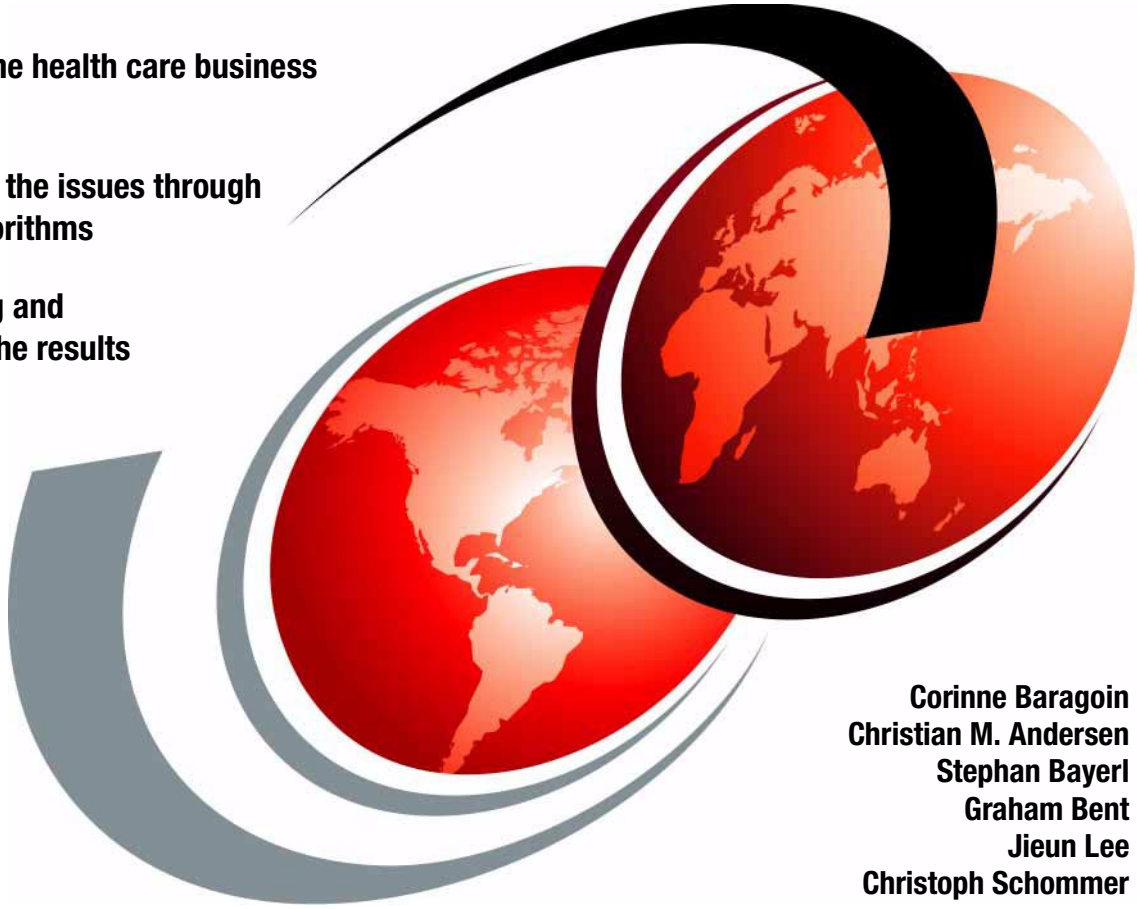
IBM

Mining Your Own Business in Health Care

Using DB2 Intelligent Miner for Data

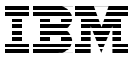
- Exploring the health care business issues
- Addressing the issues through mining algorithms
- Interpreting and deploying the results



Corinne Baragoin
Christian M. Andersen
Stephan Bayerl
Graham Bent
Jieun Lee
Christoph Schommer

Redbooks

ibm.com/redbooks



International Technical Support Organization

Mining Your Own Business in Health Care Using DB2 Intelligent Miner for Data

September 2001

SG24-6274-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 183.

First Edition (September 2001)

This edition applies to IBM DB2 Intelligent Miner For Data V6.1.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Preface

The data you collect about your patients is one of the greatest assets that any business has available. Buried within the data is all sorts of valuable information that could make a significant difference to the way you run your business and interact with your patients. But how can you discover it?

This IBM Redbook focuses on a specific industry sector, the health care sector, and explains how IBM DB2 Intelligent Miner for Data (IM for Data) is the solution that will allow you to mine your own business.

This redbook is one of a family of redbooks that has been designed to address the types of business issues that can be solved by data mining in different industry sectors. The other redbooks address the retail, banking, and telecoms sectors.

Using specific examples for health care, this book will help medical personnel to understand the sorts of business issues that data mining can address, how to interpret the mining results, and how to deploy them in health care. Medical personnel will want to skip certain sections of the book, such as “The data to be used”, “Sourcing and preprocessing the data”, and “Evaluating the data”.

