t_k}} = e_{t_k} + e_{transfer(t_k \rightarrow n_k)} + e_{V_k} + e_{n_k} + e_{etc(cooling, \dots)} \quad (7)$$

Algorithm 1 Decision Making ()

Description:

```

1: relatedNodes = CheckTaskRelationInCash( $t_i$ )
2: transfer = Distance( $n_{t_i}$ , relatedNodes)
3: monitoredVM = monitorVM(relatedNodes)
4: energyUsage = PowerProfile( $t_i$ , relatedNodes, transfer, monitoredVM)
5: weightOfNodes = AHP(relatedNodes, transfer, monitoredVM, energyUsage)
6: selectedNode = Max(weightOfNodes)
7: backupNode = nearBy(selectedNode)
8: trust = Execute( $t_i$ , selectedNode, backupNode)

```

Algorithm 2 MonitorVM ()

Description:

```

1: For (int  $i = 0$ ;  $i < \text{count}(\textit{relatedNodes})$ ;  $i++$ )
2: if (there is any online VM in relatedNodes[ $i$ ]) then
3:   monitoredVM[ $i$ ] = Count(VMon[ $i$ ])
4: end if
5: if (there is any off-line VM in relatedNodes[ $i$ ]) then
6:   monitoredVM[ $i$ ] = Count(VMoff[ $i$ ])
7: end if
8: if (there is not any off-line or online VM in relatedNodes[ $i$ ]) then
9:   monitoredVM[ $i$ ] = null
10: end if

```

3.5 Decision Making Model

The scheduling algorithm runs by a coordinator VM that decides to assign input tasks between the nodes in the cloud. Algorithm 1 shows the proposed scheduling of the parameters calculated in the previous section. Using the relationships between the tasks created by the DAG, the scheduling algorithm forms the *relatedNodes* array. This array contains all nodes that have at least one task related to t_i and a higher priority is given to the node that has the most related tasks to t_i . The scheduling algorithm forms the *transfers* array to save the current node's distance in which t_i is located to reduce the makespan and energy consumption caused by the transfer between the nodes. The *energyUsage* array also stores the energy to run t_i in each *relatedNodes*. The energy profile is queried for the t_i and its *relatedNodes* to obtain the total amount of energy of running t_i by calculating the energy consumed by each component of the cloud and determine the node with the lowest energy consumption.

To monitor the status of VMs and select the node with the most appropriate VM status, the scheduling algorithm forms the *monitoredVM* array by executing the Algorithm 2. The *monitorVM* function checks all nodes in the *relatedNodes* array and if there is an online VM on that node, the corresponding space in the *monitoredVM* array is set equal to the number of online VMs in that node, otherwise this value will be equal to the off-line VMs of this node and if all the VMs of this node are occupied, the null value will be recorded.

After obtaining the required parameters of the algorithm, the AHP decision maker is queried to calculate the weight of each node based on the four obtained parameters. The weight of each node is stored in the *weightOfNodes* array; the scheduler selects the highest weight as the main node for executing t_i and stores in the *selectedNode*. Then using the *nearBy* function the closest node to the main node is selected as the backup node so that if the main node's VMs are filled, it could run t_i in the shortest time using the least energy. After running t_i the score of the executor node increases in trust so that it is used in the subsequent queries.

4 Experimental

4.1 Startup and Settings

To simulate the proposed algorithm a set of seven nodes is used each of which is composed of four VMs. A VM is formed of 28 VMs as coordinator; the coordinator launches the intended services first and before scheduling. The coordinator receives the input tasks as a DAG graph; in order to compare the proposed algorithm tow different DAG graphs including the 700 tasks and 1400 tasks with the relationship between them were sent to the coordinator so that the proposed algorithm would be tested for different modes of the relationship between the various tasks. The coordinator calculates and stores the distances between the nodes at the startup stage.

4.2 Performance and Overhead

In the final section the scheduling algorithm's performance implementation was evaluated based on the makespan of each task. By changing the matrix priority at each scheduling algorithm's implementation each task's makespan was calculated to determine which one has higher effect on energy reduction and makespan. Figure 2 shows the makespan of 700 tasks for different priorities and Fig. 3 shows the makespan of 1400 tasks.

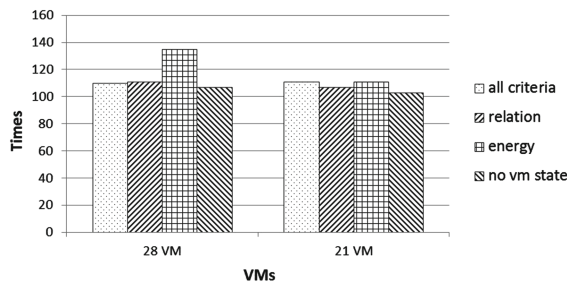


Fig. 2. Allocating tasks with 700 tasks

This comparison can show the difference in the impact of each parameter on decision making by the AHP so that the best result is obtained when the implementation of the related tasks in a node has the highest priority. In this way, the transmission parameters and energy consumption will also improve which can increase load traffic in a certain part of the cloud which has a negative effect on the *VMState* parameter.

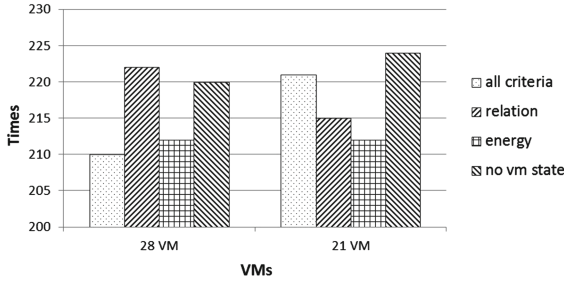


Fig. 3. Allocating tasks with 1400 tasks

5 Conclusion


In this paper task scheduling focused on energy reduction was studied. In order to achieve the best result, i.e. reducing makespan and energy consumption, the most important influencing parameters including the relationship between the tasks, transmissions required to run the tasks, VM status and the amount of energy consumption to run the tasks were determined. The scheduling algorithm chooses the best node for task execution by AHP which actually has the highest priority among the parameters. The results show that the relationship between the tasks has the greatest impact on the overall makespan of the tasks and the overall energy consumption of the cloud. As a future work the negative effects of the VMs' mode and the reduction of this effect will be examined on the resource allocation function.

References

1. Moganaragan, N., Babukarthik, R.G., Bhuvaneswari, S., Basha, M.S., Dhavachelvan, P.: A novel algorithm for reducing energy-consumption in cloud computing environment: web service computing approach. *J. King Saud Univ. Comput. Inf. Sci.* **28**(1), 55–67 (2016)
2. Yang, S., Wieder, P., Yahyapour, R., Fu, X.: Energy-aware provisioning in optical cloud networks. *Comput. Netw.* **8**(118), 78–95 (2017)

3. Dighe, S., Vangal, S.R., Aseron, P., Kumar, S., Jacob, T., Bowman, K.A., Howard, J., Tschanz, J., Erraguntla, V., Borkar, N., De, V.K.: Within-die variation-aware dynamic-voltage-frequency-scaling with optimal core allocation and thread hopping for the 80-core teraflops processor. *IEEE J. Solid-State Circuits*. **46**(1), 184–93 (2011)
4. Shamsollah, G., Othman, M., Bakar, M.R.A., Leong, W.J.: Multi-objective method for divisible load scheduling in multi-level tree network. *Future Gener. Comput. Syst.* **54**, 132–143 (2016)
5. Shamsollah, G., Othman, M.: A priority based job scheduling algorithm in cloud computing. *Procedia Eng.* **50**, 778–785 (2012)
6. Rong, H., Zhang, H., Xiao, S., Li, C., Hu, C.: Optimizing energy consumption for data centers. *Renew. Sustain. Energy Rev.* **31**(58), 674–91 (2016)
7. Singh, A., Mishra, N., Ali, S.I., Shukla, N., Shankar, R.: Cloud computing technology: reducing carbon footprint in beef supply chain. *Int. J. Prod. Econ.* **30**(164), 462–71 (2015)
8. Chen, D.R., Chiang, K.F.: Cloud-based power estimation and power-aware scheduling for embedded systems. *Comput. Electr. Eng.* **31**(47), 204–21 (2015)
9. Gerasoulis, A., Yang, T.: On the granularity and clustering of directed acyclic task graphs. *IEEE Trans. Parallel Distrib. Syst.* **4**(6), 686–701 (1993)
10. Juarez, F., Ejarque, J., Badia, R.M.: Dynamic energy-aware scheduling for parallel task-based application in cloud computing. *Future Gener. Comput. Syst.* **78**, 257–271 (2016)
11. Aupy, G., Benoit, A., Robert, Y.: Energy-aware scheduling under reliability and makespan constraints. In: 2012 19th International Conference on High Performance Computing (HiPC), 18 Dec 2012, pp. 1–10 (2012)
12. Rizvandi, N.B., Taheri, J., Zomaya, A.Y., Lee, Y.C.: Linear combinations of dvfs-enabled processor frequencies to modify the energy-aware scheduling algorithms. In: 2010 10th IEEE/ACM International Conference on InCluster, Cloud and Grid Computing (CCGrid), 17 May 2010, pp. 388–397 (2010)
13. Chase, J.S., Anderson, D.C., Thakar, P.N., Vahdat, A.M., Doyle, R.P.: Managing energy and server resources in hosting centers. *ACM SIGOPS Oper. Syst. Rev.* **35**(5), 103–16 (2001)
14. Urgaonkar, B., Shenoy, P., Chandra, A., Goyal, P., Wood, T.: Agile dynamic provisioning of multi-tier internet applications. *ACM Trans. Auton. Adapt. Syst. (TAAS)*. **3**(1), 1 (2008)
15. Saaty, T.L.: How to make a decision: the analytic hierarchy process. *Eur. J. Oper. Res.* **48**(1), 9–26 (1990)
16. Saaty, T.L. *Fundamentals of decision making and priority theory with the analytic hierarchy process*. RWS Publications, Pittsburgh (1994)
17. Saaty, T.L.: The modern science of multi-criteria decision making and its practical applications: the AHP/ANP approach. *Oper. Res.* **61**(5), 1101–1118 (2013)
18. Shamsollah, G.: Multi-criteria divisible load scheduling in binary tree network. Ph. D. Dissertation (2016)

Breast Cancer Recurrence Prediction Using Random Forest Model

Tahsien Al-Quraishi() , Jemal H. Abawajy, Morshed U. Chowdhury,
Sutharshan Rajasegarar, and Ahmad Shaker Abdalrada

Deakin University, Burwood, VIC 3125, Australia
`talqurai@deakin.edu.au`

Abstract. Breast cancer is the second most common cause of death among Australian females. To reduce the probability of death, early detection and prevention of breast cancer is a crucial factor. Evaluating the probability of breast cancer recurrence is an important act related to breast cancer prognosis. The aim of this paper is to predict the probability of breast cancer recurrence among patients. The researchers individually applied Random Forest and Deep Neural Network classifiers to increase the prediction accuracy of those models. Wisconsin Prognosis Breast Cancer dataset was obtained from UCI machine learning Repository. The results of our experiment indicate that Random Forest technique achieved the highest accuracy compared to the existing works.

Keywords: Breast cancer · Random forest · Deep neural network

1 Introduction

Breast cancer in Australia was the fourth most common cause of cancer death in 2014. It was also the second most common reason of death among females. In Australia, 2,814 females died from breast cancer in 2014. This number is predicted to increase to 3,087 females in 2017 [1]. The treatment of breast cancer recurrence can be either local, such as surgery and radiation, or systemic, involving chemotherapy, hormone therapy, and immunotherapy. Breast cancer may recur within five years after the treatment, with regional recurrence and distant metastasis included in this time. However, the probability of recovery is low once the recurrence occurs [2]. Hence, there is a need to develop an accurate approach to predict the probability of breast cancer recurrence to assist the medical physicians to propose possible treatments for breast cancer patients.

Advanced methods in the data mining area have come into existence recently, which showed better performance than statistical methods [3]. Classification is the most common and widely-used task in machine learning. It aims to predict the outcome in invisible data and depict the important data classes by extract-

ing the models [4]. Researchers [5] have obtained the knowledge from data sets by applying a single classifier. The decision tree is commonly used in medical analysis. Scholars [6] have utilized decision trees to obtain the knowledge from medical data.

Prior datasets are used to discover knowledge by developing a data mining algorithm in the classification task. The trends of each class have information, which is provided by prior data sets to predict new samples. Nowadays, classification methods cover many applications including medicine [7]. Wisconsin Prognostic Breast Cancer (WPBC) dataset [8] represents imbalanced data. The dataset is called imbalanced when the majority class exceeds the number of the minority class [9]. In classification tasks, many unpredicted mistakes and even significant outcomes are available in imbalanced data, because the classification algorithms are affected by the biased distribution of class samples, which are biased to the majority class. The imbalanced data classification problems are solved by utilizing a re-sampling technique in data level as well as by classifying the design of a high-level model in algorithm level. To increase the number of the minority class and avoid the over-fitting drawback, synthetic minority over-sampling technique (SMOTE) is proposed by researchers [10] for data pre-processing.

Predicting the recurrence probability of breast cancer is significant, because it decreases the mortality rate of patients and enhances the survival rate. The aim of our work is to predict the probability of breast cancer recurrence based on the outcome attribute within a five-year period or longer. To accommodate this issue, we estimate the performance of Random Forest (RF) and Deep Neural Network (DNN) technique. Random Forest is an effective way to balance the data and estimate the importance of features that are applied in the classification task [11] whereas, Deep Neural Network is able to obtain and simplify the imbalanced data [12]. The contribution of this paper is to enhance the accuracy of the recurrence probability among breast cancer patients compared to the existing work [13].

The rest of this paper is organized as follows: Sect. 1 introduces the basic concepts of RF and DNN classifiers. Section 2 presents previous work related to our research, while Sect. 3 explains the evaluated methodology. Section 4 describes the experiment results. In Sect. 5 the conclusion is covered.

2 Related Work

The literature review explains several studies on the breast cancer recurrence problem by applying different computational approaches and artificial neural networks. However, few papers are related to medical diagnosis and recurrence using data mining methods. In this section, we examine a number of machine learning approaches to predict breast cancer recurrence. The study of Kim, et al. [14] investigated the performance of SVMs, ANNs, and the Cox regression methods to predict the recurrence of breast cancer within a five-year period after surgery treatment. St. Gallen's guidelines Nottingham Prognostic Index (NPI), and Adjuvant! Online were selected to assess the performance of the

proposed methods. A total of 679 patients participated. In this study, there were 195 recurrence samples, and 484 no-recurrence samples. Seven attributes such as histological grade, tumour size, many metastatic lymph nodes, ER status, Lymph vascular Invasion (LVI), local invasion of a tumour, and many tumors were chosen out of 193 attributes. Holdout method (70% train–30% test data) was used to evaluate the performance of the classifiers. Accuracy, sensitivity, specificity, precision, AUC, and Negative Predictive Value (NPV) metrics were selected in this evaluation. SVMs outperformed all other approaches with 84.58, 0.89, 0.73, 0.75, 0.85, and 0.89% respectively. ANN achieved 81.37, 0.95, 0.52, 0.80, 0.80, and 0.82% while Cox performed less rates with 72.55, 0.24, 0.94, 0.63, 0.73, and 0.74%.

Salama, et al. [15] estimated breast cancer recurrence by comparing the performance of DTs, MLP, SVM, NB, and KNN approaches. In this experiment, many breast cancer imbalanced datasets were applied such as Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC), which were provided by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. In terms of classification performance, the fusion between approaches was investigated, to estimate whether the multi-classifier technique is beneficial or not. Ten-fold cross-validation technique was employed to assess the classifiers by evaluating their accuracy. SVM and DT achieved better performance compared to all other classifiers with an accuracy rate of 76.3%. MLP, KNN, and NB showed lower performance with 66.5, 64.4, and 50.5% respectively. SVM was combined individually with NB, MLP, DT, and KNN in the first fusion, while SVM and DT were combined independently with NB and MLP in the second fusion. Same accuracy rate (76.3%) resulted for all combinations. SVM-DT-MLP-KNN and SVM-DT-MLP-NB were combined in the third fusion. SVM-DT-MLP-KNN and showed best accuracy rate of 77.3%, while SVM-DT-MLP-NB achieved lower accuracy rate of 74.2%.

Tomczak [16] predicted breast cancer recurrence within a 10-year period after surgical treatment and identified the input symptoms related to breast cancer reappearance. In this experiment, Classification Restricted Boltzmann Machine (ClassRBM) learning methods such as DropOut, Drop Connect, and DropPart were applied. The performances of these methods were compared with those of Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Classification and Regression Trees (CART), and coupled with AdaBoost, Bagging, and LogitBoost classifiers. The Institute of Oncology, Ljubljana provided 949 patients and 15 features in this study, but not the breast cancer dataset. Holdout technique (70% train set and 30% test set) was employed to evaluate the performance of the classifiers, while 100 cases were used in the test data to predict from oncologists. LogitBoost + CART ensemble outperformed all other classifiers with accuracy rate of 75%, while, SVM showed deficient performance compared to other computational methods, which showed better performance than medical experts.

Chaurasia and Pal [17] assessed the performance of Dyadic Decision Trees (DDTs), Artificial Neural Network (ANN), and Logistic Regression (LR) classifiers to predict the recurrence status of breast cancer patients in a five-year period after surgical treatment. The dataset in this experiment was provided by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia, and contained 10 features and a total of 286 samples. WEKA application was used in all experiments without tuning the parameters before or during the classification process. Ten cross-validation techniques were employed to determine the classifier metrics such as accuracy, specificity, and precision for recurrence and no-recurrence classes. In this study, recurrence class is the label class and the effect of the preferred attribute on recurrence prediction was analyzed. LR outperformed all others with 74.5, 92.5, and 64.3%, respectively versus 71.3, 92, and 54.3% of DDTs, whereas ANN showed lower result rates of 73.8, 88.6, and 58.9%.

Beheshti, et al. [18] compared the performance of genetic approaches such as Centripetal Accelerated Particle Swarm Optimization (CAPSO), Particle Swarm Optimization (PSO), Gravitational Search Algorithm (GSA), and Imperialist Competitive Algorithm (ICA) with Multi-Layer Perceptron (MLP). Nine datasets of Hepatitis, Heart Disease, Pima Indian Diabetes, Wisconsin Prognostic Breast Cancer, Parkinsons disease, Echocardiogram, Liver Disorders, Laryngeal 1 and Acute Inflammations were applied to those hybrid approaches. They reported that tuning the parameters of PSO approach was time-consuming. To address this issue, fewer parameters were used and 80% train and 20% test data technique was used to evaluate the performances of the approaches. MSE, AUC, accuracy, sensitivity, and specificity metrics were employed for evaluation. CAPSO-MLP was found to be the best approach for unseen data, as it outperformed all other approaches with 0.170, 0.63, 80.3, 52.3, and 83.4% respectively. GSA-MLP performed 0.167 of MSE, 0.55% of AUC, and 79.3% of accuracy, 7.86% of sensitivity, and 80.23% of specificity. While, ICA-MLP resulted 0.177, 0.57, 78.3, 43, and 83% respectively. PSO-MLP achieved 0.173, 0.60, 78.3, 43, and 83%.

Ojha and Goel [13] investigated the probability of breast cancer recurrence among patients within a five-year period. In this study, they compared the performance of different popular clustering and classification approaches such as Wisconsin Prognostic Breast Cancer (WPBC) Data Set. This dataset is available in UCI Repository, and contains 198 samples and 35 attributes. In this experiment, many classification algorithms were applied, such as C5.0, KNN, NB, and SVM, while K-Means, EM, PAM, and Fuzzy e-means were employed as clustering algorithms. They divided the data set into 70% train data and 30% test data to evaluate the performance of the classifier. Accuracy, sensitivity and specificity metrics were involved in the evaluation process. SVM and C5.0 achieved best accuracy rate of 81.3%. While Fuzzy e-means showed lowest accuracy rate of 37%.

Table 1 shows that none of these author [14–18] have used Wisconsin Prognostic Breast Cancer (WPBC) dataset to predict breast cancer recurrence within

Table 1. Comparison of related work

Authors	Dataset(Attributes/Samples)	Method	Accuracy (%)
Kim, et al. (2012)	Tertiary teaching hospital, South Korea (195/484)	SVM	84.58
Salama, et al. (2012)	Wisconsin prognostic BC (47/151)	SVM and DT individually	76.3
Tomczak (2013)	Institute of oncology, Ljubljana (15/949)	LogitBoost + CART ensemble	75
Chaurasia and Pal 2014	Breast cancer dataset (85/201)	Logistic regression	74.5
Benetti, et al. (2014)	Wisconsin prognostic BC (47/151)	CAPSO-MLP	80.3
Ojha and Goel (2017)	Wisconsin prognostic BC (35/198)	SVM	81
Our method	Wisconsin prognostic BC (35/198)	RF	98.63 \pm 2.56

5 years period. Our study utilized this dataset and presented better accuracy compared to [13].

3 Evaluation Method

In this paper, we individually employed the most widely-used classification models such as Random Forest (RF) and Deep Neural Network (DNN) by applying holdout technique (70% train data and 30% test data). We used Wisconsin Prognostic Breast Cancer (WPBC) imbalanced dataset, to balance this data set, Synthetic Minority Over-Sampling Technique (SMOTE) was used for the training data set. A grid search was employed for both classifiers. Grid search in RF was applied to identify the best number of trees, while grid search in DNN was employed to select the best number of epochs and nodes in the hidden layers. In this study, several metrics were utilized to evaluate the test data for accuracy, precision, and recall. Moreover, the average and standard deviation were calculated by applying 20 iterations in R platform.

3.1 Dataset

We took the publicly available Wisconsin Prognostic Breast Cancer (WPBC) dataset [8]. It contains 35 attributes and 198 instances including 151 non-recr, and 47 recr cases. ID number attribute was applied to denote the number of the patient. Time attribute was ignored as in a class label, because the purpose of this study is to predict whether breast cancer will recur or not. Outcome attribute was used as a class label and represented two predicting fields such as (R, N). R represents recurrent, while N denotes non-recurrent. A total of 30 attributes with 10 real variables were calculated for each cell nucleus such as radius, texture, perimeter, area, smoothness, compactness, concavity, and concave points, symmetry and fractal dimension. The rest of the two attributes represent Tumour size and Lymph node status. We removed four instances as their Lymph nodes were missing.

3.2 Holdout Technique

Researcher [19] is required to obtain the classifier true error rate in the entire population but it is impossible to approach the entire population in the real world. However, to estimate the true error rate only a limited set of cases was obtainable. To address this issue of limited data a naive approach was applied to train the data. A serious over-fitting and optimistic error estimation would be returned while using this approach.

In the breast cancer recurrence problem, several sampling strategies are used such as holdout technique to overcome the over-fitting issue, which divides the dataset into training and test data. The training set is employed to build the model, whereas, the test set is used to evaluate the performance of the classifier. Holdout technique splits the dataset into the train and test sets, which represent many partitioning schemes such as 50–50%, 70–30% or 80–20%.

3.3 Basic Theory of Random Forest Algorithm

Random Forests [20] is also known as a random decision tree. In a random forest, the association is unavailable among the trees. Each decision tree is classified once the test data is moved into the random forest. The result is the most classified result in all decision trees. The advantage of random forest that it is applicable in several fields.

Decision tree theory Each branch in Random Forest represents a single decision tree without paper cutting. Flow chart of the tree represents the completion of decision tree building. The node represents the test of the feature, and the branch denotes the feature output. The class distribution is represented by the final tree node. The root node is the highest point of the decision tree. The main issue of the decision tree method is how to choose the feature dividing at the nodes of the tree to obtain the best feature selection. Decision tree begins from the root node, two subtrees are divided by each subtree and start to produce extra root node, which creates left and right subtrees. Each subtree recursion continues to create another subtree until the leaf node is reached. Several decision tree creation methods are available such as CLS, ID3, CART and other node dividing methods.

The steps of building a random forest algorithm The idea of building N decision tree is to obtain N training set. Each sampling method includes replacement and the non-replacement methods. In the non-replacement method, the first sample units are obtained from the whole design by including the related signs of the unit. The whole number of units is reduced when the sampling method is not iterated in the selection process. There is no possibility of the pumping repeating for each unit. The replacement sampling is the opposite of non-replacement one. When sampling process is completed, the original dataset is generated by replacing the samples into the original dataset without any changes. Two types of replacement sampling are available including the unweighted and the weighted sampling.

Table 2. Confusion matrix of binary data

Classification	Actual positive	Actual negative
Positive classified	TP	FP
Negative classified	FN	TN

Performance index of random forest The internal and external factors affect random forests, invariably causing inequitable classification performance. Several types of metrics measure the classification performance such as the classification effect, OOB, and running efficiency. The confusion matrix of binary data is shown in Table 2 to illustrate these metrics as follows:

TP represents the number of positive class samples, while TN denotes the number of negative class samples. Both TP and TN are applied in the correct classification. FP indicates the number of the positive class samples, whereas, FN represents the number of the negative class samples. FP and FN are employed in the incorrect classification. The following steps are given to evaluate the classification effect on the random forest classifier:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Step 1 explains the overall *Accuracy* of the classification measured by accuracy metrics, and higher accuracy represents the better classification performance.

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{FP + TN}} \quad (2)$$

Step 2 represents *G - mean* which measures the overall accuracy of the classification, and higher accuracy represents the better classification performance.

$$F - value = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall} \times 100\% \quad (3)$$

Fvalue in step 3 is a type of classification evaluation metrics which describes the recall and precision.

4 Experimental Results

4.1 Validating the Existing Work Experiment

Ojha and Goel [13] assessed the performance of SVM and C5.0 by splitting the Wisconsin Prognostic Breast Cancer (WPBC) dataset set [8] into 70% for the training set and 30% for the test set. Their experiment is described in the related work section. We authenticated the performance of SVM and C5.0 with the same classification technique by dividing the dataset into 70% train data and 30% test data. In this experiment, we applied 20 iterations to calculate the average and standard deviation of accuracy, precision, and recall metrics. R platform was employed to predict the probability of recurrence among breast cancer patients. We applied two packages in this experiment. Package *probsvm* [21] for SVM, and Package *C50* [22] for C5.0 model were used. Six parameters for SVM were employed such as fold, kernel, Kparam, Inum, type, and lambdas, while all default parameters for C5.0 were applied such as subset, bands, winnow, noGlobal Pruning, Confidence Factor (CF), min Cases, fuzzyThreshold, sample, early Stopping, seed, and label. We utilized the accuracy, precision, and recall measures to assess the performance of the models.

Table 3 [13] indicates that SVM model shows best accuracy rate of $77.19 \pm 0.00\%$, while C5.0 presents lower accuracy rate of $74.74 \pm 5.16\%$. C5.0 illustrates the best precision of $83.41 \pm 3.10\%$ compared to the lowest precision stated by

Table 3. Prediction summaries of SVM and C5.0 models

Classifier	Accuracy (%)	Precision (%)	Recal (%)
SVM	77.19 ± 0.00	77.19 ± 0.00	100
C5.0	74.74 ± 5.16	83.41 ± 3.10	84.09 ± 6.51

SVM, 100% is the recall rate recorded by SVM, whereas, C5.0 showed less recall rate at $84.09 \pm 6.51\%$. Again, 20 iterations were applied to calculate the average and standard deviation by utilizing grid search to optimize the parameters of SVM and C5.0 models. The values of the SVM parameters after optimization are as follows: fold = 10, kernel = radial, Kparam = 0.01, Inum = 30, type = ovo, and lambdas = 0.000976, 1, 1024, while, the parameter values for C5.0 after optimization process are as follows: subset = TRUE, bands = 0, winnow = FALSE, no-Global Pruning = FALSE, CF = 0.25, minCases = 2, fuzzyThreshold = FALSE, sample = 0, seed = sample.int (4096, size = 1) - 1L, early Stopping = TRUE, and label = “outcome”.

Table 4. Prediction summaries of SVM and C5.0 models after optimization process

Classifier	Accuracy (%)	Precision (%)	Recall (%)
SVM	76.71 ± 1.26	77.29 ± 2.12	98.62 ± 2.53
C5.0	74.21 ± 6.42	82.46 ± 4.70	84.48 ± 7.19

From Table 4, SVM showed best accuracy rate ($76.71 \pm 1.26\%$) while, C5.0 showed lowest accuracy rate ($74.21 \pm 6.42\%$). C5.0 model achieved the best precision rate ($82.46 \pm 4.70\%$) compared to lower precision rate ($77.29 \pm 2.12\%$) presented by SVM. Best recall rate ($98.62 \pm 2.53\%$) was achieved by SVM while C5.0 showed lower recall rate of $84.48 \pm 7.19\%$ (Fig. 1).

4.2 Validating Our Work

In our experiment, we independently validated the RF and DNN models by splitting the data into two groups for example train data (70%) and test data (30%). To evaluate the performance of these models, several performance metrics were employed such as accuracy, precision, and recall. Our proposed method achieved higher accuracy compared to the existing work. Figure 1 visualizes that the RF model states best accuracy rate of $98.63 \pm 2.56\%$. Although, DNN model shows lowest accuracy rate of $77.44 \pm 1.22\%$, the precision rate of RF and DNN models is $98.81 \pm 2.09\%$, $77.40 \pm 1.0\%$ respectively, while recall rate is 100, $99.45 \pm 1.73\%$ respectively.

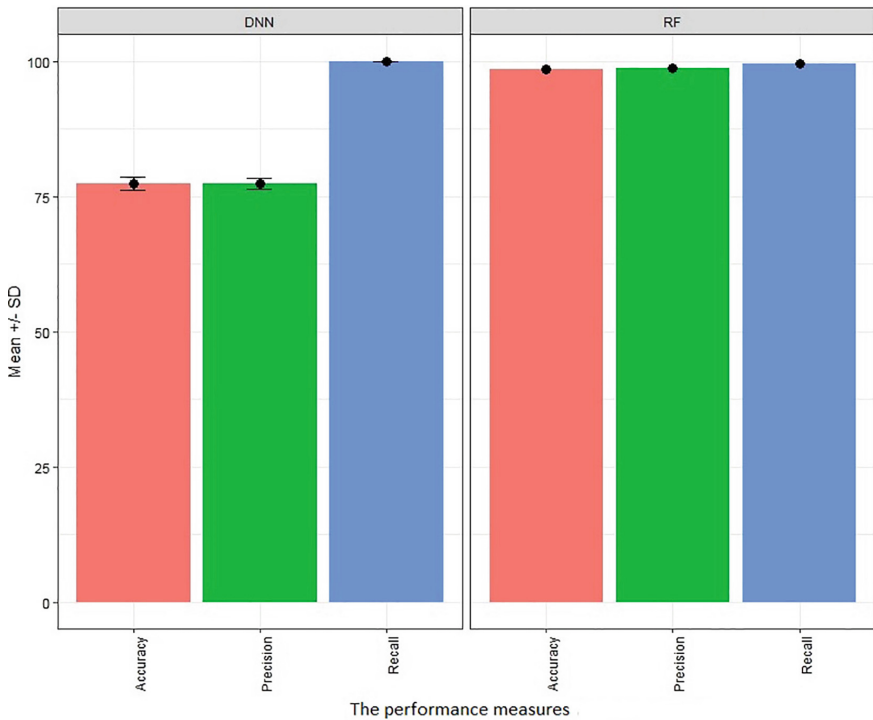


Fig. 1. Graph of prediction summaries of RF and DNN techniques

5 Conclusion

Early detection of breast cancer has become a very important topic in the field of data mining. An accurate probability of breast cancer recurrence plays a vital role in preventing the breast cancer recurrence. In this paper, Wisconsin Prognosis Breast Cancer data set was collected from UCI machine learning Repository. To balance the data, a SMOTE technique was employed. The recurrence statuses are represented by R and N . We have listed the results of extensive prediction models for assessing the probability of recurrence status of breast cancer patients. Holdout technique was applied by splitting the dataset into 70% training set and 20% for test sets. RF model showed better accuracy performance than DNN compared to the existing work.

References

1. AIHW.: Cancer in Australia (2017). <http://www.aihw.gov.au/publication-detail/?id=60129558547>
2. Mehrotra, J., Vali, M., McVeigh, M., Kominsky, S.L., Fackler, M.J., Lahti-Domenici, J., Polyak, K., Sacchi, N., Garrett-Mayer, E., Argani, P.: Very high

- frequency of hypermethylated genes in breast cancer metastasis to the bone, brain, and lung. *Clin. Cancer Res.* **10**(9), 3104–3109 (2004)
3. Ohno-Machado, L.: Modeling medical prognosis: survival analysis techniques. *J. Biomed. Inf.* **34**, 428–439 (2001)
 4. Skevofilakas, M., Nikita, K., Templealexis, P., Birbas, K., Kaklamanos, I., Bonatsos, G.: A decision support system for breast cancer treatment based on data mining technologies and clinical practice. In: *Engineering in Medicine and Biology Society*, 2005. *IEEE-EMBS 2005. 27th Annual International Conference*, pp. 2429–2432 (2005)
 5. Yi, W., Fuyong, W.: Breast cancer diagnosis via support vector machines. In: *Control Conference: CCC 2006. Chinese*, pp. 1853–1856 (2006)
 6. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intel. Med.* **34**(2), 113–127 (2005)
 7. Sobran, N.M.M., Ahmad, A., Ibrahim, Z.: Classification of imbalanced dataset using conventional naive bayes classifier. In: *International Conference on Artificial Intelligence in Computer Science and ICT*, pp. 35–42 (2013)
 8. Lichman, M.: *UCI: Machine Learning Repository* (2013). <http://archive.ics.uci.edu/ml>
 9. He, H., Shen, X.: A Ranked Subspace Learning Method for Gene Expression Data Classification, *IC-AI*, pp. 358–364 (2007)
 10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intel. Res.* **16**, 321–357 (2002)
 11. Khalilia, M., Chakraborty, S., Popescu, M.: Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inf. Decis. Mak.* **11**(1): 51 (2011)
 12. Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J.: Training deep neural networks on imbalanced data sets. In: *Neural Networks (IJCNN)*, vol. 01, pp. 4368–4374 (2016)
 13. Ojha, U., Goel, S.: A study on prediction of breast cancer recurrence using data mining techniques. In: *7th International Conference on Cloud Computing, Data Science and Engineering–Confluence*, pp. 527–530 (2017)
 14. Kim, Woojae, Sang, Kim, Ku, Lee, Jeong Eon, Noh, Dong-Young, Kim, Sung-Won, Jung, Yong Sik, Park, Man Young, Park, Rae Woong: Development of novel breast cancer recurrence prediction model using support vector machine. *J. Breast Cancer* **15**(2), 230–238 (2012)
 15. Salama, G.I., Abdelhalim, M.B., Zeid, M.A.: Experimental comparison of classifiers for breast cancer diagnosis. In: *7th International Conference on Computer Engineering and Systems (ICCES)*, pp. 180–185 (2012)
 16. Tomczak, J.M.: Prediction of Breast Cancer Recurrence Using Classification Restricted Boltzmann Machine with Dropping. [arXiv:1308.6324](https://arxiv.org/abs/1308.6324) (2013)
 17. Chaurasia, Vikas, Pal, Saurabh: Prediction of breast cancer recurrence using classification restricted boltzmann machine with dropping. *Int. J. Comput. Sci. Mob. Comput.* **3**(1), 10–22 (2014)
 18. Beheshti, Z., Shamsuddin, S.M.Hj., Beheshti, E., Yuhani, S.S.: Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis. *Soft Comput.* **18**(11), 2253–2270 (2014)
 19. De sa Marques, J.P.: *Pattern Recognition: Concepts, Methods and Applications*, Springer, Berlin (2012)
 20. Ho, T.K.: Random decision forests. In: *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282 (1995)

21. Shin, Seung Jun, Wu, Yichao, Zhang, Hao Helen: Two-dimensional solution surface for weighted support vector machines. *J. Comput. Graph. Stat.* **23**(2), 383–402 (2014)
22. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann San Mateo. CA, Google Scholar (2014)

A New Theoretical Framework for Testing Consciousness in a Machine

Azree Nazri^(✉), Abdul Azim Abd Ghani, Izuan Hafez,
and Keng-Yap Ng

Faculty of Computer Science and Information Technology, Universiti Putra
Malaysia, 43400 Serdang, Selangor, Malaysia
{azree, azim, mohdizuan, kengyap}@upm.edu.my

Abstract. The major aim of artificial general intelligence's (AGI) is to allow a machine to perform general intelligence tasks similar to human counterparts. Hypothetically, this general intelligence in a machine can be achieved by establishing cross-domain optimization and learning machine approaches. However, contemporary artificial intelligence (AI) capabilities are only limited to narrow and specific domains utilizing machine learning. Consciousness concept is particularly interesting topic to attain the approaches because it simultaneously encodes and processes all types of information and seamlessly integrates them. Over the last several years, there has been a resurgence of interest in testing theories of consciousness using computer models. The studies of these models are classified into four categories: external behavior associated with consciousness, cognitive characteristics associated with consciousness, a computational architecture correlate of human consciousness and phenomenally of conscious machine. The critical challenge is to determine whether these artificial systems are capable of conscious states by providing a measurement the extent to which the systems are succeeded in realizing consciousness in a machine. Several tests for machine consciousness have been proposed yet their formulation is based on extrinsic measurement of consciousness. Yet extrinsic measurement is not inclusive because many conscious artificial systems behave implicitly. This research proposes a new framework to test machine consciousness based on intrinsic measurement so-called *Pak Pandir* test. The framework leverages three quantum double-slit settings and information integration theory as consciousness definition of choice.

Keywords: Turing test · Artificial intelligence · Artificial general intelligence
Pak Pandir · Consciousness · Machine consciousness

1 Introduction

Making machines as intelligent as human is undeniable the most difficult problems for scientists and engineers to solve. While this is the original goal of Artificial Intelligence (AI) field, most contemporary AI studies are application-specific and domain-specialized. Artificial General Intelligence (AGI) emerges to fill in the gap by widening the gap. The aim of AGI is to create machine that capable of general-purpose intelligence comparable to that of the human mind. Machine consciousness is

important to AGI because consciousness concept provides two platforms to AGI; (1) a platform for efficient cross-domain optimization and (2) a platform for learning machine.

Domains-specific and application-specialized systems are currently the best implementation for AI. However, human is not domain-specific problem solvers. Human can concurrently execute many general tasks such as solving jigsaw puzzle while recognizing the images of the puzzles on the floor. Therefore, cross-domain optimization method is needed to achieve this general intelligence. While, Chinese Room Argument [1] raised doubt about contemporary artificial intelligent algorithms that stated the algorithms do not achieve real intelligence. The argument sternly specified that by merely manipulating encoded symbols are not the acts of intelligence. This machine learning algorithm does not produce new pattern or rules but merely finding the hidden pattern in learning data. To make a machine the ability to derive its own rules or patterns, learning machine concept is required. Learning machine contains an algorithm that can independently derive its own rules and patterns. It has been said that learning machine needs awareness to operate, in which a component of consciousness.

This research adapts information integration theory (IIT) as a consciousness concept of choice. Consciousness is stated by information integration theory is that the result of special kind of information integration [2, 3]. Moreover, IIT focuses on intrinsic characteristics of consciousness. Conscious artificial systems can have both extrinsic and intrinsic characteristics yet all machine consciousness tests are extrinsic features. Table 1 shows four categories of conscious artificial systems (CAS). It can be seen from Table 1 that machine consciousness can be measured in two ways: extrinsic or intrinsic. Extrinsic measurement is measuring system’s external behavior because it is the only guide to phenomenal states. In contrast, intrinsic measurement is measuring system’s internal state that are capable of conscious states. From Table 1, only CAS1 requires extrinsic measurements while the other CAS are measured by intrinsic measurement.

Table 1. Types of conscious artificial system (CAS) and their measurements

Type	Description	Measurement
CAS1	Artificial systems with the external behavior associated with consciousness	Extrinsic
CAS2	Artificial systems with the cognitive characteristics associated with consciousness	Intrinsic
CAS3	Artificial systems with an architecture that is claimed to be a cause or correlate of human consciousness	Intrinsic
CAS4	Phenomenally conscious machine	Intrinsic

Table 2 shows the computational models for conscious machine that fall into CAS1, CAS2, CAS3 and CAS4 include CRONOS [13], Cyberchild [15], Khepera [13, 14], global workspace models [6], agent-based conscious architecture [7–9], Haikonen’s [10], CiceroBot [12], CofAff Schema [16], LIDA [17] and CODAM [18].

Table 2. Computational model for conscious machine

Types	Artificial systems
CAS1	Cog [4, 5], global workspace models [6], agent-based conscious architecture [7–9], Haikonen’s [10], Schema-based model [11], Cicerobot [12]
CAS2	CRONOS [13], Cog [4, 5], Khepera [13, 14], global workspace models [6], Haikonen’s [10], Schema-based model [11], Cicerobot [12]
CAS3	CRONOS [13], CyberChild [15], global workspace models [6], Haikonen’s [10], Schema-based model [11], CofAff Schema [16], LIDA [17], CODAM [18]
CAS4	CRONOS [13], CyberChild [15], Khepera [13, 14], global workspace models [6], agent-based conscious architecture [7–9], Haikonen’s [10], Cicerobot [12]

CAS1 is aimed to replicate conscious human behavior that use first-order logic to generate the behavior. Conscious human behavior is important in understanding of machine consciousness because the researchers need to ascribe qualia to a machine to understand them and the only guide to these qualia is a system’s external behavior. The computational models of CAS1 are materialized in Cog [4, 5], global workspace models [6], agent-based conscious architecture [7–9], Haikonen’s [10], Schema-based model [11], Cicerobot [12]. CAS2 connects consciousness and cognitive characteristics such as imagination, emotions and a self. The simulation of the cognitive associated with consciousness has been a strong theme in machine consciousness includes CRONOS [13], Cog [4, 5], Khepera [13, 14], global workspace models [6], Haikonen’s [10], Schema-based model [11], Cicerobot [12]. CAS3 links the simulation architectures and human consciousness. The motivation of this research is to test neural or cognitive theories of consciousness includes CRONOS [13], CyberChild [15], global workspace models [6], Haikonen’s [10], Schema-based model [11], CofAff Schema [16], LIDA [17] and CODAM [18]. CAS4 is relatively controversial because CAS1, CAS2 and CAS3 do not claim about real phenomenal states but CAS4 is concerned with an artificial system that have real phenomenal experiences and is not just tools in consciousness research. These include CRONOS [13], CyberChild [15], Khepera [13, 14], global workspace models [6], agent-based conscious architecture [7–9], Haikonen’s [10], Cicerobot [12]. Most of the studies include CAS1-4 into their models to achieve synthetic phenomenology.

However, systems with real consciousness cannot be developed without methods for measuring and debugging phenomenal states. This research tries to address this problem with one question; what would be better test for machine consciousness that is defined in bits? In answering these questions, this research proposes a theoretical framework for testing machine consciousness by leveraging quantum double-slit mechanism. In summary, this research makes the following main contribution:

- As many conscious artificial systems shows non-behavioral property of consciousness, this research proposes a new theoretical framework for testing machine consciousness that evaluate non-behavioral/intrinsic property of consciousness. This new framework is called *Pak Pandir* test.

The remainder of the paper is organized as follows. Section 2 briefly reviews the related works, Sect. 3 elaborates *Pak Pandir* test for machine consciousness to measure intrinsic property of consciousness and Sect. 4 explains the expected results from the series of experiments.

2 Related Works

Machine consciousness is a widely investigated area and there are many attempts to define consciousness [19–22]. Some advocates of machine consciousness considered consciousness as causal/non-causal [22], accessible/inaccessible [23], stateless/having physical state [24], representational/non-representational [25]. Some studies refer consciousness as virtual machines that catalyze the actions of conscious entities [26]. The previous proposed computational models for conscious machine are shown in Table 2 and replicates some of the aforementioned definition of consciousness. These computational models of consciousness are yet to be tested for synthetic phenomenology. Synthetic phenomenology refers to the synthesizing of phenomenal states.

Typically, synthetic phenomenology is measured using behavioral mechanisms [27–29]. Currently, there are plenty of synthetic phenomenology tests based on their behavioral mechanisms including Glasgow Coma Scale/Simplified Motor score (GSC/SMS) [27], Private self-consciousness (PSC) [27], and mirror test [28]. However, GSC/SMS does not cover the broad range of consciousness functional components, PSC focused on specific aspects of consciousness and mirror test has been easily passed by many. Turing test is explicitly measuring synthetic phenomenology in behavioral way in which the machine has to replicate behaviors that are carried out consciously in humans. In a nutshell, none of the tests can be directly applied to intrinsic or non-behavioral aspects of consciousness. Intrinsic consciousness test is important because not all organisms and machines reveal their consciousness in the form of behaviors. Table 3 describes the contemporary consciousness tests that use behavioral mechanism.

Table 3. Test for consciousness

Test	Measurement (extrinsic/intrinsic)	References
GSC/SMS	Extrinsic (behavior)	Van de Voorde et al. [27]
PSC	Extrinsic (behavior)	Fenigstein et al. [28]
Mirror test	Extrinsic (behavior)	Lewis [29]
Turing test	Extrinsic (behavior)	

This paper hypothesizes that quantum double-slit apparatus can be utilized to test the presence of non-behavioral aspects of machine consciousness. Thus, three double-slit optical settings are proposed as a theoretical framework for testing intrinsic consciousness property.

3 *Pak Pandir* Test: A Theoretical Framework for Machine Consciousness Test

In this section, we elaborate the new theoretical framework for testing machine consciousness that based on quantum double-slit system. At first, we discuss about the result of the quantum double-slit system. Then the proposed framework is introduced together with a series of experimental setup to measure machine consciousness. Eventually, the expected outcomes of each experiment are presented.

Determining the level of consciousness of a living organism is a hard problem. Testing for consciousness remains an open problem even when it is aimed to humans or other mammals. Most contemporary neuroscientists agree that consciousness can be tackled scientifically using alternative strategies. Additionally, this research believes that establishing a framework for measuring and testing the level of consciousness of a subject is of central importance in the scientific quest for consciousness. Considering the former considerations, a theoretical framework is designed to test the level of consciousness of machine called *Pak Pandir* test.

The theoretical framework for machine consciousness test are shown in Fig. 1. To determine the presence of consciousness in machine two indicators are used as shown in Fig. 1:



Fig. 1. Double-slit experiment **a** with observers and **b** without observer

1. Before any measurement is done on the observable property of quantum system, the system is in wave-pattern form or a superposition of all possible state as shown in Fig. 1a.
2. After an observer attempts to measure the observable, the system ends up in only one of the possible states or particle pattern form and provides the observed the value of the system's property in the actual state as shown in Fig. 1b.

Figure 2 explains the connection between quantum mechanics system and conscious entities in a theoretical framework. The measuring device and lifeless quantum system in the figure is the quantum double-slit system. To measure the presence of consciousness, *Pak Pandir* test includes three experiments: which-way experiment, quantum eraser experiment and delayed quantum erasure experiment.

1. *Which-way experiment setup*: this experiment attempts to get information by determining the path of single photons when they get through screen with two slits as shown in Fig. 3. This experiment is to test the presence of weak level of consciousness. This is mainly caused by the design of the which-way experiment that

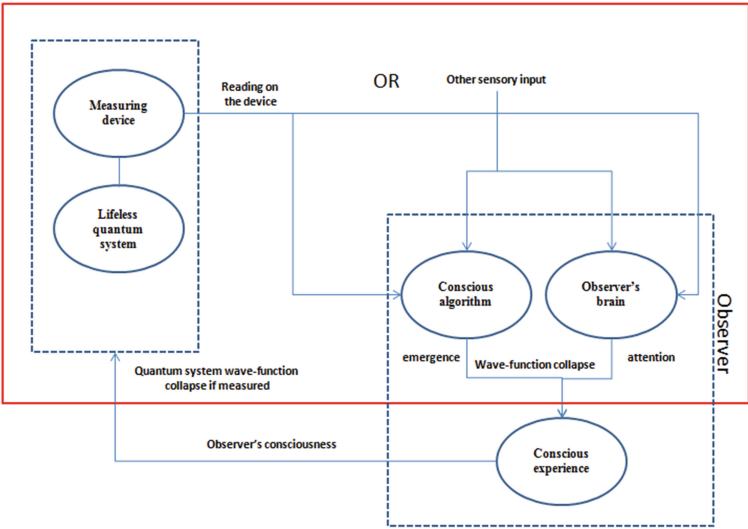


Fig. 2. The theoretical framework for machine consciousness test or *Pak Pandir* test

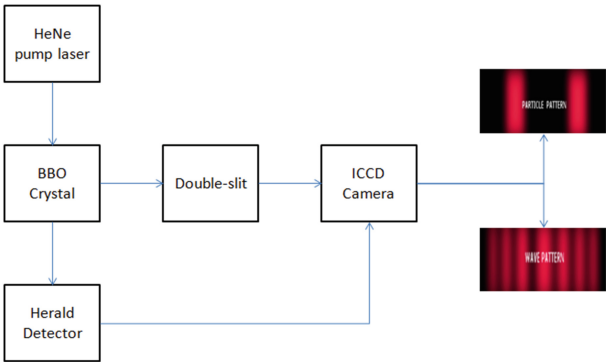


Fig. 3. Simplified schematic: which-way experiment setup

- produces experiment’s noise. The noise is caused by the way the detectors are placed, and it is still insignificant to render the experiments unreliable.
- 2. *Quantum eraser experiment*: this experiment utilizes single photons to produce entangled photon pairs each it twice the original length as shown in Fig. 4. The noises in the which-way experiment have been removed in this quantum-eraser setup. Thus, the results will to some extent accurately measures the presence of consciousness.
 - 3. *Delayed quantum erasure experiment*: This experiment incorporates concepts considered in Wheeler’s delayed choice experiment. As shown in Fig. 5, this experiment shows that delayed erasure erases *p* polarization after *s* is detected. This can be done by extending *p* distance to herald detector 2 to beyond *s* distance to herald detector 1 as shown in Fig. 5.

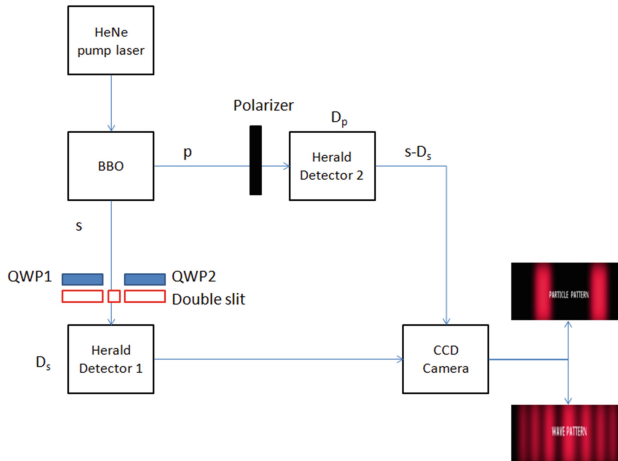


Fig. 4. Simplified schematic: quantum-eraser experiment setup

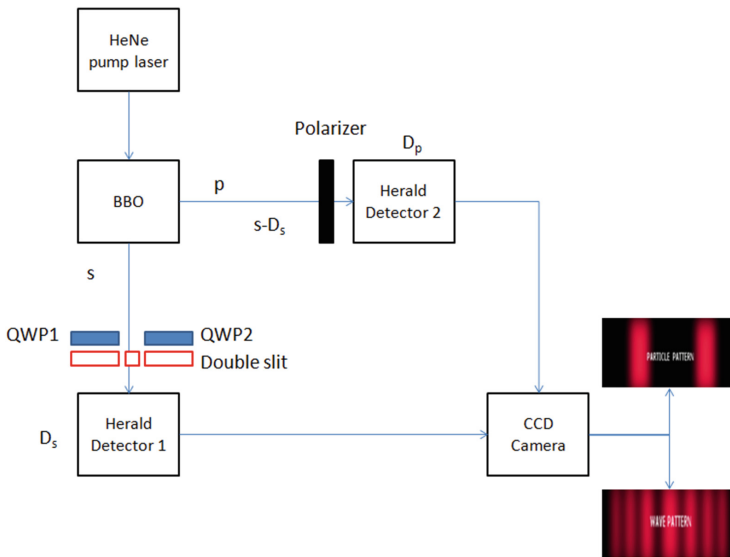


Fig. 5. Simplified schematic: delayed quantum-eraser experiment setup

4 Expected Results

In this section, we present the arrangement of *Pak Pandir* experiments and its expected results. *Pak Pandir* test will determine whether computational model reaches synthetic phenomenology. *Pak Pandir* test does not have any empirical results as it is still under development. Figure 6 shows the setup for *Pak Pandir* test. For a machine to pass *Pak Pandir* test, the results of the experiments will look like Table 4. *Pak Pandir* test has to

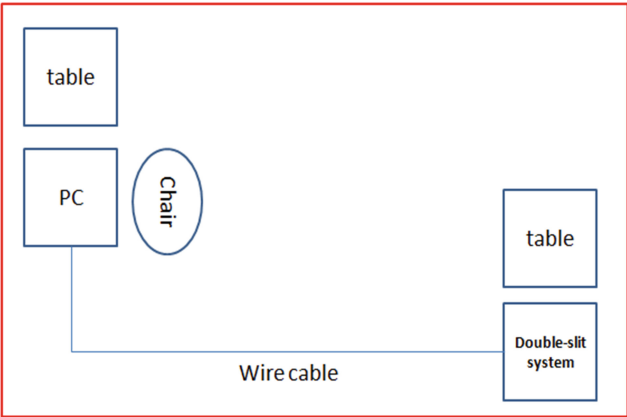








Fig. 6. The experiment will be conducted in electromagnetically shielded chamber. The PC is connected to double-slit system for the system to send signals to the PC. The human observers will sit on the chair in front of PC and the conscious algorithms will be installed on the PC

be tested on human first. The expected results from human experiments is shown in Table 4. Any algorithm or machine that can reproduce these results are considered passes *Pak Pandir* test. To declare that a computational model emerges consciousness intrinsically, the model needs to reproduce the results in Table 4.

Table 4. Quantum-wave function collapse results

<i>Pak Pandir</i>	Which-way	Erasure	Delayed
Observed			
Unobserved			

5 Conclusion

In Malay folklore, *Pak Pandir* is a common people who have been told to have low intelligence yet undisputedly is a conscious human. Thus, *Pak Pandir* framework is used to test the potential of the presence of consciousness of a given machine but not

the presence of intelligence. Identifying consciousness by means of interpreting behavior remains an open problem. However, more effort should be put in measuring the intrinsic property of consciousness concept. *Pak Pandir* test is proposed instead that is designed to measure intrinsic property of consciousness.

References

1. Aleksander, I., Burnett, P.: Thinking machines : the search for artificial intelligence. Knopf (1987)
2. Tononi, G.: An information integration theory of consciousness. *BMC Neurosci.* **5**(1), 42 (2004)
3. Tononi, G.: Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* **215**(3), 216–242 (2008)
4. Brooks, R.A., Breazeal, C., Marjanović, M., Scassellati, B., Williamson, M.M.: The Cog project: building a humanoid robot, pp. 52–87 (1999)
5. Dennett, D.C.: Consciousness in human and robot minds. *Cogn. Comput. Conscious.* **1**, 1–10 (1994)
6. Dehaene, S., Kerszberg, M., Changeux, J.-P.: A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci.* **95**(24), 14529–14534 (1998)
7. Steels, L.: Language games for autonomous robots. *Intell. Syst. IEEE* **16**(5), 16–22 (2001)
8. Steels, L.: Language re-entrance and the ‘inner voice’. *J. Conscious. Stud.* **10**(4), 173–185 (2003)
9. Clowes, R.: The problem of inner speech and its relation to the organization of conscious experience: a self-regulation model. In: *Proceedings AISB06 Symposium Integrative Approaches to Machine Consciousness*, pp. 117–126 (2006)
10. Haikonen, P.O.: The cognitive approach to conscious machines (2003)
11. Samsonovich, A., DeJong, K.: A general-purpose computational model of the conscious mind. In: *ICCM—Six International Conference Cognitive Model*, pp. 382–383 (2004)
12. Chella, A., Macaluso, I.: Sensations and perceptions in Cicerobot, a museum guide robot
13. Holland, O., Goodman, R.: Robots with internal models a route to machine consciousness? *J. Conscious. Stud.* **10**(4), 77–109 (2003)
14. Stening, J., Jacobsson, H., Ziemke, T.: Imagination and abstraction of sensorimotor flow: towards a robot model. In: *Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity and Embodiment*, pp. 50–58 (2005)
15. Cotterill, R.M.J.: CyberChild a simulation test-bed for consciousness studies. *J. Conscious. Stud.* **10**(4–5), 31–45 (2003)
16. Sloman, A., Chrisley, R.: Virtual machines and consciousness. *J. Conscious. Stud.* **10**(4–5), 133–172 (2003)
17. Baars, B.J., Franklin, S.: Consciousness is computational: the lida model of global workspace theory. *Int. J. Mach. Conscious.* **1**(1), 23–32 (2009)
18. Taylor, J.G.: Beyond consciousness? *Int. J. Mach. Conscious.* **1**(1), 11–21 (2009)
19. Chrisley, R.: Philosophical foundations of artificial consciousness. *Artif. Intell. Med.* **44**(2), 119–137 (2008)
20. Haikonen, P.O.A.: Reflections of consciousness: the mirror test. *Proc. 2007 AAAI Fall Symp. Conscious.* **9**(4), 67–71 (2007)
21. Sun, R.: Learning, action and consciousness: a hybrid approach toward modelling consciousness. *Neural Netw.* **10**(7), 1317–1331 (1997)

22. Velmans, M.: Making sense of causal interactions between consciousness and brain: reply. *J. Conscious. Stud.* **9**(11), 69–95 (2002)
23. Anderson, M.L.: Circuit sharing and the implementation of intelligent systems. *Conn. Sci.* **20**(4), 239–251 (2008)
24. Torey, Z.L.: The immaculate misconception. *J. Conscious. Stud.* **13**(12), 105–110 (2006)
25. Browne, C., Evans, R., Sales, N., Aleksander, I.: Consciousness and neural cognizers: a review of some recent approaches. *Neural Netw.* **10**(7), 1303–1316 (1997)
26. The virtues of virtual machines—Google Scholar. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=The+virtues+of+virtual+machines+&btnG. Accessed 04 Oct 2017
27. Van de Voorde, P., et al.: Assessing the level of consciousness in children: a plea for the glasgow coma motor subscore. *Resuscitation* **76**(2), 175–179 (2008)
28. Fenigstein, A., Scheier, M.F., Buss, A.H.: Public and private self-consciousness: assessment and theory. *J. Consult. Clin. Psychol.* **43**(4), 522–527 (1975)
29. Lewis, M.: The emergence of consciousness and its role in human development. *Ann. N. Y. Acad. Sci.* **1001**, 104–133 (2003)

Temporal Based Factorization Approach for Solving Drift and Decay in Sparse Scoring Matrix

Al-Hadi Ismail Ahmed Al-Qasem¹, Nurfadhlin Mohd Sharef^{2(✉)},
Sulaiman Md Nasir², and Mustapha Norwati²

¹ Amran University, Amran, Yemen
esmail.hadi2009@gmail.com

² Faculty of Computer Science and Information Technology, University Putra
Malaysia, UPM, 43400 Serdang, Selangor, Malaysia
{nurfadhlin, nasir, norwati}@upm.edu.my

Abstract. Collaborative filtering (CF) is one of the most popular techniques of the personalized recommendations, where CF generates personalized predictions in the rating matrix. The rating matrix typically contains a high percentage of unknown rating scores which is called the sparsity problem. The matrix factorization approach through temporal approaches has the accurate performance in addressing the sparsity issue but still with low accuracy. However, there are four issues when a factorization approach is adopted which are latent feedback learning, score overfitting, user's interest drifting and item's popularity decay over time. Therefore, this work introduces the temporal based factorization approach named TemporalMF++ to address all the issues. The experimental results show the TemporalMF++ approach has a higher prediction accuracy compared to the benchmark approaches. In summary, the TemporalMF++ approach has a superior effectiveness in improving the accuracy prediction of the CF by learning the temporal behaviour.

Keywords: Collaborative filtering · Matrix factorization · Temporal Drift · Decay · Bacterial foraging

1 Introduction

Recommendation system (RS) is becoming popular due to its great utility of users interests to recommend personalized items [1]. Collaborative filtering (CF) is one of the most popular techniques of the personalized recommendations, where CF generates personalized predictions based on the similarities among members in the rating matrix. However, the rating matrix contains a high percentage of unknown rating scores which lowers the quality of the prediction. The similarity evaluation among the common users will be impossible or not reliable when the percentage of sparse rating scores are high which lower the quality of the prediction [2].

Several factorization approaches have addressed the sparsity problem [3, 4]. The factorization approaches use the latent feedback of preferences, the baseline and the latent factors which are the combination between the baseline variables and the latent

feedback in several formulas. The Singular Value Decomposition (SVD) is one of the algorithms which provide the latent feedback of preferences. During streaming the rating scores of the users into the memory, some rating scores have misplaced from its appropriate cell in the rating matrix which lower the quality of the latent feedback and this limitation has been solved by the Ensemble Divide and Conquer approach [5] which is used to learn the accurate latent feedback of preferences.

The temporal recommendation system is used to recommend the items to the users at a suitable time where the time is a significant factor to learn the interest of users and the popularity of items over time [1, 3]. The time-aware approaches [6] analyzes the interest of users using their information and preferences during temporal periods via several rules which are suitable for several types of information but with high cost. Besides, the temporal based matrix factorization is used to learn the user preferences and temporal preferences to generate the accurate recommendations. The temporal based matrix factorization is one of the successful collaborative-based approaches compared to the factorization based approaches in addressing data sparsity. The Temporal Dynamics approach [7] has improved the prediction accuracy of the CF, where this approach divides the time of preferences into static numbers of bins while the time preferences are changed over time. This approach learns a global weight based on the stochastic gradient descent algorithm for minimizing the overfitting. However, the global weight has a weakness in term of personalized preferences, which motivates to find other weights based on personalized preferences and find the suitable learning algorithm for learning these weights.

The CF technique has been improved by the temporal interaction approach [8] which integrate between the long-term preferences and the short-term preferences. The short-term preferences are represented by the shrunk neighbour method. In addition, the short-term technique based on the shrunk neighbour [8, 9] gives the short-term feedback of users and incorporated it into the baseline variants which suffer from overfitting [7]. The overfitting in the predicted rating scores means a few of the predicted values are bigger than the range scale of the rating scores, e.g. the range scale of Movie Lens dataset [0–5] when the predicted rating scores are such as {5.3; 6.2; 5.5}. Although the overall performance of the prediction accuracy of the Temporal Interaction approach [8] is better than the Short-Term based Latent approach [9], the performance of the prediction accuracy is still poor and it needs more improvements. Besides, the concept drift is the most significant challenge for the temporal recommendation system where the customer interest is drifting over time [8, 10]. The prediction accuracy in the sparse rating matrix is still low by the latent feedback of the preferences due to concept drift in the users' preferences and also in the popularity of items over time. Therefore, the TemporalMF++ approach is proposed for solving the current limitations of temporal approaches.

The paper is organised as follows. Firstly, we introduce the background about the CF, the matrix factorization (MF), and the temporal-based factorization approaches. Secondly, we explain the methodology of the TemporalMF++ approach and how to integrate the long-term and short-term preferences. Finally, we describe the results of the proposed approach compared to the validation approaches.

2 Related Work

2.1 Collaborative Filtering

The CF technique generates the personalized recommendations based on the similarities of common users with the active user in the rating matrix. The personalized recommendation is becoming a common technique due to its magnificent utility for recommending the items to the users based on their interest. There are three stages in the CF technique where the first stage is computing the similarity values between the active user and all common users using the similarity functions such as Cosine similarity [11]. The second stage is using the similarity values to compute the predicted value of each item rated by the target user based on the neighbours of this item. The third stage is computing the prediction accuracy by the root mean squared error function (RMSE) [11]. The functions of Cosine, prediction and RMSE are described in [5]. In addition, The Cosine similarity, prediction and RMSE functions will be used for evaluating all the evaluation approaches and the TemporalMF++ approach.

2.2 Bacterial Foraging Optimization Algorithm

The swarming algorithms have been used in RS for learning the features of users and items such as the genetic algorithm [12] and the particle swarm optimization algorithm [13]. Therefore, the bacteria foraging optimization algorithm (BFOA) will be used for solving the challenges of RS based on CF. BFOA is an evolutionary computing method which uses the behaviours of *E. coli* bacteria during the foraging process in the human intestine, and it used for global optimization [14], and it can potentially produce the effective solutions for very large scale problems. There are three main stages during the swarming process of bacteria as follow:

1. Chemotaxis (Tumbling, Swimming, Swarming)
2. Reproduction
3. Elimination and Dispersal.

The description of each stage and more details have been described in [15].

2.3 *k*-Means Clustering Algorithm

The *k*-means is a popular clustering algorithm which is used for controlling the big area of optimization according to the personal behaviours of members. It is impossible for the huge matrix to use weight for each member. Therefore, *k*-means is used to reduce the huge dimensions for controlling the optimization area using few weights based on clustering technique. The *k*-means is one of the widely used iterative optimization algorithm, and it is observed as a popular clustering approach, due to its integrity of execution [16]. The convergence distance in this work is measured by the squared Euclidean distance. The TemporalMF++ approach relies on the *k*-means algorithm and the BFOA.

2.4 Matrix Factorization

Recently, MF has become a popular approach for CF [4] because it is one of the successful approaches that address data sparsity problem [3]. A few of the MF approaches integrate the latent feedback of the users and latent feedback of the items by several approaches such as [5]. Besides, several MF approaches integrate between the latent feedback and the baseline features of users and items by several approaches such as temporal dynamics approach [7–9]. The MF approach is used in predicting the missing values in the rating matrix as showed in Eq. 1,

$$\hat{r}_{ui} = \mu + b_u + b_i + p_u q_i^T, \quad (1)$$

where \hat{r}_{ui} represents the prediction value of the missing value, p represents the matrix of users' latent feedback, q^T represents the matrix transpose of items' latent feedback, μ represents the global average of rating scores, and b_u and b_i represent the observed deviations of user u and item i respectively. The factorization factors μ, b_u, b_i, p_u, q_i and q_i^T have integrated with other factors in several factorization approaches [4, 17, 18] and temporal approaches [8–10] to predict the sparse rating scores. For example, the Neighbours based Baseline approach [8] integrates the factors of baseline with the distance between the rating scores and the base features of the neighbours who provide the rating scores for each item as shown in Eq. 2,

$$\hat{r}_{ui} = b_{ui} + \sum_{x \in N_i} \text{sim}_x(r_{xi} - b_{xi}) / \sum_{x \in N_i} \text{sim}_x, \quad (2)$$

where N is the set of users who provide item i by rating scores, sim_x is the similarity value between user x and active user.

2.5 Temporal

Temporal RS is designed to recommend the items to the users at a suitable time where the time is an essential factor in making the final decision and it is used in several approaches to get the accurate predictions. There is two kinds of temporal preferences which are long-term and short-term preferences. The difference between long-term preferences is that it utilizes the whole recorded preferences while the short-term preferences utilizes the recorded preferences within a session (e.g. week, month, season, etc.). The long-term approach [8] integrates the long-term preferences within the baseline features. Besides, the Short-Term based Latent approach [9] and the Short-Term based Baseline approach [8] learn the short-term preferences using the latent feedback and the baseline features of neighbours respectively, where the neighbours are defined according to the rating scores which are provided during a session. In addition, more details about the matrix factorization and the temporal approaches have been described in [19].

3 Temporal-Based Factorization Approach for Recommender System

The long-term preferences is defined according to the latest timestamp value and the earliest timestamp value where each rating score has one timestamp value. The timestamp value can provide the factorization factors, by the small weights between 0 and 1. The exponential function is used to define the temporal weights of the rating matrix based the personal taste. There are temporal vectors of long-term in the time matrix. The first vector is the personal weight for each user in the timestamp matrix where can define by Eq. 3,

$$\lambda_u = \exp[-(t_e^u - t_s^u)/t_e^u], \quad (3)$$

where t_s^u and t_e^u denote to the first time and last time of providing rating scores by user u . Therefore, each row in the timestamp matrix has temporal weight λ_u for the interaction of user u . The third weight for the tasting of each item by the set of users, which proving this item by set of rating scores. This temporal weight λ_i is extracted by Eq. 4,

$$\lambda_i = \exp[-(t_e^i - t_s^i)/t_e^i], \quad (4)$$

where t_s^i and t_e^i denote to the first time and last time of the set of users proving the item i by the rating scores, and each column in the timestamp matrix has λ_i . The short-term preferences have been recognized by dealing with the timestamp convergence in the timestamp matrix using k -means algorithm. The k -means estimates k based on the number of short-term periods in the timestamp matrix, and the timestamp matrix is divided into k clusters. The periods of short-term are used in this approach to generate optimize temporal space which is integrated with the latent space for learning the temporal behaviours such as the drift in the users' preferences. The number of clusters are estimated based on the number of short-term periods as shown in Table 1.

Table 1. The periods of short-term preferences using Netflix dataset

Total no. of days	Short-term		Clusters' no. (k)	Success clustering
	Period type	Period no.		
2190 days	One year	6	6	✓
	One season	24	24	✓
	One month	73	73	✓
	Two weeks	156	156	✗

The dataset of Netflix uses three periods which are one year, one season and one month, but the period of two weeks is not a success for clustering using the k -means algorithm where the number of clusters is bigger than the number of users or the number of items in some rating matrix (clustering problem). The time convergence

among users and the time convergence among items are used to define the short-term preferences as shown in Fig. 1.

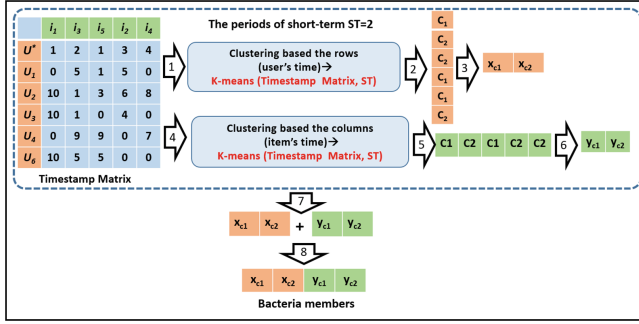


Fig. 1. An example of creating a group of bacteria in TemporalMF++

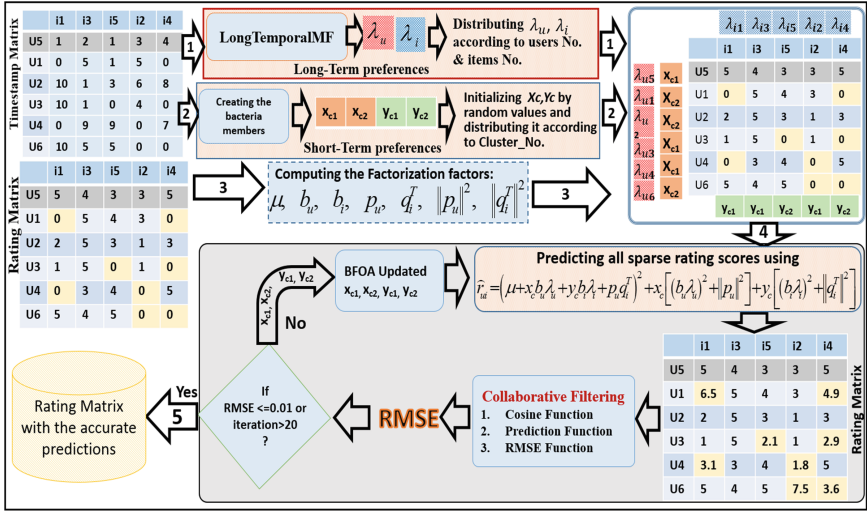


Fig. 2. The framework of TemporalMF++ approach

The weights of short-term can be integrated with the weights of long-term by the TemporalMF++ approach to learn the drift in the user's behaviors and the decay in the item's behaviors over time. Integrating the weights of short-term and long-term during the optimization process using Eq. 5,

$$\begin{aligned} \hat{r}_{ui} = & (\mu + x_c b_u \lambda_u + y_c b_i \lambda_i + p_u q_i^T)^2 \\ & + x_c [(b_u \lambda_u)^2 + \|p_u\|^2] + y_c [(b_i \lambda_i)^2 + \|q_i^T\|^2], \end{aligned} \quad (5)$$

where x_c acts as the users' weight of the cluster x_c , y_c acts as the items' weight of the cluster y_c , and the vectors of λ_u and λ_i acts as the weights of users and items based on long-term respectively. In the first part of Eq. 5, the weights of x_c and λ_u are integrated for learning the drift of users' taste which affecting on the baseline of users. Besides, the weights of y_c and λ_i are integrated for learning the time decay of the popularity of items which affecting on the baseline of items. In the second part, the weight of x_c minimizes the overfitting of the latent feedback of users, which combine between the baseline of users under the long-term effects and the norm of latent feedback of users' preferences. In the third part, the weight of y_c minimizes the overfitting of the latent feedback of items which combine between the baseline of items under the long-term effects and the latent feedback of items. Equation 5 acts as the contribution of TemporalMF++ approach. In addition, BFOA learns the significance of each short-term weight by the RMSE value which acts as the fitness value. Figure 2 shows the framework of TemporalMF++ approach (Table 2).

Table 2. The parameters' values of BFOA

Parameter	Value	Parameter	Value
No. of bacteria groups S	6	The length of a swim	4
No. of iteration	20	Reproduction steps	4
Optimum RMSE	0.01	Elimination-dispersal	4
P: cluster no. by one month	72	Probability of elimination-dispersal	0.25
P: cluster no. by one season	24	d_{attract}	0.1
P: cluster no. by one year	6	w_{attract}	0.2
Run length unit C_i	0.1	$h_{\text{repellant}}$	0.1
No. of chemotactic steps j	6	$w_{\text{repellant}}$	5

The main stages of TemporalMF++ framework are summarized as follows:

1. Preparing the rating matrix and the timestamp matrix based on the target user activities and his common users.
2. Learn the long-term preferences λ_u and λ_i by Eqs. 4 and 5 respectively.
3. Learn the short-term preferences x_c and y_c by dividing the timestamp matrix into k clusters based the periods of each short-term and creating groups of bacteria members as shown in Fig. 1.
4. Distributed the temporal weights x_c on the users and y_c on the items in of the rating matrix according to the cluster number of each.
5. Learn the factorization factors from the rating matrix.
6. Applying Eq. 5 for predicting all sparse rating scores in the rating matrix.
7. Computing the RMSE for the rating matrix based the CF technique.
8. Learn the accurate temporal weights x_c and y_c through the iteration of optimization using BFOA, and the parameters' values of BFOA in Table 3 are determined according to the experimental of several runs.
9. Stop the iteration after getting the lowest RMSE value or complete the iteration.
10. The output rating matrix will contain the accurate predicted values for all sparse rating scores.

4 Experiments and Results

The dataset of Netflix Prize is used in this experimental work. This data has been collected from the Netflix website during the six-year period from 1999 to 2005. Netflix dataset recorded the user rating about movies where this data was collected over 100 million rating scores provided by 480,189 users on 17,770 movies. Each rating score is provided with the date. The date of each rating score contains day, month and year which act the difficulties of extracting the time convergence among the rating scores. Therefore, these dates will be transferred into the timestamp values.

In this work the long-term factors are integrated with the short-term factors through integrated the long-term factors and the short-term factors within the prediction function which reflect the impact of these temporal preferences separately. The TemporalMF++ approach has been applied for three periods of short-term based the dataset of Netflix under rating scale [0–5] as mentioned in Table 1. The RMSE is used to evaluate the prediction accuracy of the CF technique where the lowest RMSE value means the best prediction accuracy. Explain here why u use RMSe instead of Top-N.

Figure 3 shows the results of temporal learning using TemporalMF++ approach where the period of one year has a higher prediction accuracy from iteration 1–5, and through the iterations 5–20 the period one year is more accurate than the period one month but less than the period one season. The temporal features are extracted by TemporalMF++ through 20 iterations to perform the accurate predictions by optimizing the weights of each period of time. The TemporalMF++ approach has improved the prediction accuracy of the CF technique by the temporal weights of season more than the temporal weights of month and year because of the season is the middle between the periods of the month and the year which has the accurate temporal behaviours.

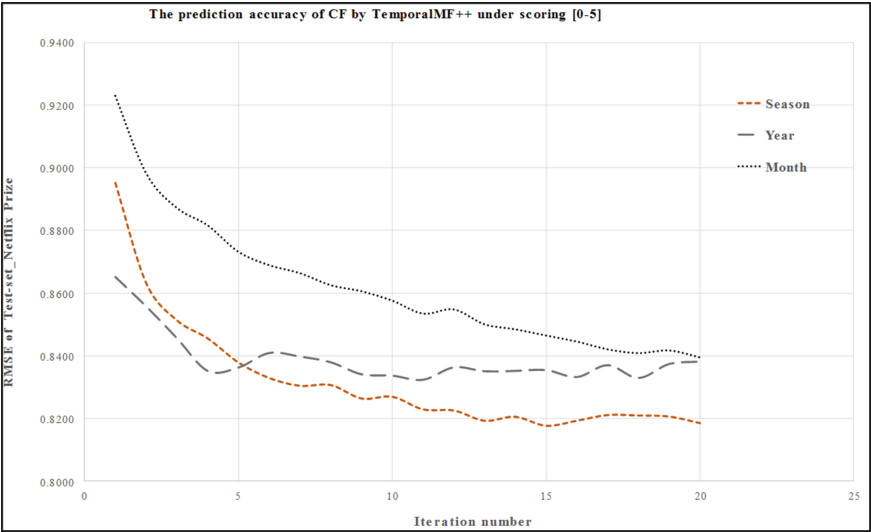


Fig. 3. TemporalMF++ learns the features of Netflix under [0–5]

There are six benchmark approaches have been implemented in this research for the purpose of evaluating the TemporalMF++ approach as shown in Table 3. In Table 3, the results show that the prediction accuracy of the CF has been improved by the temporal approaches of Short-Term based Latent [9], Long-Term [8], Short-Term based Baseline [8] and Temporal Interaction [8]. However, the evaluation temporal approaches have weaknesses in learning both the drift and the decay of items which happen either in the long-term or in the short-term. Therefore, the TemporalMF++ approach has covered the limitations of previous temporal approaches and the experimental results show the highest prediction accuracy by TemporalMF++ compared to the six benchmark approaches.

Table 3. The prediction accuracy of CF using several prediction approaches

Approach	RMSE of test-set
Collaborative filtering	0.9983
Temporal dynamics [7]	1.0173
Short-term based latent [9]	0.9982
Long-term [8]	0.9982
Short-term based baseline [8]	0.9982
Temporal interaction [8]	0.9982
TemporalMF++	0.8136

5 Conclusion and Future Works

A recommender system provides users with personalised suggestions for items based on the user’s behaviour history. These systems often use the CF for analysing the users’ preferences for items in the rating matrix. The CF technique suffers from several challenges such as data sparsity, drift in the user’s behaviours and decay in the item’s popularity over time. This research focuses on solving these challenges using the temporal based factorization approach named TemporalMF++ which integrates the long-term preferences and the short-term preferences. The experimental results show the TemporalMF++ approach has a higher prediction accuracy compared to the CF technique and to all the benchmark approaches of temporal. In summary, the TemporalMF++ approach has a superior effectiveness in improving the accuracy prediction of the CF by learning the temporal behaviour. In the future, this work can be implemented for several datasets such as Epinions, Movielens, and Yahoo! Music or learning more taxonomy features related to the user’s behaviours or the item’s popularity such as genres, album, and artist.

Acknowledgements. The publication of this paper is sponsored by the Ministry of Higher Education, Malaysia under the Fundamental Research Grant Scheme.

References

1. Hong, W., Li, L., Li, T.: Product recommendation with temporal dynamics. *Expert Syst. Appl.* **39**(16), 12398–12406 (2012)
2. Bobadilla, J., Hernando, A., Ortega, F., Gutiérrez, A.: Collaborative filtering based on significances. *Inf. Sci. (Ny)* **185**(1), 1–17 (2012)
3. Koenigstein, N., Dror, G., Koren, Y.: Yahoo! music recommendations : modeling music ratings with temporal dynamics and item taxonomy. In: *Proceedings Fifth ACM Conference on Recommended Systems*, pp. 165–172 (2011)
4. Mirbakhsh, N., Ling, C.X.: Clustering-based factorized collaborative filtering. In: *Proceedings 7th ACM Conference Recommended Systems*, pp. 315–318 (2013)
5. Al-hadi, I.A.A., Sharef, N.M., Sulaiman, N., Mustapha, N.: Ensemble Divide and Conquer Approach to Solve the Rating Scores' Deviation in Recommendation System. *J. Comput. Sci. Sci. Publ.* (2016)
6. Campos, P.G., Díez, F., Cantador, I.: Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Model. User Adap. Inter.* **24**(1–2), 67–119 (2014)
7. Koren, Y.: Collaborative filtering with temporal dynamics. *Commun. ACM* **53**(4), 89–97 (2010)
8. Ye, F., Eskenazi, J.: Feature-based matrix factorization via long-and short-term interaction. *Knowl. Eng. Manag.* 473–484 (2014)
9. Yang, D., Chen, T., Zhang, W., Yu, Y.: Collaborative filtering with short term preferences mining. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR*, vol. 2, p. 1043 (2012)
10. Koren, Y.: Factor in the neighbors: scalable and accurate collaborative filtering. *ACM Trans. Knowl. Discov. Data (TKDD)* **4**(1), 1–24 (2010)
11. Patra, B.K., Launonen, R., Ollikainen, V., Nandi, S.: A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. *Knowl. Based Syst.* **82**, 163–177 (2015)
12. Bobadilla, J., Ortega, F., Hernando, A., Alcalá, J.: Improving collaborative filtering recommender system results and performance using genetic algorithms. *Knowl. Based Syst.* **24**(8), 1310–1316 (2011)
13. Abdelwahab, A., Sekiya, H., Matsuba, I., Horiuchi, Y., Kuroiwa, S.: Feature optimization approach for improving the collaborative filtering performance using particle swarm optimization. *Comput. Inf. Syst. J.* **8**(1), 435–450 (2012)
14. Shen, H., Zhu, Y., Zhou, X., Guo, H., Chang, C.: Bacterial foraging optimization algorithm with particle swarm optimization strategy for global numerical optimization. In: *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, pp. 497–504 (2009)
15. Al-Hadi, I.A.A., Hashim, S.Z. M., Shamsuddin, S.M.H.: Bacterial foraging optimization algorithm for neural network learning enhancement. In: *2011 11th International Conference on, Hybrid Intelligent Systems (HIS)*, pp. 200–205 (2011)
16. Altıngövdé, I.S., Subakan, Ö.N., Ulusoy, Ö.: Cluster searching strategies for collaborative recommendation systems. *Inf. Process. Manag.* **49**(3), 688–697 (2013)
17. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 426–434 (2008)

18. Mirbakhsh, N., Ling, C.X.: Leveraging clustering to improve collaborative filtering. *Inf. Syst. Front.* 1–14 (2016)
19. Al-hadi, I.A.A., Sharef, N.M., Sulaiman, N., Mustapha, N.: Review of the temporal recommendation system with matrix factorization. *Int. J. Innov. Comput. Inf. Control* **13**(5), 1579–1594 (2017)

Part III

**Intelligent Human-Centred
Computing (IHCC)**

Preliminary Design of a Dual-Sensor Based Sign Language Translator Device

Radzi Ambar¹(✉), Chan Kar Fai¹, Chew Chang Choon¹,
Mohd Helmy Abd Wahab¹, Muhammad Mahadi Abdul Jamil²,
and Ahmad Alabqari Ma'Radzi²

¹ Department of Computer Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia
aradzi@uthm.edu.my

² Department of Electronic Engineering, Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn, Parit Raja, Malaysia

Abstract. There are many different types of sign languages that are used around the world which are important as the medium of conversation among hearing impaired community. However, majority of hearing people do not know or understand sign languages. Thus, communication between a hearing-impaired person and a hearing person is a difficult issue. In order to solve this problem, this project proposes a development of a dual-sensor based sign language translator. The goal of the project is to translate sign language into speech and display on screen by using the device. The device was developed in a glove-based system which was able to read the movements of every finger and arm using two (2) types of sensors, an accelerometer and five (5) units of flex sensors. This paper describes the design of the glove-based sign language translator. Subsequently, the preliminary experimental results show the usefulness of the accelerometer and flex sensors.

Keywords: Sign language translator · Accelerometer · Flex sensor · Experiment

1 Introduction

Since the 17th century, humans have started using sign languages as the medium to deliver messages and convey meanings [1]. Sign language enables a speaker to express his/her opinion using the movement of the hands and facial expressions. Sign languages have been developed around the world wherever hearing impaired communities exist [2]. Generally, every country has its own native sign language. In Malaysia it is called Malaysia Sign Language (MSL) [3]. With the tremendous evolution of technology since 19th century, computers are getting smaller and thinner. Therefore, wearable devices are able to be invented in order to help the hearing impaired people to translate their own sign language into spoken language which is able to be understood by the public. Communities with hearing disability will be able to communicate with others easily with the aid of these devices.

There are several methods to help hearing impaired people to communicate with others or hear others such as using hearing aid devices and sign languages. Hearing aid devices help those who have not lose their sense of hearing completely, while people with total hearing impairment can only depend on sign language to communicate between each other. There are several types of commercially available hearing aid devices which include behind-the-ear, in-the-ear and canal aid to help deafness and other communication disorder [4]. However, a hearing impaired person might experience some problems such as potential discomfort and annoying feedback sounds and noises with these kinds of devices. Besides that, people who use sign languages also face difficulties in letting the public understand their languages and it will make them miss a lot of chance to show their talents in particular fields. Therefore, with the existence of a sign language translator which can translate every gesture into voice for communication, the public can easily communicate with them.

Various studies have been carried out to develop sign language translator devices. Basically, there are two methods that are usually used in studies related to sign language translator which are vision-based system and wearable devices. Vision-based systems utilize image processing method by featuring extraction techniques to identify hand and finger movements [4]. Bauer and Kraiss introduced a colour video-based sign language recognition system based on Hidden-Markov Models (HMM) using subunits which could get self-organized using the derived data [5]. They pointed out that the advantage of subunits compared to models for whole signs was future reduction of necessary training material. Dreuw et al. presented a vision-based approach for a continuous automatic sign language recognition [6]. They demonstrated that appearance-based features in image recognition were suitable for the recognition of sign language. However, they concluded that there were still many works that needed to be done such as analyzing and combining new features in describing the hand and body configuration with the existing feature set. There are many other studies on sign language translation using vision-based system [7]. The advantage of vision-based systems is that the systems may not involve wearing sensory devices that can be uncomfortable and requires the user to wear it prior to communicate with a normal person. However, vision-based systems require complex and extensive computations in developing algorithms for feature and movement recognition.

Wearable devices for sign language recognition usually utilize sensors attached on the user or glove-based approach. Bui and Nguyen utilized six (6) accelerometers on a data glove for recognition of 23 Vietnamese Sign Language [8]. Su et al. developed a 3D imaging data glove that utilized 11 electromagnetic sensors to recognize sign language [9]. These devices are capable to translate sign languages with high precision. However, the systems are expensive and complex. Ahmed et al. presented a glove-based sign language translator that utilized only five (5) flex sensors [10]. The device translates simple words from sign language into speech (via a speaker) and text (displayed on an LCD). However, there were only four (4) gestures that were designed for user input which could produce speeches. There are many other studies related to wearable devices for translating sign languages which show that glove-based approach is a very popular method [11]. However, there are several challenges when using this method, such as discomfort of wearing the device, difficulty to move hands due to data

cables and costs of using the most suitable sensors to achieve the best results with limited errors.

The main objective of this project is to design a wearable device to help hearing-impaired communities communicate easily with the public. In this work, five (5) flex sensors and an accelerometer were used in the development of a glove-based sign language translation device. The hardware design of the device is also presented in this paper. The details of the experimental results on verifying the usefulness of sensors are described thoroughly. At the end of this work, the device was able to translate sign language and the results were a display on an LCD and audio sound via a speaker.

2 Description of the Sign Language Translator Device

2.1 Hardware Components and Circuit Design



Fig. 1. Overview of the hardware components used in the sign language translator

Figure 1 shows an overview of the hardware components for the proposed sign language translator device which consisted of input sensors (an accelerometer and five (5) flex sensors) and the output mediums (an LCD and a speaker). The device used an Arduino Mega for processing the data from all sensors. The figure shows that the input values from the flex sensors and the accelerometer were sent to the Arduino Mega to determine the input gesture. Then, Arduino processed the data and the output result was displayed simultaneously on an LCD as well as through audio speech using a speaker.

As shown in Fig. 1, an Arduino Mega 2560 microcontroller was used to control the whole system. In this project, Arduino Mega 2560 received and processed the input signals from the input sensors, and then provided the commands to the output component which corresponded to the input from sensor. The board contained 16 analogue input pins which were sufficient to handle a total of eight (8) analogue inputs from the flex sensor and accelerometer. Prior to selecting Arduino Mega 2560, we initially considered using Arduino Uno which is cheaper and can lower the production cost. However, Arduino Uno is not suitable for this project because Uno has only six (6) analogue input pins only. Furthermore, the device was also installed with an LCD to display the result of translation through text. Therefore, Arduino Mega 2560 provided enough analogue pins to cater for the project requirement. In this project, we used Arduino's pin 4, 5, 6, 7, 8 and 9 to connect with the LCD.

Figure 1 also shows that we used an accelerometer (GY61) to measure the acceleration of the hand. GY61 consists of ADXL335 chip that produces acceleration values that can be used to measure the angle of the 3 axis at your current position

corresponding to the origin position. The sensor is very useful to various projects such as self-balancing robots and quadcopters. This particular sensor consists of five (5) pins which are Vcc, GND, x-, y- and z-axes pins. In this project, accelerometer was used to measure the movement of arm from x-axis, y-axis, z-axis and determine whether it was a roll, pitch or yaw posture. The library file for ADXL335 was necessary to be included in the coding of Arduino in order to communicate between the accelerometer and Arduino Mega 2560. Besides, all the three output pins from the sensor were connected to the Arduino's analogue pin 5, 6 and 7 for further process of the signal values.

Five (5) flex sensors from Spectra Symbol were used in this project. The sensor was printed with a conductive particles embedded polymer ink [6]. The sensor worked when it bent with the ink on the outside of the curve. The value of the resistance did not change if it bent on the other side. As the sensor bent more, the resistance of the sensor increased as well. In this work, flex sensors were functioning as an angle displacement measurement components which could calculate the bending angle of five (5) fingers and helped to identify the gesture of a sign language. The sensor resistance range was limited from 60 to 110 K ohm, and the signals from five (5) units of flex sensors were fed into five (5) analogue pins of Arduino microcontroller.

Besides an LCD, another medium of the output was an 8 Ω audio speaker with the purpose of translating sign language through audio speech. The 0.5 W speaker was used due to its low power consumption and optimum frequency range which could play an audio between 600 Hz until 10 K Hz of frequency. The speaker was connected to the Arduino's analogue pin 11 via an LM286 amplifier circuit. When a certain movement was translated, the corresponded audio speech was played through the speaker. The amplifier circuit was to amplify the voice recorded in a micro secure digital (SD) card. An external 9 V battery was used as the power supply for this amplifier circuit. A potentiometer was used to control the volume of the audio being played.

A micro SD card reader module was used in this work to store the audio speech files. The SD card module contained six (6) pins which were CS, SCK, MOSI, MISO, VCC and GND. This module used Serial Peripheral Interface (SPI) interface as the communication interface. With the aid of In-Circuit Serial Programming (ICSP) pins on Arduino Mega 2560, it was much easier for us to communicate between the SD card reader and microcontroller. We just had to insert the SPI library and SD library in order to use the card reader module. As different type of Arduino board has different chip select (CS) pin settings, we specified the CS pin to be connected to Arduino's pin 53.

3 Preliminary Experiments and Results

3.1 Experiment of Flex Sensor

As the first step in verifying the usefulness of the components used in this project, first, we will describe two (2) simple experimental results on flex sensors.

In the first experiment, we carried out tests in order to measure the minimum and maximum resistances of all flex sensors that were used in this project. These tests were necessary in order to determine the range of resistivity for each flex sensor that was

used in developing the source code algorithm. By using a digital multi-meter, the minimum resistance (12 K ohm) was measured when the sensor was extended (180°) while the maximum resistance (40.1 K ohm) was measured when the sensor was bent to the maximum ($<45^\circ$). Similar tests were done for the other four (4) flex sensors to determine the minimum and maximum resistance values. These data were collected and used in the sign language translator algorithm.

Now, we had the minimum and maximum resistance values for all flex sensors. In the second experiment, we utilized the values into the algorithm and then, we developed the algorithm to demonstrate simple output on the LCD using only three (3) flex sensors. Figure 2a shows the experimental setup for this experiment. As shown in the figure, three (3) flex sensors were connected to three (3) analogue pins of Arduino Mega 2560.

Figure 2b shows that the output result was displayed on the serial monitor. As shown in the figure, each flex sensor starting from the left represented the index finger, middle finger and ring finger, respectively. Based on the figure, the output “Waiting...” stated that all the flex sensors were not bent which meant that all three (3) fingers were extended and spread apart. Then, when all three (3) of the flex sensors were bent ($<45^\circ$), the output showed “Zero!” which represented that all fingers were curled into a fist. Furthermore, the LCD displayed “One!” when the flex sensors on the centre and right were bent and “Two!” when only the flex sensor on the right was bent. Overall, the results obtained from these experiments were encouraging and showed that the flex sensors and algorithm for the sensors were working well.

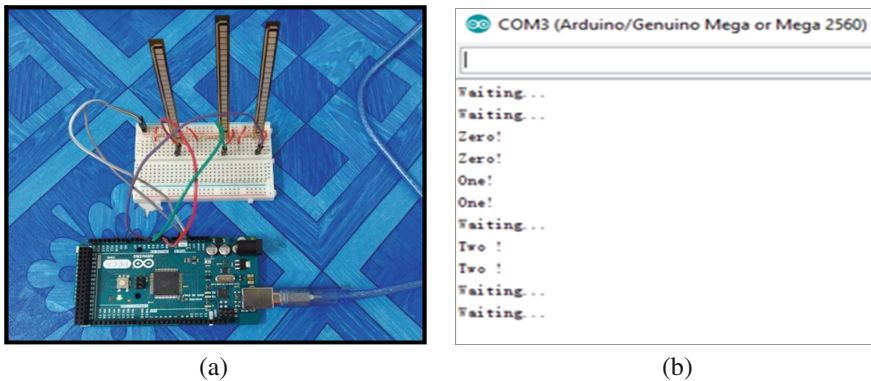


Fig. 2. **a** Experimental setup for flex sensors test, **b** results displayed on serial monitor

3.2 Simulating Flex Sensors Using Proteus

In the previous section, the tests to measure the minimum and maximum resistances of all flex sensors used in this project were demonstrated. As explained previously, these tests were necessary in order to determine the range of resistivity for each flex sensor that would be used in developing the source code algorithm. After determining the range of resistivity for all flex sensors, the values were utilized in the sign language

translator algorithm that we had developed. In order to easily test the effectiveness of the developed algorithm, we simulated the results using Proteus v7.8 software based on the work done by [10].

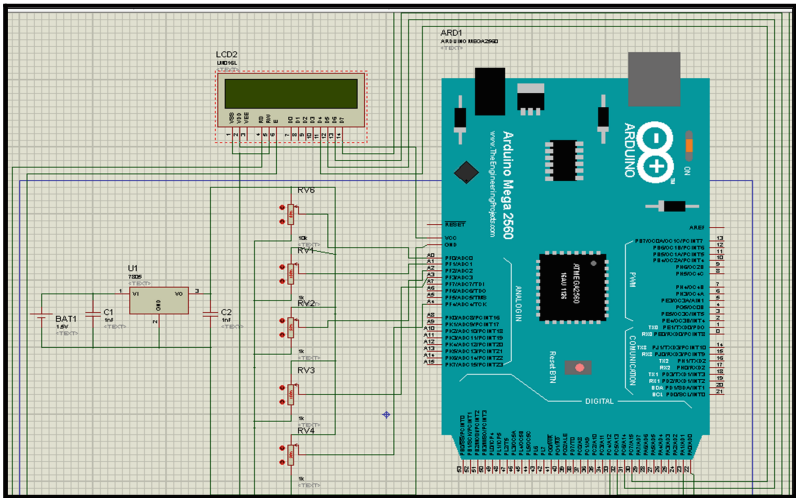


Fig. 3. Prototype of the glove-based sign language translator device

Figure 3 shows the schematic diagram designed using Proteus for simulating flex sensor, LCD and Arduino Mega 2560. However, since Proteus software did not have flex sensor in the software's inventory, it was replaced with potentiometers. Therefore, using this method, we could manually adjust the desired resistance value in order to imitate the movement of a finger. The simulation using this software provided easy and fast method for verifying the expected results of finger's bending movements. Figures 4, 5 and 6 show several results of simulation for various sign languages. As the first step for verifying the usefulness of the components used in this project, two (2) simple experimental results on flex sensors will be described.

Figure 4a shows an image of a sign language which means "Waiting", with all four fingers extended and spread apart. This condition was represented on the simulation by assigning maximum resistance value 1 K ohm to all five (5) potentiometers as shown in Fig. 4b. At this moment, the LCD was showing "Waiting..." while waiting for the changes of input as shown in Fig. 4c.

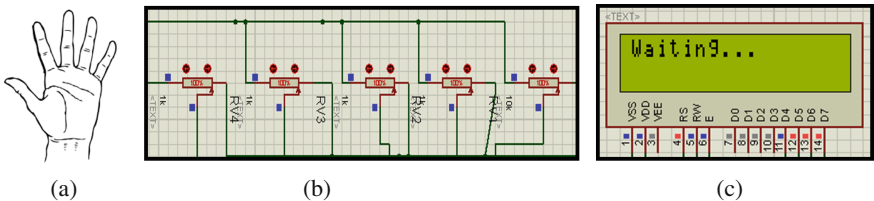


Fig. 4. a Sign language for "Waiting", b all potentiometers are in initial state of maximum resistance 1 K[Ohm] (100%), (c) LCD displays "Waiting..."

Figure 5a shows the image of a sign language which means “Good”, where only the thumb is extended, while the other four (4) fingers are curled into a fist. The condition was represented on the simulation by assigning minimum values of resistance for four (4) potentiometers as shown in Fig. 5b. Figure 5b shows that the LCD is showing “Good!” which corresponded to the sign language done by the user.

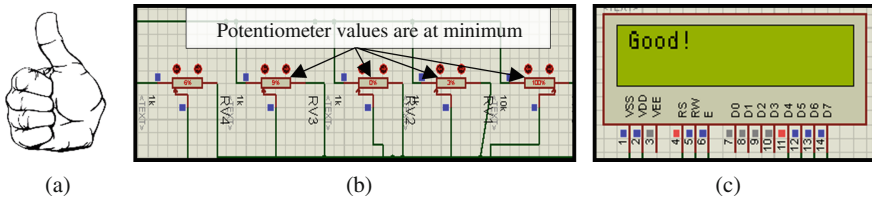


Fig. 5. **a** Sign language for “Good”, **b** last four (4) potentiometers values from right at minimum, **c** LCD displays “Good!”

Figure 6a shows the image of sign language which means “OK. I’m fine”. The figure shows only three fingers are extended and spread apart, while the tip of the thumb touches the tip of the index finger to form a loop. In the simulation, this sign language was represented by assigning the first two (2) potentiometers values from the right to minimum as shown in Fig. 6b. As a result, the LCD displayed “OK. I’m fine!” as shown in Fig. 6c.

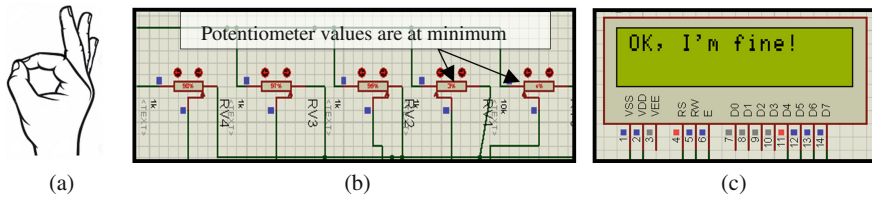


Fig. 6. **a** Sign language for “OK, I’m fine”, **b** first two (2) potentiometers values from right at minimum, **c** LCD displays “OK, I’m fine!”

From these simulation results, it showed that the microcontroller uploaded with the developed sign language recognition algorithm was able to identify the movement of fingers by reading the resistance value from flex sensors and displayed the accurate meaning of a sign language on the LCD.

3.3 Experiment on Accelerometer

We used an accelerometer (GY61) consisting of an ADXL335 chip that produced acceleration values (unit is in g). The values can be used to measure the angle of the three (3) axis at your current position corresponding to the origin position. In order to verify the usefulness and accuracy of the accelerometer, we recorded the readings from the accelerometer in six (6) different positions for calibration purposes. These values

were also used in the sign language translator algorithm. Figure 7a–f show the results from the serial monitor of Arduino IDE (lower figures) corresponding to the accelerometer positions (upper figures).

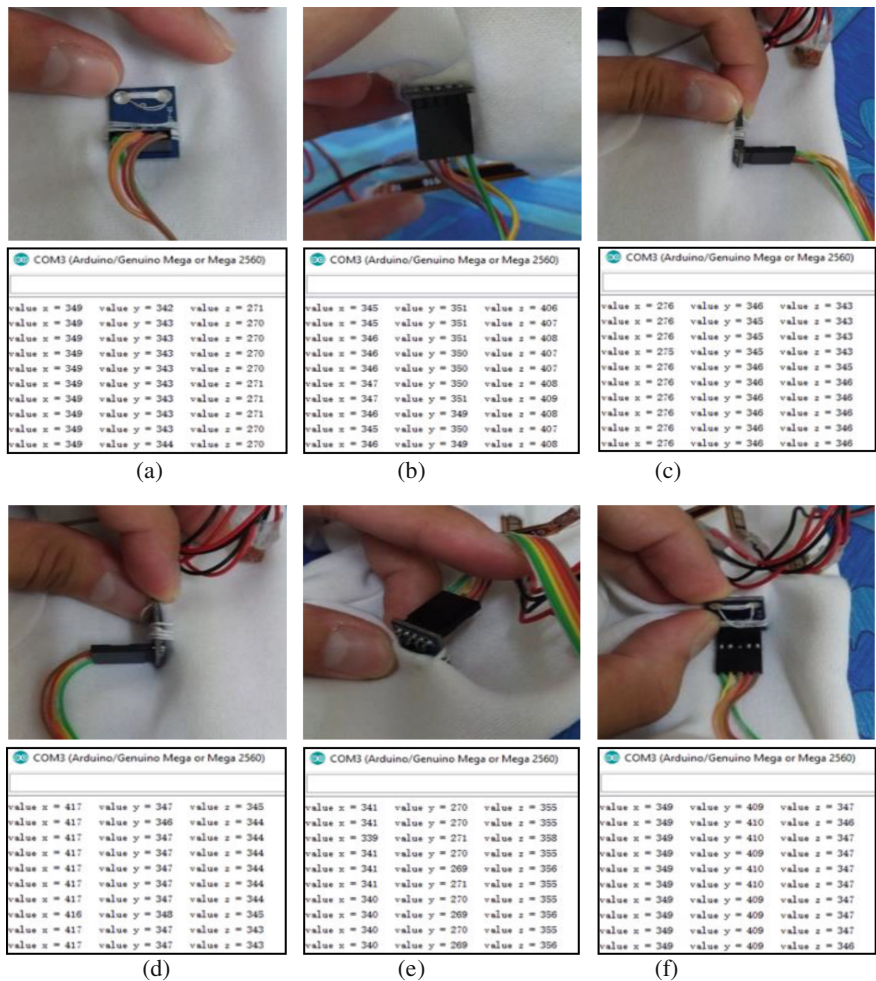


Fig. 7. Accelerometer orientations: **a** upside down **b** facing upward **c** facing to the right **d** facing to the left, **e** facing to the front, **f** facing to the back

As the method to describe the other five (5) positions was similar, only Fig. 7a is described here. Figure 7a upper and lower images show that the sensor was positioned upside down at rest on a flat surface, with analogue values of the x-, y- and z-axes on Arduino IDE's serial monitor, respectively. In this condition, the z-axis was measuring the force gravity (-1 g), while the x- and y-axes were measuring 0 g. As shown in Fig. 7a, the analogue values are: x-axis is 349 at 0 g, y-axis is 343 at 0 g and z-axis is

270 at -1 g. Therefore, based on these three values, the maximum and minimum values and also the analogue values for 1 g, -1 g and 0 g on each axis could be obtained.

Based on the acceleration analogue values from six (6) different positions, we can conclude that at 0 g, the x-, y- and z-axes analogue values were 350, while the maximum and the minimum analogue value of each axis were 420 and 270, respectively. These values are helpful when two different words have the same sign language sensor readings from the fingers but different hand orientation. Therefore, in our project, the accelerometer was useful to monitor the orientation/position of the hand.

3.4 Prototype of the Sign Language Translator Device

Figure 8 shows the first prototype of the glove-based sign language translator device. As this was the initial prototype of the device, five (5) flex sensors were attached on the glove using black insulation tape. These flex sensors were used to measure the bending movements of all five (5) fingers. These sensors were connected to an Arduino Mega 2560. However, the sign language translator algorithm is currently under development which includes the fusion of data from flex sensors and accelerometer. Therefore, the results of the work will be described in the future works.

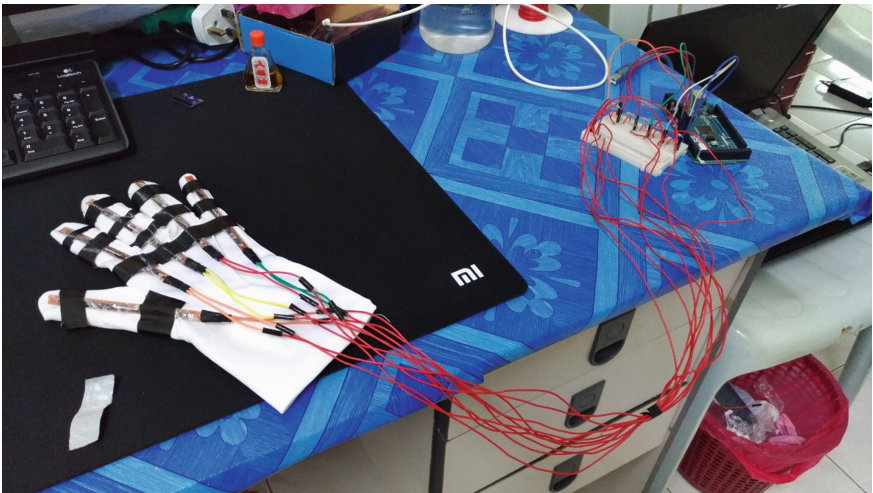


Fig. 8. Prototype of the glove-based sign language translator device

4 Conclusion

This paper describes the early development of a multi-sensor based sign language translator device. The final objective of this project is to translate sign languages into speech and also display on screen by using the device. The device was developed in a glove-based system which was able to read the movements of every finger and arm

using two (2) types of sensors, an accelerometer and five (5) units of flex sensors. In this paper, we have described the preliminary experimental results for flex sensors, accelerometer to verify the usefulness of the components. The results from these experiments will be used in the development of our sign language translator algorithm. For future work, all the components and algorithm will be integrated to complete the sign language translator device. Currently, the device is connected to the microcontroller via cables. Based on the preliminary testing of the device, the cables can disturb the hand movements. In the future, wireless connection between the device and microcontroller will be implemented to solve this problem.

References

1. Stokoe, W.C.: Sign Language Structure: An Outline of the Communicative Systems of the American Deaf. Linstock Press, Silver Spring, MD (1960)
2. Madhuri, Y., Anitha, G., Anburajan, M.: Vision-based sign language translation device. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 565–568 (2013)
3. Lewis, M.P.: Ethnologue: Languages of the World, 16th edn. SIL International, New York (2009)
4. Igari, S., Fukumura, N.: Recognition of Japanese sign language words represented by both arms using multi-stream HMMs. In: Proceedings of IMCIC-ICSIT, pp. 157–162 (2016)
5. Bauer, B., Kraiss, K.F.: Video-based sign recognition using self-organizing subunit. Proc. Int. Conf. Pattern Recogn. **2**, 434–437 (2002)
6. Dreuw, P. et al.: Speech recognition techniques for a sign language recognition system. InterSpeech-2007, pp. 2513–2516 (2007)
7. Huang, T.S., Wu, Y.: Vision-based gesture recognition: a review. In: Gesture Workshop, Gif-sur-Yvette, France, vol. 1739, pp. 103–115. LNCS (1999)
8. Bui, T.D., Nguyen, L.T.: Recognizing postures in vietnamese sign language with MEMS accelerometers. IEEE Sens. J. **7**(5), 707–712 (2007)
9. Su, Y., et al.: 3D motion system (“data-gloves”): application for Parkinson’s disease. IEEE Trans. Instrum. Meas. **52**(3), 662–674 (2003)
10. Ahmed, S., et al.: Electronic speaking system for speech impaired people: speak up. In: Proceedings of 2nd International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) (2015)
11. Dipietro, L. et al.: A survey of glove-based systems and their applications. IEEE Trans. Syst. Man Cybern. Part C (Applications and Reviews) **38**(4), 461–482 (2008)

M-DCocoa: M-Agriculture Expert System for Diagnosing Cocoa Plant Diseases

Munirah Mohd Yusof^(✉), Nur Fazliyana Rosli, Muhaini Othman, Rozlini Mohamed, and Mohd Hafizul Afifi Abdullah

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Malaysia
munirah@uthm.edu.my, faz_yana75@yahoo.com, muhaini@uthm.edu.my, rozlini@uthm.edu.my, hafizul94@gmail.com

Abstract. Major technological advancements were experienced including mobile applications in the various domain. The advancement in mobile applications not only used for our daily life and chores but it leads to more specific and technical purposes such as in medical, engineering, agriculture and education domain. This paper aims to study the implementation of mobile systems in agriculture and proposes a development of M-Agriculture that help in diagnosing cocoa plant diseases named as M-DCocoa. This application enables a user to recognize cocoa diseases afflict by the plant and provide user appropriate advice or treatments in shorter time period. The user will answer the questions based on cocoa plant condition or symptoms and the application generates the answer in form of disease and treatments. A rule-based and forward chaining inference engine has been used as part of the system development. With this application, it helps and allows the user to recognize cocoa diseases with useful treatments suggestion.