This book will also help implementers to understand how a generic mining method can be applied. This generic method describes how to translate the business issues into a data mining problem and some common data models that you can use. It explains how to choose the appropriate data mining technique and then how to interpret and deploy the results.

Although no in-depth knowledge of Intelligent Miner for Data is required, a basic understanding of data mining technology is assumed.

The team that wrote this redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Corinne Baragoin is a Business Intelligence Project Leader at the International Technical Support Organization, San Jose Center. Before joining the ITSO, she had been working as an IT Specialist for IBM France, assisting customers on DB2 and data warehouse environments.

Christian M. Andersen is a Business Intelligence/CRM Consultant for IBM Nordics. He holds a degree in Economics from the University of Copenhagen. He has many years of experience in the data mining and business intelligence field. His areas of expertise include business intelligence and CRM architecture and design, spanning the entire IBM product and solution portfolio.

Stephan Bayerl is a Senior Consultant at the IBM Boeblingen Development Laboratory in Germany. He has over four years of experience in the development of data mining and more than three years in applying data mining to business intelligence applications. He holds a doctorate in Philosophy from Munich University. His other areas of expertise are in artificial intelligence, logic, and linguistics. He is a member of Munich University, where he gives lectures in analytical philosophy.

Graham Bent is a Senior Technology Leader at the IBM Hursley Development Laboratory in the United Kingdom. He has over 10 years of experience in applying data mining to military and civilian business intelligence applications. He holds an master's degree in Physics from Imperial College (London) and a doctorate from Cranfield University. His other areas of expertise are in data fusion and artificial intelligence.

Jieun Lee is an IT Specialist for IBM Korea. She has five years of experience in the business intelligence field. She holds a master's degree in Computer Science from George Washington University. Her areas of expertise include data mining and data management in business intelligence and CRM solutions.

Christoph Schommer is a Business Intelligence Consultant for IBM Germany. He has five years of experience in the data mining field. His areas of expertise include the application of data mining in different industrial areas. He has written extensively on the application of data mining in practice. He holds a master's degree in Computer Science from the University of Saarbruecken and a doctorate of Health Care from the Johann Wolfgang Goethe-University Frankfurt in Main, Germany. (Christoph's thesis, *Konfirmative und explorative Synergiewirkungen im erkenntnisorientierten Informationszyklus von BAIK,* contributed greatly to the medical research represented within this redbook.)

Thanks to the following people for their contributions to this project:

- By providing their technical input and valuable information to be incorporated within these pages:

  Wolfgang Giere is a University Professor and Director of the Center for Medical Informatics at the J. W. Goethe University, Frankfurt am Main, Germany.

Gregor Meyer
Mahendran Maliapen
Martin Brown
IBM

- By answering technical questions and reviewing this redbook:

  Andreas Arning
  Ute Baumbach
  Reinhold Keuler
  Christoph Lingenfelder
  Intelligent Miner Development Team at the IBM Development Lab in Boeblingen

- By reviewing this redbook:

  Tom Bradshaw
  Jim Lyon
  Richard Hale
  IBM

Special notice

This publication is intended to help both business decision makers and medical personnel to understand the sorts of business issues that data mining can address and to help implementers, starting with data mining, to understand how a generic mining method can be applied. The information in this publication is not intended as the specification of any programming interfaces that are provided by IBM DB2 Intelligent Miner for Data. See the PUBLICATIONS section of the IBM Programming Announcement for IBM DB2 Intelligent Miner for Data for more information about what publications are considered to be product documentation.

IBM trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

e (logo)® 
IBM ®
AIX
AT
CT
Current
DataJoiner
Redbooks
Redbooks Logo 
DB2
DB2 Universal Database
Information Warehouse
Intelligent Miner
SP
400

Comments welcome

Your comments are important to us!

We want our IBM Redbooks to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  ibm.com/redbooks

- Send your comments in an Internet note to:

  redbook@us.ibm.com

- Mail your comments to the address on page ii.

1

Introduction

In today's dynamic business environment, successful organizations must be able to react rapidly to the changing market demands.

To do this requires an understanding of all of the factors that have an influence on your business, and this in turn requires an ability to monitor these factors and provide the relevant and timely information to the appropriate decision makers.

Creating a picture of what is happening relies on the collection, storage, processing and continuous analysis of large amounts of data to provide the information that you need. This whole process is what we call Business Intelligence (BI). BI is about making well-informed decisions, using information that is based on data. Data in itself provides no judgement or interpretation and therefore provides no basis for action. Putting data into context is what turns it into information. Connecting pieces of available information leads to the knowledge that can be used to support decisions. Where the context is well understood, BI enables the transformation from data to decision to become a routine process within your business. One of the main challenges is that increasing competitive pressures requires new and innovative ways to satisfy increasing customer demands. In these cases the context is not well understood.

Data mining provides the tools and techniques to help you *discover* new contexts and hence new things about your customers. Mining your own business will enable you to make decisions based upon real knowledge instead of just a gut feeling.

1.1 Why you should mine your own business

Increasing competitive pressures require you to develop new and innovative ways to satisfy the increasing demands your customers make. To develop these new ideas requires information about your customers and this information in turn must be derived from the data you collect about your customers. This information is not only invaluable from the perspective of your own business but is also of interest to the suppliers who manage the brands that you sell. Your data should be seen as one of the greatest assets your business owns.

The challenge that faces most health care organizations is that the volumes of data that can potentially be collected are so huge and the range of customer behavior is so diverse that it seems impossible to rationalize what is happening. If you are reading this book and you don't mind being told to "mine your own business" then you are probably already in this position. The question we want to address is, how can data mining help you discover new things about your customers and how can you use this information to drive your business forward?

The road from data to information, and finally to the decision making process itself, is not an easy one. In this book our objective is to show, through some example cases, what role data mining has to play in this process, what sorts of health care business problems you can address and what you need to do to mine your own business.

1.2 The health care business issues to address

There are a large number of medical health care questions to which data mining can provide answers, for example:

- Can we identify indicators that are mainly responsible for the occurrence of special diseases like diabetes, thrombosis or tuberculosis?
- Which symptoms are highly correlated with positive examination tests?
- Can we set up a model that can predict the patient's stay in the hospital concerning a special disease?
- Can we detect medical indicators that act as an alarm system?
- Do the doctors who make the diagnosis observe the same treatment?

The data mining techniques that we can use to obtain these answers are the subject of this book. It would take a much larger book than this one to address all of the questions that data mining can answer; therefore, we have chosen to restrict ourselves to just four specific health care issues.

Certain issues have been selected primarily to illustrate the range of the data mining techniques that we have available to us.

Therefore, in this book we look at answering specific questions to illustrate how these techniques can be applied:

- How can we perform weight rating for Diagnosis Related Groups (DRG) by using medical diagnoses?
- How can we perform patient profiling?
- Can we optimize medical prophylaxis tests?
- Can we detect precauses for a special medical condition?

In our first example we consider the question of *how to calculate weights for Diagnoses Related Groups (DRG)*. Diagnosis Related Groups are a highly discussed topic in the medical area. The reason for this is that medical services are not based on diagnoses anymore, but on combinations of medical diagnoses (for example, ICD10, International Classification of Medicine) and medical procedures (for example, ICPM, International Classification of Procedures in Medicine). The weight for DRG's, plays an important role, because medical service revenue will be defined as the product of the DRG weight and a fixed amount of money; the higher the weight, the higher the revenue and vice versa. In this chapter, we will describe a method to find combinations of medical diagnoses that form a basis for Diagnosis Related Groups (DRG).

Using *Association Discovery* we obtain associative rules that indicate combinations between medical diagnoses. The rules give you a statistically based report about current diagnosis trends and indicate which combinations of rules are more evident than others. As will be evident in the following chapter, the application of Association Discovery for medical diagnoses could become important in detecting higher and lower ranked weights for Diagnosis Related Groups.

For the second question on *how to perform profiling of patients,* we suggest that you use *clustering*. This method will be performed on patients who were tested for Deep Vein Thrombosis. All patients were diagnosed for thrombosis, where some of them had thrombosis and some of them didn’t. The challenge for this question is now to find groups of patients who share a similar behavior. We want to detect some new but useful indicators that may be derived from our analysis.

The third question is concerned with *medical patient records on how can they be used to optimize medical prophylaxis tests*. By introducing Diabetes Mellitus - actually one of the most important diseases - we will define a method using *classification* that helps us to obtain information about the relevance of different test components. Because some test components are more important than others, the correct order of these components and/or the right choice of the test components themselves may lead to a faster and more secure strategy.

Diseases are sometimes difficult to identify and often they remain undiscovered. Reasons for this are, for example, ambiguity of the diseases' symptoms, missing medical possibilities, or insufficient experience of the medical staff. New strategies and techniques that help to find precauses for diseases are therefore very appreciated.

For the fourth question, we present some strategies about *how we can find precauses for a special disease*. We use data that was recorded for patients who were tested for thrombosis. Here, we will concentrate on *time series* data and show what kind of analyses can be done in detail.

By concentrating on these questions, we hope that you will be able to appreciate why you should mine your own data with the ultimate objective of deploying the results of the data mining into your health care process.

1.3 How this book is structured

The main objective of this book is to address the above health care issues using data mining techniques.

However, to put this into context, it is first necessary to understand the context of data mining in an overall BI architecture. We have already explained that the road from data to decisions is not an easy one, and that if you are going to mine your own business you will need some guidance.