Keywords: Expert system · Mobile applications · Agriculture
Forward chaining · Rule-based · Cocoa plant

1 Introduction

Integrating agriculture with computing technology benefits both fields. Agriculture has played a major role in creating job opportunity to Malaysian. One of the agriculture resources that contribute to Malaysia economic income is the cocoa industry. Cocoa industry has been expanded in Malaysia in the late of 1970s and early 1980s. It has been the third most important agricultural export after palm oil and rubber and fifth largest grinded cocoa and product exporter [1]. There are about 16,000 ha of cocoa plantations nationwide mostly in Sarawak and Sabah, and Malaysia Cocoa Board has aimed to scale-up cocoa plantations in Malaysia to approximately 40,000 ha by 2020 [1]. To achieve this, several cocoa farms will

be opened with either highly-experienced owner or less-experienced owner which requires assistance in expertise to manage and handle the cocoa tree growth, any disease infection, and its required treatment. Through implementing computing technology such as mobile-based expert system in the area, it is believed that many agricultural problems addressed can be solved, therefore making the target by Malaysia Cocoa Board achievable.

In recent years, mobile technology has been widely adopted in Malaysia. Until March 2016 it has been reported that 20.6 million Malaysians (68% of total population) are active Internet users out of 30.5 million people, while 18 million Malaysians (59% of total population) are active mobile Internet users [2]. This figure shows tremendous usage of Internet and rapid growth of mobile users in Malaysia, thus reasonably increasing the popularity of mobile applications among the users. Various applications have been for mobile device used to check email, access the Internet, retrieve information or as reminder for important appointments. Today, mobile applications not only used for common task but also help in solving complex task in particular domain such as communication, education, finance, travel, entertainments, medical, and health [3]. In other words, mobile applications have become a part of our daily lives.

It is very important for cocoa farm owner to maintain the quality of cocoa to ensure the products able to hit the global market. Therefore, Expert System Applications for Diagnosing Cocoa Plant Diseases (M-DCocoa) is developed to help cocoa farmer, especially for new comer with the purpose to assist them in diagnosing cocoa plant diseases.

Here, we focus on the development of an expert system application for diagnosing cocoa plant diseases. 11 types of pests were considered including cocoa cushion and bark borer (*Squamura disciplaga*, *Indarbela disciplaga*), stem borer (*Zeuzera coffeae* and *Zeuzera conferta*), ring bark borer (*Endoclyta hosei*), termites (*Coptotermis curvignathus*), black cockchafers beetle (*Apogonia sp.*), brown cockchafers beetle (*Adoretus sp.*, *Lepadoretus sp.*, and *Chaetadoretus*), larra tortricid, mugly bug, midges cocoa mirids (*Helopeltis theivora* Miller), cocoa pod borer (*Conopomorpha cramerella*), and husk borer (*Conogethes/Dichocrochis punctiferalis*).

The pests infection has contributed to stunted growth of cocoa trees, degradation the fruit quality as well as the quantity of cocoa being produced. Misdiagnosing the symptoms not only result in production of low quality cocoa, but also contribute to financial loss to the cocoa farm owners. Therefore, this research is helpful to assist users to predict accurately the cocoa tree infection based on the given symptoms.

The paper is structured as follows. Section 2 provides the related work in expert system and mobile agriculture, while Sect. 3 discuss rules design and development and the system development. Section 4 performance testing. Section 5 draws the discussion and conclusions.

2 Related Works

Agriculture has been one of the important economic sector in Malaysia, providing approximately 1.7 million jobs to Malaysians and has contributed 8.9% to the country's Gross Domestic Product (GDP) in 2015 [4]. Major contributors in agriculture which has contributed includes oil palm, livestock, fishing, rubber, forestry, and logging. Information technology advancement can be utilized to further boost productivity in agriculture [5]. The primary aim is to boost the productivity in agricultural by combating pests and diseases, therefore allowing the production of new seeds and the development of products of higher quality.

To be even more competitive in agriculture sector, we have to produce a higher quality production, which means, pests and diseases are to be eliminated. Traditionally, farmers often rely on advice from agricultural specialist to make a decision-related with fertilization, pests and diseases problems.

However, these specialist and officers are not always available when needed. Hence, to avoid this problem, an expert systems were used to assist farmer in finding solution to their problems without any appointment with the specialists.

According to Darlington [7], an expert system is a program that attempts to mimic human expertise by applying inference methods to a specific body of knowledge. Rule-based expert system is the simplest form of artificial intelligence and it is being used in many fields, for example in medical, agriculture, education, engineering or industrial domain. Hence, by using rule-based, the knowledge acquire from the expert, for instances doctors, in the form of rules used for problem solving [8].

In agriculture sector, expert systems were identified as powerful tool with its ability to reduce the information that human users need to process, reduce personnel costs, increase output and performs tasks more consistently than human experts [9]. It have been an active research field in agriculture for the past 30 years and most agriculture expert system focused on diagnosis diseases or pests attack, suggesting treatment or suitable solutions, fertilization scheduling and plant care [6].

Several experts systems has been developed for diagnosing of diseases in plantations including Clove Expert System [10], Paddy Plant System [11], Expert System for Diagnosing Oyster Mushroom [12], and Expert system for diagnosing coffee plants [13].

Active research on expert systems in agriculture has been expanding for the past 30 years where most agriculture expert systems are developed as a standalone and web-based systems which focus on diagnosis, treatment, fertilization scheduling, and plant care [6]. As the ICT expand rapidly towards mobile technology, the platform of technology in agriculture domain are revolted towards mobile applications, assisting farmers to deal with diseases problems easily and faster where knowledge regarding the symptoms, treatments, and solutions will be accessible from a single application.

Some of these expert systems were a standalone system or online web-based system. As technology emerged towards mobile technology, many mobile

applications were developed to cater user needs in various domains including medical, education, engineering, and agriculture (known as M-Agriculture) [3].

Mobile applications can be defined as a software application that runs on mobile devices such as smartphones, tablets, or other electronic devices. Developed applications which run on mobile devices are used to check email, access the Internet, retrieve information, or as a reminder for important appointments. However, with some improvement, mobile applications are not only used for a common task but also help in solving complex task in domains such as communication, education, finance, travel, entertainment, medical domain, and agriculture sector [3].

In their work, Liu and Koc [14] have developed a mobile application for tractor rollover detection and emergency reporting, called the “Safe-Driving” which detects if tractors accident occurs. The application calculates the stability of a tractor using the physical parameters of the tractor using data obtained from built-in sensors of a mobile device. If an accident is detected, it will automatically send alert to e-mail and phone with the details of the occurrences. This work is a good example of application of mobile apps in agriculture field.

Another example of agriculture-related mobile application is the Agriculture Machinery Cost Analysis (AMACA) application which is used to calculate and estimate machinery use and cost for farm businesses [15]. This application can make a subsequent calculation of the sensitivity by varying some parameters (fuel price, interest rate, field capacity, and others) and presents the cost analysis result. The application is capable of providing wide-range recommendations such as to either purchase a new tractor, use own machinery, hire a service (outsourcing), and recommendation of suitable cultivation system.

A model for an expert system mobile application has been proposed and developed by Abdelhamid and El-Helly [16] for strawberry crop disease diagnosis. Users are required to choose the observable symptoms and the application will predict for any suspected disorders based on the real knowledge-base. Another example, the Agri eXpert System application by the UAS Bangalore allows users to interact with the experts by sharing questions concerned on pests and diseases affecting different crop plants in the form of voice or text messages of photos and get the feedback instantly [17]. Features on modern mobile phones such as high-resolution cameras, GPS sensor, proximity sensor, accelerometer and gyrometer, large storage, and others will certainly be helpful in introducing new capabilities for expert system applications such as image processing and pervasive computing, and may initiate for many research ideas.

Our project focuses on the development of an Expert System Application for Diagnosing Cocoa Plant Diseases (M-DCocoa) to help farmers diagnosing pests and diseases infecting cocoa plants. Rule-based technique and forward-chaining inference engine has been used in the development of the diagnosis algorithm, further discussed in Sect. 3.

3 Methodology

The MD-Cocoa application development has adapted the knowledge engineering methodology [8] which consists of six phases: (i) problem assessment, (ii) data and knowledge acquisition, (iii) prototype development, (iv) actual system development, (v) evaluation and revision, and (vi) integration and maintenance.

In problem assessment phase, an extensive and thorough review on the problem the faced by cocoa planters were executed, and a suitable approach to cater the problem is proposed and justified. We have approached an agricultural department expert in cocoa plantation to gather knowledge on diseases related to cocoa plantation, observable symptoms, and its association with the related pests during the knowledge acquisition phase. The knowledge will then be processed into set of rules for determination of pests through symptoms. A system prototype is developed based on the set of rules created, applying knowledge gained into a knowledge base. Various reliable sources were considered in effort to develop a good knowledge base, which is then verified by experts in the field. During the system development phase, knowledge is organized and re-arranged with a particular representation to ensure easy-access to the solution for related problems, where a rules-based representation method with forward-chaining for its inferential engine. Evaluation is performed to validate the performance of the proposed M-DCocoa application and the system is revised based on necessity (see Sect. 4). The final phase focuses on the integration and maintenance of the system, either in making arrangements for technology transfer or establishing an effective maintenance program.

3.1 Design of the M-DCocoa Engine Rules

M-DCocoa application is developed for diagnosing the cocoa trees which has been attacked by pests and developed diseases. The application utilizes a rule-based system and using forward-chaining technique to retrieve inferences from a knowledge base, allowing recognition of type of pests and treatments required. Forward-chaining starts with a set of known facts, in this case, are the symptoms of the disease. Set of rules are applied to generate new facts whose premises match the known facts [18]. The process is continuously executed until a pre-determined goal is reached, where a particular pests which attack cocoa plant is known, or until no further symptoms can be derived and premises match the known facts.

This research focuses on 11 types of pests. The knowledge base of M-DCocoa was developed using IF/THEN structure set of rules. The information contained in the IF clause is related to the information contained in the THEN clause. IF clause checks the facts about the symptoms faced by the cocoa plants, next THEN clause generates the corresponding result, which is the type of insect that infecting the cocoa plants. The IF/THEN structure set of rules is summarized in Eq. 1. Knowledge and facts in the knowledge-base will be translated into the form of the knowledge representation. Knowledge representation from expert helps in

the development of the rules set. Algorithm 1 shows sample set of rule-based for symptoms and pests particular for stems and branch.

$$\text{IF (symptoms)} \Rightarrow \text{THEN (infector)} \quad (1)$$

where symptoms represents list of symptoms existing in the knowledge-base and infector represents the list of pests which is associated to the infector.

Algorithm 1 Set of rules for stems and branch symptoms

```

if fras feces is found on stems and branches then
  if smaller fras, neatly arranged and fras feces is red-coloured then
    if large termites colonies nearby then
      | Termites
    else
      | Stem Borer, Cocoa Cushion and Bark Borer
  if smaller fras, neatly arranged then
    | Cocoa Cushion and Bark Borer
  else if fras faeces is red-coloured or noticeable black-coloured wooden residue then
    | Stem Borer and Ring Bark Borer
else
  if smaller fras, neatly arranged and noticeable black-coloured wooden residue then
    | Stem Borer, Cocoa Cushion, Bark Borer and Ring Bark Borer
end

```

3.2 M-DCocoa System Development

M-DCocoa application has been developed in two flavours; web-application for the administrator and Android application for the common users. The administrator has a more privileged access, allowing management of the contents to be displayed in the application, which includes creating, delete, and modifying contents such as information about pests, treatments, symptoms, and infection information.

By accessing the mobile application, general users are able to select the related symptoms, allowing them to diagnose the disease for a cocoa plant and to view the suggested treatment for the infection. In diagnosis module, the type of pests which has infected the cocoa plant is identified by answering several questions regarding the symptoms observed.

M-DCocoa has a minimalist and user-friendly interface design, hence provides clear interaction for the user. A registered user will need to logged in to the application, to access the available functionality. The main page consists of five buttons; information, diagnosis, about the system, suggestion, and log out. Meanwhile, to begin the diagnosis, users are required to choose one of the categories which are *Stems and Branch*, *Fruit*, and *Leaf and Sprouts* before proceeding with the diagnosis as shown in Fig. 1a.

A set of questions regarding cocoa plants symptoms were asked based on the category chosen. Figure 1b shows the list questions by selecting the symptom face by cocoa plants. Consequently, after answering all related questions, this system will generate diagnosis result with explanations of pests type and suggestion treatment.

**M-DCocoa:
M-Agriculture Expert System for
Diagnosing Cocoa Plant Diseases**

Please Select a Category:

STEM AND BRANCHES

FRUITS

SHOOTS AND LEAVES

<< BACK

**M-DCocoa:
M-Agriculture Expert System for
Diagnosing Cocoa Plant Diseases**

Please Select ALL Related Symptoms:

- ☒ Fras feces is found on stems and branches
- ☒ Fras is smaller, neatly arranged
- ☒ Fras found is red-coloured
- ☐ Noticeable black-coloured wooden residue
- ☐ Wilting leaves
- ☐ Large termites colonies nearby
- ☐ Young plant starts to die
- ☐ Stem wilts and drying

Fig. 1. a Selection of category. b Selection of symptoms

4 Evaluation/Results and Discussion

To comply with the requirements specified by users, a session for User Acceptance Testing (UAT) has been carried out to validate the robustness of the M-DCocoa application to handle required tasks according to a real-world scenarios. The results of the UAT is summarized in Tables 1 and 2.

The developed application was tested among 240 testers acting as general users who need to diagnose treatments for cocoa plantations. The UAT includes both user interface and user experience (UI/UX) testing and functionality testing. The scoring criteria were computed using the formula in Eq. 2:

$$Score_n = \frac{\sum(\text{Mark} \times \text{Rating})}{\text{No. of testers}} \quad (2)$$

where n is the number of question.

Table 1 shows 8 criteria of scoring UAT for UI/UX testing and their respective scores for M-DCocoa application. Overall, most of the users agree with the suitability of the application design interface and fluidity of its user experience.

Table 1. Result of user interface and user experience testing

No.	Scoring criterion	Score (%)
1	Menu arrangement	88
2	User interface is clear and relatable	90
3	Continuous and structured system flow	82
4	All buttons and menu are working	80
5	Icons are understandable by user	82
6	Suitable font-family and size is used	78
7	Suitable colour-scheme is used	84
8	System is user-friendly	80

Table 2 shows 6 criteria for scoring the application functionality testing of the M-DCocoa application. Overall, most of the users agree that the developed application performs its function as expected.

Table 2. Result of system functionality testing

No.	Scoring criterion	Score (%)
1	Login process	82
2	Login success and authentication	82
3	Registration functionality	82
4	Button and icon functionality	82
5	Suggestion module functionality	80
6	Display functionality	80

5 Conclusion and Future Works

M-DCocoa was developed using a rule-based approach which helps in recognizing and diagnosing of infected cocoa plants. Treatment is suggested immediately after a set of question is answered, providing a faster and convenient way to diagnose disease and find treatments.

Future improvement possible for M-DCocoa application includes searches for treatment for the known disease (without completing a set of questions), recognition of cocoa plant disease using image processing, and upgrading the application engine by using a modern technique such as fuzzy logic or neural networks to further improve the accuracy of diagnosis result.

Acknowledgements. This paper has been supported by Short Term Grant, Universiti Tun Hussein Onn Malaysia (Vot U539) for the financial support. This research is also supported by GATES IT Solution Sdn. Bhd. under its publication scheme.

References

1. The Star Online: potential for new cocoa plantations enormous, says board. <http://www.thestar.com.my/news/community/2014/08/12/potential-for-new-cocoa-plantations-enormous-says-board/>. Retrieved 2 Feb 2017
2. Malaysian Communications and Multimedia Commission (MCMC): Communications and Multimedia: Facts and Figures, 4Q 2016. <https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/4Q2016.pdf>. Retrieved 2 Feb 2017
3. Siuhi, S., Mwakalonge, J.: Opportunities and challenges of smart mobile applications in transportation. *J. Traffic Transp. Eng.* **3**(6), 582–592 (2016)
4. Department of Statistics Malaysia: KDNK Kaedah Pendapatan 2010-2016. <https://newss.statistics.gov.my/newss-portalx/ep/epFreeDownloadContentSearch.seam?contentId=56037>. Retrieved 10 Feb 2017
5. <https://www.oxfordbusinessgroup.com/overview/modern-field-new-technology-and-techniques-are-being-implemented-help-grow-sector%E2%80%99s-viability>. Accessed 21 Feb 2017
6. Prasad, G.N.R., Vinaya, B.D.A.: A study on various expert systems in agriculture. *Comput. Sci. Telecommun.* **4**, 81–86 (2006)
7. Darlington, K.: *The Essence of Expert Systems*. Prentice Hall, Harlow (2000)
8. Negnevitsky, M.: *Artificial Intelligence: A Guide to Intelligent Systems*, 2nd edn. Addison-Wesley, New York (2005)
9. Kaur, R.: Importance of Expert systems used in agriculture: a review. *Int. J. Enhanc. Res. Sci. Technol. Eng.* **3**(5), 256–269 (2014)
10. Hananto, P.E., Sasongko, P.S., Sugiharto, A.: Sistem Pakar Diagnosis Penyakit Tanaman Cengkih dengan Metode Inferensi Forward Chaining. *J. Inf. Technol.* **1**(3), 1–14 (2012)
11. Honggowibowo, A.S.: Sistem Pakar Diagnosa Penyakit Tanaman Padi Berbasis Web dengan Forward dan Backward Chaining. *Telkomnika* **7**, 187–194 (2009)
12. Munirah, M.Y., Rozlini M., Mariam, Y.S.: An expert system development: its application on diagnosing oyster mushroom diseases. In: 13th International Conference on Control, Automation and Systems. IEEE, Chicago (2013)
13. Suhartono, D., Aditya, W., Lestari, M., Yasin, M.: Expert system in detecting coffee plant diseases. *Int. J. Electr. Energy* **1**(3) (2013)
14. Liu, B., Koc, A.B.: SafeDriving: a mobile application for tractor rollover detection and emergency reporting. *Comput. Electron. Agric.* **98**, 117–120 (2013)
15. Sopegno, A., Calvo, A., Berruto, R., Busato, P., Bocthis, D.: A web mobile application for agricultural machinery cost analysis. *Comput. Electron. Agric.* **130**, 158–168 (2016)
16. Abdelhamid, Y., El-Helly, M.: A new approach for developing diagnostic expert systems on mobile phones. *Commun. Inf. Sci. Manag. Eng.* **3**(8), 374–384 (2013)
17. <https://apkpure.com/agri-expert-system/com.keyfalcon.expertsystem>. Accessed 24 Feb 2017
18. Al-Ajlan, A.: The comparison between forward and backward chaining. *Int. J. Mach. Learn. Comput.* **5**(2) (2015)

Dyslexia Adaptive Learning Model: Student Engagement Prediction Using Machine Learning Approach

Siti Suhaila Abdul Hamid^(✉), Novia Admodisastro^(✉),
Noridayu Manshor, Azrina Kamaruddin, and Abdul Azim Abd Ghani

University Putra Malaysia, Seri Kembangan, Malaysia
gs42041@student.upm.edu.my, {novia, ayu, azrina, azim}
@upm.edu.my

Abstract. Education barriers are synonym with people with dyslexia life experience. People with dyslexia encounter barriers such as in academic related areas, mistreated with negative reaction on their behaviour and limitation to acquire a suitable support to overcome the barriers. Therefore, this work focus on giving the support to help students with dyslexia deal with their difficulty through adaptively sense their behaviour for engagement perspective. For that reason, we apply machine learning approach that utilises Bag of Features (BOF) image classification to predict student engagement towards the learning content. The engagement prediction was relatively using frontal face of the 30 students. We used Speeded-Up Robust Feature (SURF) key point descriptor and clustered using k-Means method for the codebook in this BOF model. Then, we classify the model using 3 types of classifier which are Support Vector Machine (SVM), Naïve Bayes and K-Nearest Neighbour (k-NN) to find the best classification result. Through these methods, we managed to get high accuracy with 97–97.8%.

Keywords: Adaptive learning · Engagement · Dyslexia · Machine learning

1 Introduction

Systems with adaptive learning model have an ability to automatically change the structure, functionality or even an interface to personalise the different needs of individual [1]. The basic structure of an adaptive system comprises of domain model, user model and interaction model. Therefore, the adaptation specification in domain, user and interaction models focused on changes on features, user's characteristic and adaptation method. An adaptive interface relates with customising interface characteristic such manipulation of display, colour, font and background preferences [2]. Whereas, adaptive functionality work on self-regulation and self-modifying by analysing, evaluating, selecting and changing the mechanism of various possible output from a given input [1].

The effectiveness of a system that incorporates adaptation in the learning model was discussed among researchers in the wide area of domain. In language learning, for example, machine learning approach successfully personalised the learning content,

improve reading strategies and able to adapt suitable behaviour towards each learner for English, Chinese, Japanese and Spanish languages [3]. Besides that, the process of adaptation also covered in reading browser that adapts the difficulties faced by struggle reader [4]. As a result, the adaptive system offers a new advancement for individualising learning experience.

Dyslexia is a specific learning disability that caused a difficulty in reading, spelling and writing [5]. Students with dyslexia also face difficulty such as poor concentration, frustration, embarrassment, task-anxiety and high avoidance [6]. These difficulties are different with one another, thus it needs to be tackled personally. An adaptive system which offers a personal learning experience become a promising solution in improving dyslexia learning. However, based on the literature there is no existing work related with dyslexia that adaptively predict engagement behaviour. Current work only focus on cognitive aspect [7, 8] and solved the difficulty through games [9], music [10] and design [11].

Therefore, we will discuss the adaptive model for dyslexia based on engagement behaviour in order to propose a personalised learning experience using machine learning approach. To do so, we proposed to use frontal face image data to predict the engagement behaviour. BOF model which utilise machine learning approach works well to recognise and classify the images. BOF model has been used in a wide area for image recognition and classification for instance in batik classification [12], face recognition of the same person [13], hand-written classification [14], type of clothing classification [15] and others. Nonetheless, there is no engagement prediction that utilise BOF which limit us to compare any benchmark data and performance. For that reason, in this paper we discuss the process to do engagement prediction for dyslexia student using BOF model without comparing with any existence work in the similar domain.

In this paper, we organised as follows, in Sect. 2 presents related work of adaptive learning model, dyslexia and BOF as a machine learning approach. Section 3 describes the proposed dyslexia adaptive learning model. While in Sect. 4, explain about methodology used. Section 5 presents the result of the tested model and finally, Sect. 6 discusses the conclusions.

2 Related Work

In general, an adaptive learning model is a potential mechanism for dyslexia education when utilising machine learning approach. Therefore, in this section, discuss current adaptive systems towards dyslexia and current approach of machine learning that apply BOF for classification in developing our adaptive learning model.

2.1 Dyslexia

Dyslexia is defined as a specific learning disability that affecting reading and prone to have error in writing as well as slow progress in spelling [16]. From the behaviour perspective in class, students with dyslexia appear to easily get engaged or disengaged with the learning content. Their lack of motivation to focus become one of the causes

that affect their learning process [17]. Therefore, it is important to sense the student engagement to retain the student attention.

Student Engagement. Student engagement is the reflection of an active participation towards the learning. It is found that strong engagement can be a predictor of a good academic performance thus improved learning outcomes [18]. For dyslexia perspective, it is important to measure student's engagement in order to create an active participation towards the learning. There are many approaches used by researchers to measure engagement namely through facial expression to measure the engagement of emotion [19], off-task behaviour through mouse input [20] and engagement of gamification through serious game [21].

2.2 Adaptive System

An adaptive system is an interactive system that change automatically depends on individual user's behaviour [1]. It takes into account user profiles, characteristic and operates differently according to the information [22]. An adaptive system that adopts intelligent aspect includes adaptive serious game system and adaptive spellchecker. Adaptive serious game infers the learning disability types occurred and suggest a number of games the child should play with a suitable frequency of playing per week using Analytical Hierarchy Process (AHP) [21]. Besides that, an adaptive spellchecker called PoliSpell purposely to help people with dyslexia in writing [23]. It was designed and trained with people with dyslexia's typical error and predict the words using Hidden Markov Model (HMM) and correct the sentences.

2.3 Bag of Features Model (BOF)

In the direction of developing an adaptive system which specifically target student engagement, we use image data to make prediction. Due to a good reputation and a popular image classification algorithm, we therefore utilize BOF [24]. BOF is the process of extracting image patches using method such as key-point detector or grid point. Then, describe the features into a numerical vector and finally convert the description into a codebook for classification [25].

Traditional BOF uses Scale Invariant Feature Transform (SIFT) for features extraction, K-Means clustering method for creating a codebook and finally Support Vector Machine (SVM) for classification [24]. However, a few researchers begin to create a new algorithm on improving the accuracy of Bag of Features algorithm. For example the use of Speeded-Up Robust Feature (SURF) to speed up the feature extraction process and solved storage issues [14]. Such improvement makes BOF more reliable in making a prediction thus triggering us to apply in image prediction for engagement.

SIFT was introduced by Lowe [26] that used local features for detection and extraction of an images. The key interest point is represented using a circular with orientation. SIFT has four parameters that includes x and y coordinates (key point centre), radius of region and finally the orientation which comprises of angle in radian. Due to its invariant and robustness in illumination and distortion, SIFT became an alternative of image classification for rigid image as animals [27] to a flexible image

like geographical scene classification [28]. However, some of the researcher who have time and processing issues prefer SURF for the key interest point extraction.

SURF uses HAAR-like features that have integrated pixel values over local region. SURF was chosen to detect key interest point to create a codebook and finally using SVM for classifying type of clothing [15]. In addition, SURF descriptor was promoted to be as one of the alternative in extraction of important features in object classification and localisation within an image scene [29]. Then, K-means was used to cluster the features and classified using two classifier which are Naïve Bayes and SVM. Both researchers prefer SURF over SIFT due to its concise descriptor length and fast performance. Majority of the researchers choose K-means for clustering purposes and a variety of classifiers to find the best accuracy performance.

3 The Dyslexia Adaptive Learning Model

The adaptive learning model developed with aims to improve learning of the Malay language amongst students with dyslexia through an adaptive delivery of exercise and teaching models that consider both cognitive and engagement. This work focus on the students with dyslexia aged between 7–12 years old. The proposed model as shown in Fig. 1 comprises of five main components namely Exercise Model, Behaviour Processing Model, Student Model, Expert Model and Teaching Model [30]. However, for the purpose of this paper, the discussion is on the Behaviour Processing Model that represent student engagement prediction in the Malay language.

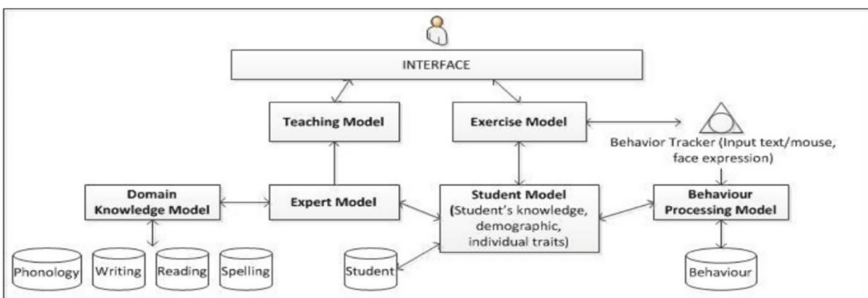


Fig. 1. Dyslexia adaptive learning model

3.1 Behaviour Processing Model

Behaviour processing model is the model that detect, process and evaluate the engagement behaviour of the student. The detection is on whether the student able to engage with the learning or not. To measure the engagement, the student were given an exercise related to the Malay language which comprises of phonology, spelling, reading and writing exercise derived from Exercise Model. While student answer the exercise, a video camera were used to record the student's behaviour using their frontal face detection. The data collected in the form of an image from the tracker then will be

analysed by the Behaviour Processing Model. The model adopts machine learning classifier to classify whether the displayed behaviour are engaged with the learning or not. The results of the model are used as an input for the Student Model.

3.2 Engagement Prediction in Behaviour Processing Model

According to Tan et al. [31], engagement is a reflection of an active involvement in learning. We utilize machine learning approach to make an engagement prediction to overcome the limitations of common systematic programming due to the uncertainty that involves vast of image patterns, gestures and interpretations. We classified the images based on the selection features to determine between engaged and disengaged using BOF.

Engaged. The prediction of engagement based on a situation where the student looks at the computer and interact with the learning content [19]. The engagement is considered when the student looks at the computer thus, the frontal face were detected. Besides that, the eyes region also can be used to determine whether eyes of the student were wide open or not. Wide open eyes signify an engaged situation.

Disengaged. On the other hand, disengagement is related with the high level of student boredom. Therefore, the prediction can be made from the situation when the student look away from computer [19] which bring the result that no face was detected. On top of that, when the student closed their eyes, it also can be concluded with disengagement thru detection of eyes region even when the frontal face was detected. Refer Table 1 for the example of engaged and disengaged face label.

Table 1. Prediction of engagement from behaviour

Engaged		Disengaged	
Situation	Features	Situation	Features
Look at the computer	Frontal face detected	Looking away from computer	Frontal face not detected
	Eyeball look at the screen		Eyeball looking left or right
	Open eyes		Close eyes

4 Methodology of Student Engagement Prediction from Image

In our methodology, we conducted a preliminary study for data collection at specific intervention centre for students with dyslexia. Then, the data were analysed to select suitable features for the classification. We use MATLAB R2016 for features extraction and classified in WEKA version 3.81. As shown in the Fig. 2, we discuss the flow of the engagement prediction. The steps include the data input acquisition, pre-processing techniques, clustering, and classification as well as performance evaluation. The

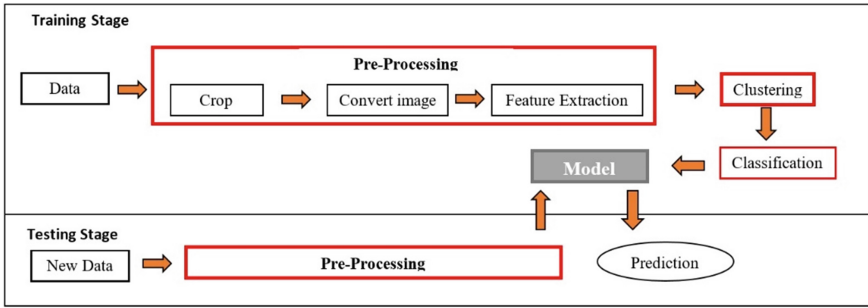


Fig. 2. Flow of the engagement prediction

performance prediction then, has been check by human coder to confirm the credibility of the prediction result but will not discuss in this paper.

4.1 Data Input

Data input collected are from preliminary study collected in observation at Dyslexia Association of Malaysia (DAM) Ampang and Bangi. A total of 30 participants comprises of 20 males and 10 females aged 7–12 years old volunteered to participate in the preliminary study. This study took place in the provided quiet room in DAM Ampang and Bangi. The setting includes usage of video camera to capture the student face while answering the Malay language question. Video camera with the tripod was placed on the table facing the subject to capture the entire face.

Using a preliminary video data input, a human coder was asked to label the frame by frame images. An expert who experienced with teaching children with dyslexia were asked to become a human coder. As a result, a total of 600 images were extracted from video data of 30 participants that comprises of 300 engaged faces and another 300 disengaged faces.

4.2 Pre-processing

The images then were cropped to an average of 600×700 pixels dimension to reduce any irrelevant background that will affect performance of the model [32]. For the next process, we convert the coloured image to grey scale due to the limitation of the SURF descriptor which works only on a grey images [33]. It is good for our model as well to avoid any skin colour differences in Malaysia multi-racial student that may affect the result of the model. Later, the extraction of the features takes place which include the detection of the key interest point. We extract the interest point from the training images using SURF descriptor both from MATLAB Computer Vision Toolbox. Local features were detected in the area of the interest point using SURF descriptor due to the speed and memory advantage [34].

In this research, our interest point is related to frontal face e.g.: eyes region and whole face area. We will not focus in depth other face features area such mouth and nose, as we want to make a prediction of engagement and disengagement using whole

frontal face not based on any specific region. We extract all features of each images without restriction of the number of strongest features because the limitation of the number of features will influenced the accuracy of the prediction. After the interest point has been identified, a feature descriptor that includes a key point of pixels region surrounding the interest point was calculated. The purpose of a descriptor is to provide a unique and robust description of a feature generated based on the area surrounding of an interest point.

4.3 Clustering

In the next process, a key point and descriptor information in the form of vector that gathered from feature extraction were grouped into N clusters of visual words using k-means [32]. Each descriptor was clustered into a similar group using Euclidean distance metric. This process also called as generating a codebook of the images. In this research, we compare the codebook with multiple size {300, 500, 800 and 1000} to see the influence of the dictionary size of this model [35]. The codebook was represented using a histogram of gradient orientations in the local neighbourhood around each key point. Every image has visual vocabularies or codebook in the form of histogram that represent the category. Therefore, BOF finds almost similar codebook to be mapped in the same categories.

4.4 Classification

In this section, there are three classifiers has been tested that include pattern recognition based classifier like SVM, probability based classifier such as Naïve Bayes and distance based classifier known as k-NN [32].

SVM. SVM analysed data and recognized pattern from the data. Classification using SVM, compares unknown target features against trained features [36]. The learning process of SVM resides on the construction of hyperplanes to distinguish between the category based on the decision boundaries [37]. We used a library for SVM packages in WEKA named as LIBSVM, based on the reputation as one of the fast and effective classifier [24]. It is important to include suitable kernel selection to spaces higher dimension. Therefore in this study, we used two types of SVM kernel that mostly used in BOF namely Radial Basis Function (RBF) and Linear [38].

Naïve Bayes. Naïve Bayes often used in categorical classification that utilise the maximum posteriori decision rule [39]. It is reported that Naïve Bayes outperforms in image classification in term of accuracy and minimum training time speed. This classifier was adopted from Bayes theorem which describes the probability of an event belong to the class can be measured based on prior knowledge of conditions that might be related to the event. Therefore it also suitable for binary classification.

k-NN. We also used k-NN because of the simplicity yet efficient machine learning algorithm [32]. This is important to find the best classifier in term of performance in this model. K-NN uses distance measure for BOF representation. From the training images, the process of classification will be based on most similar instances in each

groups. In this study, we used package of k-NN in WEKA known as IBk with k = 1 or 1NN for the classification of engaged and disengaged images.

4.5 Performance Metric

We used performance measures for binary classification which highlight the accuracy of the prediction (Eq. 1) [37]. *Accuracy* is to evaluate overall effectiveness of the model. The classification were measured based on true positive (*tp*), true negative (*tn*), false positive (*fp*) and false negative (*fn*) that matched with true label prediction. *Positive* means the classification is engaged and *negative* means disengaged. *True* on the other hand, represent an actual label and *false* for not an actual label.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

5 Result

In this section, we discuss the engagement prediction performance using our collected images data during the preliminary study. In addition, we also compare the difference size of codebook with {300, 500, 800 and 1000} and also report the comparison result of different classifiers.

The total of 600 images collected from 30 participant was divided equally with engaged and disengaged faces. Table 2 shows sample of engaged and disengaged images used in this study. While Table 3 shows, the key interest point using SURF descriptor that successfully detected on the engaged and disengaged images. From the detected interest point, SURF descriptor able to detect most of the important point for engagement element such as eyes and frontal face.

Table 2. Sample of engaged images (right) and disengaged images (left)



In comparing with accuracy of using SURF in a different codebooks size, we used SVM classifier with 10-folds cross validation due to the small samples size of our dataset [40]. LIBSVM was chosen as SVM classifier with adjustment on kernel using Linear and RBF. In addition, we also test the data with Naïve Bayes and k-NN to find

Table 3. Key-interest point detection using SURF for engaged (right) and disengaged (left)



the impact of a different classifier. Figure 3, shows the comparison of performance using multiple classifier and different size of codebooks.

The highest accuracy reported that prediction using classifier LIBSVM (Linear)

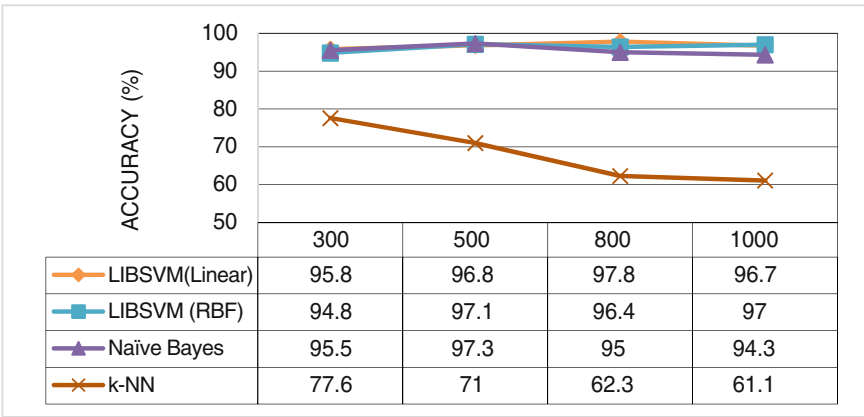


Fig. 3. Accuracy (%) of the engagement prediction using four classifiers with different size of codebooks

with 800 codebook with 97.8%, LIBSVM with RBF kernel using 500 recorded with 97.1%, and Naïve Bayes using 500 codebook achieved 97.3%. We have to exclude the k-NN classifier since it failed to predict a good result for all the codebook. We present the confusion matrix in Table 4 for these top three classifiers based on the codebook sizes.

From the confusion matrix, all of these classifiers able to distinguish engaged and disengaged images with a slight errors. False classification majority from the disengaged images. This is due to the vast description of disengaged and a variety of gestures displayed such as looking at the screen with one eyes as well as occlusion of hands and accessories.

Table 4. Confusion matrix using LIBSVM (Linear) versus LIBSVM (RBF) versus Naïve Bayes

Classifier types		Predicted			Accuracy %
LIBSVM (linear) 800 codebook	Actual		Engaged	Disengaged	97.8
		Engaged	292	8	
		Disengaged	5	295	
LIBSVM (RBF) 500 codebook		Engaged	299	1	97.1
		Disengaged	17	283	
Naive Bayes 500 codebook		Engaged	290	10	97.3
		Disengaged	8	292	

6 Discussion and Conclusion

In this paper we present, the application of machine learning approach known as bag of features which use frontal face for image classification. The experiment was conducted to find the suitable codebook to classify engaged and disengaged face image. In addition, we also intended to search for the suitable classifier that able to produced high accuracy. Majority the selection of codebook sizes depends on the selection of clas-sifier for example LIBSVM with RBF kernel and Naïve Bayes perform well when using with 500 codebook. However, LIBSVM with Linear kernel perform better with highest accuracy when using 800 size codebook. This is due to the functionality of the kernel that influence the decision boundary [41].

The challenge in this study is when there are varies of gesture that reflect engaged and disengaged. Even we have a clear rules to consider engaged through detection of frontal face with eyes open, there was a situation we failed to predict when the student yawn with face and eyes partially open. Due to our limited images of a similar situation we does not include these type of images in this paper. Nevertheless, the usage of BOF in image classification able to distinct engaged and disengaged faces when using key interest point descriptor such as SURF, clustered using k-means and finally classify using SVM and Naïve Bayes. Hence, it is suitable to be applied in our proposed Dyslexia Adaptive Learning Model.

Acknowledgements. Special thanks to Dyslexia Association Malaysia (DAM), UPM IPS grant for the research funding and university’s ethics committee who approved our application to conduct this study.

References

1. Benyon, D., Murray, D.: Applying User Modelling to Human-Computer Interaction Design (1993)

2. Siti Zulaiha, A., Nik Noor Amalina Amirah, N.L., Hawa, M.E., Arifah Fasha, R., Mohd Hafiz, I.: Bijak Membaca—applying phonic reading technique and multisensory approach with interactive multimedia for dyslexia children. In: CHUSER 2012—2012 IEEE Colloquium on Humanities, Science and Engineering Research, pp. 554–559 (2012)

3. Slavuj, V., Kova, B., Jugo, I.: Intelligent tutoring systems for language learning. In: Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE (2015)
4. Chu, C.N., Yeh, Y.M.: Adaptive reading: a design of reading browser with dynamic alternative text multimedia dictionaries for the text reading difficulty readers. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 661–664 (2010)
5. Francis, D.J., Shaywitz, S.E., Shaywitz, B.A., Stuenkel, K.K.: Developmental lag versus deficit models of reading disability: a longitudinal individual growth curves analysis. *Educ. Psychol.* **88**, 3–17 (1996)
6. Oga, C., Haron, F.: Life experiences of individuals living with dyslexia in Malaysia: a phenomenological study. *Procedia—Soc. Behav. Sci.* **46**, 1129–1133 (2012)
7. Ndombo, D.M.: An intelligent integrative assistive system for dyslexic learners. *J. Assist. Technol.* **7**, 172–187 (2013)
8. Rello, L., Ballesteros, M., Bigham, J.P.: A spellchecker for dyslexia. In: ASSETS 2015: the 17th International ACM SIGACCESS Conference of Computers and Accessibility, pp. 39–47 (2015)
9. Franceschini, S., Bertoni, S., Ronconi, L., Molteni, M., Gori, S., Facoetti, A.: “Shall We Play a Game?": Improving Reading Through Action Video Games in Developmental Dyslexia. *Current Developmental Disorders Reports* (2015)
10. Rauschenberger, M.: DysMusic: Detecting Dyslexia by Web-based Games with Music Elements. *Web4all'16* 7–8 (2016)
11. Saputra, M.R.U.: LexiPal: Design, implementation and evaluation of gamification on learning application for dyslexia. *Int. J. Comput. Appl.* **131**, 37–43 (2015)
12. Azhar, R., Tuwohingide, D., Kamudi, D., Sarimuddin, Suciati, N.: Batik image classification using SIFT feature extraction, bag of features and support vector machine. In: *Procedia Computer Science*. pp. 24–30. Elsevier Masson SAS (2015)
13. Yang, S., Bebis, G., Chu, Y., Zhao, L.: Effective face recognition using bag of features with additive kernels. *J. Electron. Imaging* **25**, 13025 (2016)
14. Al-Dmour, A., Abuhelaleh, M.: Arabic handwritten word category classification using bag of features. *J. Theor. Appl. Inf. Technol.* **89**, 320–328 (2016)
15. Surakarin, W., Chongstitvatana, P.: Predicting types of clothing using SURF and LDP based on Bag of Features. In: *ECTI-CON 2015–2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (2015)
16. Shaywitz, S.E., Shaywitz, B. a.: Dyslexia (specific reading disability). *Biol. Psychiatry* **57**, 1301–1309 (2005)
17. Sahari, S.H., Johari, A.: Improving reading classes and classroom environment for children with reading difficulties and dyslexia symptoms. In: *Asia Pacific International Conference on Environment-Behaviour Studies*. pp. 100–107. Elsevier B.V. Selection (2012)
18. Pardos, Z.A., Baker, R.S.J., Pedro, M.O.C.Z.S., Gowda, S.M., Gowda, S.M.: Affective States and State Tests: Investigating How Affect and Engagement During the School Year Predict End of Year Learning Outcomes (2012)
19. Whitehill, J., Serpell, Z., Yi-Ching Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagement from facial expressions. *IEEE Trans. Affect. Comput.* **5**, 86–98 (2014)
20. Cetintas, S., Si, L., Xin, Y.P., Hord, C.: Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Trans. Learn. Technol.* **3**, 228–236 (2010)

21. El Khayat, G.A., Mabrouk, T.F., Elmaghraby, A.S.: Intelligent serious games system for children with learning disabilities. In: *Proceedings of CGAMES'2012 USA—17th International Conference on Computer Games AI, Animation, Mobile, Interactive, Multimedia, Educational Serious Games*, pp. 30–34 (2012)
22. Magnisalis, I., Demetriadis, S.: Karakostas, a: adaptive and intelligent systems for collaborative learning support: a review of the field. *IEEE Trans. Learn. Technol.* **4**, 5–20 (2011)
23. Li, A.Q.: PoliSpell: an adaptive spellchecker and predictor for people with dyslexia. In: *Proceedings of the 21th International Conference, UMAP 2013, Rome, Italy, 10–14 June 2013*, pp. 302–309 (2013)
24. Li, K., Wang, F., Zhang, L.: A new algorithm for image recognition and classification based on improved Bag of Features algorithm. *Opt.—Int. J. Light Electron Opt.* **127**, 4736–4740 (2016)
25. Hiba, C., Hamid, Z., Omar, A.: Bag of features model using the new approaches: a comprehensive study. *Int. J. Adv. Comput. Sci. Appl.* **1**, 226–234 (2016)
26. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 (1999)
27. Mansourian, L., Abdullah, M.T., Abdullah, L.N., Azman, A.: Evaluating classification strategies in bag of SIFT feature method for animal recognition. In: *Research Journal of Applied Sciences, Engineering and Technology*. pp. 1266–1272 (2015)
28. Feng, J., Liu, Y., Wu, L.: Bag of visual words model with deep spatial features for geographical scene classification. *Hindawi Comput. Intell. Neurosci.* **2017** (2017)
29. Schmitt, D., McCoy, N.: Object Classification and Localization Using SURF Descriptors (2011)
30. Siti Suhaila, A.H., Novia, A., Abdul Azim, A.: Computer-based learning model to improve learning of the Malay Language amongst dyslexic primary school students. In: *Proceedings of the Asia Pacific HCI and UX Design Symposium*, pp. 37–41 (2015)
31. Tan, L., Sun, X., Khoo, S.T.: Can engagement be compared? Measuring academic engagement for comparison. In: *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, pp. 213–216 (2014)
32. Nasirahmadi, A., Miraei Ashtiani, S.H.: Bag-of-Feature model for sweet and bitter almond classification. *Biosyst. Eng.* **156**, 51–60 (2017)
33. Abdelkhalak, B.: Hamid Zouaki: a surf-color moments for image retrieval based on bag-of-features. *Eur. J. Comput. Sci. Inf. Technol.* **1**, 11–22 (2013)
34. Pancel, P., Pancel, S., Shah, S.: A comparison of SIFT and SURF. *Int. J. Innov. Res. Comput. Commun. Eng.* **1**, 323–327 (2013)
35. Wang, R., Ding, K., Yang, J., Xue, L.: A novel method for image classification based on bag of visual words. *J. Vis. Commun. Image Represent.* **40**, 24–33 (2016)
36. Krig, S.: *Comput. Vis. Metrics* (2016)
37. Pérez, D.S., Bromberg, F., Diaz, C.A.: Image classification for detection of winter grapevine buds in natural conditions using scale-invariant features transform, bag of features and support vector machines. *Comput. Electron. Agric.* **135**, 81–95 (2017)
38. Anthimopoulos, M., Gianola, L., Scarnato, L., Diem, P., Mougiakkou, S.: A food recognition system for diabetic patients based on an optimized bag of features model. *IEEE J. Biomed. Heal. Inform.* **18**, 1261–1271 (2014)
39. Dong, C.P.: Image classification using Naive Bayes classifier. *Int. J. Comput. Sci. Electron. Eng.* **4** (2016)

40. Karaaba, M.F., Surinta, O., Schomaker, L.R.B., Wiering, M.A.: Robust face identification with small sample sizes using bag of words and histogram of oriented gradients. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 582–589 (2016)
41. Huang, H.-Y., Lin, C.-J.: Linear and Kernel classification: When to Use Which? In: Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 216–224 (2016)

Towards Designing Tangible Interaction for Children with Dyslexia in Learning the Malay Language

Siti Nurliana Jamali^(✉), Novia Admodisastro^(✉),
Siti Suhaila Abdul Hamid, Azrina Kamaruddin,
Abdul Azim Abd Ghani, and Sa'adah Hassan

Universiti Putra Malaysia, Serdang, Malaysia
{gs45705, gs42041}@student.upm.my, {novia, azrina, azim,
saadah}@upm.edu.my

Abstract. In this paper, we provide a study of tangible interaction (TI) based on theories and related works for dyslexic children. The study is an attempt to investigate TI for dyslexic children in learning Malay language in Malaysia primary schools. TI has tremendous contribution in supporting dyslexic children to enhance their way of learning process. However, TI that were developed currently have different capabilities and purposes. For example, current works currently only developed for other languages like English, Mandarin and Dutch. The TI model for English or other languages may not be suitable to be adopted directly for the Malay language due to differences of letter sound, morphology and etc. There were nine related works reviewed in this study. Based on these previous related works and learning theories we designed a conceptual TI model for dyslexic children in learning the Malay language.

Keywords: Tangible interaction · Dyslexia · Children · Language difficulty Malay language

1 Introduction

The Malay language is a core subject in Malaysia primary syllabus that covers a period of 6 years from Standard 1 to Standard 6. The subject focuses on language skills, namely listening and speaking, reading, and writing skills. This language skill enables students to effectively use the Malay language in daily life as well as for the acquisition of knowledge. The Malay language has distinctive characteristics compare with other languages, where it has shallow alphabetic orthography (i.e. spelling-sound mappings are predictable and transparent), simple syllabic structures and transparent affixation [1]. For example, the Malay language is using affixation to form new words from a root word by using prefixes (e.g. *belajar*), suffixes (e.g. *ajaran*), circumfixes (e.g. *pengajaran*) and infixes (e.g. *kemelut* from *kelut*), while the English language emphasis on prefixes and suffixes.

The statistic by Ministry of Education (MoE) of Malaysia has revealed 65% registered disabilities in 2012 are children with learning difficulties and dyslexia prevail as

the most common specific learning difficulty [2]. Dyslexia is a language disability, affecting reading and writing, speaking and listening. It is an impairment on the use of words including when the children learn language subject such as the Malay language. Consequently, performance in every subject can be affected by dyslexia. Nevertheless, with proper help children with dyslexia can learn to read and write well. Most of the children with dyslexia need help from a teacher, tutor or therapist specially trained in using a multisensory structured language approach [3]. It is important for these children to be taught by a method that involves several senses such as seeing, hearing, touching at the same time. Links are consistently made between the visual (language we see), auditory (language we hear), and kinaesthetic-tactile (language symbols we feel) pathways in learning to read and spell.

Recent research has shown tangible interaction (TI) provides benefits for dyslexic children's learning [4, 5]. TI is a computing system which allows tangible manipulation objects in the physical world to interact and embedded with digital information in the virtual world. TI facilitate dyslexic children with physical and sensory materials, practical and concrete examples to illustrate explanations, and organising cooperative learning groups [5]. For example, physical and spatial properties of TI may be used to enable hands-on interaction with 2D or 3D letters in linear sequences in 2D space that promote effective reading acquisition [4]. Nevertheless, TI related works for dyslexia mostly designed for other languages such as English, Mandarin and Dutch and fewer were addressing letter-sound correspondences, word recognition and phonological awareness [6, 7]. In addition, the works mainly lack in utilising the viability of TI in supporting collaboration and social interaction learning which considers important for dyslexic children [5]. A study in [8] explored Dyslexia Association Malaysia (DAM) profoundly utilising multisensory approach by using materials such as flash-cards, blocks, dominoes, boards and sticks in teaching the Malay language to dyslexic children. The used of these traditional materials were less attractive to children, lack of sense, spatiality and feedback such as sound as well as rely heavily on the teacher assistance when performing the learning activities.

In this paper, we explore previous related works and provide an initial model of TI to support the Malay language learning for children with dyslexia. The rest of this section is organised as follows: the second section presents the background study, the third discusses related works, the fourth section propose TI model derived from the previous works, informed theories as well as the current teaching approach. Finally, the paper ends with a conclusion and future work.

2 Background Study

2.1 Dyslexia

Dyslexia is a language disability, affecting reading and writing, speaking and listening. According to [9] dyslexia classified into six types of difficulties: to learn language, imbalances intellectual abilities, not fluent when reading printed material, unable to write smoothly and accurately (difficult to copy words from blackboard or books), eyes become tired when focus intensely on certain words and very limited concentration in

terms of visual and hearing. Based on these different difficulties dyslexics can be categorised according to the sensory system (1) Visual Dyslexia—refers to pupils who can see well but could not interpret or remember things seen. They also have trouble pronouncing a word that has a lot of syllables. (2) Auditory Dyslexia—refers to pupils having problems to distinguish similarities and differences between the sound heard, identifying each sound in words, blending sounds to make words, and divide the word into syllables. Sometimes the children could hear the words but some of the information is lost when the brain processes the sound and sounds may be being fused, reversed or jumbled. Most dyslexics have the auditory type. (3) Visual-Auditory Dyslexia—pupils with difficulty using both senses i.e. sight and hearing. The effect will pass interference in the process of receiving information through visual and auditory.

2.2 Learning Theory for Dyslexia

Learning theories are conceptual frameworks describing how knowledge is absorbed, processed, and retained during learning. This section presents common learning theories in teaching children with dyslexia and how we could inform TI design based on these theories which will be implemented in our work.

Theory of Orton Gillingham (OG). OG theory suggested kinaesthetic-tactile reinforcement of visual and auditory associations could correct the tendency of confusing similar letters and transposing the sequence of letters while reading and writing [10]. Traditional adoption of OG theory commonly using letter tiles such as flash cards and cubes to capture children's attention to the letters-sound correspondence. The physical activity involves letter tracing which could assist dyslexic children to identify mirrored letter shapes as well as memorise letter and sounds. For example, one of the techniques the Wilson Reading System uses is a “sound-tapping” system. Students tap out each sound of a word with their fingers and thumbs to help them break the words down.

Theory of Dual Coding. The theory focuses on a relationship between a symbolic system and sensory motor system. For example, in the visual modality, there are printed words and pictures. In auditory modality, there will be spoken words and sound events. Tiger can be stored as an image of a tiger and the word tiger or both events but in different systems. Verbal and non-verbal will function independently. According to [11], verbal will process information such as text and audio meanwhile non-verbal will process visual information such as diagrams, animations and photographs.

Theory of Kinaesthetic or Tactile Learning. This theory requires children to manipulate, organise and interpret the information they have experience through tactile (touch) or kinaesthetic (movement) [12]. Dyslexic children learn well when they can perform body movements, or use their hands and provide senses of touch. Writing and drawing also can be associated with physical activities. One common kinaesthetic teaching method used with dyslexia student is ‘air writing’ whereby students say a letter out loud while simultaneously writing it in the air [13].

2.3 Tangible Interaction

For the last four decades, Human-Computer Interaction (HCI) researchers have continued to explore a wide range of interaction styles and interfaces that diverge from the Graphical User Interface (GUI), Web User Interface (WUI) and Voice User Interface (VOI) [14]. Predominantly adopted GUI is driven with working on a desktop computer, using a mouse and a keyboard to interact windows, icons, menus, and pointers (WIMP). Gradually, the concept of interaction amends from foreground activities into both foreground and background that is invisible but aware and ubiquitous [14]. Tangible User Interfaces (TUIs) has been introduced as a new interface type that interlinks the digital and physical worlds (refers Fig. 1) [15]. TI can be described as an umbrella that comprises of TUIs. It is pertaining to users' knowledge and skills of interaction with the real non-digital world. TI has shown a great potential to enhance the way in which people interact with and leverage digital information. The design principles of TI emphasise on four themes: (1) *tangible manipulation* (2) *spatial interaction* (3) *embodied facilitation* and (4) *expressive representation* [16].