To help in both of these areas:

- Chapter 2, "Business Intelligence architecture overview" on page 7 provides a BI architecture overview.
- Chapter 3, "A generic data mining method" on page 23 presents a detailed overview of what data mining describes as a generic method that can be followed.
- For the examples in the following chapters, use these methods and apply them to the business questions:
  - Chapter 4, "How to perform weight rating for Diagnosis Related Groups by using medical diagnoses" on page 47

  - Chapter 5, “How to perform patient profiling” on page 75
  - Chapter 6, “Can we optimize medical prophylaxis tests?” on page 111
  - Chapter 7, “Can we detect precauses for a special medical condition?” on page 135
- Finally in Chapter 8, “The value of DB2 Intelligent Miner for Data” on page 167 we describe the benefits of Intelligent Miner for Data (IM for Data), the data mining tool that we use in these examples.

We have provided sufficient information for you to understand how you are able to mine your own health care business without going into too many technical details about the data mining algorithms themselves. There is a difficult balance to strike here, and therefore for you to decide which sections you should read, we want to make the following comments:

We do **not:**

- Provide a user’s guide of any mining function
- Explicate any mining function in a mathematical complete way
- Deliver the basic background knowledge of a statistical introductory book
- Stress a particular data mining toolkit
- Provide a comparison of competitive mining products

Rather, we stress an operational approach to data mining by explaining the:

- Mechanics of operating a data mining toolkit
- Generic method as a guideline for the newcomer and the expert
- Technical aspects of the mining algorithms
- Necessary data preparation steps in a detailed manner
- Proven mining applications in the field
- Further steps for improvement of the mining results

It is assumed that a task like ours must remain incomplete in the sense that all examples demonstrated in this book could be copied and exploited in a short time, from several days to some weeks, while serious mining projects run from several weeks to months and longer. Therefore, the book lacks the description of the necessary bothersome and tedious mining cycles and does not offer a list of helpful tricks to simplify or overcome them totally. And of course, the approaches presented here by no means embrace all types of business issues.

1.4 Who should read this book?

This book is intended:

- To help business users figure out how data mining can address and solve specific business issues by reading the following sections in the different chapters:
  - The business issue
  - Interpreting the results
  - Deploying the mining results
- To be a guide for implementers on how to use data mining to solve business issues by explaining and detailing the generic method in each business question chapter, by providing data models to use and by including some experience-based hints and tips. It is worthwhile for implementers to progress sequentially through each business question chapter.
- To provide additional information:
  - To position data mining in the business intelligence architecture by reading Chapter 2, “Business Intelligence architecture overview” on page 7
  - To evaluate the data mining product by reading Chapter 8, “The value of DB2 Intelligent Miner for Data” on page 167

To benefit from this book, the reader should have, at least, a basic understanding of data mining.

2

Business Intelligence architecture overview

Business Intelligence (BI) covers the process of transforming data from your various data sources into meaningful information that can provide you and your company with insights into where your business has been, is today, and is likely to be tomorrow.

BI allows you to improve your decision-making at all levels by giving you a consistent, valid, and in-depth view of your business by consolidating data from different systems into a single accessible source of information — a data warehouse.

Depending on the users' needs there are different types of tools to be used to analyze and visualize the data from the data warehouse. These tools range from query and reporting to advanced analysis by data mining.

In this chapter we will describe the different components in a BI architecture. This will lead you to an overview of the architecture on which your data mining environment will be founded.

2.1 Business Intelligence

Traditionally, information systems have been designed to process discrete transactions in order to automate tasks, such as order entry, or account transactions. These systems however are not designed to support users who wish to extract data at different aggregation levels and utilize advanced methods for data analysis. Apart from these, systems tend to be isolated to support a single business system. This results in a great challenge when requiring a consolidated view of the state of your business.

This is where data warehouse and analytical tools come to your aid.

2.2 Data warehouse

Figure 2-1 shows the entire data warehouse architecture in a single view. The following sections will concentrate on single parts of this architecture and explain them in detail.

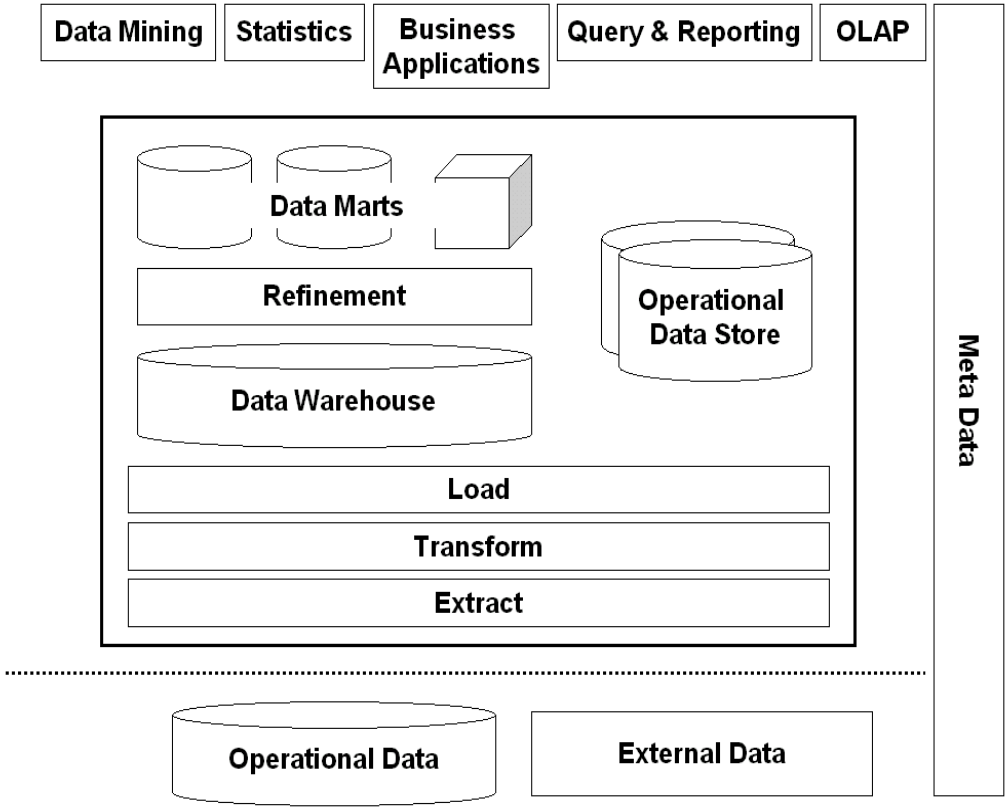


Figure 2-1 Data warehouse components

The processes required to keep the data warehouse up to date as marked are:

- Extraction/propagation
- Transformation/cleansing
- Data refining
- Presentation
- Analysis tools

The different stages of aggregation in the data are:

- On Line Transaction Programs (OLTP) data
- Operational Data Store (ODS)
- Datamarts

Metadata and how it is involved in each process is shown with solid connectors.

The tasks to be performed on the dedicated OLTP system are optimized for interactive performance and to handle the transaction oriented tasks in the day-to-day-business.

The tasks to be performed on the dedicated data warehouse machine require high batch performance to handle the numerous aggregation, pre calculation, and query tasks.

2.2.1 Data sources

Data sources can be operational databases, historical data (usually archived on tapes), external data (for example, from market research companies or from the Internet), or information from the already existing data warehouse environment. The data sources can be relational databases from the line of business applications. They also can reside on many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plain text files or pictures and other multimedia information.

2.2.2 Extraction/propagation

Data extraction / data propagation is the process of collecting data from various sources and different platforms to move it into the data warehouse. Data extraction in a data warehouse environment is a selective process to import decision-relevant information into the data warehouse.

Data extraction / data propagation is much more than mirroring or copying data from one database system to another. Depending on the technique, this process is either referred as:

- **Pulling** (Extraction of data)

  Or

- **Pushing** (Propagation of data)

2.2.3 Transformation/cleansing

Transformation of data usually involves code resolution with mapping tables, for example, changing the variable *gender* to:

- *0* if the value is *female*
- *1* if the value is *male*

It involves changing the resolution of hidden business rules in data fields, such as account numbers. Also the structure and the relationships of the data are adjusted to the analysis domain. Transformations occur throughout the population process, usually in more than one step. In the early stages of the process, the transformations are used more to consolidate the data from different sources; whereas, in the later stages, data is transformed to satisfy a specific analysis problem and/or a tool requirement.

Data warehousing turns data into information; on the other hand, **data cleansing** ensures that the data warehouse will have valid, useful, and meaningful information. Data cleansing can also be described as standardization of data. Through careful review of the data contents, the following criteria are matched:

- Replace missing values
- Normalize value ranges and units (for example, sales in the euro or dollar)
- Use valid data codes and abbreviations
- Use consistent and standard representation of the data
- Use domestic and international addresses
- Consolidate data (one view), such as house holding

2.2.4 Data refining

The atomic level of information from the star schema needs to be aggregated, summarized, and modified for specific requirements. This data refining process generates datamarts that:

- Create a subset of the data in the star schema
- Create calculated or virtual fields
- Summarize the information
- Aggregate the information

The layer in the data warehouse architecture is needed to increase the query performance and minimize the amount of data that is transmitted over the network to the end user query or analysis tool.

When talking about data transformation/cleansing, there are basically two different ways where the result is achieved. In detail, these are:

- **Data aggregation**: Changes the level of granularity in the information.

  Example: The original data is stored on a daily basis — the data mart contains only weekly values. Therefore, data aggregation results in less records.

- **Data summarization**: Adds up values in a certain group of information.

  Example: The data refining process generates records that contain the revenue of a specific product group, resulting in more records.