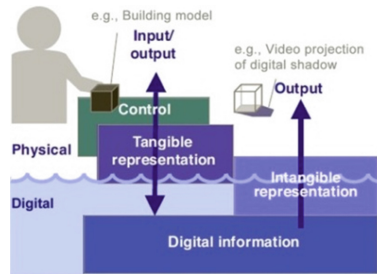


Fig. 1. Tangible user interface interaction model [15]

The *tangible manipulation* refers to bodily interaction with physical objects. These objects are embedded with computational resources that allow users to control computation. The *spatial interaction* refers to the detail that tangible interaction is embedded in real space and users make movement in real space when interacting. For example, bowling games in Wii that will use hand to throw the bowling ball in physically captured in the system. While *embodied facilitation* describes on the act or move in physical space and metaphorically in system space (software). Physical space prescribes physical structure and software defines virtual structure, determining the interaction flow that could facilitates, prohibits or hinders some actions, and allowing, directing, or limiting behaviour. Lastly, *expressive representation* describes physical representation of digital functions and data, or of other physical objects.

2.4 Tangible Interaction and Learning

While computer allows and provides various dynamic content and interactive systems there are limitations to engage children in realistic modes of representation [17]. This is due to the conventional method of the computer that does not support simultaneous

interactions and feedback as well as physical exploratory for children. Theories of learning and cognition such as [10–12] offer a compelling rationale for using tangible and embodied interaction for supporting learning and collaborative activity [18]. A number of tangible systems developed for learning have provided tremendous achievement to children’s learning [4, 5]. These systems were moving the digital interface to physical world provides multimodality interfaces using various sensory means (multisensory) [18, 19]. Similarly, [18] recommended the same strategy to enhance learning by providing concrete examples and cooperative learning environment. Meanwhile, [20] have recommended using TI for children with learning difficulties to promote children engagement and enjoyment in playful learning environments. The work by [17] also mentioned TI is more accessible to children as well as assumed to be more natural or familiar for children with learning difficulties.

In [5] using a TI approach for children with learning difficulties that reflected upon the educational resources design guidelines: *kinaesthetic method*, *modes of representations*, *collaboration* and *scaffolding*. Firstly, the *kinaesthetic method* consists of physical engagement such as moving, touching and manipulating. Secondly, *mode of representations* is where children with learning difficulties prefer concrete examples in order to make them understand abstract ideas. Adding to that these children have a higher concentration on oral interaction as well as dynamic pictorial and lower number of text. Third, *collaboration* encourages forming group work in order to provide support to one another and chance to observe others while considering individual expression. Lastly, *scaffolding* to ensure each of the children would be able to have the equal chance to participate which offers various needs and level of ability to the children.

In learning difficulty that related to language processing, several attempts in TI were introduced to improve skills of letter-sound decoding, phonological awareness, understand alphabet shape, reading etc. [6, 7, 13, 19, 21–23]. These attempts underlying children’s learning of language skills, particularly beneficial for dyslexic children. For example, [13] develop letter sequence using hands-on activity in real space for dyslexic children. The dyslexic children practice on letter tracing activities and perceive physical affordance like moving, manipulating and feeling the physical letters. Dyslexic children with different sensory classification are provided with various modalities e.g. tactile, auditory and visual. The following sections present further discussion of related works for dyslexic children.

3 Related Works

There are nine related works of TI that addresses language processing skills reviewed in this paper. These were distinctive in terms of purpose, language skill(s) addressed, type of modality(s), language support, platform, evaluation and target users.

The works in I-BLOCKS [24], has provided support for linguistic scenario and construct sentences for children with dyslexia that children can control by allowing them to have interaction with the building block. Also, emphasize on the use of physicality of the children body movements through spatial and kinaesthetic approach as well as allow rich multisensory interaction. Another work *LinguaBytes* [25] was

developed as a tangible system dedicated to encourages literacy development for Dutch preschool students. The students learn to communicate by telling a story and also combine letter sounds activities. Yet, *LinguaBytes* only provide for the beginner level which promotes one to one letter sound relations as well as having inconsistency letter sound mapping. In *Tiblo* [6] the work helped in enhancing children motor skills whereby involves connecting and attaching the blocks tangibly. Children also could improve their social collaboration where each of them will have their own representations in terms of sound, and image they created and collaborate between one another. Besides that *Tiblo* has greater significance impact where children can create storytelling and story building tools in collaborative ways. *Tiblo* provides limited support in letter-sound correspondences and mainly use to address reading difficulties. *Tiblo* provides training for dyslexic children on recognising letter sound by incorporating multisensory approach such as auditory, visual, tactile and kinaesthetic. Besides, audio is embedded in the *Tiblo* to offer feedback and increase children attention to letter sounds.

Spellbound [7] is a tangible system that helps dyslexic children in learning letter-sound correspondences which enable children to develop 2D letters by locating 2D letter on the platform to produce letter sound and image of the words. Yet the work was still in the initial stage and the system was not completely implemented. Other work in *SpellingCube* [26] provides learner in learning of letter-sound correspondences where the learner has to switch each of the cubes to choose the exact surface and combine cube together to construct a word or sentence. The system only utilises common block to symbolise letters and focus on letter sounds but does not support for letter tracing activity. In *E-Talk Pen* [22] offered the use of a special pen to support phonological awareness in Mandarin phonetic symbols and communication training for speech disorders children. The study involved three sections which are phoneme segmentation, phoneme blending as well as tone awareness. The result shown a tremendous score when using the *E-Talk Pen* compare without using it. This indicates manipulation of tangible object could give benefits the children in learning as well as improving efficacy in teaching.

Another work is *PhonoBlocks* [19] a tangible user interface to a reading system that uses dynamic colour cues embedded in 3D tangible letters to provide additional decoding information and modalities for children aged 5–8 years old, who are having difficulty learning to decode English letter-sound pairs. The work has addressed several TI design opportunities in assisting dyslexic to read such as spatiality, various types of interaction modalities, multiple ways of letter representations as well as structured procedures. *PhonoBlocks* allows concurrent use of visual, auditory as well as kinaesthetic or tactile approach in both physical and digital representations. It also focuses on the 3D design of tangible letters which facilitate dyslexic children to learn letter as a basis rather than words because they often struggling with letter-sound correspondences and mirror-letter.

TraceIt [13] is developed to helps beginner level reader to read by using multi-sensory approach. This includes kinaesthetic movements by using their body movements. This program lets the children trace the letters using physical objects that interact to the program. The program also provides audio feedback when the children trace each alphabet. Besides that, a colour based recognition is included to detect hand

motion through tracked coloured objects. The final work is Tactile Letters [23] a multimodal tangible tabletop that comprises of texture cues to teach English alphabet sound to dyslexic children aged around 5–6 years old. The prototype is developed as an instrument to examine the role of texture cues for the children to learn the alphabets. Two set of tangible letters were designed with or without texture cues and each set is comprised of 24 pieces of letter cards. The children could choose to locate the 3D tangible letters or use the letter cards on the interactive tabletop. Each of the tangible letters produced audio feedback to the children whenever they provide a correct or incorrect connection with the 3D tangible letters. Based on the researchers' findings they proposed three design goals: (1) systematic learning where learning must be initiated from simple to complex, (2) texture cues that identify letters with texture cues or no texture, (3) affordance that aid dyslexic children to differentiate mirror or flip letters.

In conclusion from the related works (refers Table 1), most of the works have utilising TI in supporting dyslexic children to decode letter correspondences, phonology and reading. The existing TI works were also only available in other languages such as English, Mandarin and Dutch. The platform used in the related works were various from educational robotics, interactive tangible alphabets, tangible blocks to educational software systems. In related to modality types, most of the related works were using tactile, visual, audio as well as kinaesthetic approach either separately or combinations. As for evaluation, common approaches used in the related works were experimentation, Wizard of Oz prototype, user acceptance testing, questionnaire and observation techniques. Lastly, the related works were developed either for dyslexic children specifically or children with difficulty in a specific sensory system. Nevertheless, works attempted in a specific sensory system in much relevant with dyslexics which facing different sensory difficulties.

4 The Tangible Interaction Model for Dyslexic Children in Learning the Malay Language

This section describes a proposed TI model to support dyslexic children learning letter-sound correspondence, word recognition, single sound values, and reading comprehension for the Malay language using 3D tangible letters (refers Fig. 2). The model intended to utilised TI in different modalities to support dyslexic children of different sensory types difficulties such as visual, auditory and kinaesthetic. The TI model implemented based on the OG theory, dual coding theory and kinaesthetic learning theory. In adopting OG theory the children are able to touch and feel the 3D tangible letters. While, in dual coding theory the children are able to see and listen to the picture with related sound play. The kinaesthetic learning theory in the TI model allows the children to use their body movement and gestures to arrange, manipulate, and rotate the 3D tangible letters.

The TI model begin with physical world that involve dyslexic children who are interacting with the 3D tangible letters by manipulating the letters and use their different sensory to connect with the digital representation in virtual world. The dyslexic children use their body movements to grasp the 3D tangible letters and arrange it on the

Table 1. Comparison of existing related works of TI for children with dyslexia

No	Model	Difficulty	Language	Platform	Modality	Evaluation	Target User
1.	I-BLOCKS	To support in linguistic problem for dyslexia and aphasia	English	Educational Robotics	Touch, Visual, Audio	Wizard of Oz, experiment	Dyslexic and aphasia children
2.	LinguaBytes	To support language development through letter sound activity	Dutch	Interactive tangible	Touch, Visual, Audio	Questionnaire, Observation	Cerebral Palsy children
3.	SpellBound	To support understanding of alphabet shape	English	Interactive electronic alphabet shape	Touch, Visual	User acceptance testing	Dyslexic children
4.	TIBLO	To support learning for English decoding letter-sound pairs	English	Interactive electronic blocks	Touch, Visual, Audio	User acceptance testing	Dyslexic children
5.	SpellingCube	To support in learning letter sound correspondence and spelling	English	Interactive blocks	Touch, Visual, Audio	Experiment, Observation	Preschool children
6.	E-Talk Pen	To support in phonology	Mandarin	Educational technology software	Touch, Visual, Audio	Experiment	Speech disorder
7.	PhonoBlocks	To support decoding problem letter-sound pairs	English	Interactive tangible	Touch, Visual, Audio	Experiment in progress project	Dyslexic children
8.	Tracelt	To support in reading through multi-sensory.	English	Educational technology software	Visual, Audio, Kinesthetic	Experiment	Dyslexic children
9.	Tactile Letters	To support English letter sound correspondence in reading	English	Interactive table associated with texture letter	Touch, Visual, Audio	Experiment	Dyslexic children aged 5–6 yrs. old

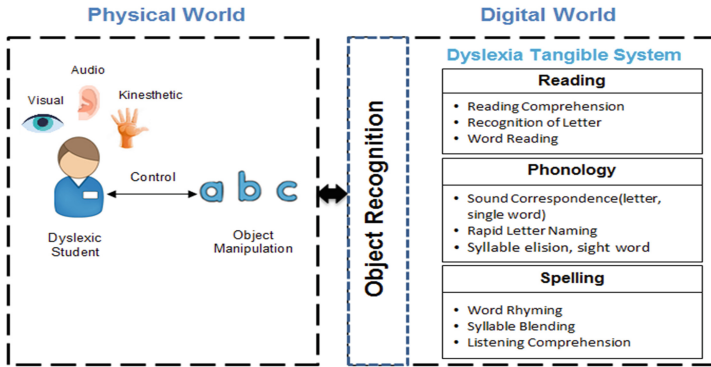


Fig. 2. The proposed TI model for dyslexic children

platform that hold the letters. The arrangement could either be a single letter or a set of letters to form a word depending on the exercises given. Once the arrangement completed, it is link to the TI program. The 3D tangible letters are recognized by the TI program using a mounted camera which capture the 3D letters and convert it to an image. The program could recognise the 3D letters using algorithm such as template matching [27–29]. In digital world, dyslexia tangible system is consisted of three main modules incorporates reading, phonology and spelling activities. The reading module support the children to perform activities such as reading comprehension, letter recognition and word reading. The children able to touch, arrange, move and trace out letters in different patterns as they wished. The phonology module applies dual coding theory where students will make sound correspondence, rapid letter naming and syllable elision as well sight word activities. Then the system is able to give audio and visual feedback to the dyslexic children by providing sound of the letter and recognize the word correctly on the screen. The spelling module is incorporating word rhyming, syllable blending and listening comprehension activities.

5 Conclusion and Future Work

The paper presented a study of existing related works of TI for dyslexic children. The study reveals the limitation of existing works due to their different capabilities and purposes. However, based on the study we have proposed a conceptual design of TI model for dyslexic children in learning the Malay language in Malay primary school. The model also has taken consideration of learning theories that related to dyslexic children. A tangible system is going to be developed based on the proposed TI model which focuses assisting dyslexic children in reading, spelling and phonology.

Acknowledgements. Special thanks to Dyslexia Association Malaysia (DAM) for great assistance in this research. We would also like to thank the University for funding the research to conduct the study.

References

1. Adelman, J.S.: Visual Word Recognition: Models and Methods, Orthography and Phonology, vol. 1. Psychology Press, Hove, East Sussex (2012)
2. UNICEF Malaysia: Children with Disabilities in Malaysia: Mapping the Policies, Programmes, Interventions and Stakeholders. UNICEF (2014)
3. International Dyslexia Association: Multisensory Structured Language Teaching (2017). Accessed from <https://dyslexiaida.org/multisensory-structured-language-teaching/>
4. Marshall, P.: Do tangible interfaces enhance learning? In: Proceedings of the International Conference on Tangible and Embedded Interaction. ACM (2007)
5. Falcão, T.P., Price, S.: Informing design for tangible interaction: a case for children with learning difficulties. In: Proceedings of the International Conference on Interaction Design and Children. ACM (2010)
6. Pandey, S., Srivastava, S.: Tiblo: a tangible learning aid for children with dyslexia. In: Proceedings of the Conference on Creativity and Innovation in Design. ACM (2011)
7. Pandey, S., Srivastava, S.: SpellBound: a tangible spelling aid for the dyslexic child. In: Proceedings of the International Conference on Human Computer Interaction. ACM (2011)
8. Hamid, S.S.A., Admodisastro, N., Ghani, A.A.A.: Computer-based learning model to improve learning of the malay language amongst dyslexic primary school students. In: Proceedings of the APCHIUX, OzCHI2015. ACM (2015)
9. British Dyslexia Association.: Assessing Reading Difficulties: A Diagnostics and Remedial Approach. Windsor: NFER-Nelson (1999)
10. Reid, G.: Dyslexia: A Practitioner's Handbook. Wiley (2016)
11. Clark, J.M., Paivio, A.: Dual coding theory and education. *Edu. Psychol. Rev.* **3**(3), 149–210 (1991)
12. Kinesthetic Learning Strategies, Kinesthetic Learning Strategies for Various Subjects. Accessed from <http://www.kinestheticlearningstrategies.com/kinesthetic-learning-strategies-for-various-subjects/> (2017)
13. Teh, T.T.L., Ng, K.H., Parhizkar, B.: TraceIt: an air tracing reading tool for children with dyslexia. In: Advances in Visual Informatics, vol. 9429. LNCS, Springer. (2015)
14. Hornecker, E.: Physicality in tangible interaction: bodies and the world. In: Position Paper of the International Workshop on Physicality. University of Lancaster (2006)
15. Ullmer, B., Ishii, H.: Emerging frameworks for tangible user interfaces. In: Carroll, J.M. (ed.) HCI in the New Millennium. Addison-Wesley Pub, Reading, MA (2001)
16. Hornecker, E., Buur, J.: Getting a grip on tangible interaction: a framework on physical space & social interaction. In: Proceedings of the ACM SIGCHI CHI. ACM (2006)
17. Zuckerman, O., Arida, S., Resnick, M.: Extending tangible interfaces for education: digital montessori-inspired manipulatives. In: Proceedings of the ACM SIGCHI CHI. ACM (2005)
18. Price, S., Sheridan, J.G., Falcao, T.P., Roussos, G.: Towards a framework for investigating tangible environments for learning. *Int. J. Arts Tech.* **1**(3–4) (2008)
19. Antle, A.N., Fan, M., Cramer, E.S.: PhonoBlocks: a tangible system for supporting dyslexic children learning to read. In: Proceedings of the International Conference on Tangible, Embedded and Embodied Interaction. ACM (2015)
20. Price, S., Rogers, Y., Scaife, M., Stanton, D., Neale, H.: Using 'Tangibles' to promote novel forms of playful learning. *J. Interact. Comput* **15**(2), 169–185 (2003)
21. Lund, H.H., Marti, P., Palma, V.: Educational robotics: manipulative technologies for cognitive rehabilitation. In: Proceedings of the International Symposium on Artificial Life and Robotics (AROB). Oita, Japan (2004)

22. Lin, C.Y., Chai, H.C.: Using an e-talk pen to promote phonological awareness on communication training. In: Proceedings of the ICCSE. IEEE (2014)
23. Fan, M., Antle, A.N.: Tactile letters: a tangible tabletop with texture cues supporting alphabetic learning for dyslexic children. In Proceedings of the Conference on Tangible, Embedded, and Embodied Interaction. ACM (2015)
24. Marti, P., Lund, H.H.: Novel tangible interfaces for physical manipulation, conceptual constructions and action composition. In: Proceedings of the Intelligent Manipulation and Grasping (IMG04) (2004)
25. Hengeveld, B., Voort, R., Hummels, C., de Moor, J., van Balkom, H., Overbeeke, K., van der Helm, A.: The development of linguabytes: an interactive tangible play and learning system to stimulate the language development of toddlers with multiple disabilities. In: Proceedings of the Advances in Human-Computer Interaction (2008)
26. Goh, W.B., Chamara Kasun, L.L., Fitriani, Tan, J., Shou, W.: The i-cube: design considerations for block-based digital manipulatives and their applications. In Proceedings of the Conference on Designing Interactive Systems Conference. ACM (2012)
27. Vijayarani, S., Sakila, M.A.: Template matching technique for searching words in document images. *Int. J. Cybern. Inform. (IJCI)* **4** (6) (2015)
28. Chantara, W., Ho, Y.S.: Object detection based on fast template matching through adaptive partition search. In: Proceedings of the JCSSE. IEEE (2015)
29. Jayanthi, N., Indu, S.: Comparison of image matching technique. *J. Latest Trends Eng. Technol.* **7**, 396–401(2016)

Comparison of Approaches Made to Enhance Pupils' Numeracy Skill

Nur Faizura Ahmad Fuadi^(✉), Muhammad Fakri Othman,
and Norhalina Senan

Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor,
Malaysia

faizurafuadi@gmail.com, {fakri,halina}@uthm.edu.my

Abstract. This paper compares between four instructional design approaches used to enhance pupils' numeracy skill. A literature review performed on four approaches that have potential in enhancing pupil's numeracy skill. Those approaches are e-Learning, Mobile Learning, Gamification and Problem Based Learning (PBL). The review focuses on design and development of the selected approaches and compares those approaches based on previous studies. The literature survey suggests that e-Learning and gamification are the most suitable instructional design tools to enhance pupils' numeracy skill.

Keywords: Instructional design approaches · Numeracy skill · Pupils

1 Introduction

Since the early 1980s schools, colleges and universities have experimented with technology for learning [1]. Cost, flexibility, and versatility are among inspirations frequently referred to for utilizing portable innovations to help to learn [2]. Furthermore, academic achievement is closely linked to learning styles and motivation [3]. Compared to all subjects they are teaching to the kids at school, Mathematics been a problematic issue.

Problem-solving is a major goal of Mathematics education and an activity that can be seen as the pith of numerical considering [4–7]. Most students have not required the skills needed in problem-solving [8–10]. In a research conducted, the result shows that when doing problem-solving questions, students are capable of performing calculations, however, do not have the capacity to interface method with their reasonable information [11, 12]. All of these be reduced if the students get early education started from kindergarten because many studies indicate that early intervention is critical for closing achievement gaps in math [13–16].

2 Approaches Through Various Ways

In this paper, few methods would be given as examples based on previous studies. Those are approaches including e-learning, mobile learning, gamification and problem-based learning. Many researchers did study things to prevent Mathematics

skills problem in schools. They then come up with approaches which they trusted can enable them to fabricate a corpus of learning in boosting engagement, enjoyment and learning experience to the pupils.

2.1 E-Learning

One of the approaches done is e-learning. Its frameworks are utilized to help the communication and appraisal forms in an online course [17] in a sophisticated way yet interesting. In different circumstances, outstandingly in the scholastic setting, engagement triggers positive passionate encounters including fun, energy, and responsibility regarding diligent work [18]. Most computer games and other educational software which currently being used in primary school Mathematics education focus on the first two aspects: number fact knowledge and operation skills (e.g., [19]).

2.2 Mobile Learning

Mobile learning is not the same as e-learning. Mobile learning more accessible anywhere at any time because the medium used to include tablets, mobile phones like androids and iPhones. The table below explained differences between e-learning and mobile learning (Table 1):

Table 1. Differences between e-learning and mobile learning

	Mobile learning	E-Learning
Aim	<ul style="list-style-type: none"> • Instant accessibility of information • Quick knowledge distribution 	<ul style="list-style-type: none"> • Understanding and retention of specific skills, or in-depth knowledge of a subject
Approach	More flexible & informal than E-Learning	Formal structure
Medium	Mobile phones and tablets like iPhones, iPad, Androids, and Blackberries	Computer or Laptop
Content and Design	<ul style="list-style-type: none"> • Easy navigation • Concise micro-lessons • Pictures, videos, and checklists 	<ul style="list-style-type: none"> • Details information and graphics • Usage of media, videos and game-based learning
Duration	3–10 minutes	20 minutes to 1 hour
User Access	Can be accessed anywhere at any time	Designed to be more static and can be accessed at your desk

Source <https://www.learndash.com/mobile-learning-versus-elearning/>

2.3 Gamification

Approaches regarding gamification can be seen in MathQuest [20]. This is an example of a gamification approach and said to be powerful learning environments for a number of reasons [21]. Gaming approach is more effective in enhancing students' knowledge of computer memory concepts and motivational compared to non-gaming approach

[22]. Thus, MathQuest is developed as a final year project to teach about numeracy and basic Mathematics operations [20].

2.4 Problem Based Learning (PBL)

This method has been used for the first time in 1996 in medical education by Barrows and gradually applied to other education fields [23]. PBL is one of the important approaches in education nowadays [24–26]. It is believed to be an effective method for bringing up qualified individuals for the needs of today's societies [23].

3 Design and Development

As for every approaches stated have differences regarding design and development, few explanations based on previous articles overview stated below:

E-learning: E-learning can be found in many online educational programs including any online forum features, WebCT, Blackboard, and MOODLE. Those programs contain different tools such as course builder, syllabus, discussion board, announcement files, grade book and calendar that can be used to set a platform for a constructivist learning environment [27].

Gamification—MathQuest: In an exploratory factor analysis of the user engagement Scale (UES), four factors solution (focused attention, perceived usability, aesthetics, and satisfaction) provided a better fit than the six factors identified in the original UES (O'Brien & Toms, 2008) [28].

As shown in Fig. 1, the user interface looks real and developer also provided the demo for the first timer. This is a game created to help pupils master their Mathematics skills and it was developed using software such as Adobe Flash, Adobe Photoshop, Audacity and Sound Forge [20]. The outline of the learning module has taken into contemplations academic issues and instructional design [20]. The development of the module is adapted from the star model [29] in Fig. 2.

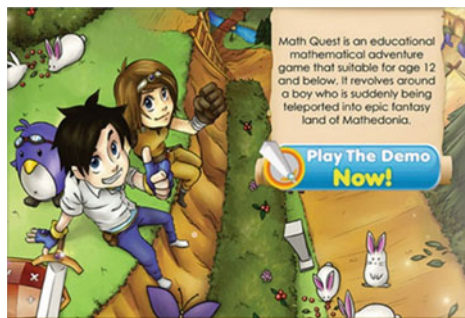


Fig. 1. MathQuest screenshot 1

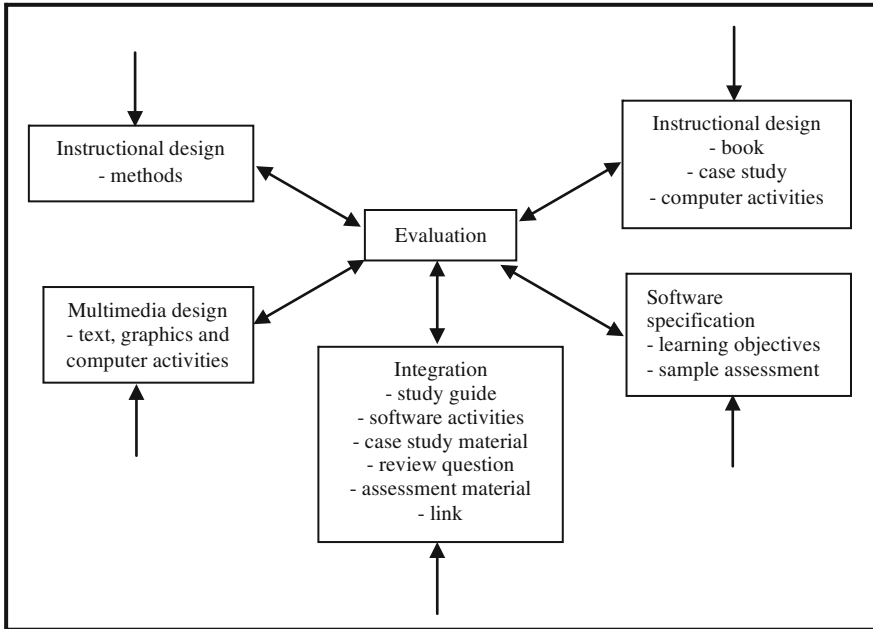


Fig. 2. Star life cycle model.

Gamification—DimensionM™: It is a set of mini-games that have their own objective. Graphics in the games are so neat and seem sophisticated, those can be seen in Fig. 3 below:



Fig. 3. DimensionM™ screenshot 1

It includes single player games named as “Evolver” “Dimenxian” while multiplayer games of “Swarm”, “Meltdown”, and “Obstacle Course”. DimensionM™ can be considered as modern game as it has advanced 3D graphics and interfaces, multiplayer options, high-speed telecommunication technologies and learner-centered approach to facilitate learning [8]. This is a good quality in DimensionM™ because several studies

have shown that game-based learning, as a student-focused learning device, can preferred pull in students over conventional educator focused learning can [30]. As far as game design elements are concerned, games that present materials in a quiz format or drill and practice format do not engage learners (e.g., [31–33]), while well-designed games can engage learners in reflective thinking [34].

Problem-Based Learning (PBL): PBL encourage pupils to face the problem in a way that can enable them to think precisely on the potential solution. Encouraging children to experience balanced-mathematics teaching may let them feel the delight and interest of mathematics [35, 36]. Educator can urge kids to communicate in their own particular words, sharing their reasoning in little gathering or entire class exercises to construct implications for mathematical idea [37]. PBL can include activities such as mathematics challenges, puzzles, and games focusing on mathematizing the play of children [38, 39] and to find solutions of problems that may occurred.

4 Discussion

Any approaches made for beneficial purposes are all counted and from all the methods involved, instructional games are thought to be effective tools for teaching procedures because they use action instead of explanation and create personal motivation and satisfaction [40, 41]. Using the ARCS model of processing, the researchers found that attention, relevance, and confidence, were significant predictors of satisfaction with the game [42]. As what Child-Computer Interaction (CCI) tend to focus on Play, Learning and Communicating in terms of PCL [43], gamification approaches are relevant to have rewards or points for players to gain as that can make them satisfy and motivate them to play, learn and achieve the highest level they could.

For example, to check the effectiveness of MathQuest, a study has been done and twelve students aged 10–11 were involved [20]. There are 5 pupils in Category 1, 5 pupils in Category 2, and 2 pupils in Category 3. The result showed the students find the game energizing, brimming with fun and will appreciate Mathematics class if directed utilizing the diversion as a helping apparatus and the result can be seen in Table 2 below:

Table 2. Evaluation of MathQuest

Questions	Responses according to category		
	Category 1	Category 2	Category 3
1. Do you think MathQuest is fun?	100%	100%	100%
2. Do you think you will enjoy a math class conducted using MathQuest?	100%	80%	50%
3. Do you understand the story?	80%	80%	100%
4. Do you think it is easy to play the game?	80%	100%	100%
5. Would you want to play MathQuest in future?	100%	80%	100%
6. What are the items you like in the game?	Combat	Combat and Graphics	Combat

Source Shafie, A., and Fatimah, W. (2010)

Throughout the study, they found that MathQuest has the potential to be a helping tool in Mathematics class [20]. Those outcomes consistent with articulations that are diversion's outline which has to impersonate firmly particular substance of the educational modules and present them in the type of what the children inclined to love is an incredible equation for progress [44, 45].

In the study of DimensionMTM's validation, 18 weeks were spent and the results showed positive game effects. Teachers also believed that the amusements have affected understudies' Mathematics class inspiration [8]. Although lecturers and classes often provide extrinsic motivation for students to learn, researchers realize that inherent inspiration is the thing that can keep students pursuing learning beyond the classroom [46]. Overall, DimensionMTM had positive effects on students who played the games for 18 weeks as they scored higher on district-wide Mathematics benchmark exam compared to students who did not play the games [8].

Moreover, perceived convenience and helpfulness as key factors in disclosing clients' expectations to utilize particular advancements [47–49] has made it clear that all games which created to give benefits to its' players must be as easy as possible to use. Alternatively, the developer could prepare instructions or demo at the beginning of the gameplay. In addition, a game for children should be very friendly and close to the kids as this can enhance children interests. In fact, CCI can better make children concentrate on interests rather than interaction with people [50].

5 Conclusion and Future Work

The investigations discussed have demonstrated the way that effective intercessions that assemble social, enthusiastic and behavioral abilities at an early age can be positive impact in the way children can solve problems and create interaction with their peers later in life [51]. All approaches stated have their own benefits to the kids regarding numeracy skill. However, for kids in elementary school, it is found not suitable for them to be using mobile phones as tools to learn because they might be distracted. Thus, it can be said desktop based is the perfect medium as it has bigger size of screen to enable teacher or parents to monitor them and ease of conveying messages and computer games can likewise be utilized for creating numerical understanding (see, e.g., [52]) [53] and it is crucial to make computer games beneficial to all including children with learning disability.

Thus, a study to identify children with learning disability numeracy skills is in progress and from this paper, e-learning and gamification approaches have the most suitable elements for kids. This leads to an educational computer game be chosen as the tool for the study to identify children with learning disability cognitive level. Furthermore, the researchers involved would be using LINUS 12 constructs in the game for reference as the study will be held in Malaysia and LINUS is the latest program implemented by Ministry of Education Malaysia (MoE) to guarantee all youngsters in Malaysia have a steady establishment in proficiency and numeracy aptitudes. Hopefully it is useful and beneficial to all especially the education system in Malaysia. The learning effectiveness of educational computer games has been subject to discussion

and debate by some scholars [54–56]. The results of the current study help researchers to reach a better conclusion on the effectiveness of educational computer games [8].

Acknowledgement. This research was supported by a Geran Penyelidikan Pascasiswazah (GPPS) from Research, Innovation, Commercialization, Consultancy Office (ORICC), Universiti Tun Hussein Onn Malaysia (vot number: U816), Short Term Grant U367 and Gates IT Solution Sdn Bhd.

References

1. Sharples, M., Taylor, J., Vavoula, G.: A theory of learning for the mobile age. In: *Medienbildung in Neuen Kulturräumen*, pp. 87–99. Springer, Wiesbaden (2010)
2. Ozdamli, F.: Pedagogical framework of m-learning. *Procedia Soc. Behav. Sci.* **31**, 927–931 (2012)
3. Tulbure, C.: Learning styles, teaching strategies and academic achievement in higher education: a cross-sectional investigation. *Procedia Soc. Behav. Sci.* **33**, 398–402 (2012)
4. NCTM.: principles and standards for school mathematics. National Council of Teachers of Mathematics, Reston (2000)
5. Kashefi, H., Ismail, Z., Yusof, Y.M., Rahman, R.A.: Promoting creative problem solving in engineering. In: *Proceeding of the 3rd International Congress on Engineering Education (ICEED)*. UiTM Publisher, Universiti Teknologi Mara Malaysia, Kuala Lumpur, Malaysia, 7–8 Nov 2011
6. Sahid: Mathematics Problem Solving and Problem-Based Learning for Joyful Learning in Primary Mathematics Instruction. Accessed on 20 July 2017 from <http://staff.uny.ac.id/sites/default/files/131930136/Mathematics%20Problem%20Solving%20and%20PBL.pdf> (2011)
7. Devrim, E.: The scale for problem solving skills in mathematics. *Procedia Soc. Behav. Sci.* **84**, 155–159 (2013)
8. Kebritchi, M., Hirumi, A., Bai, H.: The effects of modern mathematics computer games on mathematics achievement and class motivation. *Comput. Educ.* **55**(2), 427–443 (2010)
9. Mokhtar, M.Z., Tarmizi, R.A., Fauzi, A., Ayub, M.: Enhancing calculus learning engineering students through problem-based learning. *WSEAS Trans. Adv. Eng. Educ.* **7**(8), 255–264 (2010)
10. Pasmaz, A., Ozdemir, A.S.: The Effect of Web-based professional development study to mathematics teachers' problem solving strategies. *Procedia Soc. Behav. Sci.* **46**, 1380–1384 (2012)
11. Engelke, N.: A Framework to Describe the Solution Process for Related Rates Problems in Calculus. Accessed on 8 July 2017 from <http://sigmaa.maa.org/rume/crume2007/papers/engelke.pdf> (2007)
12. Martin, T.: Calculus students' ability to solve geometric related-rates problems rationale solving geometric related-rates problems. *Math. Educ. Res. J.* **12**(2), 74–91 (2000)
13. Duncan, G.J., Dowsett, C.J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., Japeli, C.: School readiness and later achievement. *Dev. Psychol.* **43**, 1428–1446 (2007)
14. Klibanoff, R.S., Levine, S.C., Huttenlocher, J., Vasilyeva, M., Hedges, L.V.: Preschool children's mathematical knowledge: the effect of teacher 'math talk'. *Dev. Psychol.* **42**, 59–69 (2006)
15. Jordan, N.C., Kaplan, D., Ramineni, C., Locuniak, M.N.: Early math matters: kindergarten number competence and later mathematics outcomes. *Dev. Psychol.* **45**, 850–867 (2009)

16. Starkey, P., Klein, A., Wakeley, A.: Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Res. Q.* **19**, 99–120 (2004)
17. Mohd Alwi, N., Fan, I.: Information security in e-learning: a discussion of empirical data on information security and e-learning. In: *Proceedings of the European Conference on e-Learning*, pp. 282–290 (2010)
18. Niemi, H., Harju, V., Vivitsou, M., Viitanen, K., Multisilta, J., Kuokkanen, A.: Digital storytelling for 21st-century Skills in virtual learning environments. *Creative Educ.* **5**, 657–671 (2014)
19. Mullis, I.V.S., Martin, M.O., Foy, P., Arora, A.: *TIMSS 2011 International Results in Mathematics*. TIMSS & PIRLS International Study Center, Chestnut Hill, MA (2012)
20. Shafie, A., Fatimah, W.: Design and heuristic evaluation of mathquest: a role-playing game for numbers. *Procedia Soc. Behav. Sci.* **8**, 620–625 (2010)
21. Oblinger, D.: The next generation of educational engagement. *J. Interact. Media Educ.* **8**, 1–18 (2004)
22. Papastegiou, M.: Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Comput. Educ.* **52**, 1–12 (2009)
23. Ari, A.A., Katranci, Y.: The opinions of primary mathematics student-teachers on problem-based learning method. *Procedia Soc. Behav. Sci.* **116**, 1826–1831 (2014)
24. George, D., Mallery, P.: *SPSS for Windows Step by Step: A Simple Guide and Reference*, 4th edn. Allyn & Bacon, Boston (2003)
25. Abdullah, N.I., Tarmizi, R.A., Abu, R.: The effects of problem based learning on mathematics performance and affective attributes in learning statistics at form four secondary level. *Procedia Soc. Behav. Sci.* **8**, 370–376 (2010)
26. Akinoglu, O., Tandogan, R.O.: The effects of problem-based active learning in science education on student's academic achievement, attitude and concept learning. *Eurasia J. Math. Sci. Technol. Educ.* **3**(1), 71–81 (2007)
27. Mohd Hilmi, M.R., Irfan, N.U.: Students' levels of knowledge construction and cognitive skills in an online forum learning environment. *Procedia Soc. Behav. Sci.* **197**, 1983–1989 (2015)
28. Wiebe, E.N., Lamb, A., Hardy, M., Sharek, D.: Measuring engagement in video game based environments: investigation of the user engagement scale. *Comput. Human Behav.* **32**, 123–132 (2014)
29. Engler, L., Jeschke, S., Ndjeka, E.M., Seiler, R., Steinmiller, U.: MEMBERS The impact of eLTR-Technologies on Mathematical Education of Non-Native Speakers (2005). <http://prints.mulf.tuberlin.de/48/01/Members.pdf>
30. Garris, R., Ahlers, R., Driskell, J.E.: Games, motivation and learning: a research and practice model. *Simul. Gaming* **33**(4), 441–467 (2002)
31. Lester, J.C., Spires, H.A., Nietfeld, J.L., Minogue, J., Mott, B.W., Lobene, E.V.: Designing game-based learning environments for elementary science education: a narrative-centered learning perspective. *Inf. Sci.* **264**, 4–18 (2014)
32. Ruggiero, D., Watson, W.R.: Engagement through praxis in educational game design common threads. *Simul. Gaming* **45**(4–5), 471–490 (2014)
33. Squire, K.: Video games in education. *Int. J. Intell. Games Simul.* **2**(1), 49–62 (2003)
34. Johnson, C.I., Mayer, R.E.: Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Comput. Hum. Behav.* **26**(6), 1246–1252 (2010)
35. Clements, D.H., Sarama, J.: *Learning and teaching early math: The learning trajectories approach*, 2nd edn. Routledge, New York, NY (2014)

36. Stipek, D.: Mathematics in early childhood education: Revolution or evolution? *Early Educ. Dev.* **24**(4), 431–435 (2013)
37. Association of Mathematics Teacher Educators.: AMTE standards for mathematics teacher preparation. Raleigh. AMTE, NC (2017)
38. Agodini, R., Harris, B., Seftor, N., Remillard, J., Thomas, M.: After Two Years, Three Elementary Math Curricula Outperform a Fourth. National Center for Education Evaluation and Regional Assistance, Washington, DC (2013)
39. Clements, D. H., Sarama, J.: Building Blocks, vol 1 and 2. McGraw-Hill Education, Columbus, OH (2013)
40. Charles, D., McAlister, M.: In: Rauterberg, M. (ed), Integrating Ideas About Invisible Playgrounds from Play Theory into Online Educational Digital Games, pp. 598–601. ICEC 2004, LNCS 3166, Accessed 24 Jul 2017 from [http://www.springerlink.com.ucfproxy.fcla.edu/\(coci1u55qul21e55wlk1aomj\)/app/home/contribution.asp?referrer=parent&backto=searcharticlesresult,4,4](http://www.springerlink.com.ucfproxy.fcla.edu/(coci1u55qul21e55wlk1aomj)/app/home/contribution.asp?referrer=parent&backto=searcharticlesresult,4,4)
41. Holland, W., Jenkins, H., Squire, K.: In Perron, B., Wolf, M. (eds) Video Game Theory. Routledge. Accessed 21 June 2017 from <http://www.educationarcade.org/gtt/>
42. Huang, W.H., Huang, W.Y., Tschopp, J.: Sustaining iterative game playing processes in DGBL: The relationship between motivational processing and outcome processing. *Comput. Educ.* **55**(2), 789–797 (2010)
43. Read, J.C., Bekker, M.M.: The nature of child computer interaction. In: Proceedings of the 25th BCS Conference on Human-Computer Interaction (BCS-HCI'11), pp. 163–170. Swinton, UK (2011)
44. Chong, S.H.: Learning mathematics through computer games. In: Proceeding of 14th Asian Technology Conference in Mathematics. Beijing, China. (2009)
45. Singh, J., Wei, L.L., Shanmugam, M., Gunasekaran, S.S., Dorairaj, S.K.: Designing computer games to introduce programming to children. In: Proceedings of the 4th International Conference on Information Technology and Multimedia (ICIMU 2008). Malaysia. (2008)
46. Schmal, V., Grabinski, C.J., Bowman, S.: Use of games as a learner-centered strategy in gerontology, geriatrics, and aging-related courses. *Gerontol. Geriatr. Educ.* **29**(3), 225–233 (2008)
47. Bourgonjon, J., Valcke, M., Soetaert, R., Schellens, T.: Students' perceptions about the use of video games in the classroom. *Comput. Educ.* **54**(4), 1145–1156 (2010)
48. Bourgonjon, J., De Grove, F., De Smet, C., Van Looy, J., Soetaert, R., Valcke, M.: Acceptance of game-based learning by secondary school teachers. *Comput. Educ.* **67**, 21–35 (2013)
49. Park, E., Baek, S., Ohm, J., Chang, H.J.: Determinants of player acceptance of mobile social network games: an application of extended technology acceptance model. *Telematics Inform.* **31**(1), 3–15 (2014)
50. Black, M., Chang, J., Chang, J., Narayanan, S.: Comparison of child-human and child computer interactions based on manual annotations. In: Proceedings of the Workshop on Child, Computer, and Interaction. Cambridge, USA (2009)
51. Hazar, M.: Development of learning and social skills in children with learning disabilities: an educational intervention program. *Procedia Soc. Behav. Sci.* **209**, 221–228 (2015)

52. Van Borkulo, S., Van den Heuvel-Panhuizen, M., Bakker, M., Loomans, H.: One mini-game is not like the other: different opportunities to learn multiplication tables. In: De Wannemacker, S., Vandercruysse, S., Clarebout, G. (vol. eds.) *Communications in Computer and Information Science*, vol. 280. *Serious games: The challenge*, pp. 61–64. Berlin. (2012)
53. Jonker, V., Wijers, M., Van Galen, F.: The motivational power of mini-games for the learning of mathematics. Paper presented at the European conference on game-based learning. Graz, Austria (2009)
54. Hays, R.T.: The effectiveness of instructional games: a literature review and discussion. In: Naval Air Warfare Center Training System Division (No. 2005-004). Accessed 20 June 2017 from <http://www.dtic.mil/cgi-bin/GetTRDoc?AD%4ADA441935&Location%4U2&doc%4GetTRDoc.pdf> (2005)
55. Mitchell, A., Savill-Smith, C.: The use of computer games for learning. Accessed 20 June 17 from <http://www.m-learning.org/archive/docs/The%20use%20of%20computer%20and%20video%20games%20for%20learning.pdf> (2004)
56. Randel, J.M., Morris, B.A., Wetzel, C.D., Whitehill, B.V.: The effectiveness of games for educational purposes: a review of recent research. *Simul. Gaming* **23**(3), 261–276 (1992)

Part IV

Web Mining and Content Analytics (WMCA)

Stock Market Prediction Using Keywords from Expert Articles

Ko Ichinose and Kazutaka Shimada^(✉)

Department of Artificial Intelligence, Kyushu Institute of Technology, 680-4 Kawazu
Iizuka, Fukuoka 820-8502, Japan
shimada@pluto.ai.kyutech.ac.jp

Abstract. The market analysis is one of the important tasks for text mining. In this situation, Web news has an important role to predict stock prices. In this paper, we propose a method to predict the Nikkei Stock Average, which is one of the most important stock market indexes. We extract viewpoints from experts' articles for analyzing Web news. The extracted words are index words in the vector space of a machine learning technique. We also incorporate word embedding and bootstrap approaches into our method. It predicts "UP" or "DOWN" of the next day by using the articles of a day. We also evaluate our method with not only one-day prediction but also simulated trading. The experimental result shows that index words based on expert articles were effective for both one-day prediction and simulated trading.

Keywords: Stock market prediction · Experts' articles · Bootstrap · Feature extraction

1 Introduction

Online stock trading on the Web is increasing dramatically. Active trading contributes significantly to economic growth and leads to promotion of economic efficiency. However, it is said that more than 90% of newcomers as personal investors withdraw within one year because of the difficulty of the trading. Therefore it is important for the personal investors to support the trading. Natural language processing is one useful approach for the task. Many researchers have proposed methods using text information for analyzing the market [1–3]. They generated models to predict the stock market, such as "UP" or "DOWN" of stock prices, by using texts and machine learning techniques.

In this paper, we describe some feature sets to predict one-day stock price performance from Web news. The target is the Nikkei Stock Average, which is one of the most important stock market indexes. One simple approach for the prediction task is to generate a model with Bag-of-words (BOW) features from news articles. However, all news articles are not always suitable for generating a

prediction model. In addition, not all words in articles contribute to generating a better model. Therefore, we focus on the point of view from experts of the stock market. We compare features using experts' articles with a normal BOW feature set.

The accuracy of the one-day classifier is one of the important criteria. It is, however, unclear whether the improvement of a classifier, such as "UP" or "DOWN" of a stock price of each day, contributes to the real trading. The most important thing is that the prediction model contributes to making a profit. Therefore we also evaluate our method through simulated trading on the Nikkei Stock Average.

2 Related Work

Lee et al. [2] have proposed a system that forecasts companies' stock price changes in response to financial events. They reported that predicting next day's price movement was improved by 10% if text is considered. The study handled financial documents. On the other hand, some researchers focused on Social Medias, such as Message boards and Twitter. Nguyen and Shirai [3] have proposed a stock market prediction model with a topic-sentiment feature and a new topic model. By using the feature and topic model, the accuracy of the proposed method increased by 6% as compared with some baselines. Bollen et al. [1] have proposed a method with 6 mood dimensions on Twitter for the stock market prediction. They obtained the high accuracy rate; 86.7%. These studies only focused on the accuracy rates of the classification models. However, there was no mention whether the improvement of the accuracies contributed to making a profit. One of the most important points for personal investors is whether they eventually gain a profit by using the prediction model.

Izumi et al. [4] have proposed a text-mining method for long-term market analysis using market reports published by financial professionals and institutions on the web. For Japanese government bond 2-year, 5-year, and 10-year markets, the proposal method forecast in higher accuracy about both the level and direction of long-term market trends. The task was forecasting long-term market trends of Japanese government bond. On the other hand, the target and resources of our task are the Nikkei Stock Average and news articles on the Web. Schumaker and Chen [5] have proposed a SVR-based method that made a discrete prediction of what +20 min stock should be. The method, AZFinText, used financial news articles and stock price quotes. They compared AZFinText against the top 10 quants for one year. As a result, AZFinText outperformed six of the top 10 quant funds. They considered the effectiveness of the method in terms of real profit-and-loss.

One simple approach for the prediction task is to generate a model with Bag-of-Words features (BoW). However, the simple features, BoW, can not capture structured information of texts. Ding et al. [6] have proposed a prediction model that handled structured information. Xie et al. [7] have applied semantic frames to their prediction model. We also incorporate a relation between words, namely dependency, as one approach of feature extraction.

Peng and Jiang [8] have proposed a method using a word embedding approach [9]. They selected bag-of-keywords by using nine seed words and a bootstrap method. We also incorporate the idea, namely word embedding and bootstrap, into our method. Here we focus on experts' knowledge for the bootstrap approach. Bar-Haim et al. [10] have described that it is beneficial to distinguish expert users from non-experts for stock prediction. We use experts' article to extract features, namely index words, for the vector space of a machine learning method.

3 Data Set

3.1 Web News

We use articles in Yahoo Japan finance news¹ for generating prediction models. We collected articles that were posted from 3 p.m to 11:40 p.m. In other words, we collected articles after closing the stock market in Japan. We predict the next day's state, "UP" or "DOWN", by using the articles. The data set consists of articles for 212 days (from January 5th, 2015 to December 13th, 2015 and weekdays only). Each day has some articles. The total number of articles about Web news is 44,164. For generating a prediction model, we handle articles in one day as one document for each brand. More properly, we extract only sentences containing each brand name from each articles. Here the brands denote company names related to the Nikkei Stock Average, namely 225 brands.

3.2 Experts' Articles

We also collected experts' articles for extracting the point of view of stock market experts. We use articles in Yahoo finance stock prediction.² On the web sites, many experts analyze several targets, such as Japanese stocks, Nikkei Stock Average and Dow Jones Industrial Average, from their own point of view, and then distribute the information irregularly. We collected 1,000 articles related to the Nikkei Stock Average (from June 30th, 2015 to May 24th, 2016).

4 Proposed Method

We explain some index words, namely features of a machine learning method, for the prediction task. First, we explain four feature sets, and then classifiers with the feature sets.

¹ <https://news.finance.yahoo.co.jp/>.

² <https://info.finance.yahoo.jp/kabuyoso/>.

4.1 Feature Selection

For the prediction, we need to extract features from articles. In this section, we explain one baseline feature set and three proposed features. The purpose of this process is to extract index words for the feature vector space. Each vector in our method is binary.

- **Bag-of-Words from Web News (BoW)**

This is a baseline. The BoW consists of nouns, verb and adjectives extracted from Web news articles themselves. We use a Japanese morphological analyzer, MeCab [11], for extraction of words.

- **Bag-of-Words from Expert Articles (BoW-E)**

The BoW-E consists of nouns, verb and adjectives extracted from expert articles, as index words for the vector space. Then, we construct the vector space from target Web news articles on the basis of the index words from expert articles.

- **Bag-of-Keywords from Expert Articles (BoK-E)**

The BoK-E is based on [8]. They applied word embedding and bootstrap approaches for the features. The method collected important words for the prediction from some seed words by a bootstrap method from target data, namely Web news articles in this paper. On the other hand, we apply this method into experts' articles for extracting index words of the vector space because expert articles tend to contain more important information and points of view for prediction as compared with normal Web news articles. We can obtain more appropriate index words by applying a word embedding approach to expert articles. The process is as follows:

1. generate a word embedding model from experts' articles by using Word2Vec [9],
2. select some seed words,
3. compute the cosine similarity measure between seed words and words in Web news articles on Word2Vec,
4. add the top 10 words with high scores,
5. iterate the step 3 and 4 until the number of words in the list exceeds the threshold.

We use the following 10 words as the seeds; risk, economic indicator, sharp drop, signal, sell, buy, brisk, up trend, down trend and sense of anticipation. The threshold is 1200 words.³ These seed words and the threshold were determined heuristically.

- **Bag-of-Tuple from Expert Articles (BoT-E)**

Word-based features are often insufficient because of a lack of a relation between words. Therefore, we apply dependency relations into the vector space. The BoT-E consists of dependency word pairs, namely tuples, from expert articles, as index words for the vector space. We use Cabocha [12] as a dependency parser. We extract verbs in each sentence first. Then, we extract

³ Since the number of BoWs is approximately 10,000, BoK-E is approximately 10% of the BoW.

nouns that contain dependency relations with the verbs. For example, we obtain two tuples, [Tokyo market, improve] and [drop-off, improve], from the sentence “The Tokyo market was improved from the drop-off”. Finally, we construct the vector space from target Web news articles on the basis of the index tuples from expert articles. For example, if the tuple [Tokyo market, improve] appears in a Web news article, the value of the vector [Tokyo market, improve] becomes 1. Otherwise it is 0.

4.2 Classifier

We generate a one-day classifier based on the index words in Sect. 4.1. The purpose of the one-day classifier is to predict “UP” or “DOWN” of the next day by using the articles of a day. The target value of the classifier is +1 (rise in stock prices of the next day) or −1 (drop in stock prices of the next day). We use Support Vector Machines [13] implemented on the machine learning tool Weka⁴ for the one-day classifier.

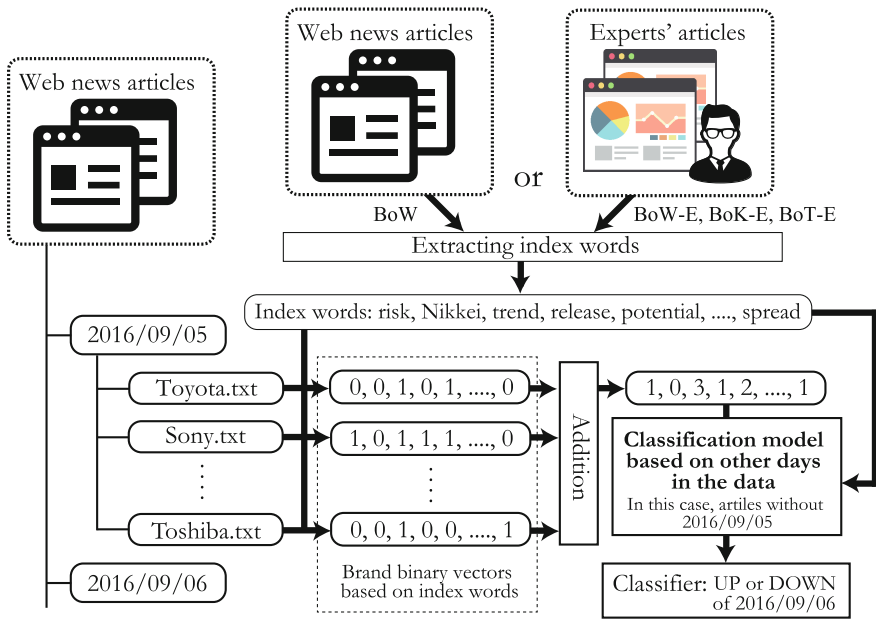


Fig. 1. The vectorization from Web news articles

We generate a one-day classifier from Web news without a target day on the basis of index words. Figure 1 shows an example of the vectorization from Web news and experts’ articles. For example, the one-day classifier to evaluate

⁴ <http://www.cs.waikato.ac.nz/ml/weka/>.

the stock market on January 5th is generated from Web news⁵ from January 6th to December 13th. The one-day classifier for January 6th is also generated from Web news of January 5th and from January 7th to December 13th, namely without January 6th. As mentioned in Sect. 3.1, data in a day are separated into each brand. First, we construct each vector space for each brand, and then add them into one vector space on the basis of a simple addition function.

5 Experiment

5.1 Accuracy of One-Day Classifier

In this experiment, we compared four feature sets explained in Sect. 4.1. We obtained 1202 index words for the BoK-E. We evaluated each method on the leave-one-out cross-validation for 212 days in the data. For example, our method generated a one-day classifier from articles without a target day, and then predicted “UP” or “DOWN” by using articles of the target day.⁶

Table 1. The one-day results

Features	Class	P	R	F	Acc
Bow (baseline)	DOWN	36.0	20.7	26.3	-
	UP	57.4	74.4	64.8	-
	Ave	48.6	52.4	49.0	52.3
BoW-E	DOWN	48.4	35.6	41.1	-
	UP	62.2	73.6	67.4	-
	Ave	56.5	58.0	56.6	58.0
BoK-E	DOWN	53.5	52.9	53.2	-
	UP	67.5	68.0	67.7	-
	Ave	61.7	61.8	61.8	61.8
BoT-E	DOWN	49.3	41.4	45.0	-
	UP	63.3	70.4	66.7	-
	Ave	57.6	58.5	57.8	58.5

The experimental result is shown in Table 1. In the table, P and R are precision and recall rates and F is the harmonic mean. Ave is the weighted average of class “DOWN” and “UP” for each criteria. Acc is the accuracy rate. Three feature sets from experts’ articles outperformed the simple BoW from Web news articles. This result shows the effectiveness of experts’ articles for extracting

⁵ Note that we never use experts’ articles for the vectorization. We just use expert’s articles for extracting index words for the vector space.

⁶ More precisely, the model predicts the next day’s “UP” or “DOWN” from the target day.

index words. The BoK-E produced the best performance in terms of all criteria. This result shows the effectiveness of the word embedding and bootstrap methods using experts' articles.

For the BoK-E, seed words and the threshold in the iteration have significant roles for generating a classifier with high accuracy. We compared two different types of seed words; *tfidf*-based seeds and randomly-selected seeds. For the *tfidf*-based seeds, we applied the *tfidf* calculation to words that appear in both Web news and experts' articles. Then we selected 10 words from the top 30 words manually. For the randomly-selected seeds, we selected 10 words from words that appear in both Web news and experts' articles randomly. As a result, the accuracy rates of the methods with two seed words decreased by 10%. For the threshold in the iteration, we also compared different settings based on the number of iteration counts from 3 to 9. As a result, the BoK-E in Table 1 also outperformed the settings. These results show the effectiveness of the current BoK-E setting and the importance of the seed word selection.

Table 1 shows the results based on each feature set, namely single feature sets. Combination of feature sets often leads to the improvement of the accuracy of a classifier. Therefore we also evaluated some combinations of features, e.g., BoK-E + BoT-E. However, there is no improvement as compared with a single model of BoK-E. We need to consider incorporating other features for the improvement of the accuracy.

5.2 Simulated Trading by One-Day Classifier

We evaluated our method (BoK-E) with simulated trading using real stock prices. For the evaluation, we need to set a virtual market and a virtual investor with a strategy.

- Virtual market: We use the past data on the Nikkei Stock Average and the Nikkei Leveraged Index. Each datum of a day consists of the opening price, closing price, highest price, and lowest price.
- Virtual investor: This is the classifier with BoK-E that is the best one-day classifier in Sect. 5.1. We set the initial money of the investor to 2 million yen (1 million yen for Spot buying and 1 million yen for Short selling). The investor buys stocks if the classifier answers that the price of the next day will rise and the investor sells stocks if the classifier answers that the price of the next day will drop.

There are many strategies for the buying and selling. We apply a simple strategy in this paper. The strategy is as follows:

- *Buying*: This is the process in the case that the output of the one-day classifier is "UP". The investor buys stocks by using all his/her money in the opening. Then, the investor sells all stocks in the closing; provided, however, that the investor carries out loss-cutting⁷ by the lowest price if the difference between the lowest price and the opening price is more than 0.8%.

⁷ Loss-cutting is a financial word. It is an order to buy or sell stocks automatically in the case that unrealized capital losses become larger.

- *Selling*: This is the process in the case that the output of the one-day classifier is “DOWN”. The investor sells stocks in the opening (short selling⁸). Then, the investor buys all stocks that sold in the opening in the closing⁹; provided, however, that the investor carries out loss-cutting by the highest price if the difference between the highest price and the opening price is more than 0.8%.

Table 2. The results of simulated trading by BoK-E

Target	Final money	Earning rate (%)
Nikkei SA	2,329,084	+16.5
Nikkei LI	2,537,400	+26.9

The result of the simulated trading is shown in Table 2. The earning rate is computed as follows:

$$\text{Earning rate} = \frac{\text{The final money at the end of the simulated trading}}{\text{The initial money}} - 1 \quad (1)$$

Our method with BoK-E generated high earning rates in the simulated trading situation.¹⁰ This result indicates not only the effectiveness of our method in terms of one-day prediction but also the potential efficacy in the real trading.

6 Conclusions

In this paper, we compared four types of index words to predict one-day stock price performance from Web news. The index words were extracted from experts’ articles. We also applied word embedding and bootstrap approaches into our method, namely BoK-E. The method based on BoK-E outperformed a simple BoW from Web news articles themselves, a simple BoW based on index words extracted from experts’ articles (BoW-E) and the vector space based on dependency relations (BoT-E). In a simulated trading task, the earning rate of the method was +26.9%.

We obtained a good result from our method with BoK-E through the one-day classification and the simulated trading. However, there are some concerns about the experiment. We evaluated the one-day classification task by the leave-one-out cross-validation for 212 days. In other words, our method used information about future articles. Although each classifier of each day is independent, the experimental setting might influence the result of the accuracy. Therefore, we

⁸ Short selling is a financial word and the practice of selling stocks that are not currently owned by borrowing them from a securities company.

⁹ More properly, the borrowed stocks are paid back to the securities company. The balance is the benefit of the day.

¹⁰ Although we evaluated other methods, such as the method with BoT-E, the BoK-E also produced the best performance in terms of this simulated trading.

need to verify how much effect this situation has on the accuracy. For example, we need to generate a classifier from articles of a year, and then evaluate the classifier with another year. As is the case in the classification, expert articles for index word extraction also contained future articles in the one-day classification task. Although expert articles were just used for the index word extraction, they also might influence the result of the accuracy. The verification is also important future work. We need a larger and consistent data set to dispel these concerns.

References

1. Bollen, J., Mao, H., Zeng, X.-J.: Twitter mood predicts the stock market. *J. Comput. Sci.* **2**(1), 1–8 (2011)
2. Lee, H., Surdeanu, M., MacCartney, B., Jurafsky, D.: On the importance of text analysis for stock price prediction. In: *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC)*, pp. 1170–1175 (2014)
3. Nguyen, T.H., Shirai, K.: Topic modeling based sentiment analysis on social media for stock market prediction. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pp. 1354–1364 (2015)
4. Izumi, K., Goto, T., Matsui, T.: Trading tests of long-term market forecast by text mining. In: *Proceedings of 2010 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 935–942 (2010)
5. Schumaker, R.P., Chen, H.: A discrete stock price prediction engine based on financial news. *Computer* **43**(1), 51–56 (2010)
6. Ding, X., Zhang, Y., Liu, T., Duan, J.: Using structured events to predict stock price movement: an empirical investigation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1415–1425 (2014)
7. Xie, B., Passonneau, R.J., Wu, L., Creamer, G.G.: Semantic frames to predict stock price movement. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 873–883 (2013)
8. Peng, Y., Jiang, H.: Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In: *Proceedings of NAACL-HLT 2016*, pp. 374–379 (2016)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Proceedings of NIPS* (2013)
10. Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., Goldstein, G.: Identifying and following expert investors in stock microblogs. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1310–1319 (2011)
11. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237 (2004)
12. Kudo, T., Matsumoto, Y.: Japanese dependency analysis using cascaded chunking. In: *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69 (2002)
13. Vapnik, V.N.: *Statistical Learning Theory*. Wiley, New York (1999)

Sarcasm Detection Using Features Based on Indicator and Roles

Satoshi Hiai^(✉) and Kazutaka Shimada

Department of Artificial Intelligence, Kyushu Institute of Technology, 680-4 Iizuka,
Fukuoka 820-8502, Japan
{s_hiai, shimada}@pluto.ai.kyutech.ac.jp

Abstract. Sarcasm is a non-literalistic expression and presents a negative meaning with positive expressions. Sarcasm detection is a significant challenge for sentiment analysis which is to analyze documents with opinions. In this study, we propose a method of sarcasm detection on Twitter. We focus on two kinds of feature words. One is words modified by the indicator “(皮肉)”. The other is words expressing a role. First, we extract these words from tweets. Next, our method uses the lists of these words for a machine learning approach to detect sarcastic tweets. The lists of extracted words are used as features in our method. In the experiment, we compare our method with a baseline based on the features in previous studies. The experimental result shows the effectiveness of our method.

Keywords: Sarcasm · Sentiment analysis · Opinion mining
Microblogging · Classification

1 Introduction

Sarcasm is defined as “the activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry.”¹ For example,

Example 1 “政治家は口先だけで仕事ができ素晴らしいね。”
(Politicians are great because they just run their big mouth.)

In this sentence, the writer uses the positive expression “素晴らしい (great)”. However, the intention of the writer is to criticize “政治家 (politician)”.

Sarcasm detection is a significant challenge for sentiment analysis which is to analyze documents with opinions [1]. Many researchers have studied sarcasm detection as a task that classifies a text into sarcastic or non-sarcastic [2–5]. They used Twitter data as target texts. Twitter is a microblogging service where users

¹ <http://www.macmillandictionary.com>.

post short messages (tweets). Some of the tweets contain user-defined hashtags to categorize topics. Sarcastic tweets in English can be obtained by collecting tweets tagged with “# sarcasm”. In this paper, we handle Japanese tweets as target texts. The Japanese expression of “sarcasm” is “皮肉”. Therefore, we attempted to obtain sarcastic tweets by collecting tweets tagged with “皮肉”. However, there were not enough tweets with the “#皮肉” tag. On the other hand, some sarcastic tweets in Japanese contain the indicator “(皮肉)”. Sarcastic tweets with the indicator “(皮肉)” are as below:

Example 2 先生方、大量の宿題を出してくれてありがとうございます。(皮肉)
(Thank you for teachers who give me tons of homework. (sarcasm))

Tweets with the indicator “(皮肉)” are posted more frequently than tweets with “#皮肉”. Therefore, we collect sarcastic tweets with the indicator “(皮肉)”. We create training data and test data by using the tweets.

In past studies, researchers used machine learning approaches to classify each tweet. Bag-of-Words is a commonly used feature for machine learning. In addition, feature words of sarcastic tweets such as positive and negative words in target texts (e.g. “great” in *Example 1*) are useful.

In this paper, we focus on two kinds of feature words of sarcastic tweets. One is words modified by the indicator “(皮肉)”. We also use the indicator to identify feature words as well as to create a dataset. This indicator often appears in the middle of sentences. For example,

Example 3 見事 (皮肉) な部屋の散らかりようですね！
(How beautiful(sarcasm) your messy room is !)

In this sentence, “(皮肉)” modified “見事 (beautiful)”. We focus on the feature words such as “見事 (beautiful)”. Past studies relied on positive and negative words to classify tweets. On the other hand, by using the indicator “(皮肉)”, our method detects a sarcastic sentence without any sentiment words because the indicator implies sarcastic likelihood of the adjoining word. For example,

Example 4 居眠りが彼の仕事 (皮肉) みたいです。(皮肉)
(It seems that his job(sarcasm) is to fall asleep on his desk.)

In this example, “(皮肉)” modified “仕事 (job)”. Although this example contains no sentiment words, our method with features based on words modified by the indicator “(皮肉)” can correctly treat this example as sarcasm. Thus the indicator “(皮肉)” helps to recognize sarcasm. The other is words expressing a role. The target of sarcasm is a crucial component of sarcasm [6]. Words expressing a role indicate targets of sarcasm. In *Example 1*, “politician” indicates a target of sarcasm. Words such as “politician” are role expressions. A parallel relation between a writer and a target of sarcasm often appear in sarcastic tweets. In

Example 1, the writer criticizes “politician” as the target of sarcasm from the point of view of “commoner”. Therefore, we focus on role expressions with a parallel relation, such as “politician and commoner”, as a target of sarcasm.

The contributions of this paper are:

- We introduce new features: words modified by the indicator “(皮肉)” and words expressing a role.
- We verify the effectiveness of new features in our classification task.

First, we extract and select two kinds of feature words of sarcastic tweets (Sect. 3). Second, we propose features based on the feature words (Sect. 4). Finally, we verify the effectiveness of the features (Sect. 5).

2 Related Work

Many researchers used machine learning approaches to classify a text. Bag-of-Words and the length of a text are used as basic features [3, 7]. In addition to these features, linguistic features of sarcasm, such as positive words, negative words and a target of sarcasm, were also used. For example,

Example 5 無視するなんて、君はいい性格してる。
(Oh, I love being ignored.)

This sarcastic sentence contains the positive word “love” and the negative word “ignored”. Joshi et al. [3] have focused on positive words and negative words in sarcastic sentences. They used features such as the number of positive and negative words and the number of times that a word is followed by a word with opposite polarity. We also use these features for the proposed method. Karoui et al. [4] have used the presence of named entities and personal pronouns as features. These features are related to a target of sarcasm. However, named entities are too rare to appear in a training set. Moreover, personal pronouns are too abstract. In this paper, we focus on common nouns. Words expressing a role such as “politician” are often used as a target of sarcasm. The words are common nouns. Common nouns appear more frequently than named entities, and appear less frequently than personal pronouns. Therefore, common nouns are effective features for classification to express a target of sarcasm.

3 Extraction and Selection of Feature Words

In this section, we explain extraction and selection of feature words. The outline is shown in Fig. 1. We focus on (1) words modified by the indicator “(皮肉)” and (2) words expressing a role. First, we extract these words from different datasets and obtain two kinds of lists. Next, we select feature words that appear more frequently in sarcastic tweets than those in non-sarcastic tweets from these lists. We explain each step in detail.

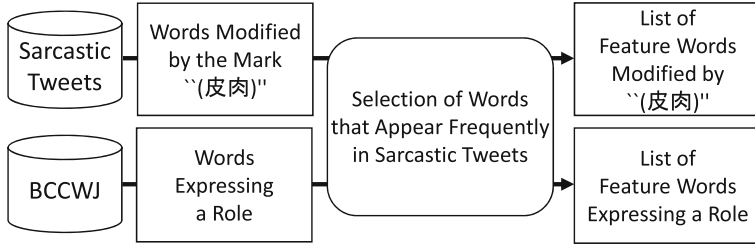


Fig. 1. Extraction of feature words

3.1 Extraction of Words Modified by the Indicator

“(皮肉)”

The details of this extraction process are shown in Fig. 2. We extract words modified by the indicator “(皮肉)” from posts on Twitter. We collected 3,000 tweets with “(皮肉)” as the sarcastic tweets for extraction. In *Example 3*, the word “見事 (beautiful)” is modified by “(皮肉)”. Thus the indicator “(皮肉)” highlights words as sarcasm. In this section, We extract highlighted words from the sarcastic tweets. However, “(皮肉)” at the end of a tweet often modifies the entire tweet, e.g. *Example 2*. In this situation, it is not clear which words are modified by the indicator “(皮肉)”. It leads to extraction of noise feature words for a sarcasm classifier. Therefore, we extract words only from the tweets not ending with “(皮肉)”, e.g. *Example 3*. The number of tweets except for the tweets ending with “(皮肉)” in the 3,000 tweets for extraction was 924. We obtained the list of 366 words modified by the indicator “(皮肉)” from the tweets.

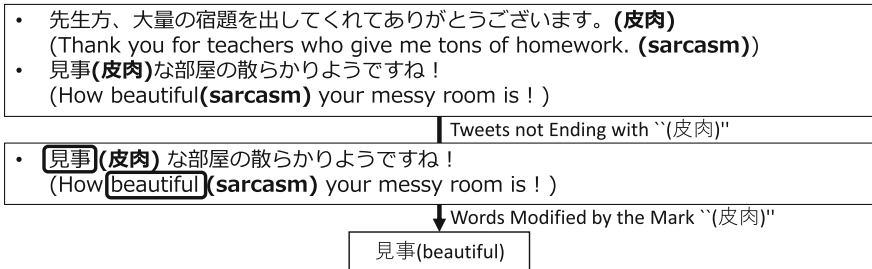


Fig. 2. Extraction of words modified by the indicator

3.2 Extraction of Words Expressing a Role

We explain extraction of words expressing a role, such as “politician” in *Example 1*. The details of this extraction process are shown in Fig. 3. Writers use these

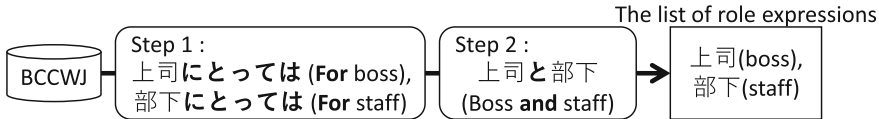


Fig. 3. Extraction of words expressing a role

expressions to criticize a target from a point of view of a parallel role to the target in sarcastic tweets.

We extract words expressing a role with a two-step process using two phrases. The phrases are “にとっては (for)” and “と (and)”. The extraction process is as follows:

- Step 1. A word that adjoins “にとっては (for)” tends to express a role. We extract nouns that adjoin “にとっては (for)”. We regard each extracted word as a role expression.
- Step 2. A parallel relation is expressed before and after the Japanese particle “と (and)”, e.g. “政治家 と 一般人 (politician and commoner)”. When the phrase “A と B” appears, we regard that A and B are a parallel role. For the phrase “A と B”, if a word extracted in Step 1 appears as A or B, we regard the word as a role with a parallel role. We discard words that do not appear even once before or after “と”.

As with Sect. 3.1, we attempted to extract role expressions from Twitter data. However, the number of appearance of the phrase “にとっては” was not enough because Twitter users use spoken language in tweets and the phrase “にとっては(for)” is likely to be written language. In this section, our aim is extraction of role expressions. It is not equal to extract feature words for the classification. We need to extract role expression candidates from a rich data set for a sarcasm classifier. Therefore, we extract role expressions from BCCWJ.² BCCWJ is a corpus created for the purpose of attempting to grasp the breadth of contemporary written Japanese. The corpus contains written Japanese texts comprised of about 100 million words in various domains. It seems that the phrase “にとっては(for)” appears more frequently in the data. Therefore, we use BCCWJ to obtain the list of words expressing a role. We obtained the list of 1,784 words expressing a role from the data.

3.3 Selection of Feature Words

In Sects. 3.1 and 3.2, we extracted the words that modified by the indicator “(皮肉)” and role expressions. However, these words are not always sarcastic because they are often used in non-sarcastic tweets. Therefore, to select feature words that appear more frequently in sarcastic tweets, we compare the frequency of each word in sarcastic tweets and non-sarcastic tweets. We collected additional

² <http://pj.ninjal.ac.jp/corpus.center/bccwj>.

3,000 tweets without “(皮肉)” as non-sarcastic tweets. For each extracted word, we compare the number of times that the word appears in sarcastic tweets and the number of times that the word appears in non-sarcastic tweets. We discard words that appear more frequently in non-sarcastic tweets than in sarcastic tweets from the list. We also discard that words appear only one time in sarcastic tweets since they are too sparse. As a result, we obtained the list of the 257 words modified by the indicator “(皮肉)” and the list of the 472 words expressing a role.

4 Proposed Method

The outline of our method is shown in Fig. 4. The proposed method uses two classifiers. In this section, first, we explain the features to train the classifiers in detail. Next, we explain the framework of the proposed method, which combines two classifiers.

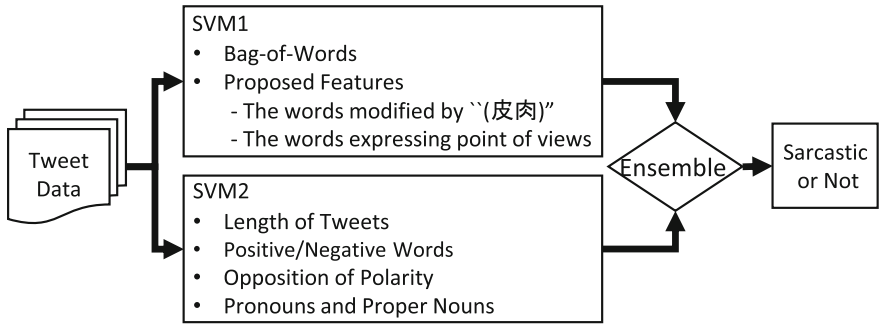


Fig. 4. The outline of our method

4.1 Features

We use features in previous studies and proposed features. We explain the features in detail.

Features in previous studies We explain features in previous studies.

- Bag-of-Words
This feature is the presence of unigrams in the training data. This is a commonly used feature in text classification tasks.
- The number of words in tweets
This is the length of tweets.
- Positive/negative words, opposition of polarity
Many previous studies used positive/negative words [2–5, 8]. Joshi et al. [3] used these features: the number of positive words, the number of negative

words and the number of times of a word followed by a word of opposite polarity in tweets, namely opposition of polarity. In our method, a Japanese Sentiment Polarity Dictionary (Volume of Verbs [9] and Adjectives [10]) and a Polar Phrase Dictionary [11] are used as the lists of positive and negative words.

- Pronouns and proper Nouns
Karoui et al. [4] used the features of pronouns and named entities. In our method, we use the number of pronouns and proper nouns as features since we focus on a target of sarcasm.

Proposed features We explain proposed features.

- Words modified by the indicator “(皮肉)”
We use the list of the 257 words modified by the indicator “(皮肉)” that we obtained in Sect. 3.1. This feature is the presence of each word in the list.
- Words expressing a role
We use the list of the 472 words expressing a role that we obtained in Sect. 3.2. This feature is the presence of each word in the list.

4.2 The Framework of Proposed Method

In the proposed method, classifiers are trained with the features in previous studies and proposed features. We use SVM classifiers. SVM classifiers are used in many previous studies [3–5, 8]. Tunghamthiti et al. [8] used two SVM classifiers with two different feature sets. Then they combined the outputs of classifiers. More reliable output between two outputs was chosen as the final result. They used two feature sets separately: one is the features in previous studies and the other is their proposed features with various ranges of the feature value. An SVM is sensitive to features with a large range value. For example, when an SVM classifier is trained by binary features such as Bag-of-Words and features with a large range value such as the number of words in documents, features with a large range value tend to greatly affect a classification. Although we use features with various range values, we want to equally handle the binary features and large range features. Therefore, we classify features on the basis of the range of feature values. To reduce effects of difference between ranges, we classify features into two classes: one is binary features such as Bag-of-Words and the other is numeric features such as the number of words in documents. The framework is shown in Fig. 4. Therefore, features for SVM1 in Fig. 4 are as below:

- Bag-of-Words
- Words modified by the indicator “(皮肉)”
- Words expressing a role

Features for SVM2 in Fig. 4 are as below:

- The number of words in tweets
- Positive/negative words, opposition of polarity
- Pronouns and proper nouns

We used linear-kernel SVM classifiers implemented in scikit-learn library [12]. In the classification, if the outputs of two SVMs are the same, the outputs become the final classification result. If they are not the same, we compare the distance from a separating hyperplane of both classifiers. We choose the output with a larger distance as the final result.

5 Experiment

In this section, we describe the experiment to evaluate our method.

5.1 Data

We created training data and test data by using the tweets with “(皮肉)”. We collected the tweets from 4/25/2016 to 12/16/2016.

- Training data
For training data, we collected 3,120 tweets with “(皮肉)” as sarcastic tweets, and 3,120 tweets without “(皮肉)” as non-sarcastic tweets. Then, we removed the indicator “(皮肉)” in the tweet texts because the presence of the indicator was unfair for the generation of a classifier.
- Test data
For test data, we annotated tweets manually. We collected 3,000 tweets with “(皮肉)”. Then we also removed the indicator “(皮肉)” in the tweet texts for the same reason as the training data construction. We divided the 3,000 tweets into 3 parts of 1,000 tweets. Three annotators annotated the different dataset of 1,000 tweets. The annotators judged 109 tweets of 3,000 tweets as sarcastic. We used the 109 tweets as sarcastic tweets. We also collected 109 tweets without “(皮肉)” as non-sarcastic tweets.

5.2 Evaluation

We prepared two baseline methods, and compared our method with them. The first baseline method (Baseline 1) was based on a single SVM classifier trained with features of previous studies. The features for Baseline 1 are as follows:

- Bag-of-Words
- The number of words
- The number of positive words
- The number of negative words
- The number of a word followed by a word of opposite polarity
- The number of proper nouns
- The number of pronouns

We used the same dictionary as in Sect. 4.1 to identify positive words and negative words in tweets. There were two differences between Baseline 1 and our method. One was that our method used an ensemble of two classifiers. The other was two proposed features: words modified by the indicator “(皮肉)” and words

expressing a role. In addition to Baseline 1, we prepared the second baseline method to verify the effectiveness of two proposed features. Baseline 2 used an ensemble of two classifiers without proposed features. In addition, we evaluated two proposed features and the combination as follows:

- Classification using an ensemble and the feature of words modified by the indicator “(皮肉)” (Method 1).
- Classification using an ensemble and the feature of words expressing a role (Method 2).
- Classification using an ensemble and both of two proposed features (Method 3).

Table 1 shows the results. P, R and F in the table indicate precision, recall and F-score respectively on the test data. We applied a one-sided sign test to evaluate significant differences statistically. Baseline 2 outperformed Baseline 1 on all criteria; precision, recall and F-score. This result shows the effectiveness of the classification using an ensemble approach. Method 1–3 outperformed Baseline 2. This result shows the effectiveness of each proposed feature. However, there was no significant difference between the F-score of Method 3 using both of the proposed features and the F-scores of Method 1 and 2. Therefore, the effectiveness of the classification using both of the proposed features was not clear.

Table 1. Experimental result

Method	P	R	F
Baseline 1	0.74	0.74	0.74
Baseline 2 (Ensemble)	0.82	0.81	0.81 ^a
Method 1 (Ensemble and words modified by the indicator “(皮肉)”)	0.86	0.85	0.85 ^{a, b}
Method 2 (Ensemble and words expressing a role)	0.85	0.84	0.84 ^{a, b}
Method 3 (Ensemble and both of proposed features)	0.85	0.84	0.84 ^{a, b}

^a indicates $p < 0.01$ from Baseline 1

^b indicates $p < 0.05$ from Baseline 2

6 Discussion

We extracted words modified by the indicator “(皮肉)” as features. The effectiveness of the features was based on an assumption that words modified by the indicator “(皮肉)” were more important than the other words in a tweet. To verify the assumption, we extracted the list of words randomly selected from entire tweets, and exchanged words modified by the indicator “(皮肉)” for the list. Then, in the same manner as Sect. 3.1, we extracted words that appears more frequently in sarcastic tweets than in non-sarcastic tweets from the list and obtained 257 words as the features. We evaluated the random-selected method.

As a result, the F-score was 0.84. The F-score for Method 3 using the feature of words modified by the indicator “(皮肉)” in Sect. 5.2 was 0.85. Although the F-score of Method 3 was slightly better than that of the random method, there was not a significant difference between these results. Therefore, the importance of words modified by the indicator “(皮肉)” is not always clear, and further analysis is needed.

In addition, our method used an ensemble of two classifiers. Since these classifiers were trained with different features, it seems that the importance of each classifier was different. However, we assigned an equal weight to each classifier. Therefore, some weighting scheme is needed to improve our method.

7 Conclusion

In this paper, we proposed a method to detect sarcasm. We focused on words modified by the indicator “(皮肉)” and words expressing a role. We proposed features based on the lists of the words. For the classification, we applied an ensemble of two SVM classifiers trained with different features. In the experiment, we compared our methods with baseline methods. The result showed that the two proposed features were effective. In the future work, analysis of the importance of feature words and using a weighting scheme for classifiers are needed to improve our method.

Acknowledgements. This work was partially supported by JSPS KAKENHI Grant Number 17H01840.

References

1. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: sentiment analysis of figurative language in twitter. In: Proceedings 9th International Workshop on Semantic Evaluation (SemEval2015), Co-located with NAACL, pp. 470–478 (2015)
2. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.* **47**(1), 239–268 (2013)
3. Joshi, A., Sharma, V., Bhattacharyya, P.: Harnessing context incongruity for sarcasm detection. In: Proceedings of ACL-IJCNLP, pp. 757–762 (2015)
4. Karoui, J., Farah, B., Moriceau, V., Aussenac-Gilles, N., Hadrich-Belguith, L.: Towards a contextual pragmatic model to detect irony in tweets. In: Proceedings of ACL-IJCNLP, pp. 644–650 (2015)
5. Riloff, E., Qadir, A., Surve, P., Silva, L.D., Gilbert, N., Huang, R.: Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of EMNLP 2013, pp. 704–714 (2013)
6. Campbell, J.D., Katz, A.N.: Are there necessary conditions for inducing a sense of sarcastic irony? *Lang. Resour. Eval.* **49**(6), 459–480 (2012)
7. Hernandez-Farias, I., Benedi, J.M., Rosso, P.: Applying basic features from sentiment analysis on automatic irony detection. In: Proceedings of 7th ibPRIA, pp. 337–344 (2015)

8. Tungthamthiti, P., Shirai, K., Mohd, M.: Recognition of sarcasm in tweets based on concept level sentiment analysis and supervised learning approaches. In: Proceedings of the 28th PACLIC, pp. 404–413 (2014)
9. Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., Fukushima, T.: Collecting evaluative expressions for opinion extraction. In: Proceedings of IJCNLP-04, pp. 584–589 (2004)
10. Higashiyama, M., Inui, K., Matsumoto, Y.: Acquiring noun polarity knowledge using selectional preferences. In: Proceedings of the 14th Annual Meeting of the Association for NLP, pp. 584–587 (2008)
11. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of html documents. In: Proceedings of EMNLP-CoNLL, pp. 1075–1083 (2007)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)

Reducing Computational Effort for Plagiarism Detection with Approximate String Matching

Tetsuya Nakatoh¹(✉) and Toshiro Minami²

¹ Research Institute for Information Technology, Kyushu University,
744 Motoooka Nishi-Ku, Fukuoka 819-0395, Japan
nakatoh@cc.kyushu-u.ac.jp

² Kyushu Institute of Information Sciences, 6-3-1 Saifu Dazaifu,
Fukuoka 818-0117, Japan
minamitoshiro@gmail.com

Abstract. Currently, a large number of documents are created as digital material and distributed world-wide. Digital materials are easy to publish and copy at a remarkably low cost. As a result, many documents are copied illegally, and this practice is spreading, making plagiarism a significant social issue. Therefore, the need to develop systems that detect plagiarism is very high. We have developed a new plagiarism detection method that compares documents by using approximate string matching to detect plagiarism. We have also developed a technique that reduces the computational time of the comparison method. In this paper, we demonstrate our proposed method's usefulness through experiments and through the measuring indexes of precision and recall.

Keywords: Plagiarism detection · Approximate string matching · FFT

1 Introduction

In recent years, most documents are being created as electronic documents. Advances in internet technology have also made it easier to access these electronic documents. For convenience, many electronic documents can be browsed on the internet. Because electronic documents are easy to duplicate, this convenience also creates a problem as copyright infringement and plagiarism can be easily carried out. This flood of unauthorized copies will negatively affect the interests and rights of the copyright owner. As a result, there is a possibility that publishing excellent documents may be suppressed or publishing costs will be driven upwards. In order to solve this problem, it is necessary to protect the rights of the copyright owner, and a technique to search for and detect plagiarism at high speed from many documents is required.

Many studies on copied document detection are summarized in Lukashenko et al. [1]. Existing research is basically based on similarity calculation of documents using statistical properties of words. Apart from those studies, there is a study based on similarity calculation using the Hamming distance between two documents [2]. Although this method requires a great deal of computational complexity, it also proposes a technique to reduce computations [3].

In this paper, we will examine the effectiveness of the proposed idea [3] to reduce computational complexity by using a test bed for plagiarism detection. First, we show the criteria for similarity of character strings for plagiarism detection. These criteria, based on new ideas, greatly reduce computation time. We will also examine the accuracy of plagiarism detection using this standard. The rest of this paper is organized as follows. Section 2 introduces our method for plagiarism detection based on approximate string matching. Section 3 presents the experiments that evaluate our method in terms of the accuracy of plagiarism detection, and we evaluate the experimental results and discuss future directions. In Sect. 4, we discuss related work. Finally, in Sect. 5, we present our conclusions.

2 The Plagiarism Detection Method

In this section, we present our proposed method of plagiarism detection. First, in Sect. 2.1, we define the string matching problem and string matching problem with mismatches. In Sect. 2.2, we demonstrate our idea for accelerating the algorithm.

2.1 Algorithms for String Matching with Mismatches

2.1.1 String Matching Problem

Let Σ be the set of an alphabets, and let $T = t_1, t_2, \dots, t_n$ and $P = p_1, p_2, \dots, p_m$ be strings over Σ , where $n > m$. We name T as the longer string, i.e., a text, and P as the shorter one, i.e., a pattern. The string matching problem [4–6] poses a problem in detecting all of the occurrences of pattern P in the text T . This problem can be solved using an algorithm with complexity $O(n)$.

2.1.2 String Matching with Mismatches Problem

This is a string matching problem that allows replacement only for editing. The distance is defined as the Hamming distance in this type of matching. The matching score is defined by the difference between the length of a pattern and the Hamming distance. As the score is calculated at every position of the text T , this problem can be considered as a score vector $C(T, P) = (c_1, c_2, \dots, c_{n-m+1})$ of the matching score of pattern P at every available position in the text T . Note that c_i for $i = 1, 2, \dots, n - m + 1$ is the number of matching symbols between the substring $(t_i, t_{i+1}, \dots, t_{i+m-1})$ of T and P , and when $c_i = m$, the length of the pattern P , the pattern P itself appears at the i -th position in the text T .

The simple algorithm for calculating the score vector is to compare symbols m times each for $i = 1, 2, \dots, n - m + 1$, which has a computational complexity of $O(mn)$.

Fischer et al. [7] determined that it is possible to calculate matching scores by using convolution. Gusfield presented an algorithm to calculate scores with computational complexity of $O(|\Sigma|n \log m)$ using the Fast Fourier Transform (FFT), on the basis of Fischer's findings [4]. As improvements to this algorithm, randomized algorithms that approximate the solution have been proposed [8–12]. Furthermore, Nakatoh et al. provided optimistic solutions for this type of randomized algorithm [13].

An algorithm for calculating string matching with mismatches is as follows:

1. Convert the strings with lengths n and m to sequences of complex numbers using a mapping function.
2. Convert the two sequences of complex numbers by discrete Fourier transform (DFT). The computational complexity of DFT is $O(kn \log m)$, where k is the number of samples in the stochastic (or randomized) algorithm. We consider k as a constant in the following steps.
3. Multiply the elements of the two sequences that have been converted by DFT.
4. Calculate the score vector by converting the obtained sequence inversely using Inverse DFT (IDFT).

2.2 Application to Plagiarism Detection and a Technique for Improving Speed

Many string matching algorithms are difficult to apply to cases where mismatching occurs many times. Our algorithm, on the other hand, is applicable to such cases. We can detect plagiarism by detecting partial matches by directly comparing two documents as strings, using this property.

However, the computational complexity of our algorithm is of the order $O(m \log m)$, and thus the computational complexity rises to $O(dm \log m)$ when comparing d papers, which implies high computational complexity in comparison with statistical methods. We discuss our idea for reducing computational complexity to approximately $O(dm)$ below.

In order to compare a new paper to d papers with average length m ,

- Apply the DFT to d papers in advance.
- Filter the paper that is possibly plagiarized in advance, before applying the IDFT.

As an effect of prefiltering, if the number of target papers is r , the computational complexity becomes $O(dm + rm \log m)$ instead of $O(dm \log m)$, and then the processing time of our method can be estimated as $O(dm)$ if $r \ll d$.

This prefiltering is possible by discriminating vectors before calculation using the IDFT. Figure 1 shows a sample vector by comparing articles with similar contents, and Fig. 2 shows a sample vector by comparing articles with different contents.

3 Experiments

We conducted experiments on plagiarism detection using the proposed method described in the previous section. We applied the method to a dataset and investigated its accuracy.

3.1 Method

We used the Training Corpus dataset for Text Alignment Task of Plagiarism Detection available from (PAN2013). The dataset contains 1,000 pairs of documents with plagiarism (positive pairs) and 1,000 pairs of documents with no plagiarism (negative pairs).

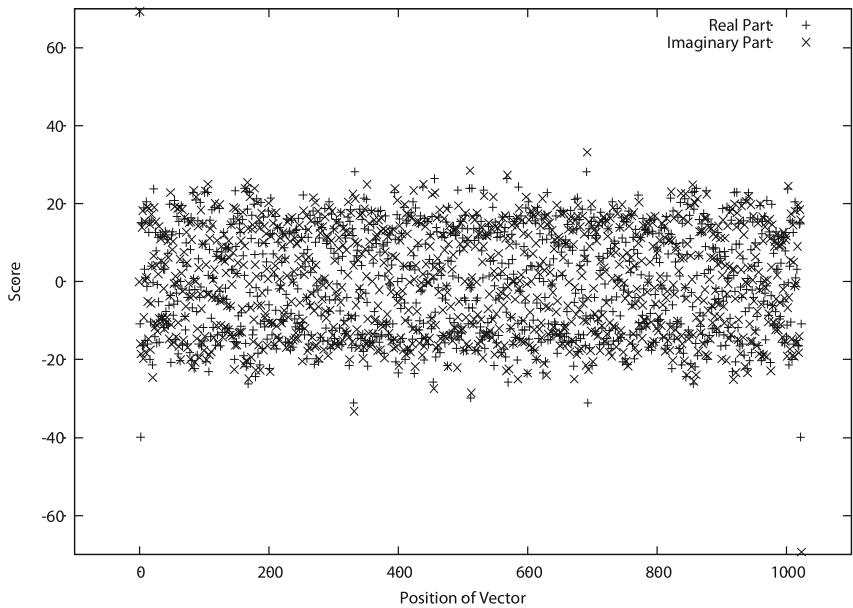


Fig. 1. Sample vector by comparing articles with similar contents

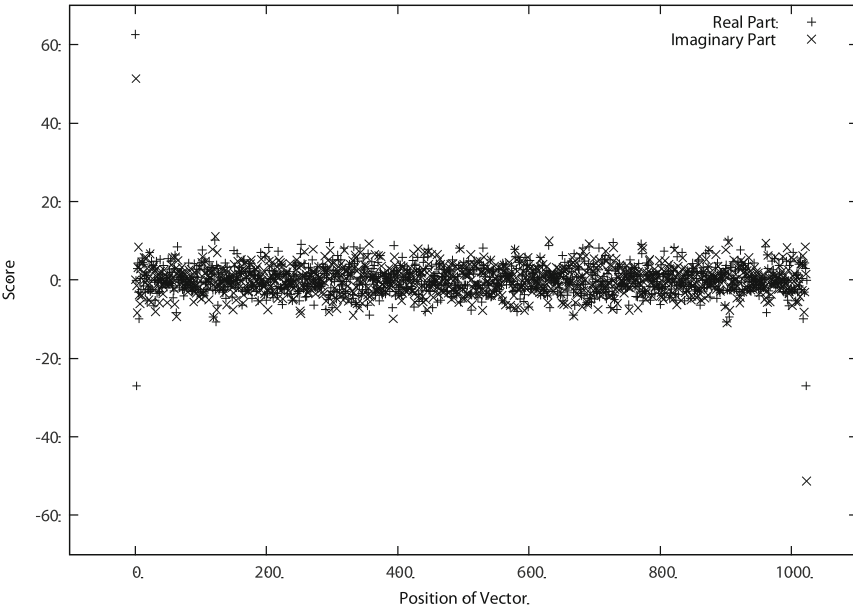


Fig. 2. Sample vector by comparing articles with different contents

We applied the proposed method to the dataset and investigated the following accuracy measures for plagiarism detection.

Let the number of the pairs predicated by a detection be as follows:

- Positive in positive pairs (True Positive, or TP).
- Positive in negative pairs (False Positive, or FP).
- Negative in positive pairs (False Negative, or FN).

Then, the precision and recall of the detection are defined as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$, respectively.

3.2 Result

Figure 3 shows the precision and the recall with different thresholds of the variance. Plagiarism was detected by the proposed method with relatively low accuracy. The highest value of the F-measure (the harmonic mean of the precision and the recall) was 0.73.

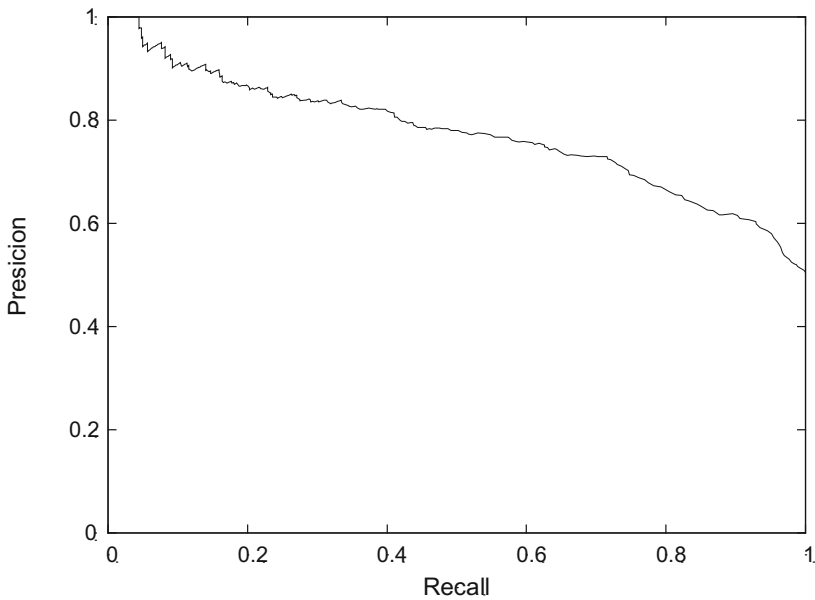


Fig. 3. Precision and recall of plagiarism detection

3.3 Discussion

Figure 3 shows the possibility of plagiarism detection using the proposed method. However, the accuracy is not high. To actually use this method, an improvement in accuracy is necessary. For example, in the experiment, parameters in approximation string matching were not adjusted, and each letter of a text was used directly as a unit of

the matching process. Using a word or a phrase as the unit may improve accuracy. On the other hand, all the samples available were used in the randomized algorithm; therefore, it is also necessary to evaluate the accuracy based on the number of samples compared.

4 Related Work

Plagiarism is normally detected by text alignment and matching. A text alignment task is included in the plagiarism detection task presented at the international conference of PAN@CLEF, where many research outcomes [14–16] are reported. The most important outcome in PAN2012 was provided by Kong et al. [17]. They used term frequency-inverse document frequency (TF-IDF) and cosine similarity in their paper. Torrejón et al. [18] achieved the highest performance in PAN2013. They combined more than one method using n -grams. Sanchez-Perez et al. [19] were the winners in PAN2014. They used cosine similarity and a Dice coefficient for TFIDF representations.

These algorithms are based on one-to-one comparison. One-to- N and N -to- N document comparisons lead to computational complexity issues. Our approach presents a new challenge regarding investigating an algorithm having improved computational complexity. We use vectors which take characters as their elements in this paper. Other types of vector representations useful in other works are also possible using our algorithm. Our future work will investigate the application of our algorithm to other works.

5 Conclusion

In this paper, we evaluated the possibility of reducing computational effort for plagiarism detection experimentally. It was confirmed that the intermediate data of approximate string matching produces a difference in the numerical values regarding plagiarism. It demonstrates that plagiarism detection is possible by using the proposed reduction, although improved precision is necessary for practical use. Further research is required on how the symbol for string matching is determined and how other parameters are set.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 15K00426.

References

1. Lukashenko, R., Graudina, V., Grundspenkis, J.: Computer-based plagiarism detection methods and tools: an overview. In: Proceedings of the 2007 International Conference on Computer Systems and Technologies. ACM, 2007, pp. 1–6
2. Nakatoh, T., Baba, K., Yamada, Y., Ikeda, D.: Partial plagiarism detection using string matching with mismatches. In: Proceedings of Informatics Engineering and Information Science, Springer CCIS254, pp. 265–272 (2011)
3. Nakatoh, T., Baba, K., Yamada, Y., Ikeda, D.: Speed improvement of the plagiarism detection method. In: Proceedings of IIAI International Conference on Advanced Information Technologies 2013 (IIAI AIT 2013), 2013, P38.pdf

4. Gusfield, D.: Algorithms on Strings, Trees, and Sequences. Cambridge University Press, New York (1997)
5. Crochemore, M., Rytter, W.: Text Algorithms. Oxford University Press, Inc., New York, NY, USA (1994)
6. Crochemore, M., Rytter, W.: Jewels of Stringology. World Scientific Publishing Company (2002)
7. Fischer, M.J., Paterson, M.S.: String-matching and other products. In Proceedings of the SIAM-AMS Applied Mathematics Symposium. Massachusetts Institute of Technology Cambridge, MA, USA, pp. 113–125 (1974)
8. Atallah, M., Chyzak, F., Dumas, P.: A randomized algorithm for approximate string matching. *Algorithmica* **29**(3), 468–486 (2001)
9. Baba, K., Shinohara, A., Takeda, M., Inenaga, S., Arikawa, S.: A note on randomized algorithm for string matching with mismatches. *Nordic J. Comput.* **10**, 2–12 (2003)
10. Nakatoh, T., Baba, K., Ikeda, D., Yamada, Y., Hirokawa, S.: An efficient mapping for computing the score of string matching. *J. Automata Lang. Combinat.* **10**(5/6), 697–704 (2005)
11. Schoenmeyr, T., Zhang, D.Y.: FFT-based algorithms for the string matching with mismatches problem. *J. Algorithms* **57**(2), 130–139 (2005)
12. Baba, K., Tanaka, Y., Nakatoh, T., Shinohara, A.: A generalization of FFT algorithm for string matching. In: Proceedings of International Symposium on Information Science and Electrical Engineering, pp. 191–194 (2003)
13. Nakatoh, T., Baba, K., Ikeda, D., Mori, M., Hirokawa, S.: Accuracy evaluation of FFT-based randomized algorithms for string matching with mismatches (in Japanese). *IPSP Trans. Databases (TOD)* **2**(4), 24–31 (2009)
14. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection in Working Notes Papers of the CLEF 2012 Evaluation Labs, eds. by Forner, P., Karlgren, J., Womser-Hacker, C., Sept. 2012
15. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection in Working Notes Papers of the CLEF 2013 Evaluation Labs, eds. by Forner, P., Navigli, R., Tufis, D. Sept. 2013
16. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection in Working Notes Papers of the CLEF 2014 Evaluation Labs, ser. CEUR Workshop Proceedings, L. Cappellato, N. Ferro, M. Halvey, W. Kraaij, Eds. CLEF and CEUR-WS.org, Sep. 2014
17. L. Kong, H. Qi, S. Wang, C. Du, S. Wang, and Y. Han, “Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection-Notebook for PAN at CLEF 2012,” in *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, 17–20 September, Rome, Italy, P. Forner, J. Karlgren, and C. Womser-Hacker, Eds., Sept. 2012
18. Rodríguez Torrejón, D., Martín Ramos, J.: Text alignment module in CoReMo 2.1 plagiarism detector-notebook for PAN at CLEF 2013. In: CLEF 2013 Evaluation Labs and Workshop Working Notes Papers, 23–26 September, Valencia, Spain, eds. by Forner, P., Navigli, R., Tufis, D. Sept. 2013
19. Sanchez-Perez, M., Sidorov, G., Gelbukh, A.: A winning approach to text alignment for text reuse detection at PAN 2014 Notebook for PAN at CLEF 2014. In: CLEF 2014 Evaluation Labs and Workshop—Working Notes Papers, 15–18 September, Sheffield, UK, ser. CEUR Workshop Proceedings, eds. by Cappellato, L., Ferro, N., Halvey, M., Kraaij, W., CEUR-WS.org, Sept. 2014

Weighting of Noun Phrases Based on Local Frequency of Nouns

Yasuhiro Yamada¹(✉), Yuusuke Himeno², and Tetsuya Nakatoh³

¹ Interdisciplinary Graduate School of Science and Engineering, Shimane University,
1060 Nishikawatsu-cho, Matsue-shi, Shimane 690-8504, Japan

yamada@cis.shimane-u.ac.jp

² Interdisciplinary Faculty of Science and Engineering, Shimane University, 1060
Nishikawatsu-cho, Matsue-shi, Shimane 690-8504, Japan

³ Research Institute for Information Technology, Kyushu University, 744 Motooka,
Nishi-ku, Fukuoka 819-0395, Japan

nakatoh@cc.kyushu-u.ac.jp

Abstract. The tf-idf is a well-known weighting measure for words in texts. It measures both the frequency and the locality of words. It is often used for information retrieval and text mining. However, a lot of infrequent words have the same tf-idf value. In this study, the words are noun phrases. This paper proposes a novel weighting measure for noun phrases in texts by using the local frequency of nouns that construct a noun phrase. The proposed measure is calculated by combining the tf-idf of a noun phrase and the average of the difference between its frequency and the frequency of nouns within the phrase. The proposed measure was evaluated in experiments on the datasets of 19,997 newsgroup texts written in English and 206 Wikipedia pages written in Japanese. The experiments showed that the number of noun phrases with the same proposed measure is less than the number of noun phrases with the same tf-idf.

Keywords: Term weighting · Noun phrase · Information retrieval
Text mining

1 Introduction

Term weighting is used to quantify the importance of words in each text of a set of given texts. Each text is expressed by a vector of the importance of words. This model is called a vector space model. Term weighting is often used as preprocessing in information retrieval and text mining. A well-known weighting measure is the tf-idf measure [1], which is calculated based on the frequency and the locality of words. Given a set of texts, the tf-idf of a word in a text is high if the word frequently appears in a small number of texts.

Zipf's law states that the frequency rank of words in natural language texts is inversely proportional to the frequency of words [2]. This law means that the frequency of most words in the texts is quite low. Many words appear only once or twice in the texts. Such words have the same tf-idf value because the frequency of the words and the document frequency are the same. Therefore, we cannot distinguish the importance of these words by the tf-idf. Infrequent words are often ignored as trash in information retrieval and text mining. However, many rare words, such as technical words in a particular domain, are useful for understanding the texts, and so a method is needed to extract such words in the texts.

The target of this paper is the weighting of noun phrases instead of words in texts. This paper defines a noun phrase as nouns appearing successively.¹ Phrases tend to appear infrequently because they are longer than nouns. Therefore, the above problem about words having the same tf-idf value also affects noun phrases.

We propose a novel weighting measure for noun phrases based on the local frequency of the nouns composing a noun phrase. The idea is simple: a noun phrase is important if the nouns within the noun phrase appear only in the phrase and do not appear in other parts of the texts. Such nouns are considered to be strongly related with only the phrase. We calculate the average of the difference between the frequency of a noun phrase and the frequency of nouns within the phrase. The novel weighting measure is calculated by combining the tf-idf of the noun phrase and the average of the difference. The measure considers the locality of the nouns in addition to the locality of the noun phrase in the tf-idf measure.

We conducted experiments on two datasets. One is called the 20-newsgroups dataset, which is 19,997 English texts for 20 different newsgroups [3]. The other is 206 Japanese Wikipedia pages describing information about countries, e.g., the USA and Japan.

2 Related Work

2.1 Term Weighting

Term weighting is used to quantify the importance of words in each text in a set of given texts. Term weighting is often calculated in advance in information retrieval and text mining. In a vector space model, each text is expressed by a vector of the weights of words.

As stated above, the tf-idf is a traditional and well-known weighting measure [1]. We introduce the definition of the tf-idf. Let $D = \{d_1, d_2, \dots, d_m\}$ be a set of texts, and $W = \{w_1, w_2, \dots, w_n\}$ be a set of words. The frequency of $w_i (1 \leq i \leq n)$ in $d_j (1 \leq j \leq m)$ is denoted by $freq(w_i, d_j)$. The total occurrence of all words in d_j is denoted by $t(d_j)$. The total frequency $tf(w_i, d_j)$ of w_i in d_j is defined as follows: $tf(w_i, d_j) = freq(w_i, d_j)/t(d_j)$.

¹ The original definition of noun phrases is complex compared to the definition used in this paper.

Next, the document frequency of w_i in D is denoted by $df(w_i)$. The inverse document frequency $idf(w_i)$ of w_i in D is defined as follows: $idf(w_i) = \log(m/df(w_i)) + 1$ where m is the number of texts in D .

Finally, the tf-idf $tfidf(w_i, d_j)$ of w_i in d_j is defined as follow:

$$tfidf(w_i, d_j) = tf(w_i, d_j) \times idf(w_i).$$

From the definition of tf-idf, the tf-idf of a word is high if the word appears many times in a small number of texts. On the other hand, the tf-idf is small if the word appears fewer times or appears in a large number of texts. The tf-idf considers both the frequency of the occurrence of words and the locality of the words in the texts. It should be noted that other definitions of the tf-idf can be given [4, 5]. In Sect. 3, we point out a problem about the tf-idf of infrequent words.

Similar to the tf-idf, Okapi BM25 is a weighting measure for words which appear in both a query and a document in information retrieval [6]. Some researches have discussed about the BM25 [7, 8].

2.2 Noun Phrase Extraction

The purpose of this paper is to quantify the importance of all noun phrases in given texts. Other studies have proposed methods for extracting noun phrases in a text or a set of texts [9, 10].

Kita et al. proposed a statistic to extract collocations [9]. A collocation is defined as a word sequence in this study. They compared the difference of the frequency between two kinds of collocations: one is a word sequence composition s_1 , and the other is a word sequence, which is a concatenation of s_1 and a word. Frantzi and Ananiadou also proposed a statistic to extract collocations [10]. They compared the difference of the frequency in the same way as [9].

Noun phrase extraction is often utilized in domain-specific terms extraction from domain-specific texts [11, 12]. Kathait et al. proposed an unsupervised method to extract important noun phrases [13]. In advance, they determined words that appear as the first word of a noun phrase and appear as the successive word of the phrase. Then, the method extracts a word sequence using the above words as a noun phrase. The method extracts noun phrases that include keywords calculated from the co-occurrence of words.

3 Novel Weighting Measure for Noun Phrases

This section describes the novel weighting measure for noun phrases. We calculate the weight of noun phrases instead of words in texts. As stated above, this paper defines a noun phrase as nouns appearing successively in texts. For example, “information retrieval” is a noun phrase consisting of two nouns “information” and “retrieval.”

3.1 Tf-Idf of Noun Phrases

A problem of calculating the tf-idf of nouns phrases in a text is that the frequency of most phrases is quite low. This problem causes the phrases to have the same tf-idf. Figure 1 shows the total frequency and the tf-idf of noun phrases in the texts of the 20-newsgroups dataset.² Here, “alt.atheism” and “comp.graphics” are group names. The vertical axis is a logarithmic scale.

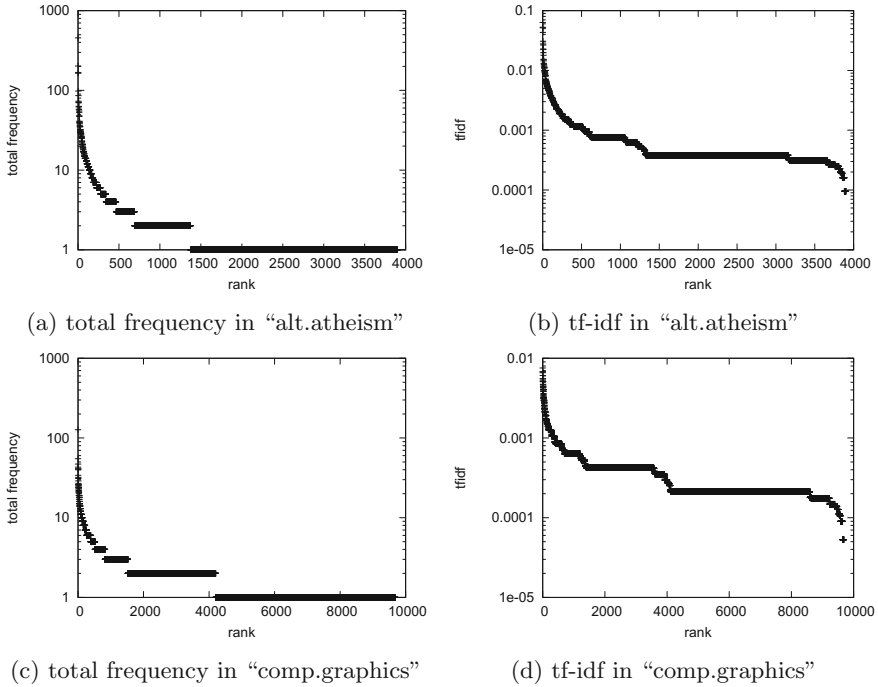


Fig. 1. Total frequency and the tf-idf of noun phrases in the 20-newsgroups dataset

We see that the total frequency of most phrases is less than 10 in Fig. 1a, c. The frequency of more than half of the phrases is 1. In Fig. 1b and d, we also see that many phrases have the same tf-idf. In particular, the tf-idf of noun phrases that appear once is the same because the total frequency of the phrases and the document frequency are 1. Many of the infrequent noun phrases have the same tf-idf, so we cannot distinguish the importance of such phrases by tf-idf.