Data preparation for mining is usually a very time consuming task, often the mining itself requires less effort. The optimal way to do data preprocessing for data mining is typically very dependent on the technology used and the current skills, the volume of data to be processed and the frequency of updates.

2.2.5 Datamarts

Figure 2-1 shows where datamarts are located logically within the BI architecture.

A datamart contains data from the data warehouse tailored to support the specific requirements of a given business unit, business function or application.

The main purpose of a data mart can be defined as follows:

- To store pre-aggregated information
- To control end user access to the information
- To provide fast access to information for specific analytical needs or user group
- To represent the end users view and data interface of the data warehouse
- To create the multidimensional/relational view of the data

The database format can either be multidimensional or relational.

When building data marts, it is important to keep the following in mind:

- Data marts should always be implemented as an extension of the data warehouse, not as an alternative. All data residing in the data mart should therefore also reside in the data warehouse. In this way the consistency and reuse of data is optimized
- Data marts are typically constructed to fit one requirement, ideally. However, you should be aware of the trade-off between the simplicity of design (and performance benefits) compared to the cost of administrating and maintaining a large number of data marts.

2.2.6 Metadata

The metadata structures the information in the data warehouse in categories, topics, groups, hierarchies and so on. They are used to provide information about the data within a data warehouse, as given in the following list (also see Figure 2-2):

- Metadata are “subject oriented” and are based on abstractions of real-world entities, for example, “project”, “customer”, or “organization”.

- Metadata define the way in which the transformed data is to be interpreted, for example, “5/9/99” = 5th September 1999 or 9th May 1999 — British or US?
- Metadata give information about related data in the data warehouse.
- Metadata estimate response time by showing the number of records to be processed in a query.
- Metadata hold calculated fields and pre-calculated formulas to avoid misinterpretation, and contain historical changes of a view.

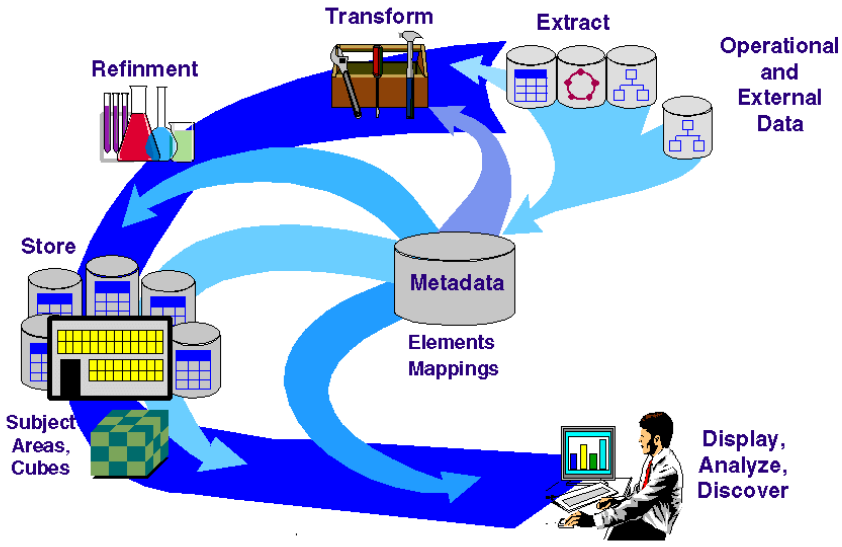


Figure 2-2 Metadata with a central role in BI

The data warehouse administrator’s perspective of metadata is a full **repository** and documentation of all contents and processes within the data warehouse; from an end user perspective, metadata is the **roadmap** through the information in the data warehouse.

Technical versus business metadata

Metadata users can be broadly placed into the categories of business users and technical users. Both of these groups contain a wide variety of users of the data warehouse metadata. They all need metadata to identify and effectively use the information in the data warehouse.

Therefore, we can distinguish between two types of metadata that the repository will contain technical and business metadata:

- Technical metadata
- Business metadata

Technical metadata provides accurate data in the data warehouse. In addition, technical metadata is absolutely critical for the ongoing maintenance and growth of the data warehouse. Without technical metadata, the task of analyzing and implementing changes to a decision support system is significantly more difficult and time consuming.

The business metadata is the link between the data warehouse and the business users. Business metadata provides these users with a road map for access to the data in the data warehouse and its datamarts. The business users are primarily executives or business analysts and tend to be less technical; therefore, they need to have the DSS system defined for them in business terms. The business metadata presents, in business terms, what reports, queries and data are in the data warehouse; location of the data; reliability of the data; context of the data, what transformation rules were applied; and from which legacy systems the data was sourced.

Types of metadata sources

There are two broad types of metadata sources — formal and informal metadata. These sources comprise the business and technical metadata for an organization.