Generally speaking, infrequent noun phrases (or words) are often ignored in the studies on information retrieval and text mining. This is because infrequent

² The 20-newsgroups dataset is a set of 19,997 newsgroups texts written in English [3]. The dataset has 20 different groups. We concatenated texts in the same group into a text. Therefore, the number of texts is 20 in the experiments of this paper.

noun phrases or words are not useful to these studies and they make the dimension of the vector of a text quite large. However, many infrequent words, such as technical words in a specific domain, are useful for understanding the texts. Consequently, we need a method to extract such phrases in the texts.

3.2 Purity

We propose a novel weighting measure for noun phrases in this section. We look at the difference between the frequency of a noun phrase and the frequency of nouns within it. A similar idea is described in [14], in which Yamada et al. proposed statistics for quantifying the peculiarity of a pattern. The pattern is a substring, not a word or a noun phrase in a text [14]. The peculiarity of a pattern is called the purity of it.

We introduce the definition of the purity in [14]. Let x be a string. A substring from position i to j on x is denoted by $x[i : j]$. Let s be a substring of x . The component of s with respect to x , denoted by $com(s)$, is a set of all pairs $\langle i, j \rangle$ s.t. $i' \leq i \leq j \leq j'$, where i' (j') is the starting (ending) position of s on x . The average of the difference of frequency for s with respect to x is calculated as follows:

$$diff(s) = \frac{1}{|com(s)|} \sum_{\langle i, j \rangle \in com(s)} (freq(x[i : j]) - freq(s))$$

where $freq(*)$ is the frequency of a substring $*$ on x . The average $diff(s)$ is called the purity of s . The purity is 0 if all substrings of s appear within s and they do not appear within other substrings in x .

It is natural that the frequency of substrings of s is higher than the frequency of s itself because s includes the substrings. However, if a substring s does not satisfy this natural assumption, the substring is extraordinary and valuable in x .

We apply this natural assumption to the weighting of noun phrases in texts. Let n_1, n_2, \dots, n_m be nouns, and let $p = n_1 n_2 \dots n_m$ be a noun phrase, where “ $_$ ” means a space.³ If the nouns n_1, n_2, \dots, n_m appear within only the noun phrase p , then the nouns are strongly associated with only the phrase. The frequency of n_1, n_2, \dots, n_m and p is the same when this situation occurs.

As shown in [14], it is natural that the frequency of nouns n_1, n_2, \dots, n_m is higher than the frequency of a noun phrase p because the phrase includes the nouns. If the noun phrase does not satisfy this assumption, then it is extraordinary and valuable in the texts.

From the above considerations, we define the purity of a noun phrase. Let $D = \{d_1, d_2, \dots, d_x\}$ be a set of texts, and $P = \{p_1, p_2, \dots, p_y\}$ be a set of noun phrases. First, the average of the difference between the frequency of a noun phrase $p = n_1, n_2, \dots, n_m$ and the frequency of nouns n_1, n_2, \dots, n_m in a text d_j is defined as follows:

$$d(p, d_j) = \frac{1}{m} \sum_i (freq(n_i, D) - freq(p, d_j)) \quad (1)$$

³ In Japanese, a noun phrase is expressed by $p = n_1, n_2, \dots, n_m$ without spaces.

Table 1. Number of noun phrases with the same tf-idf and the same purity

File name	# of noun phrases	With the same tf-idf	With the same purity
alt.atheism	3,902	3,833	1,619
comp.graphics	9,684	9,631	6,312
USA	2,664	2,621	499
Japan	4,926	3,864	1,677

where $freq(*, D)$ is the total frequency of $*$ in D .

Then, the purity of p_i in d_j , i.e. the weight of p_i in d_j , is defined by combining the tf-idf of p_i and the above average as follows:

$$purity(p_i, d_j) = tfidf(p_i, d_j) \times \frac{1}{d(p_i, d_j) + 1}. \quad (2)$$

The value of $1/(d(p_i, d_j) + 1)$ in the formula (2) is 1 if the above average is 0, that is, the frequency of p_i and the frequency of all nouns within p_i are the same. The value is 0.1 if the average is 9. The larger the average is, the smaller the purity is.

Given a set of texts, we calculate $purity(p_i, d_j)$ for all noun phrases in all the texts.

4 Experiment

We use two kinds of datasets. The first dataset is the 20-newsgroups with 19,997 newsgroups texts written in English. As in Fig. 1, we concatenated texts in the same group into a text. Therefore, the number of texts is 20. The sum of the file size of the texts is about 44 MB, and the average file size is about 2.2 MB. We used TreeTagger⁴ for morphological analysis of the English texts. In TreeTagger, a word whose part of speech starts with “N” is a noun.

The second dataset is 206 texts of Wikipedia that describe information about countries.⁵ The texts are written in Japanese. The sum of the file sizes of the texts is about 10 MB, and the average file size is about 50 KB. We used MeCab⁶ for morphological analysis of the Japanese texts.

Figure 2a, b show the tf-idf and the purity of noun phrases in the files “alt.atheism” and “comp.graphics” of the 20-newsgroups dataset. Figure 2c, d show the tf-idf and the purity of noun phrases in the files “USA” and “Japan” of the Japanese Wikipedia dataset. Both axes in the figures are logarithmic scales.

⁴ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.

⁵ The texts were collected from a Wikipedia page written about a list of countries on June 22, 2017. The URL of the page is <https://ja.wikipedia.org/wiki/%e5%9b%bd%e3%81%ae%e4%b8%80%e8%a6%a7>.

⁶ <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>.

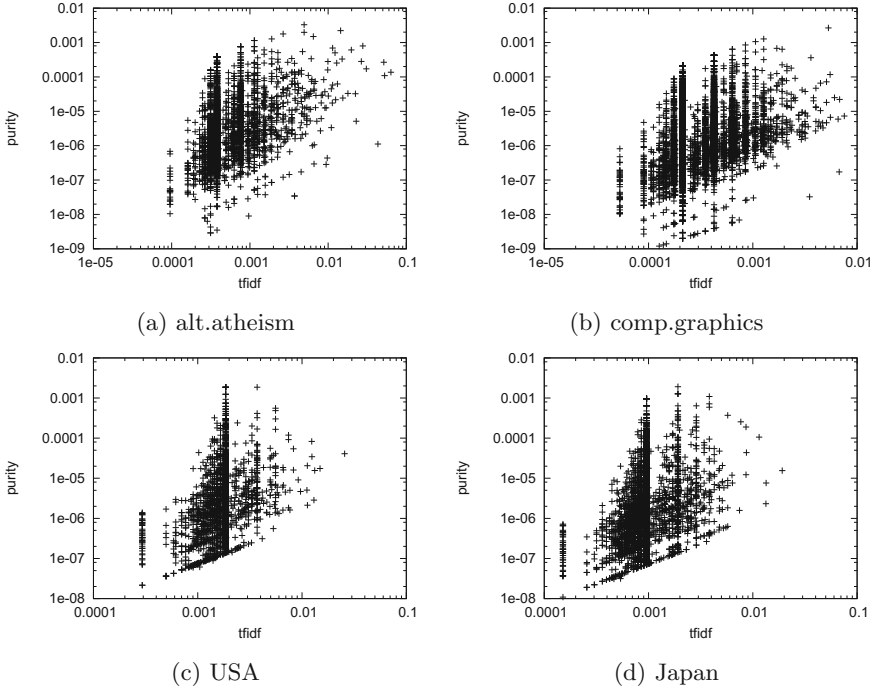


Fig. 2. Tf-Idf and purity

The correlation coefficient of “alt.atheism” between the tf-idf and the purity is 0.183, and that of “comp.graphics” is 0.137. The correlation coefficient of “USA” is 0.0481, and that of “Japan” is 0.0825. We see no correlation between the tf-idf and the purity in these datasets.

We counted the number of noun phrases with the same tf-idf and the same purity (see Table 1). The third column in Table 1 is the number of noun phrases with the same tf-idf as some other phrase. For example, 14 noun phrases have a tf-idf of 0.266×10^{-2} and 115 phrases have a tf-idf of 0.114×10^{-2} in “alt.atheism”. The third column is the number of such noun phrases. Similarly, the fourth column is the number of noun phrases with the same purity as some other phrase. Table 1 shows that the number of different noun phrases with the same purity is less than the number of noun phrases with the same tf-idf. Therefore, the purity can distinguish more noun phrases as compared to the tf-idf.

In the 20-newsgroups dataset, Table 2 shows the top 15 noun phrases with high purity in the file “alt.atheism”, and Table 3 shows the top 15 noun phrases with high purity in the file “comp.graphics”. The third column “freq.” means the frequency of a noun phrase in the file. The fourth column “doc. freq.” means the document frequency of a noun phrase in all 20 files. The fifth column $d(p_i, d_j)$ is calculated by the formula (1). A person name is replaced by “<name>.” We see many person names in the results. The document frequency of the top 15

Table 2. Top 15 noun phrases with high purity in the file “alt.atheism” of 20-newsgroups. The string “<name>” is the name of a person

Noun phrase	Purity	Freq	Doc. freq	$d(p_i, d_j)$
<name>	0.330×10^{-2}	13	1	0.500
<name>	0.222×10^{-2}	38	1	5.50
<name>	0.198×10^{-2}	13	1	1.50
acceptable _behavior_	0.178×10^{-2}	7	1	0.500
<name>	0.149×10^{-2}	17	1	0.333
Salaam a-laikum	0.127×10^{-2}	5	1	0.500
Alaikum Wassalam	0.114×10^{-2}	3	1	0.000
Sociologist sub-branch	0.114×10^{-2}	3	1	0.000
<name>	0.793×10^{-3}	73	1	34.0
AAH EXAMINER	0.761×10^{-3}	2	1	0.000
<name>	0.761×10^{-3}	2	1	0.000
<name>	0.761×10^{-3}	2	1	0.000
<name>	0.761×10^{-3}	2	1	0.000
nanny-nanny-boo-boo TBBBBBBBT'T'T'TTHHHHH	0.761×10^{-3}	2	1	0.000
oher S.Am	0.761×10^{-3}	2	1	0.000

Table 3. Top 15 noun phrases with high purity in the file “comp.graphics” of 20-newsgroups

Noun phrase	Purity	Freq	Doc. freq	$d(p_i, d_j)$
<name>	0.265×10^{-2}	25	1	1.00
<name>	0.127×10^{-2}	6	1	0.000
<name>	0.106×10^{-2}	5	1	0.000
<name>	0.849×10^{-3}	4	1	0.000
<name>	0.679×10^{-3}	8	1	1.50
<name>	0.636×10^{-3}	3	1	0.000
hacked-over v1	0.636×10^{-3}	3	1	0.000
<name>	0.636×10^{-3}	6	1	1.00
Delaunay triangulation	0.636×10^{-3}	9	1	2.00
<name>	0.566×10^{-3}	4	1	0.500
<name>	0.530×10^{-3}	5	1	1.00
.r.c V.t.ell	0.424×10^{-3}	2	1	0.000
<name>	0.424×10^{-3}	2	1	0.000
B-9000 Gent	0.424×10^{-3}	2	1	0.000
<name>	0.424×10^{-3}	2	1	0.000

Table 4. Top 15 noun phrases with high purity in the file “USA” of Japanese Wikipedia

Noun phrase	Purity	Freq	Doc. freq	$d(p_i, d_j)$
Granted patents	0.185×10^{-2}	1	1	0.000
Roaring twenties	0.185×10^{-2}	1	1	0.000
Teriyaki	0.185×10^{-2}	1	1	0.000
Self checkout	0.185×10^{-2}	1	1	0.000
The whirlwind of Red Scare	0.185×10^{-2}	1	1	0.000
The space of The Virginia Gazette	0.185×10^{-2}	1	1	0.000
Daddy-Long-Legs	0.185×10^{-2}	2	1	1.00
<name>	0.124×10^{-2}	1	1	0.500
Incandescent lamp	0.124×10^{-2}	1	1	0.500
<name>	0.124×10^{-2}	1	1	0.500
Tammany Hall	0.928×10^{-3}	1	1	1.00
Trout Fishing	0.743×10^{-3}	1	1	1.50
<name>	0.743×10^{-3}	1	1	1.50
Common bean	0.743×10^{-3}	1	1	1.50
Admiral Matthew Perry	0.743×10^{-3}	1	1	1.50

phrases is 1. We also see that some noun phrases that appear twice or three times in the file are included in the top 15 phrases. The value $d(p_i, d_j)$ of most noun phrases except for the ninth phrase in Table 2 is quite small. The purity of the phrase is high because its frequency in the file is high.

In the Japanese Wikipedia dataset, Table 4 shows the top 15 noun phrases with high purity of the file “USA.” The noun phrases in the table were originally written in Japanese. We see that not only the frequency of a noun phrase but also its document frequency is quite low. The value $d(p_i, d_j)$ of all of the noun phrases is also quite small. Some phrases related to Japan are in the result of “USA,” such as “teriyaki” and “Admiral Matthew Perry,” because the file is written in Japanese.

5 Conclusion

We proposed a novel weighting measure for noun phrases based on the local frequency of nouns constructing a noun phrase. We focus on the average of the difference between the frequency of a noun phrase and the frequency of nouns within the phrase. In contrast to the natural assumption about the frequency of noun phrases and the nouns, the noun phrases are extraordinary and interesting if the average is small. The proposed purity measure is calculated by combining the tf-idf of a noun phrase and the average. Experiments showed that the number of noun phrases with the same purity is less than the number of noun phrases

with the same tf-idf. The experiments also showed that most of noun phrases with high purity are infrequent phrases.

Our future work is to apply the proposed measure to keyword extraction from texts, domain-specific term extraction, information retrieval, clustering of texts, and so on. From the viewpoint of these applications, we need to examine the definition of purity and the experimental results in detail.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Numbers 15K00426.

References

1. Salton, G., McGill, J.M.: Introduction to Modern Information Retrieval. McGraw-Hill Inc, New York (1983)
2. Zipf, G.K.: The Psychobiology of Language. Routledge, London (1936)
3. Home Page for 20 Newsgroups Data Set. <http://qwone.com/~jason/20Newsgroups/>. Accessed 28 June 2017
4. Manning, D., Raghavan, P., Shütza, H.: An Introduction to Information Retrieval. Cambridge University Press (2008)
5. Rousseau, F., Vazirgiannis, M.: Composition of TF normalizations: new insights on scoring functions for ad hoc IR. In: 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 917–920. ACM, New York (2013)
6. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: 3rd Text REtrieval Conference, pp. 109–126 (1994)
7. Trotman, A., Puurula, A., Burgess, B.: Improvements to BM25 and language models examined. In: 2014 Australasian Document Computing Symposium, pp. 58–65. ACM, New York (2014)
8. Lipani, A., Lupu, M., Hanbury, A., Aizawa, A.: Verboseness fission for BM25 document length normalization. In: 2015 International Conference on The Theory of Information Retrieval, pp. 385–388. ACM, New York (2015)
9. Kita, K., Kato, Y., Omoto, T., Yano, Y.: A comparative study of automatic extraction of collocations from corpora: mutual information vs. cost criteria. *J. Nat. Lang. Process.* **1**(1), 21–33 (1994)
10. Frantzi, K.T., Ananiadou, S.: Extracting nested collocations. In: 16th Conference on Computational Linguistics, vol. 1, pp. 41–46. Association for Computational Linguistics, Stroudsburg (1996)
11. Li, S., Li, J., Song, T., Li, W., Chang, B.: A novel topic model for automatic term extraction. In: 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 885–888. ACM, New York (2013)
12. Astrakhantsev, N.A., Fedorenko, D.G., Turdakov, DYu.: Methods for automatic term recognition in domain-specific text collections: a survey. *J. Program. Comput. Softw.* **41**(6), 336–349 (2015)
13. Kathait, S.S., Tiwari, S., Varshney, A., Sharma, A.: Unsupervised key-phrase extraction using noun phrases. *Int. J. Comput. Appl.* **162**(1), 1–5 (2017)
14. Yamada, Y., Nakatoh, T., Baba, K., Ikeda, D.: Mining pure patterns in texts. In: 2012 IIAI International Conference on Advanced Applied Informatics, pp. 285–290 (2012)

Multi-layers Convolutional Neural Network for Twitter Sentiment Ordinal Scale Classification

Muath ALALI, Nurfadhlinah Mohd Sharef^(✉), Hazlina Hamdan,
Masrah Azrifah Azmi Murad, and Nor Azura Husin

Faculty of Computer Science and Information Technology, Intelligent
Computing Research Group, Universiti Putra Malaysia UPM Serdang, 43400
Selangor, Malaysia

baniatamuath@gmail.com, {nurfadhlinah, hazlina, masraha,
n_azura}@upm.edu.my

Abstract. Twitter sentiment analysis according to five points scales has attracted research interest due to its potential use in commercial and public social media application. A multi-point scale classification is a popular way used by many companies to evaluate the sentiment of product reviews (e.g. Alibaba, Amazon and eBay). Most of the classification approaches addressed this problem using traditional classification algorithm that requires expert knowledge to select the best features. Even though deep learning has been utilized, most of them employed a simple structure that not enough to capture the important features. In this paper, a complex structure of convolutional neural network (CNN) is proposed to classify the tweet into five-point scale and obtain a more several tweet representation. After a series of experiments with CNN including different hyperparameters and pooling strategies (Max and Average), we found that the best structure for our model is three convolutional layers, each one followed by average pooling layer. The proposed multi-layers convolutional neural network (MLCNN) model achieve the lowest Macro average mean absolute error (MAE^M) and outperforms the state-of-the-art approach on tweet 2016 dataset for Ordinal classification. Experimental results show the ability of average pooling to preserve significant features that provide more expressiveness to ordinal scale.

Keywords: Deep learning · Convolutional neural network (CNN) · Twitter sentiment analysis (TSA) · Ordinal classification · Text classification
Sentiment analysis (SA)

1 Introduction

The rapid growth of web technology and social media applications (e.g. Twitter, Facebook, google+, and several blogs), become the common platform for users to publish their thoughts, opinions, and reviews on any topic in a simple way, with Twitter that has become the most dominant microblogging site. The available information has become crucial for the government business, and individuals to gain

efficient understanding of the public opinions about a concept. Sentiment analysis (SA) [1] is an approach that could be applied for this means as it can analyze and detect the sentiment expressed in the texts. Twitter sentiment analysis is considered as an area of natural language processing (NLP) to classify the sentiment polarity of a Twitter message. Typically, a message can be polarized according to two or binary point scale (e.g., dislike, like), three-point (e.g., dislike, neutral, like), or five-point (e.g., strongly dislike, dislike, neutral, like, strongly like) [2–4].

Most approaches used to address Twitter sentiment analysis (TSA) based on sentiment lexicons and handcrafted and tweet specific features [5, 6]. These features are used as input for ML algorithm such as Random Forest (RF), Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), Maximum Entropy (MaxEnt). However, these methods require expert knowledge and are time-consuming.

The increasingly popular deep learning approaches for sentiment analysis, on the other hand, provide more robustness and adaptability by automatically extract features. Deep neural network approaches have been shown to produce state-of-the-art discovery [3, 5] and has the ability for capturing salient features.

Particularly the most successful networks for SA were Convolutional Neural Network (CNN). CNNs using pre-trained word embedding as input have been shown to produce state of the art for SA in two and three points scale [2–4]. Newly TSA tasks are extended to cover five-point scale classification, where a few of researchers tackled this task using deep learning and traditional ML methods to classify the tweet according to five points scale {Highly Positive, Positive, Neutral, Negative, and Highly Negative} [7]. However, the existing works in deep learning methods mainly used a simple structure of CNN (one convolutional layer with max pooling) based on methods that utilized for Two and Three scales to fit the ordinal classification problem. In the case of ordinal scale required to exploit other different structure of CNN which can obtain several tweets representation with pooling strategy that can preserve most significant information from the sentence. The main contributions of this paper are as follow:

- A complex CNN model is developed to extract a comprehensive representation in tweets that lead to extract an essential semantics in a tweet and to enhance the learning capacity.
- Determine and investigate the performance between max and average pooling that retrieve and preserve the most significant features for ordinal scale.
- Experiment results show that the proposed *MLCNN* model can achieve lower Macro average mean absolute error (MAE^M) compared with benchmark approach.

In this paper, we propose a novel CNN architecture to classify the tweet into five points scale. Our model consists of multiple convolutional and pooling layers trained on top of word embedding [8, 9]. We studied the impact of pooling strategies and filter size on the performance. The proposed model outperforms the benchmark approach for TSA according to an ordinal scale by [10].

The rest of this paper as follows. Section 2 gives a review of related work. Section 3 the proposed *MLCNN* model is described in details. The experiment results in Sect. 4. The conclusion in Sect. 5.

2 Related Work

Deep neural network methods have achieved effective results in SA especially CNN [2–4]. Newly TSA task is extended to cover ordinal scale from highly negative to highly positive. This task was proposed by *SemEval2016* [7], where fewer researchers tackled this problem compared to the three-point scale. State of the art in this task was achieved by [10] where they treated the problem as ordinal classification. They extracted several features including N-gram, Character gram, Tweet specific feature, and Lexicons. They evaluated two ways to represent the features bag of words and hashing function, and they found the performance using hashing representation is better than the bag-of-words method to classify the tweet they used logistic regression (LR) to reduce the multinomial loss across the classes.

Ensemble method is introduced in [11] to decompose the ordinal classification into several binary problems. The algorithms used are Random Forest (RF), SVM, and Gradient Boosting Trees (GBT) as a Base classifier. After performing the feature analysis, they found using character-grams better than word n-grams and noted using negated character-grams, and character-gram lexicons could enhance the performance. Furthermore, in [12] have used Balanced Binary Tree for Ordinal Regression (BBTOR) with different features including N-gram, Character gram, and tweet specific feature. Each tweet is transformed into its vectorial and weighted by TF-TDF, while [13] classify the tweet into ordinal scale using Bayes classifier combined with Alchemy-API based on different feature including N-gram and Tweet Specific Features.

Most of the methods that mentioned above focus on feature analysis to select the best features for five points polarities. Other works investigate the effectiveness of word embedding and local representation output from deep learning model against NLP features and the combination of them on the performance. For example, [14] have compared between typical NLP features and word embedding feature with four ML algorithm. They found that the traditional NLP features with LR classifier produce better results than Word Embedding with LR, SVM, and RF for Ordinal classification problem. Also, they found that using n-gram, lexicons, negation, and tweet specific feature such as punctuation, all-caps, hashtag, and emoticons are the best features. Meanwhile [15] extracted the hidden value from CNN and concatenated them with linguistic features and then used as features for SVM. However, these combinations of low and high variance features decrease the performance of the classification. In addition, the processes of selecting best features take a lot of time and arduous to define and may lead to over specific or incomplete feature.

On the other hand, [16] applied deep learning approach using the combination of CNN and a gated recurrent neural network. Both neural networks used pre-trained word embedding. Neural networks are used to obtain different tweet representation .the feature representation output from the networks are fused and fed to Softmax Regression classifier. Similarly, [17] used Convolutional Neural Network and long short-term memory to classify the Tweets according to Five Points Scale. The model combines the CNN and LSTM to capture the sentiment aspect of words. CNN is used to reduce the dimension of sentence Matrix Followed by LSTM layer to represent the sentence, then a linear regression layer to fit the sentiment of the sentence. Another

deep learning architecture is [18] using Convolutional Neural Networks (CNN) for ordinal classification. The CNN trained on top of word embedding. They found fine tuning embedding on distance supervision did not improve the performance and noted that additional training data could improve the performance.

The CNN approaches are applied to this task, based on distance supervision and a single convolutional layer with Max pooling to enhance the performance. However, distance supervision did not contribute to the performance and the structure of CNN is not enough to capture the structure of the problem. Furthermore, they utilized Max pooling regarding other researches we mentioned earlier in SA, that achieve state of the art. However, we are dealing with ordinal classification problem which requires appropriate pooling strategies that preserve the most important information in a sentence and provide more expressiveness to CNN. The performance of existing ordinal classification approaches in TSA could be enhanced since the MAE^M is still relatively high, further investigation of the different structure of CNN and pooling strategies that can extract more silent feature and improve the performance should be conducted. Table 1 summarizes the approaches that applied on TSA for five-point scale, where it is shown that the TwiSE model is the best existing work and the benchmark for the proposed MLCNN.

Table 1. Summary of approaches that addressed TSA according to five points scale

Model	Method	MAE^M
<i>TwiSE</i> [10]	<i>Logistic regression</i>	0.719
<i>ECNU</i> [14]	<i>Logistic regression</i>	0.806
<i>PUT</i> [11]	<i>Ensemble</i>	0.860
<i>LyS</i> [15]	<i>CNN + SVM</i>	0.864
<i>FINKI</i> [16]	<i>Fusion</i>	0.869
<i>INSIGHT-1</i> [18]	<i>CNN</i>	1.006
<i>ISTI-CNR</i> [12]	<i>SVM</i>	1.074
<i>YZU-NLP</i> [17]	<i>CNN and LSTM</i>	1.111
<i>Sentimentalists</i> [13]	<i>naïve bayes</i>	1.148

3 The Proposed MLCNN Model

The objectives of this paper are to address the identification of the best hyperparameters (number of convolutional layers and filter size) combination for the CNN implementation and to investigate the performance between max and average pooling. The CNN approaches are applied to this task using an extension of simple structure of CNN which achieves state of the art SA for two and three scales [2, 4, 19]. However, experiment in computer vision researches suggests that using multiple convolutional layers are important to achieve the best performance [20]. In this paper, our model for ordinal classification Twitter sentiment analysis consists of three convolutional layers with 100 filters with window size (2, 3, 4) and each one is followed with average

pooling trained on top of word embedding. The model is implemented using Keras¹ library on a Theano backend.

3.1 Pre-trained Word Embeddings

Recent researches have shown the word embeddings (WE) are more powerful for deep learning methods; these embeddings encode the semantic and syntactic features of the word. As we mentioned, the proposed CNN consist of word embedding. Each word in the tweet is mapped to a word embedding representation. For the proposed model, we used the publicly available GloVe² Embeddings [9] pre-trained on 2B tweets to initialize our word embedding with 200-dimensional Glove vector. For this phase, we define a lookup table, where each token is associated with WE representation and For tokens not appear in lookup table set to Zero. We do not update the WE during training.

3.2 MLCNN

The CNN take the input as text, which is padded to length n . We represent each sentence as a concatenation of WE for its tokens. Let w be a word of a vocabulary V , the distributed representation for that word as low dimensional vector $v \in \mathbb{R}^k$. So, we can create a matrix of embeddings $E \in \mathbb{R}^{v \times k}$ to be the input layer for CNN.

Consequently, given a tweet $s = [w_1, w_2, w_3, \dots, w_s]$, after initializing the input layer we will obtain a matrix $S \in \mathbb{R}^{s \times k}$ that will be the input for the convolutional layer. The sentence length is set to 50, taking the first 50 tokens if the sentence longer and padding with zeros if it is shorter. In our model, we use three convolutional with 100 filters with filter size 2, 3, 4, respectively. Convolutional layer utilizes the local correlation between the words in the sentence. Convolutional operation with f convolutional filter with size m applied to the embeddings input, to generate M feature maps.

Formally, let $S \in \mathbb{R}^{s \times k}$ the embeddings input matrix and let $W \in \mathbb{R}^{h \times k}$ be a filter, and the filter convolve over the word embeddings matrix to generate feature maps M_i for a window of h words defined as:

$$M_i = f(w \cdot S_{i:i+h-1} + b) \quad (1)$$

Where b is a bias, and f is a non-linear activation function *ReLU* [21], while $S_{i:i+h-1}$ is a concatenation vector which represents a filter of h words from a position i to position $i + h - 1$. The operation of the filter of h words in the sentence, obtaining the following feature maps:

$$M = [m_1, m_2, \dots, m_{n-h+1}] \quad (2)$$

Then, the average pooling operation applied over the feature maps which take the average value average $[m_1, m_2, m_3 \dots, m_{n-h+1}]$. The output is a fixed size vector and

¹ <http://keras.io>.

² <https://nlp.stanford.edu/projects/glove/>.

naturally deal with variable length input. The output of pooling layer will be the input to the next convolutional layer. The same operation will be applied to all convolutional and pooling layers. The final output from all convolutional layers is passed to a fully connected layer then to softmax layer to produce a probability distribution output overall output classes.

3.3 Dataset and Evaluation Metric

The model is trained on benchmark dataset provided by SemEval challenge 2016³ [7], where the dataset was divided into training, development, development test and test. We exclude the development set and use the training and development test set for training and test set for validation. The dataset annotated according to a five-point scale (Highly Positive, Positive, Neutral, Negative, HighlyNegative), Table 2 summarizes the class distribution among five class [-2, -1, 0, 1, 2]. Since we are dealing with ordinal classification and the dataset is highly imbalanced, we use to evaluate our model the Macro average mean absolute error (MAE^M). The MAE^M defined as:

$$MAE^M(h, Te) = \frac{1}{|c|} \sum_{j=1}^C \frac{1}{Te_j} \sum_{Xi \in Te_j} |h(Xi) - Yi|. \quad (3)$$

where X_i , $h(X_i)$ is the predicted table and Y_i is a true label, and Te_j is a set of test document whose true class is c_j .

Table 2. Summary of data statistics

Name	2	1	0	-1	-2	Total
Train	437	3154	1654	668	87	6000
DEVTEST	148	1005	583	233	31	2000
Test	382	7830	10081	2201	138	20632

3.4 Hyperparameters

The hyper-parameters of CNN are chosen based on the performance on the validation set. We applied different filter size (2, 3, 4) separately, and the combination of them with different pooling technique (Max and Average). We find the following parameter that produces the best performance for Ordinal Classification: 3 convolutional layers, filter size (2, 3, 4), the number of filters 100, maximum tweet length of 50, embedding size 200 dimension and average pooling. Since we use multiple convolutional layers, the CNN suffer from overfitting. For that, we utilize dropout [22] which omit a proportion of hidden units randomly in each training iteration. The CNN is trained using Adam for 10 epoch and batch size 100.

³ <http://alt.qcri.org/semeval2016/task4/index.php?id=data-and-tools>.

4 Result

The performance of our model outperforms state of the art on Benchmark Dataset. Table 3 shows the results of the proposed CNN model for MAE^M . Since we use word embeddings that encode the semantic and syntactic but do not express the semantic of the sentiment scale, we do not use any technique that takes the order information of classes. However, the *MLCNN* model yield the lower MAE^M than the benchmark model. Using multiple convolutional layers help to extract more silent features and produce different tweet representation that can increase the expressiveness of ordinal scale. Average pooling contributed on CNN performance. Table 4 shows the impact of pooling strategies.

Max pooling returns the largest value of feature maps and can capture the most significant feature. However, the Max pooling ignore other features values by taking the max value. Hence, it lost the position and intensity information of feature. On the other hand, the average-pooling taking the average value of feature maps that preserve more information and allow all features contribute to the classification. In our model average pooling always outperform the Max pooling.

Table 3. Results of the proposed CNN

Model	Method	MAE^M
TwisE	Logistic regression	0.719
MLCNN	CNN	0.617

Table 4. The impact of pooling strategies

Filter size	MAE^M	
	Max-pooling	Avg-pooling
2	0.768	0.763
3	0.782	0.780
4	0.803	0.772
2, 3	0.735	0.676
3, 4	0.766	0.715
2, 3, 4	0.628	0.617

5 Conclusions

The existing approaches for TSA into five points scale used traditional machine learning algorithms and deep learning approaches. In traditional ML algorithm the work focus on feature selection that requires expert domain knowledge and no guarantee to find the best feature. Also, CNN approaches based on fine tuning for word embedding with one convolutional layer with max pooling and do not exploit other

structure of CNN and pooling techniques. Moreover, the features selection at pooling layer plays a critical role in classification, Max pooling by taking the Max value lead to lose the position of feature and intensity information for the features we cannot know whether the feature occurs multiple time or once.

Therefore, in the proposed *MLCNN* model, we employed three convolutional layers with average pooling that can capture the structure of the problem and detect the sentiment aspect of the word with more expressiveness for ordinal scale. In our experiment, average pooling outperforms the max pooling in the case of ordinal classification problem. We present a novel multi-layer's convolutional neural network (*MLCNN*) model to classify the Tweet into five-point scale. Experimental results show that our model improve the state of the art based on Macro average mean absolute error with 10.2% difference.

Future works will focus on the investigation of other techniques based on the ordinal classification theoretical framework and others pooling strategies that can improve the performance.

References

1. Pang, B., Lee, L.: Opinion Mining And Sentiment Analysis, vol. 1, no. 2 (2008)
2. Kim, Y.: Convolutional Neural Networks for Sentence Classification (2014). [arXiv:1408.5882](https://arxiv.org/abs/1408.5882)
3. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 655–665 (2014)
4. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15), pp. 959–962 (2015)
5. Mohammad, S.M., Kiritchenko, S., Zhu, X.: NRC-Canada: building the state-of-the-art in sentiment analysis of tweets (2013)
6. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, pp. 1320–1326 (2010)
7. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F.: SemEval-2016 Task 4: sentiment analysis in Twitter (2016)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. NIPS, pp. 1–9 (2013)
9. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference On Empirical Methods In Natural Language Processing, pp. 1532–1543 (2014)
10. Balikas, G., Amini, M.: TwiSE at SemEval-2016 Task: Twitter sentiment classification. In: International Workshop on Semantic Evaluation 2016, pp. 85–91 (2016)
11. Lango, M., Brzezinski, D., Stefanowski, J.: PUT at SemEval-2016 Task 4 : The ABC of Twitter sentiment analysis. In: International Workshop on Semantic Evaluation 2016, pp. 131–137 (2016)
12. Esuli, A., Faedo, I.A., Moruzzi, G.: ISTI-CNR at SemEval-2016 Task 4: quantification on an ordinal scale. In: International Workshop on Semantic Evaluation 2016. Accept., pp. 92–95 (2016)

13. Florean, C., Bejenaru, O., Apostol, E., Ciobanu, O., Iftene, A., Trandab, D.: SentimentalTsts at SemEval-2016 Task 4: building a Twitter sentiment analyzer in your backyard, pp. 248–251 (2016)
14. Zhou, Y., Zhang, Z., Lan, M.: ECNU at SemEval-2016 Task 4 : An empirical investigation of traditional NLP features and word embedding features for sentence-level and topic-level sentiment analysis in Twitter. In: International Workshop on Semantic Evaluation 2016, pp. 256–261 (2016)
15. Vilares, D., Doval, Y., Alonso, M.A., Carlos, G.: LyS at SemEval-2016 Task 4 : exploiting neural activation values for twitter sentiment classification and quantification. In: International Workshop on Semantic Evaluation 2016, pp. 79–84 (2016)
16. Stojanovski, D., Strezoski, G., Madjarov, G., Dimitrovski, I.: Finki at SemEval-2016 Task 4: deep learning architecture for Twitter sentiment analysis. In: International Workshop on Semantic Evaluation 2016. Accept., pp. 154–159, (2016)
17. He, Y., Yu, L., Yang, C., Lai, K.R., Liu, W.: YZU-NLP team at SemEval-2016 Task 4: ordinal sentiment classification using a recurrent convolutional network, pp. 256–260 (2016)
18. Ruder, S., Ghaffari, P., Breslin, J.G.: INSIGHT-1 at SemEval-2016 Task 4: convolutional neural networks for sentiment classification and quantification. In: International Workshop on Semantic Evaluation 2016. Accept., pp. 178–182 (2016)
19. Collobert, R., et al.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
20. Yann, L., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015)
21. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning, no. 3, pp. 807–814 (2010)
22. Srivastava, N., Geoffrey, H., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout : a simple way to prevent neural networks from overfitting, vol. 15, pp. 1929–1958 (2014)

Instance-Based Ontology Matching: A Literature Review

Mansir Abubakar, Hazlina Hamdan^(✉), Norwati Mustapha,
and Teh Noranis Mohd Aris

Faculty of Computer Science and Information Technology, University Putra
Malaysia, Seri Kembangan, Malaysia
abubakar.mansir@auk.edu.ng, {hazlina, norwati, noranis}
@upm.edu.my

Abstract. The volume of research articles published today associated to instance-based ontology matching is significant and it is thought to reflect the growing interest of ontology matching research community. Nonetheless, for new researchers in the field of instance-based ontology matching, this amount of information seems to be devastating. Therefore, the aim of this study is to assist researchers and practitioners to get a broad idea on the state-of-the-art instance-based ontology matching and to determine potential research directions in the areas of matching different ontologies in order to represent a single real world object. We performed an intensive literature review in the field of ontology matching, instance-based matching and Semantic Web. Our study shows that there is need for research attention on instance-based matching than usual concentration on conceptual-based matching of two or more ontologies. We also highlighted some important areas that require research attentions.

Keywords: Semantic web · Ontologies · Ontology matching · Instance-based ontology matching

1 Introduction

The exponential growth of web data in terms of its volume, variety, and velocity across organizations necessitated the concern over issues linked to ontology design, schema matching as well as recent interest on ontologies instance level matching. Existing studies have reached a certain point to find solutions to the heterogeneity issues associated with web data with regard to its semantics, popularly known as semantic heterogeneities [1]. These solutions address a quite number of different problems, such as the application of different ontology specification languages, details in describing the domain of interest [2], high level of semi-automation as well as possible matching solutions in both schema and instance level [3].

Today, the high expressiveness of semantic information enables the automatic or semi-automatic processing of web resources. Industries and academia have discovered that Semantic Web can ease the interoperability and integration of both Intra and Inter-business processes [4]. To achieve the effectiveness of information systems, some systems have to undergo reuse via integration performed if not all but in some certain

features of the systems [5]. Many matching systems have been studied in the different areas of computer science such as Artificial Intelligence, database systems, as well as in some areas of information systems [6]. Furthermore, matching algorithms and matching tools becomes essential in the Semantic Web systems framework which did not only concentrate on schema-based matching but also Instance-based matching to support data discovery and management of instances (individuals) representing the same real-world object or entity [7].

This study is aimed to investigate the ontology matching at instance level which will focus on techniques and tools towards the large-scale instance matching approach. This interest has a twofold motivation: (i) associated with distinct studies on the research area to determine its maturity. Thus, Ontology matching is regarded as the joint research field that already has many grounded type of research in existence most especially on schema level [8–10] as well as in [11, 12]. (ii) With respect to schema matching, instance-based matching is considered to be new research area which requires attention and contribution towards its maturity in order to attain better classification of the instance matching techniques [13], matching tools [14] as well as proposed frameworks for large scale matching approaches [15]. Additionally, this study refers to Ontology Alignment Evaluation Initiative (OAEI) instance matching track participants of 2010 through 2016. This track was introduced into OAEI campaign in 2009 to evaluate the performance of instance matching tools.

The rest of the paper is organized as follows: In Sect. 2, the aspect of semantic web, Sect. 3, ontologies and Sect. 4, ontology matching, while in Sect. 5, the main discussion on instance-based ontology matching. A brief discussion about OAEI and some result presentation in Sect. 6. In Sect. 7, an outline of some notable issues associated to instance-based ontology matching and finally, Sect. 8 conclude the paper.

2 Semantic Web

The notion of Semantic Web is simply to define the possible path to enable information on the web to be utilized by computers not just for the display of information, but also for information integration and interoperability between applications and systems. An important way for facilitation of machine-to-machine communication as well as auto-processing has been to provide computer-understandable information.

New languages and standards that provide meaning to the Web information are being studied and developed; things like *Resource Description Framework (RDF)* and the *Web Ontology Language (pronounced OWL) (W3C)*,¹ have emerged. These languages enable a characterization of individual information on the Web also allow for relationships identification between the information sources. Figure 1 depicts the evolution of the Semantic Web mark-up languages.

Furthermore, it enables the development of document sharing facilities that make searching and information reuse simple. According to [3], the presence of ontologies and semantic resources on the web improves the web's interoperability. Semantic Web

¹ <https://www.w3.org/>.

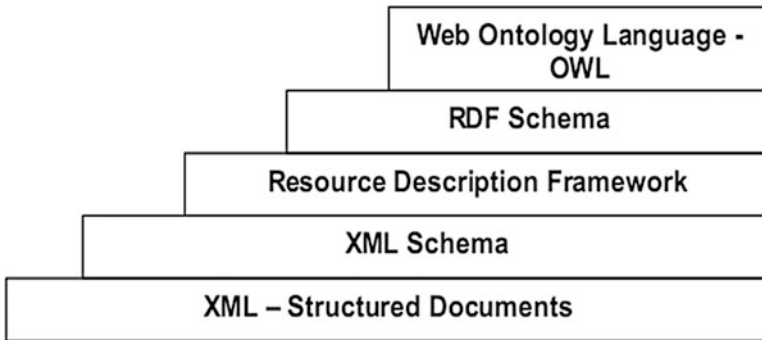


Fig. 1. Markup language pyramid (W3C)

technology has become an emerging domain in the areas of computer science and information communication technology that supports sharing of information across distinct domains of discourse [16]. As a result of the web's flexibility, different people can define Semantic Web in a different domain of interest which may lead to the increase in semantic heterogeneity. Furthermore, good nature of the Web provided an avenue that individuals may publish their online resources when needed thereby increasing the semantic heterogeneity among the web resources.

To achieve mapping, matching and heterogeneous data integration, Semantic Web can be used. The presence of Semantic Web proposed a global avenue of information and knowledge exchange. It can assist to retrieve appropriate information and semantic knowledge discovery effectively. Semantic Web assures machine intelligence which can support various tasks such as search improvement and question answering.

3 Ontologies

The popular and most cited definition of ontology describes ontologies as “*explicit, formal representations of concepts and their relationships within a domain of knowledge*” [2]. The idea is originated from the Greek philosophy; it denotes the study of the existence of things in the world (in Greek *ontos*-to be). It pins down how an object exists in relation to other objects and classes or categories of objects: its study centered in the areas of Metaphysics, like Epistemology (how do we know anything at all), and Semantics (how to communicate the meaning of things) [8].

Ontologies are applied in different areas of studies like medical information modeling, management of a knowledge, geographic information representation, user profile matching, and web mining. We can realize the potential of Semantic Web when programs are created which collect the content of the Web from different sources, this information will be processed and exchanged with another program [17]. Ontology is the knowledge representation technology that describes important concepts in a certain relationship which include: rules, concepts, and properties. The major benefit of ontology is its provision for the organization of information according to own entities and attributes that describe the relationship between concepts and classes. They share the common indulgent structures among software agent and people [18].

To achieve the desired objective of ontologies interoperability, it is obvious that the process involves conflict resolution popularly known as ontology matching. Although in some situations conflict may occur only when the ontologies represent the same real word. However, ontologies conflicts in both same and different domain need to be considered, different level of clashes at *concept level*, *instance level*, and *relation levels*, as well as different approaches toward the conflicts resolution at each of these levels need to be taking into consideration in addressing the matching problem.

4 Ontology Matching

Ontology matching can simply refer as finding correspondences between semantically related objects or entities of distinct ontologies in order to represent one real-world object [19]. Ontology matching provides an environment for data sharing and knowledge expression in different format and languages [20]. The inputs of ontology matching are in most cases two ontologies and produce an alignment as an output of the matching process. Furthermore, expressing entities in ontology matching can be complex enough, like concept definitions, queries, and formulas. Many matching techniques have been proposed across the field of computer science and engineering, such as artificial intelligence, information systems and database systems [21]. Table 1 shows notable techniques applicable to ontology matching.

Table 1. Notable techniques of ontology matching and their methods

Techniques	Method	Source
Record linkage	This method is applied to identify records belonging to the same object from several datasets. Scholars usually influence key and inverse functional property for classifying elements when instances comparison	[22–25]
Trust propagation	This approach provides a framework that deduces the correct and incorrect relations with trust metrics on each user who offers assessed relations	[26–30]
Statistical	This technique find out frequently-appeared attribute correspondences in assessed correct relations. Statistical technique also acts as an auxiliary when other techniques are used to link RDF data sets	[31–33]
Schematic matching	This provides an environment for data sharing and knowledge expression in different format and languages. The technique relies solidly on matching ontologies concepts or schema	[8, 9, 11, 34]
Machine learning	This technique is applied for matching through constructing and improving the mapping pattern classifier. Through learning the assessed links, the mapping pattern can be improved so as to generate more correct relations. The technique of evaluating the mapping patterns is to connect the ontologies with the mapping patterns and compute the Precision, Recall and F-measure of the generated related sets	[31, 35–41]

4.1 Matching Technique Workflow

The general workflow of a matching system involves taking two input ontologies by specifying one as a source and another one as target ontology respectively. Figure 2 shows details on the workflow of matching process. The matcher sub-workflow may work in three different ways; sequential, parallel and the combination of the first two as well.

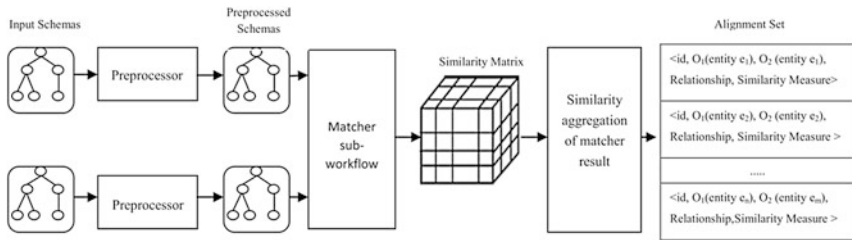


Fig. 2. Matching techniques workflow [51]

5 Instance-Based Ontology Matching

The plain impression about the Instance-based matching is regarded as the high the significant of overlap in similar instances of two objects the higher the relationship of these objects. The issue here is that how one can define the degree of significance of that overlap? Figure 3 depicts the general architecture of instance-based ontology matching.

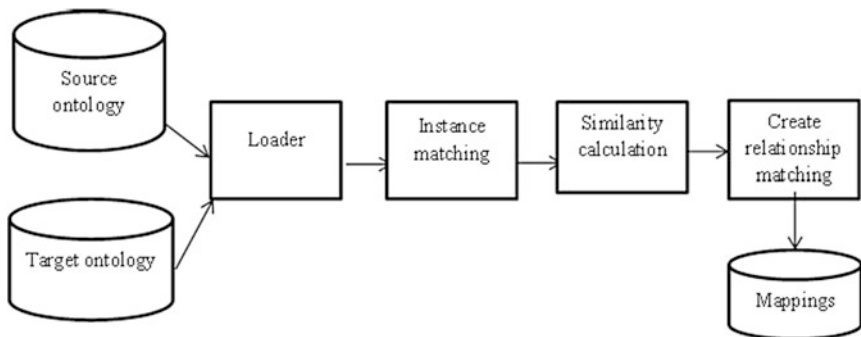


Fig. 3. Instance matching architecture [55]

Ontology Instance matching is a matching which compares two or more sets of individuals of objects or classes so as to decide whether or not they can represent a real-world object. They combine items into a single form and then produce the result as

a mapping alignment. Instance matching is an important aspect of ontology integration as it groups all important points of instances for better interoperability among different information sources [7].

There are many pieces of evidence on why one real-world entity is presented in different sources. In an open and social source of data, anyone has ample right to published data and/or information, and simply adheres to representation that suits his application. Another reason may be as a result of different data acquirement methods. Furthermore, entities are dynamic in nature; they change over time resulting in the frequent update which is often either impossible or found to be hard enough. Lastly, if data is being integrated from multiple sources, the integration process is bound to involve irregular data.

5.1 Instance Matching Techniques

Instance-based matching techniques are widely adapted from the platform of record linkage in databases. Record linkage technique allow for the determination and identification of records of correspondences that corresponds to the single real-world object with a very good precision. Figure 4 depicts an extended classification of similarity-based matching techniques proposed by [14]. These techniques are categorized into two groups: Value-oriented and Record oriented techniques that corresponds to levels of granularity.

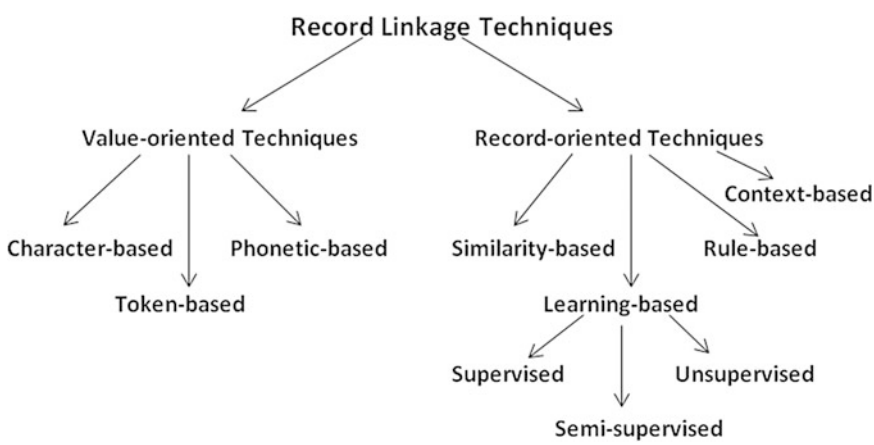


Fig. 4. Extended classification of instance-based matching techniques adopted from record linkage techniques

- i. **Value-oriented techniques:** In these techniques, a similarity of two records can be obtained by matching their comparable attributes under the assumption of granularity value. Furthermore, value-oriented matching mostly concentrated on the calculating string attributes similarity because they are most regularly applied data

types and the source of the real-world description of the object or entity. Identification of a correct threshold is the main setback of similarity-based techniques in which differentiating matching from non-matching records is rational.

- ii. **Record-oriented techniques:** These techniques are regarded as techniques in form of similarity-based. As in value oriented, the similarity value is allocated to every pair record distinct from similarity-based techniques but they created a binary output (value 1 if inputs refer to a particular real-world entity and value 0 if otherwise). The main impression of these approaches is that even with a non-available major feature, a set of attributes can be identified to differentiate each record. These techniques provide a precise alignment; however, they are domain independent and determining good heuristic rules for the choosing domain may be difficult.

5.2 State-of-the-Art Instance-Based Ontology Matching Systems

According to the definition of ontology instance matching, instances are compared among each other within same or different ontology with the goal of identifying the same real-world objects. Yet, numerous kinds challenges in heterogeneity namely value transformation, structural and logical heterogeneity arises in order to weigh the level of similarity pairs of instances across heterogeneous knowledge sources to determine whether they can represent the same real-world entity in a given domain [42].

In [43], an algorithm for ontology matching, *Automated Semantic Matching of Ontologies with Verification (ASMOV)* was developed to handle matching by combining both sets of element-level with structural-level similarity measures with the method formal semantics, to verify the compliance of ontologies correspondences with preferred properties. Formal semantics shows adaptability features of the ASMOV via the use thesaurus API adapter. However, the performance of this approach is limited in the pre-processing of ontologies to demonstrate more information for semantic verification which is part of the comments of Instance Matching (IM) track at OEAI2010 test on this technique. Furthermore, the system could not handle large-scale ontology matching.

To improve the capability of the instance matching to match a lot of instances, [44] proposed a matching technique known as *Logical and Numerical Reference Reconciliation (LN2R)* that focus on the reference reconciliation approaches that are informed. These informed techniques identified knowledge stated in the ontology to reconcile data. The LN2R was tested in an Instance Matching track at OEAI2010 campaign as an unsupervised (linear classifier) knowledge-based, and it is based on two approaches, L2R, and N2R respectively. The main strength of this approach is the ability to minimize comparisons number through its step for filtering which helped to improve the execution time. However, the performance of this technique was affected due to the absence of knowledge of the functionality of properties, it is also less capable of handling large-scale ontologies that involve the application of a large-scale data sets.

Another approach in [45] proposed an efficient matching system that exploits distributed framework to index and process Semantic Web resources as well as employed scalable matching strategies for similarities computation between two

ontologies at their instance-level. This system demonstrated a high sense of robustness and interoperability through its capability to accept large-scale real-word data. However, basic raw data may defect. Moreover, instance matching systems are ever expected to exploit very much information as it could for better discriminability. Therefore, this constrained the performance of the similarity computation algorithm.

In [46], an effort has been made to implement highly scalable/optimized ontology instance matching technique called LogMap. These structures calculated an initial anchor mapping (closely exact correspondences) and allocated confidence value to each of the mappings. The scalable ontology reasoner implemented by this approach alongside a greedy diagnosis algorithm allowed it to locate and repair non-satisfiable classes in the course of matching process. This technique succeeded in presenting a scalable matching tool that matches large-scale ontologies with thousands of classes. However, it has been in an early-stage of scalability, more effort is needed to provide a tool that can scale richer ontologies than in the case of LogMap possibly by applying a technique that can allow the partitioning and clustering of correspondences in order to achieve high scalability/optimization. The approach is similar to the work called Anchor-flood which can measure equivalent one-to-one alignment [47]. In this study, no similarity metrics model of instance incorporated for classification of an instance pairs as whether or not the matching in some instances had performed.

Another approach in [48], proposed a new matching method based on searching Wikipedia pages associated with the ontology terms. The classes extracted in these pages use the graphs organized in matching ontology terms. This work is an extension of AgreementMaker² ontology matching system. A major drawback of this approach is that it allows lot of comparisons among the concepts of both source ontology and target ontology. Therefore, advancement in the scalability/optimizations approach is required to minimize the degree of these comparisons, such as identifying only those instances of the ontologies that need to be aligned.

An instance matching technique that is large-scale in nature called *Large-scale instance matching via multiple indexes (VMI)*, which utilizes many indexes as well as to enable selection of candidate entities was proposed [45]. The objective of this approach is to be able to align large-scale instances of the candidate ontologies with a better matching result. Precisely, this technique adapted *vector space model* to describe information about the instances. This approach escapes pair-wise comparison and improves the matching efficiency by reducing greatly the matching space. The study has limited entity classification capability as its algorithm did not employ meta-classification strategies in the instance matching process.

A *schema-independent linked data interlinking* called *SLINT+* was proposed in order to perform predicate matching [49]. The approach is capable of interlink various data sources together and it is not schema-dependent for a data sources. The aim of the SLINT+ was to perform predicate selection and produce predicate alignment specifying the similar entities of instances. The goal of this method was the implementation of graph matching algorithm with due consideration to the graphical nature of linked data, thereby influencing linking features among instances which resulted to satisfiable

² <http://www.agreementmaker.org>.

matching quality. However, this approach needs to be improved to fully scalable system.

In [50], similarity measures in both lexical and structure is combined in order to compute a distance of information between entities to produce an alignment with a high level of scalability. The study focusses on aligning ontologies that are generally conceptualized. In a situation where entities are not complete, the method may perform poorly, particularly, when similarity measure did not gratify the two unequal conditions.

An approach based on *Information Retrieval (IR) techniques* and an indexing strategy called ServOMap was proposed to address the issue of scalability and efficiency of matching techniques [51]. One of the originalities of the approach is reduction of search space through the use of efficient searching strategy over the built indexes to be matched. The study showed that ServOMap was among the best matching systems that applied the standard benchmarks with matching problems provided by OAEI. It also indicated that the introduction of a background knowledge and ML-based strategy for similarity computing contextually has a positive effect on F-score while increasing the execution time. This approach is deeply based on lexical similarity calculation. This is a serious drawback in such a way that when dealing with ontologies with poor lexical descriptions, this approach may perform poorly in its recall. Moreover, ServOMap is able to provide only equivalence mappings, which is a setback when dealing with some matching, because recall could be affected negatively if reference alignment includes subsumption and disjoint relations.

Automatic instance-based ontology matching systems InsMT and InsMTL was proposed to align ontologies [52]. These systems output alignments of two matched ontologies at instance level which consists of correspondences that are semantical as a whole.