- Formal metadata sources are those sources of metadata that have been discussed, documented and agreed upon by the decision-makers of the enterprise. Formal metadata is commonly stored in tools or documents that are maintained, distributed and recognized throughout the organization. These formal metadata sources populate both technical and business metadata.
- Informal metadata consist of corporate knowledge, policies and guidelines that are not in a standard form. This is the information that people already know. This type of information is located in the “company consciousness” or it could be on a note on a key employee's desk. It is not formally documented or agreed upon; however, this knowledge is every bit as valuable as that in the formal metadata sources. Often, informal metadata provides some of the most valuable information, because it tends to be business related. It is important to note that in many cases much of the business metadata is really informal. As a result, it is critical that this metadata is captured, documented, formalized and reflected in the data warehouse. By doing this you are taking an informal source of metadata and transforming it into a formal source. Because every organization differs, it is difficult to say where your informal sources of metadata are; however, the following is a list of the most common types of informal metadata:

- Data stewardship
- Business rules
- Business definitions
- Competitor product lists

2.2.7 Operational Data Store (ODS)

The operational data source can be defined as an updateable set of integrated data used for enterprise-wide tactical decision making. It contains live data, not snapshots, and has minimal history that is retained. Below are some features of an Operational Data Store (ODS):

An ODS is **subject oriented:** It is designed and organized around the major data subjects of a corporation, such as “customer” or “product”. They are not organized around specific applications or functions, such as “order entry” or “accounts receivable”.

An ODS is **integrated:** It represents a collectively integrated image of subject-oriented data which is pulled in from potentially any operational system. If the “customer” subject is included, then all of the “customer” information in the enterprise is considered as part of the ODS.

An ODS is **current valued:** It reflects the “current” content of its legacy source systems. “Current” may be defined in various ways for different ODSs depending on the requirements of the implementation. An ODS should not contain multiple snapshots of whatever “current” is defined to be. That is, if “current” means one accounting period, then the ODS does not include more than one accounting period’s data. The history is either archived or brought into the data warehouse for analysis.

An ODS is **volatile:** Because an ODS is current valued, it is subject to change on a frequency that supports the definition of “current.” That is, it is updated to reflect the systems that feed it in the true OLTP sense. Therefore, identical queries made at different times will likely yield different results, because the data has changed.

An ODS is **detailed:** The definition of “detailed” also depends on the business problem that is being solved by the ODS. The granularity of data in the ODS may or may not be the same as that of its source operational systems.

The features of an ODS such as subject oriented, integrated and detailed could make it very suitable to mining. These features alone do not make an ODS a good source for mining/training, because there is not enough history information.

2.3 Analytical users requirements

From the end user's perspective, the presentation and analysis layer is the most important component in the BI architecture.

Depending on the user's role in the business, their requirements for information and analysis capabilities will differ. Typically, the following user types are present in a business:

- The "non-frequent user"

  This user group consists of people who are not interested in data warehouse details but have a requirement to get access to the information from time to time. These users are usually involved in the day-to-day business and do not have time or any requirements to work extensively with the information in the data warehouse. Their virtuosity in handling reporting and analysis tools is limited.

- Users requiring up-to-date information in predefined reports

  This user group has a specific interest in retrieving precisely defined numbers in a given time interval, such as:

  "I have to get this quality-summary report every Friday at 10:00 AM as preparation to our weekly meeting and for documentation purposes."

- Users requiring dynamic or ad hoc query and analysis capabilities

  Typically, this is the business analyst. All the information in the data warehouse may be of importance to these users, at some point in time. Their focus is related to availability, performance, and drill-down capabilities to "slice and dice" through the data from different perspectives at any time.

- The advanced business analyst — the "power user"

  This is a professional business analyst. All the data from the data warehouse is potentially important to these users. They typically require separate specialized datamarts for doing specialized analysis on preprocessed data. Examples of these are data mining analysts and advanced OLAP users.

Different user-types need different front-end tools, but all can access the same data warehouse architecture. Also, the different skill levels require a different visualization of the result, such as graphics for a high-level presentation or tables for further analysis.

In the remainder of this chapter we introduce the different types of tools that are typically used to leverage the information in a data warehouse.

2.3.1 Reporting and query

Creating reports is a traditional way of distributing information in an organization. Reporting is typically static figures and tables that are produced and distributed with regular time intervals or for a specific request. Using an automatic reporting tool is an efficient way of distributing the information in your data warehouse through the Web or e-mails to the large number of users, internal or external to your company, that will benefit from information.

Users that require the ability to create their own reports on the fly or wish to elaborate on the data in existing reports will use a combined querying and reporting tool. By allowing business users to design their own reports and queries, a big workload from an analysis department can be removed and valuable information can become accessible to a large number of (non-technical) employees and customers resulting in business benefit for your company. In contrast to traditional reporting this also allows your business users to always have access to up-to-date information about your business. This thereby also enables them to provide quick answers to customer questions.