InsMT utilizes string-based algorithms; these algorithms computed the similarities that are represented in a matrix at terminological level. In contrast to InsMT, InsMTL approach computes similarities within instances at both terminological and linguistic (i.e., external resource) level. InsMTL provided a step toward aggregation method by introducing linguistic matcher to achieve scalability, even though they require more scalability enrichment to select the best aggregation and filtering strategy. An improvement over InsMT system which participated in IM@OAEI2014 was presented [53]. This method applies another string-based system with a local filter. This version illustrates better results than the initial version but still did not guarantee scalable performance.

The study conducted in [54] proposed a system to solve the instance matching problem automatically. The system called STRIM work by pointing out all information to be matched from two candidate's ontologies and use the neuro-linguistic programming (NLP) technique to normalize them, and then compute the similarities between the normalized information by applying edit distance matcher. This system appears for the first time in IM@OAEI2015³ evaluation campaign and it presents very good results in both recall, precision, and f-measure which testified its effectiveness and

³ Instance Matching Track at OAEI annual campaign held in 2015.

deficiency. Conversely, more effort needs to be done to achieve high scalability by combining different scalability approaches considering the exponential growth of Web data in terms of volume, variety and velocity (3Vs).

In order to achieve high efficiency in instance matching, another RiMOM family system with an improvement over initial ontology matching system, called RiMOM-IM was proposed [15]. The core notion in its improvement was to exploit the use of different existing matching data and information. This method presents a new blocking algorithm that increases matching efficiency and applies “*exponential function based similarity aggregation*” method for better precision. This system contributes immensely in providing efficient system that can accommodate instance matching task. In large-scale approach, it performs matching iteratively, it select candidate instance pairs by utilizing predicates and different object features, thereby reducing the execution time and avoid tempering the accuracy. Lastly, a “*weighted exponential function based aggregation method “ExpAgg”*” proposed for aggregating the similarity. This system is among the recent instance matching techniques that achieved high

Table 2. Some selected instance-based matching tools (extracted from Sect. 5.2)

Tool	Technique(s)/algorithm	Languages supported	Support large-scale matching?	OAEI participation
ASMOV	Similarity-based context-based	RDF, OWL, UMLS	No	OAEI 2010
LN2R	Unsupervised	RDF, OWL	No	OAEI 2010
ObjectCoref	Semi-supervised	RDF	No	OAEI 2010
SERIMI	Unsupervised/string matching algorithm	RDF	No	OAEI 2010
LogMap	Context-based/Dowling-Gallier algorithm	OWL, UMLS	Yes	OAEI 2011
Anchor-flood	Similarity-based/SOIM	RDF, OWL	No	OAEI 2012
SBUEI	String-based, token-based	RDF, XML		OAEI 2012
IM-VMI	Machine learning	RDF	Yes	OAEI2013
FITON	Machine learning	RDF, OWL	Yes	None
ServOMap	Lexical and machine learning-based contextual similarity	RDF, OWL	Yes	None
InsMT and InsMTL	String-based algorithms	RDF, OWL	Yes	OAEI 2014
STRIM	String-based	RDF	Yes	OAEI 2015
RiMOM-IM	Blocking algorithm	RDF		OAEI 2016

performance with an additional validation process and iterative matching scheme. However, the parameters configuration was performed manually which identified as a drawback of the system and also the system assumes the predicates have already been aligned in its process, therefore incorporating the capability of aligning predicates by the system is also important to empower the system.

Table 2 depicts some popular Instance-based ontology matching tools showing the techniques used, supported languages and other necessary information.

6 Ontology Alignment Evaluation Initiatives (OAEI)

Due to the rapid increase in the amount of methods being use for ontology matching, the need for the evaluation of these methods necessitated the formation of *OAEI* organization. The task of *OAEI* is to coordinate international initiative campaign to confront ontology matching challenges.

With due consideration to the amount of interest on instance based ontology matching, *OAEI* in 2009 introduced ontology instance matching track that aims at evaluating the performance of matching tools which aimed to detect the degree of similarity and overlap between pairs of instances expressed in the form of *OWL*. Figure 5 shows performance of some instance matching track participants of 2010–2016.

F-measure is the weighted harmonic of precision and recall. More specifically presented as:

$$F\text{-measure} = 2(\text{precision})(\text{recall}) / \text{Precision} + \text{Recall} \quad (1)$$

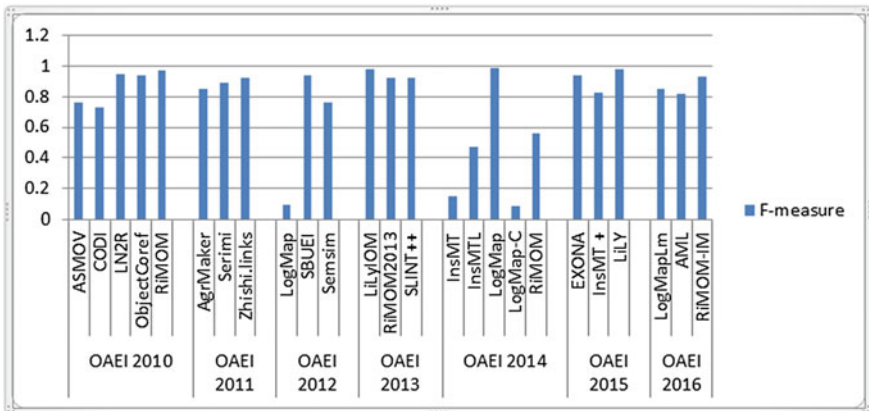


Fig. 5. Best performed instance matching track participants 2010–2016

7 Research Issues in Instance-Based Ontology Matching

Some issues are discovered to be some challenges that hinder the achievement of large-scale ontology instance matching of similar entities between candidates' ontologies. To produce better alignments, more researches need to be carried out on but not only the under-listed research directions with regard to ontology matching at instance level.

1. Instance matching scalability resolution issues
2. Standardized instance matching framework
3. Standardized instance-based matching tools
4. Well-defined matching methodology
5. Automatic parameter configuration in matching ontologies
6. Large-scale ontologies matching evaluation strategy

8 Conclusion

This review studied the current situation about the ontology matching with specific consideration to ontology instance-based matching rather than the popular conceptual (schema) matching. The study involves a review of state-of-the-art ontology instance-based matching; it is techniques as well as some notable existing instance-based matching systems. We have realized that all instance-based matching systems follows the same pattern. In this review, we emphasized on the studies that participated in an instance matching track between 2010 and 2016 annual OAEI competition. The numbers of participants as well as the annual evaluation results obtained by each participant are summarized in order to understand the current state of the performance for further improvements. Finally, some research issues in ontology instance-based matching are identified for further study.

The participant's results and available publications shows that ontology instance-based matching systems demonstrated significant improvement in its performance over the period of six years, even though, more efforts needs to be done in some areas highlighted to achieve better matching performance.

References

1. Maree, M., Belkhatir, M.: Knowledge-based systems addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. *Knowl.-Based Syst.* **73**, 199–211 (2015)
2. Gruber, T.R.: A translation approach to portable ontology specifications. *Knowl. Acquis.* **5**, 199–220 (1993)
3. Warren, P.: Knowledge management and the Semantic Web: from scenario to technology. *IEEE Intell. Syst.* **21**, 53–59 (2006)
4. Abanda, F.H., Tah, J.H.M.: Trends in built environment Semantic Web applications: where are we today? *Expert Syst. Appl.* **40**, 5563–5577 (2013)

5. Hakimpour, F., Geppert, A.: Resolving semantic heterogeneity in schema integration. *Ontol. Inf. Syst.* IOS Press, 297–308 (2001)
6. Shvaiko, P.: A survey of schema-based matching approaches. *J. Data Semant.* **3730**, 146–171 (2005)
7. Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: An empirical study of instance-based ontology matching. In: *Belgian/Netherlands Conference on Artificial Intelligence*, 317–318 (2008)
8. Euzenat, J., Shvaiko, J.P.: *Ontology Matching*, vol. 18, pp. 333. Springer, Heidelberg (2007)
9. Link, S., Nikovski, D., Esenther, A., Ye, X.: Matcher composition methods for automatic schema matching. *Enterp. Inf. Syst.* **141**, 108–123 (2013)
10. Ding, G., Dong, H., Wang, G.: Appearance-order-based schema matching. *Database systems for advanced applications. Lecture Notes in Computer Science*, vol. 7238, pp. 79–94. Springer, Berlin, Heidelberg (2012)
11. Sabbah, T., Selamat, A., M., Ashraf, Herawan, T.: Effect of thesaurus size on schema matching quality. *Knowl.-Based Syst.* **71**, 211–226 (2014)
12. Suresh kumar, G., Zayaraz, G.: Concept relation extraction using Naive Bayes classifier for ontology-based question answering systems. *J. King Saud Univ. Comput. Inf. Sci.* **27**, 13–24 (2015)
13. Castano, S., Ferrara, A., Lorusso, D., Montanelli, S.: On the ontology instance matching problem. In: *Proceedings of International Workshop on Database and Expert Systems Applications, DEXA*, pp. 180–184 (2008)
14. Castano, S., Ferrara, A., Montanelli, S., Varese, G.: Ontology and instance matching. *Lecture Notes in Computer Science (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6050, pp. 167–195 (2011)
15. Shao, C.L., Hu, M., Li, J.Z., Wang, Z.C., Chung, T., Xia, J.B.: RiMOM-IM: a novel iterative framework for instance matching. *J. Comput. Sci. Technol.* **31**, 185–197 (2016)
16. Budura, A., Sebastian, M., Philippe, C.: *European Semantic Web Conference* (2009)
17. Decker, S., et al.: The Semantic Web—on the respective roles of XML and RDF. *IEEE Internet Comput.* **4**, 19 (2000)
18. Nacer, H., Aissani, D.: Semantic Web services: standards, applications, challenges and solutions. *J. Netw. Comput. Appl.* **44**, 134–151 (2014)
19. Akbari, I., Fathian, M.: A novel algorithm for ontology matching. *J. Inf. Sci.* **36**, 324–334 (2010)
20. Horrocks, I.: Description logic: A Formal Foundation for Ontology Languages and Tools. *Methods Cell Biol.* **78**, 765–775 (2007)
21. Zang, B., Li, Y., Xie, W., Chen, Z., Tsai, C.F., Laing, C.: An ontological engineering approach for automating inspection and quarantine at airports. *J. Comput. Syst. Sci.* **74**, 196–210 (2008)
22. Gu, L., Baxter, R.: Record linkage: current practice and future directions. In: *17th International Conference on Database Systems for Advanced Applications*, pp. 03–83 (2003)
23. Freire, S.M., de Almeida, R.T., Cabral, M.D.B., de Assis Bastos, E., Souza, R.C., da Silva, M.G.P.: A record linkage process of a cervical cancer screening database. *Comput. Methods Programs Biomed.* **108**, 90–101 (2012)
24. Ong, T.C., Mannino, M.V., Schilling, L.M., Kahn, M.G.: Improving record linkage performance in the presence of missing linkage data. *J. Biomed. Inform.* **52**, 43–54 (2014)
25. Goldstein, H., Harron, K.: Record linkage: a missing data problem. *J. Methodol. Dev. Data Link.* 109–124 (2016)
26. Liu, X., Wang, Y., Zhu, S., Lin, H.: Combating web spam through trust-distrust propagation with confidence. *Pattern Recognit. Lett.* **34**, 1462–1469 (2013)

27. Wang, X., Su, J., Wang, B., Wang, G., Leung, H.F.: Trust description and propagation system: semantics and axiomatization. *Knowl.-Based Syst.* **90**, 81–91 (2015)
28. Jiang, C., Liu, S., Lin, Z., Zhao, G., Duan, R., Liang, K.: Domain-aware trust network extraction for trust propagation in large-scale heterogeneous trust networks. *Knowl.-Based Syst.* **111**, 237–247 (2016)
29. Wu, J., Xiong, R., Chiclana, F.: Uninorm trust propagation and aggregation methods for group decision making in social network with four tuple information. *Knowl.-Based Syst.* **96**, 29–39 (2016)
30. Xiong, F., Liu, Y., Cheng, J.: Modelling and predicting opinion formation with trust propagation in online social networks. *Commun. Nonlinear Sci. Numer. Simul.* **44**, 513–524 (2017)
31. Goetz, J.N., Brenning, A., Petschko, H., Leopold, P.: Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **81**, 1–11 (2015)
32. Zhuhadar, L.: A synergistic strategy for combining thesaurus-based and corpus-based approaches in building ontology for multilingual search engines. *Comput. Hum. Behav.* **51**, 1107–1115 (2015)
33. Kocbek, S., et al.: Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. *J. Biomed. Inform.* **64**, 158–167 (2016)
34. Gal, A., Roitman, H., Sagi, T.: From diversity-based prediction to better schema matching. In: *International World Wide Web Conference on Communication*, pp. 1145–1155 (2016)
35. Nejehadi, A.H., Shadgar, B., Osareh, A.: Ontology alignment using machine learning techniques. *Int. J. Comput. Sci. Inf. Technol.* **3**, 139–150 (2011)
36. Aher, S.B., Lobo, L.M.R.J.: Combination of machine learning algorithms for recommendation of courses in E-Learning system based on historical data. *Knowl.-Based Syst.* **51**, 1–14 (2013)
37. Wang, S., Li, D., Petrick, N., Sahiner, B., Linguraru, M.G., Summers, R.M.: Optimizing area under the ROC curve using semi-supervised learning. *Pattern Recognit.* **48**, 276–287 (2015)
38. Vock, D.M., et al.: Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *J. Biomed. Inform.* **61**, 119–131 (2016)
39. Ichise, R.: Machine learning approach for ontology mapping using multiple concept similarity measures. In: *Seventh IEEE/ACIS International Conference on Computer and Information Science (ICIS 2008)*, pp. 340–346 (2008)
40. Souza, A.H., Corona, F., Barreto, G.A., Miche, Y., Lendasse, A.: Minimal learning machine: a novel supervised distance-based approach for regression and classification. *Neurocomputing* **164**, 34–44 (2015)
41. Cerón-Figueroa, S., et al.: Instance-based ontology matching for e-learning material using an associative pattern classifier. *Comput. Hum. Behav.* **69**, 218–225 (2017)
42. Gracia, J., Mena, E.: Semantic heterogeneity issues on the web. *IEEE Internet Comput.* **16**, 60–67 (2012)
43. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. *J. Web Semant.* **7**, 235–251 (2009)
44. Sa, F.: LN2R—a knowledge based reference reconciliation system : OAEI 2010 results (2010)
45. Li, J., Wang, Z., Zhang, X., Tang, J.: Large scale instance matching via multiple indexes and candidate selection. *Knowl.-Based Syst.* **50**, 112–120 (2013)

46. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: logic-based and scalable ontology matching. *Lecture Notes in Computer Science (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7031, no.1, pp. 273–288. LNCS (2011)
47. Deb Nath, R., Seddiqui, P.H., Aono, M.: Resolving scalability issue to ontology instance matching in Semantic Web. In: *Proceeding of 15th International Conference on Computer and Information Technology (ICCIT 2012)*, pp. 396–404 (2012)
48. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The agreementmakerlight ontology matching system. *Lecture Notes in Computer Science (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8185, pp. 527–541. LNCS (2013)
49. Nguyen, K., Ichise, R.: SLINT+ results for OAEI 2013 instance matching (2013)
50. Jiang, Y., Wang, X., Zheng, H.: A semantic similarity measure based on information distance for ontology alignment. *Inf. Sci. (Ny)* **278**, 76–87 (2014)
51. Diallo, G.: An effective method of large scale ontology matching. *J. Biomed. Semant.* **5**, 44 (2014)
52. Khiat, A., Benaissa, M.: InsMT/InsMTL results for OAEI 2014 instance matching (2014)
53. Khiat, A., Benaissa, M.: InsMT+ results for OAEI 2015 instance matching, no. 1 (2015)
54. Khiat, A., Benaissa, M., Belfedhal, A.: STRIM results for OAEI 2015 instance matching evaluation. *Ontology alignment evaluation initiative* (2015). <http://oaei.ontologymatching.org>
55. Sai Baba, R. M., Meenachi, M. N., Balasubramanian P.: Instance Based Matching System for Nuclear Ontologies. **4**(1), 10–13 (2016)

Part V

Web Mining, Services and Security (WMSS)

Maximum Attribute Relative Approach of Soft Set Theory in Selecting Cluster Attribute of Electronic Government Data Set

Deden Witasryah Jacob¹(✉), Iwan Tri Riyadi Yanto²,
Mohd Farhan Md Fudzee³, and Mohamad Aizi Salamat³

¹ Department of Industrial Engineering, Telkom University, Bandung, West Java, Indonesia

dedenw@telkomuniversity.ac.id

² Department of Information Systems, University of Ahmad Dahlan, Kampus III UAD, Jalan Prof Dr Soepomo, Yogyakarta, Indonesia

yanto.itr@is.uad.ac.id

³ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia

{farhan,aizi}@uthm.edu.my

Abstract. Electronic government (e-government) is the use of information and communication technology to provide information and services for the citizen. Many researchers argue that it is very important to know what the dominant variable influences citizen in using e-government. A number studies have used empirical approach to know the variables, but very rarely use other technique such as data mining. One of the powerful data mining technique is Maximum Attribute Relative (MAR), the technique is based on a soft set theory by introducing the concept of the attribute relative in information systems. Therefore, we present the applicability of MAR for clustering attribute selection. The real data set is taken from a survey at Bandung District in Indonesia. A total 200 participants have jointed in this survey. Most respondents are female i.e. 105 persons and the rest are male i.e. 95 persons with alpha score yielded 0.846. At this stage of the research, we show how MAR can be used to select the best clustering and found that Facilitating Condition (FC) is the highest variable of the citizen behavior in adopting e-government service. Furthermore, the result also may potentially contribute to decision maker how to design a good e-government in order to reduce bureaucracy and further to improve public services.

Keywords: Clustering · Soft set theory · Maximum attribute relative (MAR) e-Government · Facilitating condition · Effort expectancy · Performance expectancy

1 Introduction

The development that occurred in the field of Information and Communication Technology (ICT) has brought a significant impact on human life. Changes occurred in various fields, which initially used traditional methods or methods to change their service method to ICT-based, including public service sector managed by the government. With e-Government, public services can be done for 24 h, anytime and anywhere without having to meet directly with the officer. This of course provides convenience for people who need certain services quickly without having to waste time by coming directly to the location of the service.

E-Government is the use of information technology by the government to provide information and services for the community [1]. E-government is expected to increase efficiency, comfort, and better accessibility of government services. Carter [2] stated that one important factor for the success of e-government services is the acceptance and willingness of people to use e-government services. Meanwhile, the other scholar [3] stated that with the imperative of e-government for better transparency, accountability and public services, the problem of low-level citizen adoption of e-government services has been recognized in developed and developing countries.

According to Heeks the failure rate of e-government implementation in developing countries reaches 85% [4]. The failure rate, 35% is classified as total failure (e-government is not implemented at all or implemented shortly and then rejected), 50% is classified as a partial failure (unattainable goal or unexpected benefit). Furthermore, Zhao et al. investigated some of the challenges and issues related to the use and development of e-government in Dubai [5]. They encounter problems such as the use of language on the website, lack of integration, low use of e-government services, the quality of existing websites and the utilization of e-services. Meanwhile, Ray examined the challenges in developing e-government services for people in developing countries and found several influencing factors such as lack of awareness, organizational ambiguity, operational relationships, interests and top management support. Taking a lesson from the failure and problems of e-government implementation and challenges, it is important to understand factors of e-government implementation to reduce the risk of failure.

The traditional main objectives of citizen in adopting e-government are to deal with the uncertainty due to design framework, to reduce failure and further to improve awareness citizen in adopting e-government service. To achieve this objective, certain clustering techniques are also being applied. Clustering of data is a method to create a group of objects where each cluster contains data objects that are similar to each other [6, 7]. However, In any case, some approach is not appropriate for dissecting vulnerability as exhibited in the past work [8–10]. Next, the finding should solve the uncertainty and imperfection of the accuracy. One of the solutions is to use MAR approach [11]. We have summarized the main contributions to solve the above-mentioned problems is We apply the MAR based on the soft set theory model to obtain a maximum variable in adopting e-government and elaborate the approach on

experiment tests through e-government data sets. Finally, this work is organized as follows; Sect. 1 presents an introduction, Sect. 2 describes the methodology, furthermore Sect. 3 talks about result and the last section present discussion and conclusion.

2 Related Work

Some research on e-government adoption and satisfaction has been widely used, but research on the adoption or public satisfaction of e-government in developing countries like Indonesia is still very rare. The following are some of the studies that become the reference in this study. The research model was formed from several approaches of several models obtained from previous literature studies. This study uses the basic model of Unified Theory of Acceptance and Use of Technology (UTAUT).

Venkatesh et al. stated that performance expectancy is the most powerful variable affecting one's intention to use information systems. He found that the variable expectancy has a positive relationship with behavioral intention variables in accordance with previous studies [12, 13].

In his work, a few authors stated that the higher a person feels that a system is easy to use and does not require a hard effort to use it the higher the intention of the person to use the system later. This relationship is consistent with previous studies which suggest that effort expectancy is positively associated with behavioral intention [14–18].

The authors suggested that facilitating conditions have a positive effect on the use of behavior, where higher people believe that organizations support them to use e-government then it will increase the use of e-government [19–21]. Meanwhile, to solving the decision-making problem in adopting e-government service Soft set theory can be applied [22–25].

3 Proposed Method

The earlier idea of soft-set is presented in the work of Pawlak [26], where the Pawlak's concept of the soft-set theory is a unified view of the classical set, rough set, and fuzzy set. However, today's soft-set theory is a result of Molodtsov's work [27] where the notion of soft-set theory has been defined. Molodtsov's notion of soft-set theory is a general method for dealing with uncertain that is free from the inadequacy of the parameterization tools. Next subsections describe how soft-set theory was implemented in a group of data set.

3.1 Soft Set Theory

Definition 1 A pair (F, A) is called a soft set over U , where F is a mapping given by $F : A \rightarrow P(U)$. In other words, a soft set (F, A) over U is a parameterized family (subset) of the universe U . For $\alpha \in A$, $F(\alpha)$ may be considered as the set of α elements of the soft set (F, A) or the set α -approximate elements of the soft set (F, A) . Clearly, a soft set is not a (crisp) set.

Example 1 Let a universe $U = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}\}$ be a set of candidates and a set of parameters $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ be a set of soft skills which stand for the parameters “communicative”, “critical thinking”, “teamwork”, “information management”, “entrepreneurship”, “leadership” and “moral”, respectively. Consider F is be a mapping of E into the set of all subsets of the set U as $F(e_1) = \{c_1, c_2, c_4, c_5\}$, $F(e_2) = \{c_3, c_8, c_9\}$, $F(e_3) = \{c_6, c_9, c_{10}\}$, $F(e_4) = \{c_2, c_3, c_4, c_5, c_8\}$, $F(e_5) = \{c_2, c_5, c_6, c_7, c_8, c_9, c_{10}\}$, $F(e_6) = \{c_6, c_9, c_{10}\}$ and $F(e_7) = \{c_6, c_9, c_{10}\}$. Now consider a soft set (F, E) , which describes the “capabilities of the candidate for hire”. Based on the data collected, the soft set (F, E) is following Fig. 1.

$$(F, E) = \left\{ \begin{array}{l} \text{communicative} = \{c_1, c_2, c_4, c_5\}, \\ \text{critical thinking} = \{c_3, c_8, c_9\}, \\ \text{team work} = \{c_6, c_9, c_{10}\}, \\ \text{information management} = \{c_2, c_3, c_4, c_5, c_8\}, \\ \text{entrepreneurship} = \{c_2, c_5, c_6, c_7, c_8, c_9, c_{10}\}, \\ \text{leadership} = \{c_6, c_9, c_{10}\}, \\ \text{moral} = \{c_6, c_9, c_{10}\} \end{array} \right\}$$

Fig. 1. The soft set based on data collected

Obviously, the soft set (F, E) is not a crisp set and (F, E) is a parameterized family $\{F(e_i), i = 1, 2, 3, \dots, 7\}$ of subsets of the set U that have two parts of approximation: predicate (p) and value (v). For example, for the approximation $\text{moral} = \{c_6, c_9, c_{10}\}$, p is moral and “ $v = \{c_6, c_9, c_{10}\}$ ”.

Definition 2 Let $S = (U, A, V_{\{0,1\}}, f)$ be an information system. If $V_a = \{0, 1\}$, for every $a \in A$, then $S = (U, A, V_{\{0,1\}}, f)$ is called a Boolean-Valued information system.

Proposition 1 Each soft set can be considered as a Boolean-valued information system Proof: Let (F, E) be a soft-set over the universe U , $S = (U, A, V, f)$ be an information system. Obviously, the universe U in (F, E) can be considered as the universe U , the parameter set E may be considered as the attributes A . Then, the information function f is defined by

$$f = \begin{cases} 1, & h \in F(e) \\ 0, & h \notin F(e) \end{cases} \quad (1)$$

That is, when $h_i \in F(e_j)$ where $h_i \in U$ and $e_j \in E$, then $f(h_i, e_j) = 1$, otherwise $f(h_i, e_j) = 0$. To this, we have $V(h_i, e_j) = \{0, 1\}$. Therefore, a soft set (F, E) can be considered as a Boolean-valued information system where $S = (U, A, V_{\{0,1\}}, f)$ and a soft set (F, E) can be represented in the form of the Boolean table. From Proposition 1, a soft set in (1) can be easily represented in the Boolean table as follow (Table 1).

Table 1. Tabular representation of soft set (F, E) in [11]

U/E	e ₁	e ₂	e ₃	e ₄	e ₅	e ₆	e ₇
c ₁	1	0	0	0	0	0	0
c ₂	1	1	0	1	1	0	0
c ₃	0	0	0	1	0	0	0
c ₄	1	0	0	1	0	0	0
c ₅	1	0	0	1	1	0	0
c ₆	0	0	1	0	1	1	1
c ₇	0	0	0	0	1	0	0
c ₈	0	1	0	1	1	0	0
c ₉	0	1	1	0	1	1	1
c ₁₀	0	0	1	0	1	1	1

3.2 MAR Algorithm

The function of MAR technique based on the soft set theory is to specify the partition attribute of the given information system [9]. Figure 2 shows the pseudo-code of the MAR algorithm.

```

Algorithm: MAR
Input: data-set of Categorical-valued
Output: A Clustering attribute
Begin
1. Develop the estimation of the multi-soft set
2. Count the Max and Min Support
for i = all categories
  for j = all categories
    intersection = Data(i) And Data(j)
    Sup(i,j) = Intersection/Data(j)
    if Sup (i,j) = 1 then
      MaxSup(i) = MaxSup(i) + Sup(i,j)
    else
      MinSup(i) = MinSup(i) + Sup(i,j)
    End
  end
end
3. Finding Clustering Attribute
if Mode(Max(MaxSup(data(1)..data(n)))) = 1 then
  Clustering Attribute = Max(MaxSup(Data(1)...Data(n)))
else
  Clustering Attribute = Max(MinSup(Data(1)...Data(n)))
end
End
  
```

Fig. 2. The MAR algorithm

3.3 The Studies of E-government Data Set

A survey was conducted in Bandung District Indonesia to know the variables in influencing citizens in using an e-government service. Most respondents are female (103 persons) and the rest are male (97 persons), so a total of respondents is 200 persons. The variables namely (1) Performance Expectancy (PE), (2) Effort Expectancy (EE) and (3) Facilitating Condition (FC). Furthermore, PE is defined as the level of individual belief that using the system will help it to achieve its job performance. Meanwhile, EE is a level of ease of use of the system that will be able to reduce the effort (energy and time) of individuals in doing their work, and the last one, FC refers to a level of confidence in which an individual believes that an organization and a technical infrastructure are available to support the use of the system [12]. Detail description for every variable includes the numbering of indicators can see in Table below (Table 2).

Table 2. Measurement scales and items wording

Variable	Description	Indicator
Performance expectancy	Using e-gov allows me to access more quickly	PE1
	Using e-gov adds to my effectiveness	PE2
	Using e-gov is an efficient way to manage my time	PE3
	Using e-gov makes it easier	PE4
	I can save money if using e-gov to access government	PE5
	E-gov allows me to access (24 h/day, 7 days/week)	PE6
	E-gov provides equal opportunities to the whole community	PE7
Effort expectancy	Learning how to use e-gov is very easy for me	EE1
	E-gov service is very flexible to use	EE2
	I think using e-gov services is easy	EE3
	Working with e-gov is difficult to understand.	EE4
	Using e-gov services takes too much time	EE5
	The length of time I need to learn to use e-gov services is worth the effort	EE6
	It's easy for me to make the e-gov service system do what I want to do	EE7
	For me the e-gov service is very easy to use	EE8
Facilitating condition	I have the resources needed to use e-gov services	FC1
	I have the necessary knowledge to use e-gov services	FC2
	Based on the resources, opportunities, and knowledge needed to use the system, it would be easy for me to use e-gov	FC3
	Guidance is available for using the e-gov system	FC4
	Special instructions are available on e-gov systems	FC5
	There is a helpful officer in case of difficulty accessing the e-gov service	FC6
	I think, using e-gov services is in line with what I want	FC7

Coefficient Cronbach's Alpha is a reliability test used to measure quality a questioner. By looking at the value of alpha obtained, it will be known consistency between indicators used. The standard alpha value used is 0.6, so if the value obtained below 0.6 can be said that the measuring tool is not reliable. From Table 3, it could be seen that all variables have a coefficient value of Cronbach's Alpha large alpha of 0.6. It means that all variables can be said reliable. Furthermore, the collected data is then recapitulated and obtained a descriptive analysis of each variable as the table below (Table 4).

Table 3. The reliability test of questioners

Variable	Coefficient Cronbach's alpha
PE	0.802
EE	0.886
FC	0.850

Descriptive analysis was conducted using SPSS to find the mean and standard deviation for each variable as the table below.

Table 4. Mean and standard deviation of variables

	PE	EE	FC
Mean	3.53	3.05	3.58
Standard deviation	1.02	0.97	0.94

4 Result of the Experiment

In order to apply the MAR technique, the study used MATLAB version 7.6.0.324. The clear stage of the result as follows;

4.1 Performance Expectancy (PE)

The five highest of MAR results are shown in Table 5. The selected attribute is "Using e-government is an efficient way to manage my time." with max support 4 and min support 6.441224. The visualization of the clusters is presented in Fig. 3.

Table 5. MAR results of performance expectancy

Soft Set	Max support	Min support	Categorical rank	Attribute rank
PE13	3	2.370321	2	4
PE14	2	17.07952	3	
PE33	4	6.441224	10	1
PE52	3	5.361692	19	2
PE63	3	2.076141	24	3

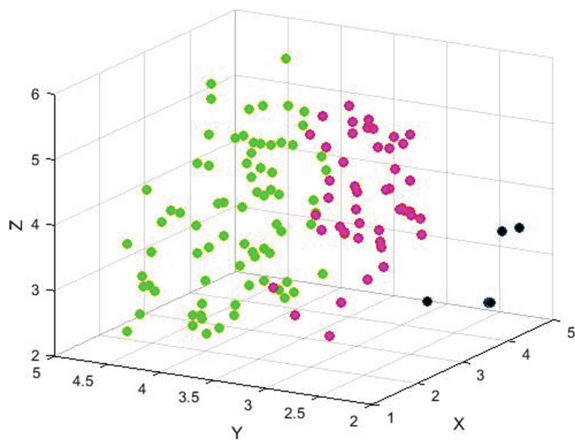


Fig. 3. Cluster visualization of PE data set

4.2 Effort Expectancy (EE)

The Five highest of MAR results are shown in Table 6. The selected attribute is “The length of time I need to learn to use e-government services is worth the effort”, with max support value is 4 and min support value is 15.70014. The visualization of the clusters is captured in Fig. 4.

Table 6. MAR results of effort expectancy

Soft set	Max support	Min support	Categorical rank	Attribute rank
EE 22	4	1.532815	6	3
EE 44	4	14.17021	16	2
EE 64	4	15.70014	24	1
EE 73	3	4.994116	27	5
EE 84	3	15.80228	32	4

4.3 Facilitating Condition (FC)

The five highest of MAR results are shown in Table 7. The selected attribute is “I think, using e-government services is in line with what I want”, with max support value is 4 and min support value is 1.178318. The visualization of the clusters is captured in Fig. 5.

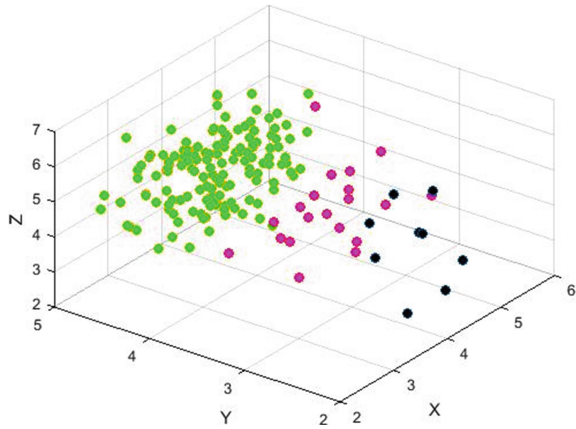


Fig. 4. Cluster visualization of EE data set

Table 7. MAR results of facilitating condition

Soft set	Max support	Min support	Categorical rank	Attribute rank
FC 21	3	1.040197	6	3
FC 35	2	6.321259	14	5
FC 52	3	1.632763	20	2
FC 64	2	12.97404	26	4
FC 71	4	1.178318	28	1

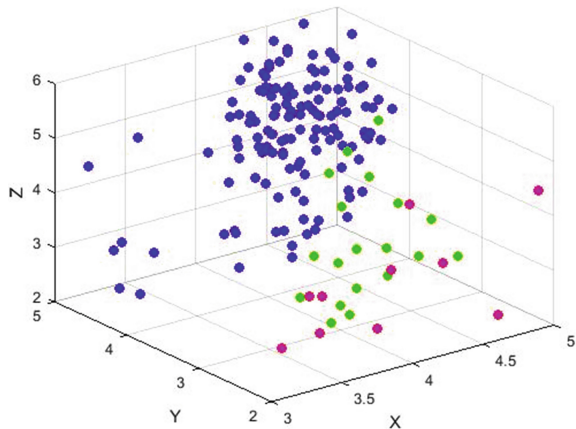


Fig. 5. Cluster visualization of FC data set

5 Discussion

Based on Fig. 6, Facilitating Condition (FC) is the highest variable in influencing the people to use e-government, followed by EE and PE. In summary, MAR has successfully cluster the data-set into their corresponding cluster. A short explanation about how the result founded is on the first run, MAR technique has chosen specific attribute and categorical variable '0' as the partition attribute. As the result, the data set is clustered into two (2) partitions. the first partition contains all results with categorical variable not equals to zero (0), while the second partition contains all results with categorical variable equals to zero (0).



Fig. 6. Clustering result of E-government data set

6 Conclusion

In this paper, the MAR has been used as attribute selection to e-government data set. The concept of attribute relative is based on the technique where the attributes comparison is made by considering the relative of the attribute at the category level. Furthermore, the work integrated the technique approach through three of studies e-government data set among citizens at Bandung District, i.e., performance expectancy, effort expectancy and facilitating condition. The results show that maximum attribute relative can be used to groups' citizens in each study's e-government adoption and can give the recommendation for the stakeholders.

References

1. Deden, W.: The critical factors affecting e-government adoption in Indonesia: a conceptual framework. *Int. J. Adv. Sci. Eng. Inf. Technol.* **7**(1), 160–167 (2017)

2. Carter, L., Weerakkody, V.: E-government adoption: a cultural comparison. *Inf. Syst. Front.* **10**(4), 473–482 (2008)
3. Al-hujran, O., Al-debei, M.M., Chatfield, A., Migdadi, M.: Computers in human behavior the imperative of influencing citizen attitude toward e-government adoption and use. *Comput. Human Behav.* **53**, 189–203 (2015)
4. Heeks, R.: Most e-government-for-development projects fail: how can risks be reduced?. Institute for Development Policy and Management University of Manchester, Manchester (2003)
5. Zhao, F., José Scavarda, A., Waxin, M.-F.: Key issues and challenges in e-government development: An integrative case study of the number one eCity in the Arab world. *Inf. Technol. People* **25**(4), 395–422 (2012). <https://doi.org/10.1108/09593841211278794>
6. Parmar, D., Wu, T., Blackhurst, J.: MMR: An algorithm for clustering categorical data using rough set theory. *Data Knowl. Eng.* **63**, 879–893 (2007)
7. Wang, F.H., and Hung, S.W.: On application of rough data mining methods to automatic construction of student models. In: Cheung, D., Williams, G.J., Li, Q. (eds.) *PAKDD 2001. Lecture Notes On Artificial Intelligence*, pp. 161–166. Springer-Verlag, Berlin, Heidelberg (2001)
8. Jacob, D.W., Fudzee, M.F.M., Salamat, M.A., Saedudin, R., Abdullah, Z., Herawan, T.: Mining significant association rules from on information and system quality of indonesian e-government dataset. In: Herawan, T., Ghazali, R., Nawi, N., Deris, M. (eds.) *Recent Advances on Soft Computing and Data Mining SCDM 2016. Advances in Intelligent Systems and Computing*, vol. 549. Springer, Cham (2017)
9. Jacob D.W., Fudzee M.F.M., Salamat M.A., Saedudin R.R., Yanto I.T.R., Herawan T.: An application of rough set theory for clustering performance expectancy of Indonesian e-government dataset. In: Herawan, T., Ghazali, R., Nawi, N., Deris, M. (eds.) *Recent Advances on Soft Computing and Data Mining SCDM 2016. Advances in Intelligent Systems and Computing*, vol. 549. Springer, Cham (2017)
10. Yanto, I.T.R., Vitasari, P., Herawan, T., Deris, M.M.: Applying variable precision rough set model for clustering student suffering studys anxiety. *Expert Syst. Appl.* **39**(1), 452–459 (2012)
11. Mamat, R., Herawan, T., Deris, M.M.: MAR: maximum attribute relative of soft set for clustering attribute selection. *Knowl.-Based Syst.* **52**, 11–20 (2013)
12. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. *MIS Q.* **27**(3), 425–478 (2003)
13. Weerakkody, V., El-Haddadeh, R., et al.: Examining the influence of intermediaries in facilitating e-government adoption: An empirical investigation. *Int. J. Inf. Manage.* **33**(5), 716–725 (2013)
14. Lozanova-belcheva, E.: The Impact of Information Literacy Education for the Use of e-Government Services : e-government Services Usage—Reasons Why Citizens Are, pp. 155–161. (2013). https://doi.org/10.1007/978-3-319-03919-0_19
15. Williams, M.D., Rana, N.P., Dwivedi, Y.K.: The unified theory of acceptance and use of technology (UTAUT): a literature review. *J. Enterp. Inf. Manag.* **28**(3), 443–488 (2015)
16. Rana, N.P., Dwivedi, Y.K., Williams, M.D.: A meta-analysis of existing research on citizen adoption of e-government. *Inf. Syst. Front.* **17**(3), 547–563 (2015). <https://doi.org/10.1007/s10796-013-9431-z>
17. Jacob, D.W., Md Fudzee, M.F., Salamat, M.A., Kasim, S., Mahdin, H., Ramli, A.A.: Modelling end-user of electronic-government service: the role of information quality, system quality and trust. *IOP Conf. Ser. Mater. Sci. Eng.* **226**, 12096 (2017)

18. Witarsyah, D., Fudzee, M.F., Salamat.: A conceptual study on generic end users adoption of e-government services. *Int. J. Adv. Sci. Eng. Inf. Technol.* **7**(3), 1000–1006 (2017) [Online] Available: <http://dx.doi.org/10.18517/ijaseit.7.3.1654>
19. Athmay, A.L., Al, A.A., Fantazy, K., Kumar, V.: E-government adoption and users satisfaction: an empirical investigation. *EuroMed J. Bus.* **11**(1), 57–83 (2016). <https://doi.org/10.1108/EMJB-05-2014-0016>
20. Ghalandari, K.: The effect of performance expectancy, effort expectancy, social influence and facilitating conditions on acceptance of e-banking services in Iran: the moderating role of age and gender. *Middle-East J. Sci. Res.* **12**(6), 801–807 (2012). <https://doi.org/10.5829/idosi.mejsr.2012.12.6.2536>
21. Tarhini, A., El-Masri, M., Ali, M., Serrano, A.: Extending the UTAUT model to understand the customers acceptance and use of internet banking in Lebanon. *Inf. Technol. People* **29**(4), 830–849 (2016). <https://doi.org/10.1108/ITP-02-2014-0034>
22. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. *Comput. Math Appl.* **44**, 1077–1083 (2002)
23. Ma, X., Norrozila, S., Qin, H., Herawan, T., Zain, J.M.: A new efficient normal parameter reduction algorithm of soft sets. *Comput. Math. Appl.* **62**, 588–598 (2011)
24. Feng, F., Jun, Y.B., Liu, X.Y., Li, L.F.: An adjustable approach to fuzzy soft set based decision making. *J. Comput. Appl. Math.* **234**, 10–20 (2010)
25. Feng, F., Li, Y.M., Cagman, N.: Generalized uni-int decision making schemes based on choice value soft sets. *Eur. J. Oper. Res.* **220**(1), 162–170 (2012)
26. Pawlak, Z.: Rough sets. *Int. J. Comput. Inf. Sci.* **11**, 341–356 (1982)
27. Molodtsov, D.: Soft set theory—first results. *Comput. Math. Appl.* **37**(4), 19–31 (1999)

Android Malware Detection Based on Network Traffic Using Decision Tree Algorithm

Aqil Zulkifli¹, Isredza Rahmi A. Hamid^{1(✉)}, Wahidah Md Shah²,
and Zubaile Abdullah¹

¹ Information Security Interest Group (ISIG), Faculty of Computer Science
and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Johor,
Malaysia

aqilzulkifli720@gmail.com, {rahmi, zubaile}@uthm.edu.my

² Faculty of Information Technology and Communication, Universiti Teknikal
Malaysia Melaka, 76100 Durian Tunggal, Malaysia
wahidah@utem.edu.my

Abstract. Android mobile operating system has well developed and gained absolute popularity among user. Although android is an open source operating system, it fits user daily life requirement nowadays. However, this is the reason why android malware keep on increasing every year. There are various method used to detect the occurrence of android malware such as based on static or dynamic analysis. Static analysis is favourable approach because it is quick and inexpensive. However, the static analysis unable to monitor the malicious application behavior during runtime. Therefore, we proposed a dynamic detection technique based on network traffic which records the application behavior during runtime. We consider seven network traffic features extracted from Drebin and Contagiodumpset dataset. The Drebin dataset achieved higher accuracy value with 98.4% as compared to Contagiodumpset dataset when tested using J48 decision tree algorithm.

Keywords: Android · Malware · Decision tree algorithm

1 Introduction

Android was created on 2007 and become one of the most popular operating system on mobile communication platform. Android operating system has modern technology features which can ease human daily activities such as online banking, online shopping, social media and other utilities applications. Since android has been used to store human private information, there are many android malware created to exploit user's information. More than 100 kind of malware [1] created with functionality such as Botnet, gain root access, download malicious program through third party market, send

Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

malicious Short Messaging Service (SMS), steals location information and also act as a Trojan. Some malware send users information to remote server and remotely controlled by the server. The objectives of this research are as follows:

- (a) To design an android malware detection model based on dynamic features.
- (b) To detect android malware based on network traffic features using Decision Tree Algorithm.
- (c) To evaluate the performance metric in network traffic features tested on J48 Decision Tree Algorithm based on Accuracy rate, True Negative (TN), True Positive (TP), Error rate and Receiver Operating Characteristic (ROC).

The rest of the paper is organized as follows: Sect. 2 will discuss about related works of android malware detection. Then, Sect. 3 discussed about detection methodology which has been used in this paper. Section 4 will discuss about the result from the experiment and finally, Sect. 5 will discuss about the conclusion and future work.

2 Related Works

Android malware is capable of selling user information, stealing user credentials, making premium-rate calls and SMS, SMS spam, search engine optimization and ransom. Most of analyzed android malware dataset shows that the common malicious applications collected user information and send premium rate SMS message. Zhou and Jiang [2] have more than 1,200 datasets of Android malware found in the Google Play market and other third party markets. Based on the analysis, 1,200 of android malware dataset have the android malware features as stated in Table 1. In general, there are two basic ways to analyse android malware which are static and dynamic analysis.

Table 1. Android malware feature

Feature	Description
Privilege escalation	Application gains higher privileges on the phone than necessary to perform otherwise unauthorized actions
Remote control	Allows application to control the device without informing the user
Financial charge	Application uses services that charge the users without their knowledge such as send message and outgoing call
Information collection	Collect information in user's phone such as contact numbers and other account information

2.1 Static Analysis

Static analysis provides a method to analyse the application without installation. In static analysis, there is no execution of the applications where only code and other components like manifest file are analyzed. Therefore, it is a quick and inexpensive

approach. However, the static analysis unable to monitor the malicious application behavior on the run time process since there is no executable process occurs [3].

Shabtai et al. [4] conducted a research to classify Android applications through static analysis. The objectives of the experiments is to classify application such as tools, games and other normal applications that does not contain android malware. They looked at the different elements in Android Manifest XML files, the classes.dex file and aspects of the .apk file, including file size and methods. To classify the applications, they utilized classifiers such as Decision Trees, Naive Bayes, Bayesian networks, Partial Decision Trees (PART), boosted Naive Bayes (NB), boosted Decision Trees (DT), Random Forest, and Voting Feature Intervals.

Wu et al. [5] built the framework called DroidMat to detect malware on Android. Droidmat looked at elements in the AndroidManifest.xml file and API calls to classify the applications. DroidMat used 238 malicious applications from Contagiodumpset Mobile and 1,500 applications from the Android Market. By using k-means algorithm, the framework was able to classify datasets with 97.87% accuracy rate.

Suarez-Tangil et al. [6] is a detection framework for android malware based on static analysis. This framework is fast, accurate and resilient to obfuscation. First, DroidSieve decide either the application is malware or not. Then, it will classify the application into related malware family that have obfuscation technique. DroidSieve exploits the obfuscation-invariant features and artifacts introduced by obfuscation mechanisms used in malware. The accuracy of this framework is 99.82% with zero false positives and 99.26% for obfuscated malware.

2.2 Dynamic Analysis

Dynamic analysis is a detection technique that evaluate malware by executing the application in a real environment. The main advantage of this technique is, it detects dynamic code loading and records the application behavior during runtime. Afonso et al. [7] introduced a dynamic analysis technique which records the frequency of system calls and Application Program Interface (API) calls to detect the malware and normal applications. However, this technique has its own drawback that it only detect the malware if the application meets certain API level.

Andronamaly [8] used 88 dynamic features including memory page activity, SMS message events, Central Processing Unit (CPU) usage, network usage, touch screen pressure, binder information and battery information. They evaluated different combinations of features using Information Gain and Fisher scores. Then, features with the best scores were selected. Next, they applied several classifiers including Decision Trees, Naive Bayes, Bayes Nets, Histograms, K-Means and logistic regression. Four artificial malware were used and the best configuration was able to achieve approximately 88% accuracy.

Credroid [9] used network traffic as their analysis features. It identifies malicious application on the basis of their Domain Name Server (DNS) queries. They also analysed network traffic logs of the data transmits to remote server in offline mode. These semi-automated approaches have observed that 63% of the malware applications generated network traffic. Our study differs from the previous dynamic analysis approach [7–9] in several ways. We propose a dynamic analysis using network traffic

features extracted from Drebin and Contagiodumpset dataset that are Average Packet Size, Average Number of Packets Sent per Flow, Average Number of Packets Received per Flow, Average Number of Bytes Sent per Flow, Average Number of Bytes Received per Flow, Ratio of Incoming to Outgoing Bytes and Average Number of Bytes Received per Second. Then, we use J48 Decision tree algorithm as our classifier.

3 Android Malware Detection Model

Figure 1 shows the proposed android malware detection model. The explanations about every step in the framework are as follows.

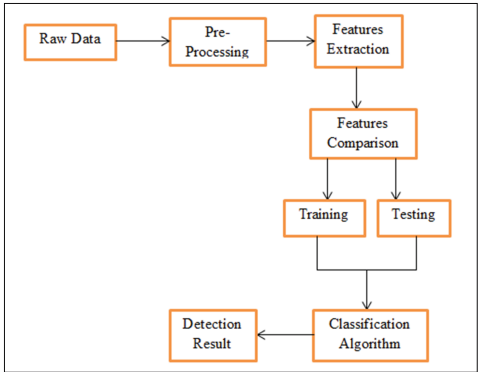


Fig. 1. Proposed android malware detection model

3.1 Raw Data

Raw data is known as original data which means that data is not been processed for use. This raw data will undergo several process such as selective extraction, organization, analysis and formatting before being used to another step or process in the framework. In this research, the malware datasets are taken from Drebin and Contagiodumpset websites. Whilst the normal datasets such as gaming application, dictionary and calculator are collected as well. A total of 700 datasets have been analysed which contains 500 normal datasets and 200 malware datasets from Drebin and Contagiodumpset.

3.2 Pre-processing

The pre-processing is used for transforming raw data into a readable format. Most of the raw data is incomplete, inconsistent, have noise, error and lack in some behavior. Hence, the data pre-processing is introduced to encounter this problem by performing the pre-processing task as shown in Table 2.

Table 2. Pre-processing task

Pre-processing	Description
Data cleaning	Smoothing the noisy data, fill in missing values and resolve inconsistencies in the data
Data integration	Resolved the conflicts within the data by putting the data with different representations
Data transformation	Data is normalized, generalized and aggregated
Data reduction	Reduced the representation of data but producing the same analytical result
Data discretization	The range of attribute interval is divided to reduce the number of values of a continuous attribute

3.3 Feature Extraction and Feature Comparison

After the data pre-processing, the clean datasets is executed on actual Android mobile phones running on Android version 6.0 (Android Marshmallow). A packet capture application named tpacketcapture was run during the execution to capture the network traffic generated by the datasets. The captured file was saved in .pcap format. Then, the behavior of the datasets will be analysed by using wireshark software.

In features comparison process, the network traffic features of malware and normal dataset are compared. Next, the range value for each network traffic features as shown in Table 3 is obtained and saved in .xlms format. Finally, the dataset is converted to .arff file format in order to classify it using machine learning algorithm that run in Waikato Environment for Knowledge Analysis (WEKA) tools.

Table 3. List of traffic features [10]

Traffic features	Description
Average packet size	The average number of packet size for the whole runtime session of the datasets
Average number of packet sent per flow	The average number of packet sent per flow between the host and client during runtime
Average number of packets received per flow	The average number of packet receives per flow between the host and client during runtime
Average number of bytes sent per flow	The average number of bytes of the packet sent per flow during runtime session of the datasets
Average number of bytes received per flow	The average number of bytes of the packet receives per flow during runtime session of the datasets
Ratio of incoming to outgoing bytes	Ratio of incoming to outgoing bytes of the packet during the runtime session
Average number of bytes received per second	The average number of bytes of the packet received per second during the runtime session

3.4 Training and Testing

Machine learning deals with algorithm which learned from datasets. Training data set is for machine learning algorithm to perform correlational task such as classifying, clustering and learning the attributes. Testing data set is a set of data for testing the machine learning algorithm after it learned from the training data set. In testing data set, the outcome is already known and from that the accuracy of machine learning and others calculation can be performed. In this research, the dataset is split into 60:40 ratios that is 60% for training data, while 40% is used for testing data. The training and testing data is tested on J48 decision tree algorithm using WEKA tools.

3.5 Classification Algorithm

Basically, Decision Tree is a form of tree structure. It breaks down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Based on the simple decision rules inferred from the tree, the goal of decision tree algorithm is to construct a model that can predict the value of testing application. We selected J48 as the classification algorithm because this algorithm constructed and trained the decision tree fast and easy. Moreover, the J48 decision tree algorithm can predict faster based on the nodes of the tree and manage to handle irrelevant attributes easily. The J48 Decision tree algorithm [10] is as shown in Fig. 2.

```

Input: training dataset set T, the collection of candidate
attribute attribute_list
Output: a decision tree.
1. Create a root node N;
2. If T belongs to the same category C, then return N as a leaf
   node, and mark it as class C;
3. If attribute_list is empty or the remainder datasets of T is
   less than a given value, then return N as a leaf node, and mark
   it as the category which appears most frequently in
   attribute_list, for each attribute, calculate its information
   gain ratio
4. Suppose test_attribute is the testing attribute of N, then
   test_attribute= the attribute which has the highest information
   gain ratio in attribute list:
5. If testing attribute is continuous, then find its division
   threshold;
6. For each new leaf node grown by node N
   {
   a) Suppose T is the dataset subset corresponding to the leaf
      node.
   b) If T has only a decision category, then mark the leaf node
      as this category;
   c) Else continue to implement J48_Tree (T', T'_attributelist)
   }
7. Calculate the classification error rate of each node and then
   prune the tree.

```

Fig. 2. J48 decision tree algorithm

4 Experimental Setup

The experimental setup started by collecting the dataset from Drebin and Contagiodumpset. For normal dataset it was downloaded from Google Playstore. Then, all datasets are converted into .apk format and run one by one on an actual mobile phones. During its runtime, we run the network traffic capture application (tcpdump) to capture the network traffic file of each datasets and saved in .pcap format. These pcap files is analysed on Wireshark software to extract the network traffic features value. Then, the extracted value of network traffic features are saved in .xlsx format.

Most of the dataset have different range of value corresponding to their traffic features as shown in Table 4. The malware datasets from Drebin and Contagiodumpset shows that the malware network traffic features values is lesser than the normal network traffic features. Data cleaning is performed to combine all extracted features that contain malware and normal dataset. After that, the dataset is divided into training and testing dataset with 60:40 ratios and saved into .csv format. Each training and testing dataset consist of malware and normal datasets as shown in Table 5. Next, both dataset (Drebin and Contagiodumpset) are classified using J48 Decision Tree algorithm. Finally, the performance for both dataset tested on machine learning algorithm is obtained and compared.

Table 4. Network traffic features with extracted value of malware and normal datasets

Traffic features	Malware traffic	Normal traffic
Average packet size, bytes	90–40	1000–20000
Average number of packets sent per flow	0–15	20–100
Average number of packets received per flow	10–30	20–150
Average number of bytes sent per flow	500–2000	5000–50000
Average number of bytes received per flow	500–10000	10000–50000
Ratio of incoming to outgoing bytes	0.5–10	20–50
Average number of bytes received per second	15–2000	5000–20000

Table 5. The number of training and testing dataset

Dataset	Drebin	Contagiodumpset	Normal	Total
Training	75	75	300	450
Testing	25	25	200	250

4.1 Performance Metric

In order to measure the effectiveness of the proposed network traffic features using J48 Decision Tree Algorithm, we used five performance metrics. These metrics are:

- (1) Accuracy (Acc): How many android malware are correctly predicted by the classification algorithm?
- (2) True Negatives (TN): How many normal application is classified as normal?

- (3) True Positives (FP): How many android malware is classified as malware?
- (4) Error rate (Err rate): How many android malware are wrongly predicted by the classification algorithm?
- (5) Receiver Operating Characteristic (ROC): is where the classification made and algorithm used can be determine it certainty. The best result of ROC value should be close to one.

4.2 Result and Discussion

This section discussed result of the experiment and performance metric of the detection model for Drebin and Contagiodumpset dataset tested on J48 Decision Tree algorithm.

4.2.1 Accuracy

Figure 3 shows the accuracy result between Drebin and Contagiodumpset. Contagiodumpset dataset has lower accuracy than Drebin dataset tested on J48 Decision Tree algorithm. The accuracy of Contagiodumpset dataset is 97.6% while the accuracy of Drebin Dataset is 98.4%. This shows that Drebin dataset network traffic feature detects android malware more accurate than Contagiodumpset dataset network traffic feature when tested using J48 Decision Tree algorithm.

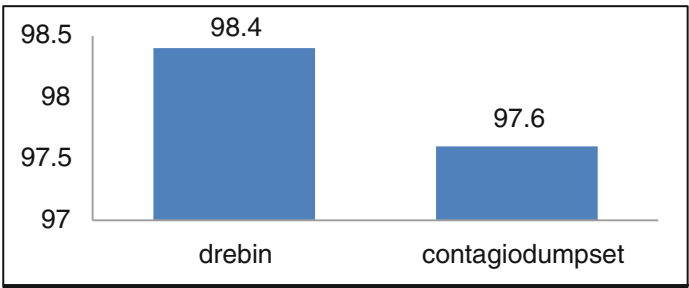


Fig. 3. Accuracy value for Drebin and Contagiodumpset dataset

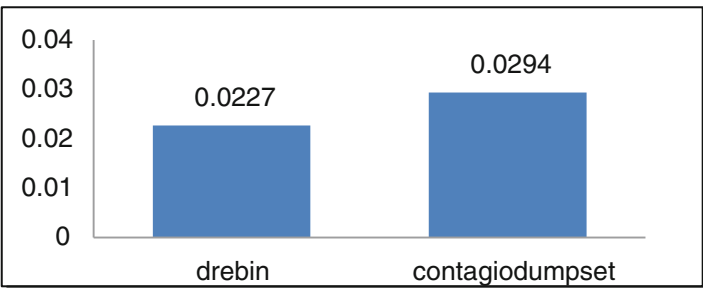


Fig. 4. Mean absolute error value for Drebin and Contagiodumpset dataset

4.2.2 Mean Absolute Error

Figure 4 shows the mean absolute errors for both datasets. Drebin dataset have lower mean absolute error value than Contagiodumpset dataset with 0.0227 and 0.0294 respectively tested on J48 Decision Tree algorithm. This shows that Drebin datasets predicted correctly with its outcome as compared to Contagiodumpset dataset.

4.2.3 Receiver Operating Characteristic (ROC)

Figure 5 shows that the Drebin dataset achieved the best ROC value as compared to Contagiodumpset dataset with 0.954 and 0.932 respectively.

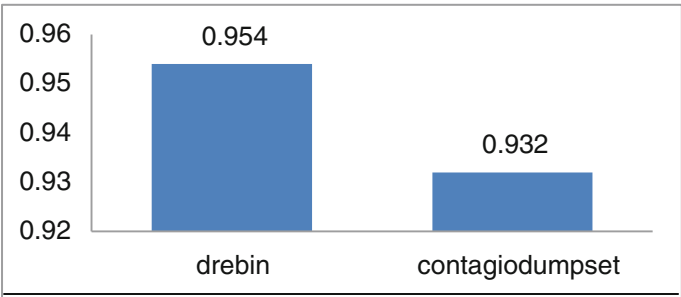


Fig. 5. ROC value for Drebin and Contagiodumpset

4.2.4 True Positive (TP) and False Positive (FP) Rate

Table 6 shows the TP rate and FP rate for Drebin and Contagiodumpset dataset tested on J48 Decision Tree algorithm. Overall, both datasets achieved same value for TP and FP with 0.92 and 0.80 respectively.

Table 6. TP and FP of Drebin and Contagiodumpset

Drebin		Contagiodumpset	
TP	FP	TP	FP
0.920	0.80	0.920	0.080

5 Conclusions

Generally, there are two basic ways to analyse android malware which are static and dynamic analysis. Static analysis gives on what the malware is coded for. Moreover, static analysis is limited to source code and permission analysis of the malware since the actual behaviour of the malware is not observed. This is because the malware dataset does not run on actual mobile phones or any android emulator. Some have evasion technique which can hide their malicious code and download it during runtime. This limitation gives an advantage to dynamic analysis since it analyse android malware during its runtime which based on specific features used by the researcher.

Therefore, we utilized the network traffic features extracted from Drebin and Contagiodumpset for malware dataset and Google Playstore for the normal dataset. Then, seven network traffic features were used which will be divided into training and testing dataset with 60:40 ratio. These features are tested on J48 Decision Tree algorithm run on WEKA tool. The performance metric of the detection model is analysed based on its accuracy, mean absolute error, ROC and TP and FP rate. Overall, the Drebin dataset give higher accuracy than Contagiodumpset with 98.4%.

Acknowledgements. The authors express appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM). This research is supported by Short Term Grant vot number U653 and Gates IT Solution Sdn. Bhd. under its publication scheme.

References

1. Bisson, D.: Trojan found in more than 100 Android apps on Google Play Store. Cluley Associates Ltd. <https://www.grahamcluley.com/advertising-trojan-100-android-apps-google-play-store/>
2. Nazish: Dissecting android malware: characterization and evolution summarized by: Nazish Asad. 4 (2011)
3. Burguera, I., Zurutuza, U., Nadjm-Tehrani, S.: Crowdroid: behavior-based malware detection system for android. In: Proceedings of 1st ACM Work. Security and Privacy in Smartphones and Mobile Devices—SPSM 2011, p. 15 (2011)
4. Shabtai, A., Fledel, Y., Elovici, Y.: Automated static code analysis for classifying android applications using machine learning. In: Proceedings of the 2010 International Conference on Computational Intelligence and Security, pp. 329–333 (2010)
5. Wu, D.J., Mao, C.H., Wei, T.E., Lee, H.M., Wu, K.P.: DroidMat: Android malware detection through manifest and API calls tracing. In: Proceedings of 2012 7th Asia Jt. Conference Information Security Asia JCIS 2012, pp. 62–69 (2012)
6. Suarez-Tangil, G., Dash, S.K., Ahmadi, M., Kinder, J., Giacinto, G., Cavallaro, L.: DroidSieve: fast and accurate classification of obfuscated android malware. In: Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy (CODASPY 2017), pp. 309–320 (2017)
7. Afonso, V.M., de Amorim, M.F., Grégio, A.R.A., Junquera, G.B., de Geus, P.L.: Identifying Android malware using dynamically obtained features. *J. Comput. Virol. Hacking Tech.* **11** (1), 9–17 (2015)
8. Shabtai, A., Kanonov, U., Elovici, Y., Glezer, C., Weiss, Y.: Andromaly: A behavioral malware detection framework for android devices. *J. Intell. Inf. Syst.* **38**(1), 161–190 (2012)
9. Malik, J., Kaushal, R.: CREDROID: Android malware detection by network traffic analysis. In: Proceedings of the 1st ACM Workshop on Privacy-Aware Mobile Computing (PAMCO 2016), pp. 28–36 (2016)
10. Sharma, D.: Android malware detection using decision trees and network traffic. **7**(4), 1970–1974 (2016)

An Improved Low Contrast Image in Normalization Process for Iris Recognition System

Abdulrahman Aminu Ghali^(✉), Sapiee Jamel,
Kamaruddin Malik Mohamad, Shamsul Kamal Ahmad Khalid,
Zahraddeen Abubakar Pindar, and Mustafa Mat Deris

Faculty of Computer Science and Information Technology, Universiti Tun
Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
aminuabdulrahman81@yahoo.com, {sapiee, malik, shamsulk,
mmustafa}@uthm.edu.my, deenpindar@gmail.com

Abstract. Iris recognition system is one of the most predominant methods used for personal identification in the modern days. Low quality iris image such as low contrast and poor illumination presents a setback for iris recognition as the acceptance or rejection rates of verified user depend solely on the image quality. This paper presents a new method for improving histogram equalization technique to obtained high contrast in normalization process thereby reducing False Rejection Rate (FRR) and False Acceptance Rate (FAR). The proposed technique is developed using C++ and tested using four datasets CASIA, UBIRIS, MMU and ICE 2005. The experimental results show that the proposed technique has an accuracy of 95%, as compared to the existing techniques: CLAHE, AHE, MAHE and HE which have an accuracy of a 93.0, 85.7, 92.8 and 90.71% respectively. Hence it can be concluded that the proposed technique is a better enhancement technique compared to the existing techniques for image enhancement.

Keywords: Iris recognition · Histogram equalization · Image enhancement
Normalization

1 Introduction

Human identification using biometric technologies has attracted more attention in security applications. The security applications are access control, forensic, border control and banking. Among various biometric technologies, such as iris, fingerprints, face, ears, retina and hand geometry [1]. Iris recognition system is considered the most high reliability for personal identification [2, 3]. The uniqueness of the iris pattern remains unchanged despite the aging process, and if modified may affect one's health

[4]. Essentially, in iris recognition system low quality iris image such as poor illumination and low contrast present a big issue [5]. These two weaknesses indirectly increase false rejection rate (FRR) and false acceptance rate (FAR) for an image, thereby minimizing the performance of iris system. Shivakumara et al. [6] defined low contrast as a dim of intensity value. While high contrast refers to sharpness of the iris image.

Essentially, iris recognition systems consist of four different stages namely: segmentation, normalization, feature extraction and matching [3]. Segmentation is to detect inner and outer boundary of iris circle. Normalization is used to standardize the iris size inconsistencies. Feature extraction is to extract significant features from normalized iris image. Matching stage compares between two templates. A user is accepted if the template matches and rejected if the templates do not match [7]. The histogram equalization (HE) is a prominent technique used for enhancing image's contrast by transforming the image into uniform histogram. The challenge of the technique reduces the local details of an image [5].

This paper proposes an enhanced histogram equalization technique to obtained high contrast in normalization stage, thereby addressing poor illumination and low contrast of an iris image which automatically reduces (FRR) and (FAR) in an iris system. The proposed technique is validated using four datasets: Chinese Academy of Sciences Institute of Automation (CASIA), Unconstrained Biometric Iris (UBIRIS), Multimedia University (MMU) and Iris Challenge Evaluation (ICE 2005). The results obtained were compared with different enhancement techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE), Adaptive Histogram Equalization (AHE), Multi-scale Adaptive Histogram Equalization (MAHE) and Histogram Equalization (HE).

The paper is organized as follows: Sect. 2 related work. Section 3 describes the image enhancements process. Section 4 discusses experimental results. Lastly, Sect. 5 covers the conclusion of this paper.

2 Related Work

Many techniques for enhancing image quality have been proposed in iris recognition system. These techniques can be classified as Adaptive Histogram Equalization (AHE), Multi-scale Adoptive Histogram Equalization (MAHE), Dynamic Histogram Equalization Technique (DHE), Contrast Limited Adaptive Histogram Equalization (CLAHE), Brightness Preserving Bi-histogram Equalization (BBHE), Histogram Equalization (HE) Dualistic Sub-image Histogram Equalization (DSIHE) etcetera [5]. The iris of an individual is captured in different size due to poor illumination and low contrast of the iris image. However, the occurring problem affects the performance (accuracy) of the iris system. In addition, to achieve more accuracy in recognition result

it is necessary to address these problems: iris size, poor illumination and low contrast. Furthermore, Daugman proposed a method [8–10] for addressing the iris size inconsistency by converting iris circle to Cartesian coordinate to a polar coordinate system in normalization stage. Hence, the new coordinate system represent as rectangular block of constant dimension. The Daugman's method for converting Cartesian coordinate to polar coordinate does not only normalizes the iris size, but also simplifies subsequent processing of feature extraction and matching. Such processing is performed using Rubber Sheet Model [11].

On the other hand, Ma [11] and Zhu [12] proposed a new method to improves Daugman's method using local histogram analysis to reduce the effect of poor illumination and low contrast. But the methods suffer a setback from low contrast iris image at normalization stage. Hence, there are criteria used for evaluating iris recognition performance. These evaluating method includes false rejection rate (FRR), false acceptance rate (FAR), equal error rate (EER) and receiver operating characteristics (ROC) [13]. Ghali et al. [14] define FRR as type I error it measures the number of times when legitimate user is denied access into the system. This type of error is often due to insufficient quality of an iris image. The false acceptance rate (FAR), is also known as Type II error, this error measures the number of times when impostor wrongly admitted into an iris system due low image quality. This type of error is considered as significant error to be considered in iris recognition. In addition, the equal error rate (EER) occurs when two error rates of FRR and FAR intersect each other. The error intersects between FRR and FAR at a specific point called EER. The EER is served as indicator for iris performance. For instance, a EER of 2% gives higher accuracy compared with a EER of 6% [13]. This indicates that choosing intersection point of the two error curves as threshold determining the performance of the system. In this research, the EER value of the enhanced technique is 6% which considered as successful. The significant of selecting HE as yardstick it enhances quality image and tested in many applications as posits in Siti et al. [15]. It also solved various problems in microstructure, medical, satellite, radar and biometrics [16].

3 Image Enhancement

The main objective of image enhancement is to improve image quality that can be suitable for specific application. In this process, the image attributes (pixel intensities) is modified to achieve the desire aim [17, 18]. Figure 1 described the framework of our proposed technique (E-HEN).

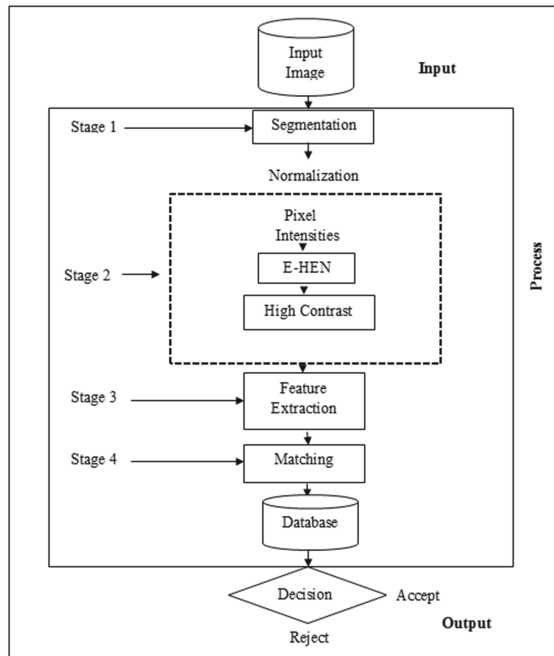


Fig. 1. Framework of proposed technique (E-HEN)

As discussed in Sect. 2 the iris size is normalized using Rubber Sheet Model devised by Daugman's methods and the iris image suffers a setback from low contrast. The proposed E-HEN addresses low contrast issue at normalization stage as shown in Fig. 1. Essentially, the enhancement is processed in spatial domain method to allow pixel intensities to manipulate directly. This process can be expressed using Eq. 1.

$$g(x, y) = T[f(x, y)] \quad (1)$$

where $f(x, y)$ is the input iris image and $g(x, y)$ is the processed image through transformation function $T()$. The transformation function T is known as intensity operator [17]. If pixel in the image f and g happens to signify by r and s the Eq. 1 can be clearly written as:

$$s = T(r) \quad (2)$$

The following examples bellow illustrates the steps for improving low contrast iris image after improving histogram equalization technique.

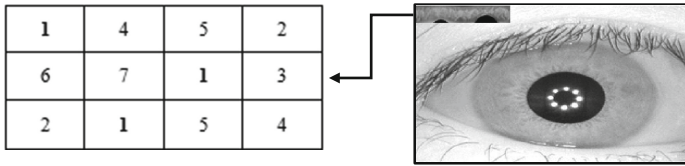


Fig. 2. Low contrast matrix

Intensity (n)	0	1	2	3	4	5	6	7
Number of pixels (y_n)	0	3	2	1	2	2	1	1

Fig. 3. Intensity values

Example 1 Let f be the low contrast iris image represent as $m \times n$ matrix of integer as shown in Fig. 2, the intensity values ranges between 0 and $L - 1$ as depicted in Fig. 3. L represents the number of intensity values often 256. Similarly y is the normalized histogram of image f with each possible intensity.

$$y_u(u = \frac{n}{k})n = 0 \dots L - 1 \quad (3)$$

where y_u is the normalized histogram and n is the number of pixels with intensity often from 0 to 256, k represents the number of pixels.

Figure 2 represents the pixel values obtained from low contrast image at normalization stage. As shown in Fig. 3 the intensity values ranges from 0 to 255. The frequency of each intensity value is calculated, for instance, intensity 1 from Fig. 3 has the frequency of 3 as can be visualize in Fig. 2. Similarly, intensity values 2 and 3 have frequency 2 and 1 respectively. The frequency of all the intensity values can be verified in Fig. 2. Sequel to obtaining the number of pixels, each value in the number-of-pixels column is divided by the total number of pixels to obtain the normalized value and the result is further given in Fig. 4.

(y_n)=Total number of pixels	0/12	3/12	2/12	1/12	2/12	2/12	1/12	1/12
Result	0	0.25	0.166	0.083	0.166	0.166	0.083	0.083

Fig. 4. Normalization method

The histogram equalized image f generated from this process is shown in Fig. 5, and the equation is express as:

$$f_{i,j} = ghal((L - 1) \sum_{n=0}^{f_{i,j}} y_n) \quad (4)$$

where $ghal()$ represents function name that rounds down the nearest integer number. This is similar to transformed pixels intensity k in f using Eq. 5 [19].

$$T(K) = ghal(L - 1) \sum_{n=0}^k y_n \quad (5)$$

Results	0	0.25	0.166	0.083	0.166	0.166	0.083	0.083
	\times 255	\times 255	\times 255	\times 255	\times 255	\times 255	\times 255	\times 255
Transformed image	0	64	42	21	42	42	21	21

Fig. 5. Transformation method

Based on Fig. 5, the result is multiplied by 255 to obtain transform image. The matrix of transformed image is shown in Fig. 6.

64	42	42	42
21	21	64	21
42	64	42	42

Fig. 6. Transformed image

The giving example shows how E-HEN obtained high contrast in the iris image.

Example 2 Let x be the transformed image obtained from Eq. 5, the transformed image is further enhanced using the proposed Equation in Eq. 6.

$$y_{ij} = x_{ij} + a \quad (6)$$

where y is the enhanced image, x is the transformed image and a adjust the pixels intensities by incrementing one value for each pixels in an image, ij represents the intensity values in the two images (y and x). Figure 7 illustrates the high contrast image.

65	43	43	43
22	22	65	22
43	65	43	43

Fig. 7. High contrast image

As shown in Fig. 7, the image is enhanced using Eq. 6. The effect of proposed technique using Matlab is shown in Fig. 8.

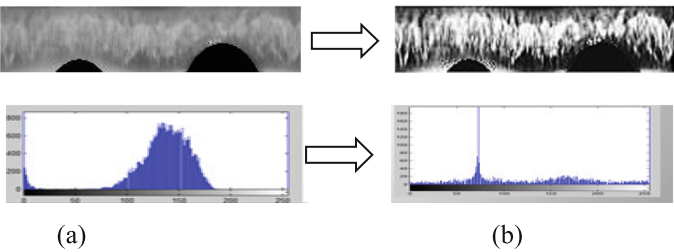


Fig. 8. Effect of the proposed method

From Fig. 8a represents the original image before enhancement Fig. 8b represent the image after enhancement.

4 Experimental Results

This research employed MATLAB R2013a (8.1.0.604) of MathWork, Inc. In Windows 7 operating system with Intel (R) Core i7 processor, 3.40 GHz and 4 GB RAM to achieve the proposed technique.

4.1 Result from CASIA Database

For CASIA iris dataset, as many as 200 images were used from CASIA iris-interval and another 100 images from CASIA iris-lamp Version 3. Both sets of images were

Table 1. Recognition results for CASIA datasets based on FRR and FAR

Input image	FRR	FAR	Database
S1001L01	0.05	0.01	CASIA
S1001L02	0.05	0.01	CASIA
S1001L03	0.15	0.02	CASIA
S1001L04	0.02	0.03	CASIA
S1001L05	0.27	0.05	CASIA

used for training and testing purposes. The results obtained from CASIA dataset is based on FRR and FAR, as shown in Table 1.

4.2 Result from UBIRIS Database

In this section, the results were obtained from UBIRIS Version 1 dataset to validate the performance of the proposed technique. Essentially, 100 images from UBIRIS were used for training and testing. The findings from UBIRIS are presented in Table 2.

Table 2. Recognition results for UBIRIS datasets based on FRR and FAR

Input image	FRR	FAR	Database
Img_1_1_1	7.52	0.00	UBIRIS
Img_1_1_2	1.47	1.35	UBIRIS
Img_1_1_3	1.30	1.16	UBIRIS
Img_1_1_4	1.25	1.10	UBIRIS
Img_1_1_5	1.18	1.02	UBIRIS

4.3 Result from MMU Database

In this section, as many as 300 images were used for training and testing using MMU Version 2 dataset. In addition, the test was carried out to validate the performance of proposed the technique. Table 3, illustrates the MMU's results.

Table 3. Recognition results for MMU datasets based on FRR and FAR

Input image	FRR	FAR	Database
a	2.47	1.12	MMU
b	2.38	6.75	MMU
c	2.29	6.44	MMU
d	2.19	6.12	MMU
e	2.09	5.80	MMU

4.4 Result from ICE 2005 Database

Additionally, in this section, ICE 2005 datasets was used to validate the performance of the proposed technique. Essentially, a total of 40 images were tested from the datasets whilst 20 images dataset were used for training and 20 images dataset for testing. Table 4 shows the ICE 2005's results.

Table 4. Recognition results for ICE 2005 datasets based ON FRR and FAR

Input image	FRR	FAR	Database
001	6.00	0.00	ICE 2005
002	5.76	0.02	ICE 2005
003	5.41	0.01	ICE 2005
004	4.93	0.01	ICE 2005
005	4.28	0.00	ICE 2005

Figure 9, further illustrates the exchange of the two error rates according to degree of matching.

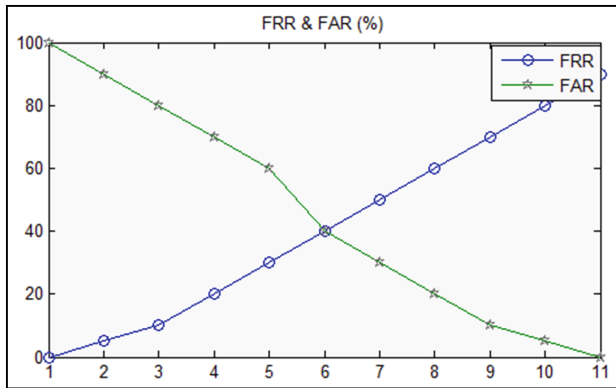


Fig. 9. Change of two error rates based on degree of matching

The exchange of the two error rates was based on the degree of matching and selecting proper threshold. By selecting intersection point of the two error curves as a threshold it reduces the FRR and FAR significantly. In essence, when the threshold was at 6, the recognition will be 95% accuracy. The lower the threshold, the more optimal performance was achieved, as discussed in Sect. 2.

Table 5. Recognition accuracy for various enhancement techniques

Method	FRR	FAR	Accuracy %
CLAHE	0.15	0.02	93.0
AHE	2.11	0.06	85.7
MAHE	1.15	0.02	92.8
HE	0.28	0.06	90.7
PROPOSED E-HEN	0.05	0.01	95.0

Table 5 summarizes the recognition performance of various enhancement techniques. The comparison has shown that the proposed technique performs the best in iris recognition with about 95% matching accuracy with less FAR 0.01 and FRR 0.05, followed by CLAHE at 93% matching accuracy with FAR 0.02 and FRR 0.15. The enhancement techniques that produced low performance might come from the higher error rates.

5 Conclusion

In this paper, an efficient method for enhancing low contrast image for iris recognition system is presented. Based on the result in Table 5, the proposed technique has the highest recognition performance. Hence, the method provides consistency in iris enhancement without affecting the iris image regions or reducing the local details of the image.

6 Discussion

In summary, iris recognition system suffers a set back from low quality image as the acceptance or rejection rates of the user being verified depends solely on the image quality. Low quality image such as poor illumination and low contrast increases false rejection rate (FRR) and false acceptance rate (FAR) thus decreases the performance of iris recognition system. Essentially, in our proposed work when the threshold was at 6 it reduces the error rates simultaneously. Therefore, the lower the value of the error rates FRR and FAR the better the performance and the higher the value of the accuracy the result better.

Acknowledgements. This research was fully sponsored by the Office for Research, Innovation, Commercialization and Consultancy (ORICC), with VOT No U614. The authors fully acknowledge Universiti Tun Hussein Onn Malaysia (UTHM) for the financial support which has made this research possible.

References

1. Othman, M.F.: Fusion techniques for iris recognition in degraded sequences, pp. 4–5, (2016)
2. Li, P., Ma, H.: Iris recognition in non-ideal imaging conditions. *Pattern Recognit. Lett.* **33** (8), 1012–1018 (2012)
3. Vatsa, M.: Comparison of iris recognition algorithms. In: *Proceedings of International Conference on Intelligent Sensing and Information Processing*, pp. 354–358 (2004)
4. Alvarez-Betancourt, Y., Garcia-Silvente, M.: A keypoints-based feature extraction method for iris recognition under variable image quality conditions. *Know.-Based Syst.* **92**, 169–182 (2015)
5. Sanpachai, H., Malisuwan, S.: A study of image enhancement for iris recognition. *J. Ind. Intell. Inf.* **3**(1), 61–64 (2015)

6. Shivakumara, P., Huang, W., Phan, T.Q., Tan, C.L.: Accurate video text detection through classification of low and high contrast images. *Pattern Recogn.* **43**, 2165–2185 (2010)
7. Othman, N., Dorizzi, B., Garcia-Salicetti, S.: OSIRIS: an open source iris recognition software. *Pattern Recogn. Lett.* **82**, 124–131 (2016)
8. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal. Mach. Intell.* (1993)
9. Daugman, J., Downing, C.: Epigenetic randomness, complexity and singularity of human iris patterns. *Proc. R. Soc. B: Biol. Sci.* 1737–1740 (2001)
10. Daugman, J.: How iris recognition works. *IEEE Trans. Circ. Syst. Video Technol.* **14**(1), 21–30 (2004)
11. Ma, T., Tan, T., Wang, Y., Zhang, D.: Personal identification based on iris texture analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(12), 1519–1533 (2003)
12. Zhu, Y., Tan, T., Wang, Y.: Biometric personal identification based on iris patterns. In: *Proceedings of the 15th International Conference on Pattern Recognition*, pp. 1–4 (2000)
13. Sanderson, S., Erbetta, J.: Authentication for secure environments based on IRIS scanning technology. In: *IEEE Colloquium on Visual Biometrics* (2000)
14. Ghali, A., Jamel, S., Pindar, Z., Disina, A., Deris, M.: Reducing error rates for iris image using higher contrast in normalization process. *IOP Conf. Ser. Mater. Sci. Eng.* **266** (2017)
15. Ahmad, S.A., Taib, M.N., Elaiza, N., Khalid, A., Taib, H.: An analysis of image enhancement techniques for dental X-ray image interpretation. *Int. J. Mach. Learn. Comput.* **2**(3), 292–297 (2012)
16. Santhi, K., Banu, R.W.: Adaptive contrast enhancement using modified histogram equalization. *Optik-Int. J. Light Electron Opt.* **126**(19), 1809–1814 (2015)
17. Das, A.: Image enhancement in spatial domain. In: *Guide to Signals and Patterns in Image Processing* (2015)
18. Maini, R., Aggarwal, H.: A comprehensive review of image enhancement techniques. *J. Comput.* **2**(3), 8–13 (2010)
19. Gonzalez, R.C., Woods, R.E.: Histogram equalization. *Digital Image Processing*, pp. 1–3 (2008)

Presenting a New Method of Authentication for the Internet of Things Based on RFID

Farshad Asadpour^{1,2(✉)} and Shamsollah Ghanbari^{1,2(✉)}

¹ Department of Computer Science, Islamic Azad University, Ashtian Branch, Iran
Asadpoor.f@gmail.com

² Iranian Non-profit Association of Distributed Computing and Systems, Qom, Iran
myrshg@gmail.com

Abstract. Nowadays, the Internet of Things is significantly used in the world of information technology. One of the most important things for the Internet of Things is RFID tags that send information out of the environment. Forasmuch as the security process is important, the network authentication should be done with great accuracy. In this study, we propose a cryptographic authentication method that can increase the security of Internet of Things against various attacks.

Keywords: RFID · Internet of things · Authentication

1 Introduction

In recent years, a new concept known as “Radio-Frequency Identification (RFID) systems” has been developed with the advancement of technology. The radio frequency identification system is one of the most widely used automatic identification technologies that retrieve and store data from remote devices. This technology is used to wirelessly detect data stored on the microchip through radio waves [1]. RFID systems have the same function as the barcode system, with the difference that they can read and write tags at any angle and between objects and there is no need for a straight line between tag and tag readers [2]. Since RFID tags are relatively small and inexpensive and can simultaneously detect information from several tags using radio frequency communication, it is expected that RFID systems will replace barcode systems in the near future [3,4]. The RFID system is a contactless automatic identification system that has attracted much attention recently [4]. It contains tag components, reader tags and databases. RFID tags include unique identification information that can be connected to animate or inanimate objects and communicate with the tag reader. Tags include two main parts of the chip and antenna. The chip is used to maintain and provide memory and the antenna is used to amplify the signals in the environment for sending information.

The tag reader can acquire information through low-frequency radio frequency communication. In other words, tag readers will send or receive data from labels, and they are one of the components for databases. The tag reader transmits the identified information to the database and manages the database of identified information inside the tags [5].

RFID tags are divided into three categories based on the energy source: active, passive, semi-active. Active tags include batteries that supply them with energy. Passive tags receive their energy from the signals sent by the tag reader. Semi-active tags use an internal battery to benefit from the energy transmitted by the tag reader [3,4]. So far, various technologies have been designed and implemented for automatic identification, such as bar codes, smart cards, voice recognition, optical character recognition, and radio frequency identification [6].

Nowadays, RFID is used in many applications to identify and manage tools and equipment (asset tracking). In addition, it is used for the Internet of Things. In this technology, security is also very important [6].

Many researchers pay special attention to the security process of the Internet of Things. Thus, many methods have been proposed, each of which has advantages and disadvantages.

However studies continue on this field. The following items should be considered to evaluate the effectiveness of the security process [7]:

- Confidentiality means ensuring that only password holders or authorized persons have access to information.
- Authentication or assurance that the identity of the second party is proven.
- Integrity or the information does not change during the transmission.
- Access control or the lack of the unauthorized use.

Authentication refers to a process in which the sender or the recipient of the information provides each other with information to ensure the claim. According to authentication, it is checked that if the person is the one who claims. If the sender or receiver of the information cannot be properly authenticated for each other, there is no confidence that they can exchange information with each other. Authentication is a process which can be very simple or very complex. The text authentication is a password used to authenticate on different systems. However, authentication involves several factors that must be taken into account in order to increase the level of security [8].

In this study, we use the technology of authentication in RFID devices for the security of the Internet of Things. The purpose of this study is to provide access to the Internet of Things for legal users and prohibit the entry of illegal users.

2 Related Work

Countermeasures against security threats of RFID systems can be divided into two groups: cryptographic algorithms and non-cryptographic algorithms. The primary application of hash functions in cryptography is message integrity. The

hash value provides a digital fingerprint of a message's contents, which ensures that the message has not been altered by an intruder, or other means [9]. The hash functions are used for security protocols and privacy protection of RFID systems. Since the connection between the tag and the tag reader occurs in an open environment through radio signals, a mechanism for validating and identifying messages on both sides is required, which is called authentication [10]. Several authentication protocols are proposed based on cryptographic techniques as follows:

In [1], a simple, scalable, and low-priced design has been proposed based on the hash operation to solve security and privacy problems, called the SRFID protocol. This design provides a two-way authentication between the database and tag and does not require a secure channel between the tag reader and the database to complete the authentication process. The database stores all information associated with tags. Tags content is indexed by a unique identifier. The tag transmits its current ID to the tag reader, which the tag reader also sends it to the database as a database index. Authentication is based on the sharing of two sets of values between tags and databases. After the successful mutual authentication, the tag identifier is updated with the tag and database that provides security for the system. This security design prevents attacks, such as eavesdropping, resend attacks, tag replication, tag tracking, denial of service, Man-in-the-middle attack (MITM).

According to [12], a hash-based authentication protocol has been proposed as a solution to privacy issues and data tampering. This protocol has been designed to send a random quantity generated by the database tagging and without disclosure. In this protocol, the random quantity is replaced with a secret and hidden value and is used in a response message. The feature of the proposed protocol is the production of fixed response messages without interfaces from the expected production requests by the enemy. This protocol is safe against attacks such as eavesdropping, retransmission, tag replication, data tampering, MITM, and especially traffic analysis attack.

In [13], the Elliptic Curve Discrete Logarithm Problem (ECDLP-based) authentication protocol was presented. In this protocol, the values s and $s_P - Z$ were private and public keys, respectively. According to this protocol, when the server challenges the tag, it generates r_1 and resend's $r_1P = M_1$ to the server. The server receives M_1 and generates r_2 for sending to the tag. The tag receives r_2 and resend's $r_1 + s_r2 = M_2$ to the server. The server calculates $M_2P + r_2Z$ and examines the equivalence of this value with M_1 . If the equivalence is correct, the tag will be recognized. This protocol can prevent counterfeit attacks, but it suffers from physical and man-in-the-middle attacks.

In [14], the RFID-based authentication protocol was presented. In this protocol, the values (s_2, s_1) and $(s_2P_2 - s_1P_1 - Z)$ were private and public keys, respectively. According to this protocol, when the server challenges the tag, it generates r_1 and r_2 and resends $r_1P = M_1$ to the server. The server receives M_1 and generates r_3 for sending to the tag. The tag receives r_2 and resends $r_1 + s_1r_3 = M_2$ to the server. The server calculates $r_3Z + M_3P_2 + M_2P_1$ and examines the equivalence of this value with M_1 . If the equivalence is correct, the tag will be recognized. This protocol suffers from counterfeit, physical, and man-in-the-middle attacks.

In [15], a validation protocol was provided for mobile nodes. In this protocol, mobile nodes were validated by the selected cluster. A valid request message was sent and a valid validation message was received. Differential formulation was used to prove the privacy properties in this method. After running, it was found that this method has less communication overhead and greater privacy and security than the protocols, such as hash functions and OSK.

In [16], a communication protocol was proposed for RFID systems on the Internet of Things, which security is provided by a random oracle. In this model, objects had the unique electronic product code (EPC). The proposed method was also called SPAP. This method used symmetric encryption, one-way hash function, and XOR. It created two-way validation and internal security, and resisted some basic attacks.

In [17], an ABC-based validation method was proposed for the perception of the Internet of Things. In this architecture, the user is the supervisor of the perception layer, and for devices such as mobile phones and smart computers. This method had better performance over sensor nodes than the other protocols.

In [18], the Hierarchical Access Control (HAC) and Resilient Access Control (RAC) protocols were used. In this method, access to the tag was controlled by locking or unlocking through the one-way hash function.

In [19], the location tag was used for authentication. Other functions used in this design included one-way hash function and binary operations.

In [20–22], the reverse-tracking was used to protect the system against tracking. The tag identifier was updated using a low cost hash-chain mechanism.

3 Proposed Method

3.1 Notations and Definitions

Table 1 indicates the basic notations used in this paper.

The proposed method consists of the two following phases.

Table 1. Definitions and notations

Notation	Description
G	group of the q order on an elliptical bend
P	elemental member of the G group
ID_i	i^{th} tag identifier
S	private key of the server
D_i	private key for the i th tag
Y	public key of the server
h	Hash function that maps
H	Hash function that maps
t_i	two timestamps
R_t, r_s	two random numbers in Z_q
R_i	A random member in G

3.2 Initialization Phase

The server generates the random number s and calculates $Y = sP$ and then sets the value s as the private key and Y as the public key. The server generates the random number d_i and sets it as the private key for the i th tag. It calculates the value of $ID_i = d_iP$ for each tag, and then distributes the values of P, Y, t_i, ID_i through a secure channel between tags.

3.3 Authentication Phase

The server performs the following tasks to authenticate an anonymous tag:

- Step 1: The server generates the random number r_s and the time stamp t_i , and sends them to the i th tag.
- Step 2: The tag receives r_s and t'_i and checks that $t'_i > t_i$, otherwise the server does not recognize the tag. Given the ascendancy of the uniformity of the time stamp, the tag generates the random number r_t and chooses the random member R_i from group G to calculate $A_1 = r_s r_t P$ and $A_2 = r_s r_t P + R_i$. Now, it calculates $X_1 = h(R_i, A_1)$ and $X_2 = h(R_i, A_2)$. If $L = \gcd(x_2, x_1)$, the tag will change the values of x_2, x_1 to x_1 / L and x_2 / L and calculate the values $B_2 = x_2 H(ID)$, $B_4 = x_4 H(ID_i)$ and $B_1 = h(R_i, A_4, rS)$, finally, send the values of A_1, A_2, B_1, B_2, B_3 to the server.
- Step 3: The server obtains the above values to calculate $R_i = A_2 - sA_4$. Now it calculates $h(R_i, A_4, r_s)$ and matches it to B_3 . In case of mismatch, the server rejects the tag. In case of equality for $x_1 = h(R_i, A_1)$, the server calculates $x_2 = h(R_i, A_2)x$. If $\gcd(x_4, x_2) = L$, the server will change x_1, x_2 to x_1/L and x_2/L . Now, K_1 and K_2 can be calculated using the Euclidean algorithm;

therefore, $L = k_1x_1 + k_2x_2 \cdot H(ID_i)$ and ID_i can be obtained by calculating $k_4 B_4 + k_2 B_2 = k_4 x_4 H(ID_i) + k_2 x_2 H(ID_i) = (k_4 x_4 + k_2x_2) H(ID_i)$. Therefore, the tag is authenticated.

4 Evaluation

We used some criteria for evaluating the proposed method. The results were compared with the results obtained in [23]. we used omnet++ to simulate and compare the simulator. The simulated scenario is that a number of nodes of Internet of things distributed in the network are randomly, several random nodes was selected for check the authentication, and the main operation for this nodes Is done. False Match Rate (FMR): The expected probability that a sample will be falsely declared to match a single randomly-selected “non-self”. A false match rate is sometimes called false accept rate (FAR). In a simpler sense, this criterion means the possibility that the effect of person B is falsely known as the effect of person A. In Fig. 1, FMR is evaluated for the proposed method and comparison method.

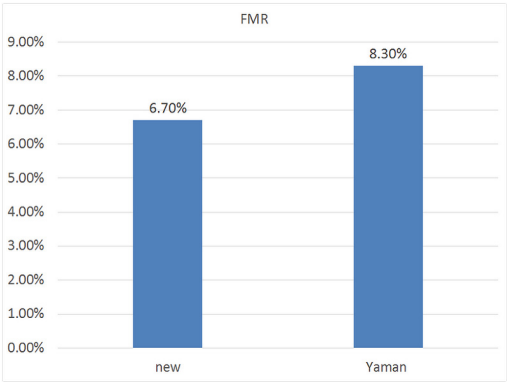


Fig. 1. Comparison of FMR (False Match Rate)

The simulation result for this criterion shows that the proposed method has better performance. False non-match rate (FNMR, also called FRR = False Reject Rate) is the probability that a legal person is falsely rejected. The simulation result for the above criterion is shown in Fig. 2.

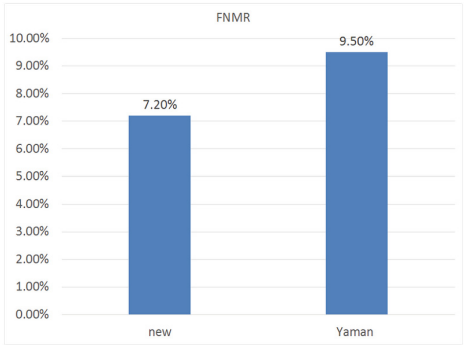


Fig. 2. Comparison of FNMR

According to the simulation result of the proposed method, FNMR occurs in 6% of cases and its performance is better than the previous method.

Accuracy: This criterion shows how well the proposed method can correctly represent the identity. In Fig. 3, the accuracy of the proposed method for 100 identity effects is compared with the other method.

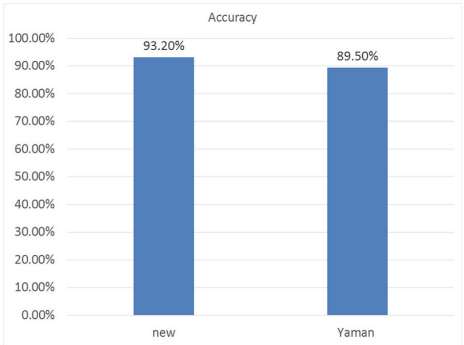


Fig. 3. Comparison of the accuracy

The simulation result shows that the proposed method has higher accuracy. The false rejection rate (FRR) is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user. FRR is sometimes called type-I error rate and represents the number of times that mismatching occurs (Fig. 4).

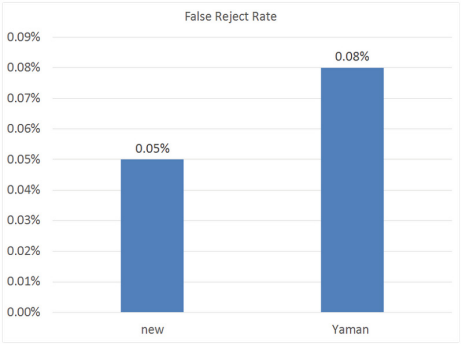


Fig. 4. FRR (False Reject Rate)

The proposed method has very good performance because of using the randomized encryption.

Equal error rate (EER): Decreasing the rate of false acceptance leads to an unintended increase in the rate of false rejection. EER is the rate at which false acceptance is equal to the rate of false rejection. The lower the amount of this parameter, the better sensitivity and good balance of the system will be. The comparison of this criterion can be seen in Fig. 5.

Runtime: it represents the length of time to complete the authentication requests. As shown in Fig. 6, the number of requests is increased in each step. Figure 7 shows the amount of memory used for authentication operations.

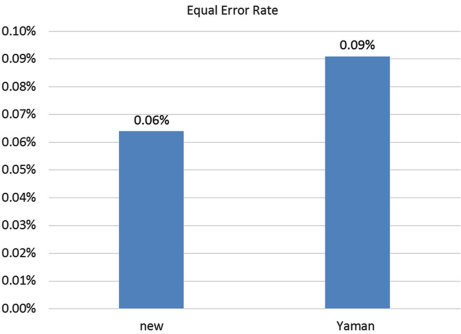


Fig. 5. EER (Equal error rate)

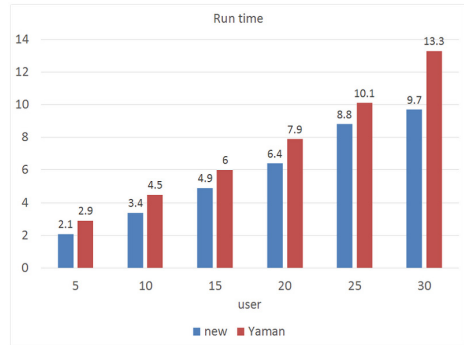


Fig. 6. Authentication duration

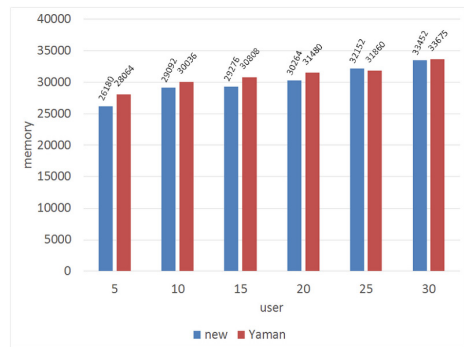


Fig. 7. Amount of memory for authentication

5 Conclusion

Considering the increasing development of this technology and the need for human use of it, we need to think about solutions for the security and privacy of this technology and provide consumers with cheap systems. If we want to use cryptographic algorithms in these systems, their price will increase. Therefore, it is necessary to use the strong authentication protocols that will have the most resistance against an attacker. We can maintain the security of these systems using the safer protocols and logical and mathematical operations, and making them more complicated.

References

1. Zhang, S., Chen, G., Zhou, Y., Li, J.: Enhanced-Bivium algorithm for RFID system. *Math. Probl. Eng.* **2015**(616182), 6 (2015). <https://doi.org/10.1155/2015/616182>
2. Alizadeh, M., Zamani, M., Shahemabadi, A.R., Shayan, J., Azanik, A.: A survey on attacks in RFID networks. *Open Int. J. Inform.* **1**(1) (2013)

3. Mitrokotsa, A., Rieback, M.R., Tanenbaum, A.S.: Classifying RFID attacks and defenses. *Inf. Syst. Front.* **12**(5) (2010)
4. Khedr, W.I.: SRFID: a hash-based security scheme for low cost RFID systems. *Egypt. Inform. J.* **14**(1) 2013
5. Cho, J.S., Yeo, S.S., Kim S.K.: Securing against brute-force attack: a hash-based RFID mutual authentication protocol using a secret value. *Comput. Commun.* **34**(3) (2011)
6. Moosavi, S.R., Nigussie, E., Virtanen, S., Isoaho, J.: An elliptic curve-based mutual authentication scheme for RFID implant systems. *Procedia Comput. Sci.* **32**, 198–206 (2014)
7. Farash, M.S.: Cryptoanalysis and improvement of an efficient mutual authentication RFID scheme based on elliptic curve cryptography. *J. Supercomput.* (2014)
8. He, D., Kumar, N., Chilamkurti, N., Lee, J.H.: Lightweight ECC based RFID authentication integrated with an ID verifier transfer protocol. *J. Med. Syst.* **38**, 1–6 (2014)
9. Liu, Z., Liu, D., Li, L., Lin, H., Yong, Z.: Implementation of a new RFID authentication protocol for EPC Gen2 standard. *IEEE Sens. J.* **15**(2) (2015)
10. Khatwani, C., Roy, S.: Security analysis of ECC based authentication protocols. In: *Proceedings of the 2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 1167–1172. Jabalpur, India, 12–14 December 2015
11. Liu, D., Liu, Z., Yong, Z., Zou, X., Cheng, J.: Design and implementation of an ECC-based digital baseband controller for RFID tag chip. *IEEE Trans. Ind. Electron.* **62**(7) (2015)
12. Benssalah, M., Djeddou, M.: Design and Implementation of a New Active RFID Authentication Protocol Based on Elliptic Curve Encryption, *SAI Computing Conference 2016*, London, UK, 13–15 July 2016
13. Chuang, Y.H., Hsu, C.L., Shu, W., Hsu, K.C., Liao, M.W.: A Secure Non-interactive Deniable Authentication Protocol with Certificates Based on Elliptic Curve Cryptography. *New Trends in Intelligent Information and Database Systems*, pp. 183–190. Springer, Berlin (2015)
14. Jin, C., Xu, C., Zhang, X., Zhao, J.: A secure RFID mutual authentication protocol for healthcare environments using elliptic curve cryptography. *J. Med. Syst.* **39**, 1–8 (2015)
15. Ye, N., Zhu, Y., Wang, R.C., Malekian, R., Min, L.Q.: An efficient authentication and access control scheme for perception layer of internet of things. *Int. J. Appl. Math. Inf. Sci.* **8**, 1617–1624 (2014)
16. Mahalle, P.N., Prasad, N.R., Prasad, R.: Object classification based context management for identity management in internet of things. *Int. J. Comput. Appl.* **63**, 1–6 (2013)
17. Roman, R., Zhou, J., Lopez, J.: On the features and challenges of security and privacy in distributed internet of things. *Comput. Netw.* **57**, 2266–2279 (2013)
18. Vasilomanolakis, E., Daubert, J., Luthra, M., Gazis, V., Wiesmaier, A., and Kikiras, P.: On the Security and Privacy of Internet of Things Architectures and Systems, In: *International Workshop on Secure Internet of Things (SIOT)* 2015
19. Gope, P., Hwang, T.: A realistic lightweight authentication protocol preserving strong anonymity for securing RFID system. *Comput. Secur.* **55**, 271–280 (2015)
20. Gope, P., Hwang, T.: Enhanced secure mutual authentication and key agreement scheme preserving user anonymity in global mobile networks. *Wirel. Pers. Commun.* **82**, 2231–2245 (2015)

21. Ghanbari, S., Othman, M., Abu Bakar, M.R., Leong, W.J.: Multi-objective method for divisible load scheduling in multi-level tree network. *Future Gener. Comput. Syst.* **54**, 132–143 (2016)
22. Ghanbari, S., Othman, M.: A priority based job scheduling algorithm in cloud computing. *Procedia Eng.* **50**, 778–785 (2012)
23. Sharaf-Dabbagh, Y., Walid, S.: On the authentication of devices in the Internet of things. In: *IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM) 2016*, pp. 1–3. IEEE, 2016

Author Index

- Abawajy, Jemal H., [135](#), [318](#)
Abdalrada, Ahmad Shaker, [135](#), [318](#)
Abdul Hamid, Siti Suhaila, [372](#)
Abdullah, Mohd Hafizul Afifi, [181](#), [363](#)
Abdullah, Noryusliza, [105](#)
Abdullah, Zubaile, [485](#)
Abubakar, Mansir, [455](#)
Adeleke, Abdullahi O., [282](#)
Admodisastro, Novia, [372](#), [385](#)
Aghababaeipour, Ali, [308](#)
Ahmad Fuadi, Nur Faizura, [396](#)
Ahmad, Afandi, [272](#)
Ahmad, Mohd Sharifuddin, [43](#), [213](#)
Ahmadian, A., [53](#)
ALALI, Muath, [446](#)
Ali, Ashikin, [225](#)
Ali, Rabei Raad, [33](#)
Al-Qasem, Al-Hadi Ismail Ahmed, [340](#)
Al-Quraishi, Tahsien, [135](#), [318](#)
Ambar, Radzi, [353](#)
Andrawina, Luciana, [124](#)
Arbaiy, Nureize, [82](#), [115](#)
Aris, Teh Noranis Mohd, [455](#)
Asadpour, Farshad, [506](#)
Astorga, Gino, [3](#)
Azmi Murad, Masrah Azrifah, [446](#)

Basir, Nurlida, [95](#)
Basri, Sri Mazura Muhammad, [200](#)
Berahim, Mazniha, [298](#)

Cheng, Shi, [24](#)
Choon, Chew Chang, [353](#)
Chowdhury, Morshed U., [135](#), [318](#)
Clemente, Filipe Manuel, [191](#)
Crawford, Broderick, [3](#)

Damayanti, Dida Diah, [124](#)

Daniel, Basil David, [105](#)
Darman, Rozanawati, [105](#)
Deris, Mustafa Mat, [95](#), [243](#), [495](#)

Efendi, Riswan, [243](#)

Fai, Chan Kar, [353](#)
Ferdiana, Ridi, [147](#)

García, José, [3](#)
Ghali, Abdulrahman Aminu, [495](#)
Ghanbari, Shamsollah, [308](#), [506](#)
Ghani, Abdul Azim Abd, [330](#), [372](#), [385](#)
Ghazali, Rozaida, [14](#), [234](#)

Hafez, Izuan, [330](#)
Hamdan, Hazlina, [446](#), [455](#)
Hamid, Isredza Rahmi A., [171](#), [485](#)
Hamid, Norhamreeza Abdul, [200](#)
Hamid, Siti Suhaila Abdul, [385](#)
Hassan, Rohayanti, [72](#)
Hassan, Sa'adah, [385](#)
Haviluddin, , [252](#)
Heng, Fatin Nabila Rafei, [95](#)
Hiai, Satoshi, [418](#)
Himeno, Yuusuke, [436](#)
Hirokawa, Sachio, [261](#)
Husin, Nor Azura, [446](#)
Hussain, Kashif, [24](#)

Ichinose, Ko, [409](#)
Insap Santosa, P., [147](#)
Iqbal, Umer, [234](#)
Ismail, F., [53](#)

Jacob, Deden Witarasyah, [473](#)
Jamali, Siti Nurliana, [385](#)
Jamel, Sapiee, [33](#), [495](#)

- Jamil, Muhammad Mahadi Abdul, 353
 Jelinek, Herbert F., 135
 Jumarni, Ratih Fitria, 161
- Kamaruddin, Azrina, 372, 385
 Kasim, Shahreen, 72
 Kasinathan, Vinothini, 64
 Khaleefah, Shihab Hamad, 43
 Khalid, Shamsul Kamal Ahmad, 33, 495
 Kontagora, Ibrahim Umar, 171
- Lashari, Saima Anwar, 252
 Lin, Pei-Chun, 115
 Lin, Yao, 261
- Ma'Radzi, Ahmad Alabqari, 353
 Mahdin, Hairulnizam, 72, 105
 Mahmoud, Moamin A., 213
 Mamat, Mustafa, 200
 Manshor, Noridayu, 372
 Md Fudzee, Mohd Farhan, 473
 Medi, Imran, 64
 Mendes, Bruno, 191
 Minami, Toshiro, 429
 Mine, Tsunenori, 261
 Mohamad, Kamaruddin Malik, 33, 495
 Mohamed, Rozlini, 181, 363
 Mohamed, Siti Aisyah, 181
 Mohammed, Mazin Abed, 43
 Mohd Sharef, Nurfadhlin, 340, 446
 Mohd. Rahman, Hamijah, 82
 Mostafa, Salama A., 43, 213
 Muhammad, Arshad, 24
 Mustapha, Aida, 43, 64, 191, 282
 Mustapha, Norwati, 455
- Nakatoh, Tetsuya, 429, 436
 Naseem, Rashid, 24
 Nasir, Sulaiman Md, 340
 Nathan, Shelena Soosay, 298
 Nawi, Nazri Mohd, 200, 282
 Nazri, Azree, 330
 Ng, Keng-Yap, 330
 Norwati, Mustapha, 340
 Nugroho, Lukito Edi, 147
- Othman, Muhaini, 181, 363
 Othman, Muhammad Fakri, 396
- Owen, Adam, 191
- Pindar, Zahradeen Abubakar, 495
- Rajasegarar, Sutharshan, 135, 318
 Reichert, Jelle, 191
 Ribeiro, João, 191
 Rosli, Nur Fazliyana, 363
- Saedudin, Rd Rohmat, 252
 Saedudin, Rd. Rohmat, 72
 Salahshour, S., 53
 Salamat, Mohamad Aizi, 473
 Salleh, Mohd Najib Mohd, 24
 Salleh, Shahril Nazim Mohamed, 105
 Samsudin, Noor Azah, 282, 298
 Santosa, Budi, 124
 Sari, Juni Nurma, 147
 Senan, Norhalina, 225, 396
 Senu, N., 53
 Setyawan, Erlangga Bayu, 124
 Shah, Habib, 14, 234
 Shah, Wahidah Md, 485
 Shi, Yuhui, 24
 Shimada, Kazutaka, 409, 418
 Soto, Ricardo, 3
 Surin, Ely Salwana Mat, 105
 Sutoyo, Edi, 72
- Tukiran, Zarina, 272
- Ullah, Ghufuran, 24
 van der Linden, Cornelis M.I. (Niels), 191
- Wahab, Mohd Helmy Abd, 353
 Wahid, Fazli, 14
 Wen, Chuah Chai, 82
- Yamada, Yasuhiro, 436
 Yamaguchi, Kohei, 261
 Yanto, Iwan Tri Riyadi, 72, 252, 473
 Yusof, Munirah Mohd, 181, 363
 Yusoff, Mohd Zaliman M., 213
- Zamri, Nurnadiah, 161
 Zulkifli, Aqil, 485