As the reports are based on the data in your data warehouse they supply a 360 degree view of your company's interaction with its customers by combining data from multiple data sources. An example of this is the review of a client’s history by combining data from: ordering, shipping, invoicing, payment, and support history.

Query and reporting tools are typically based on data in relational databases and are not optimized to deliver the “speed of thought” answers to complex queries on large amounts of data that is required by advanced analysts. An OLAP tool will allow this functionality at the cost of increased load time and management effort.

2.3.2 On-Line Analytical Processing (OLAP)

During the last ten years, a significant percentage of corporate data has migrated to relational databases. Relational databases have been used heavily in the areas of operations and control, with a particular emphasis on transaction processing (for example, manufacturing process control, brokerage trading). To be successful in this arena, relational database vendors place a premium on the highly efficient execution of a large number of small transactions and near fault tolerant availability of data.

More recently, relational database vendors have also sold their databases as tools for building data warehouses. A data warehouse stores tactical information that answers “who?” and “what?” questions about past events. A typical query submitted to a data warehouse is: “What was the total revenue for the eastern region in the third quarter?”

It is important to distinguish between the capabilities of a data warehouse from those of an On-Line Analytical Processing (OLAP) system. In contrast to a data warehouse — that is usually based on relational technology — OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis.

OLAP enables analysts, managers, and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information. OLAP transforms raw data so that it reflects the real dimensionality of the enterprise as understood by the user.

While OLAP systems have the ability to answer “who?” and “what?” questions, it is their ability to answer “what if?” and “why?” that sets them apart from data warehouses. OLAP enables decision making about future actions.

A typical OLAP calculation is more complex than simply summing data, for example: “What would be the effect on soft drink costs to distributors if syrup prices went up by $.10/gallon and transportation costs went down by $.05/mile?”

OLAP and data warehouses are complementary. A data warehouse stores and manages data. OLAP transforms data warehouse data into strategic information.

OLAP ranges from basic navigation and browsing (often known as “slice” and “dice”) to calculations, to more serious analyses, such as time series and complex modeling. As decision makers exercise more advanced OLAP capabilities, they move from data access to information to knowledge.

2.3.3 Who uses OLAP and why?

OLAP applications span a variety of organizational functions. Finance departments use OLAP for applications, such as budgeting, activity-based costing (allocations), financial performance analysis, and financial modeling. Sales analysis and forecasting are two of the OLAP applications found in sales departments. Among other applications, marketing departments use OLAP for market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation. Typical manufacturing OLAP applications include production planning and defect analysis.

Important to all of the above applications is the ability to provide managers with the information they need to make effective decisions about an organization's strategic directions. The key indicator of a successful OLAP application is its ability to provide information as needed, that is, its ability to provide “just-in-time” information for effective decision-making. This requires more than a base level of detailed data.

Just-in-time information is computed data that usually reflects complex relationships and is often calculated on the fly. Analyzing and modeling complex relationships are practical only if response times are consistently short. In addition, because the nature of data relationships may not be known in advance, the data model must be flexible. A truly flexible data model ensures that OLAP systems can respond to changing business requirements as needed for effective decision making.

Although OLAP applications are found in widely divergent functional areas, they all require the following key features:

- Multidimensional views of data
- Calculation-intensive capabilities
- Time intelligence

Multidimensional views

Multidimensional views are inherently representative of an actual business model. Rarely is a business model limited to fewer than three dimensions. Managers typically look at financial data by scenario (for example, actual versus budget), organization, line items, and time; and at sales data by product, geography, channel, and time.

A multidimensional view of data provides more than the ability to “slice and dice”; it provides the foundation for analytical processing through flexible access to information. Database design should not prejudice which operations can be performed on a dimension or how rapidly those operations are performed. Managers must be able to analyze data across any dimension, at any level of aggregation, with equal functionality and ease. OLAP software should support these views of data in a natural and responsive fashion, insulating users of the information from complex query syntax. After all, managers should not have to understand complex table layouts, elaborate table joins, and summary tables.

Whether a request is for the weekly sales of a product across all geographical areas or the year-to-date sales in a city across all products, an OLAP system must have consistent response times. Managers should not be penalized for the complexity of their queries in either the effort required to form a query or the amount of time required to receive an answer.

Calculation-intensive capabilities

The real test of an OLAP database is its ability to perform complex calculations. OLAP databases must be able to do more than simple aggregation. While aggregation along a hierarchy is important, there is more to analysis than simple data roll-ups. Examples of more complex calculations include share calculations (percentage of total) and allocations (which use hierarchies from a top-down perspective).

Key performance indicators often require involved algebraic equations. Sales forecasting uses trend algorithms, such as moving averages and percentage growth. Analyzing the sales and promotions of a given company and its competitors requires modeling complex relationships among the players. The real world is complicated — the ability to model complex relationships is key in analytical processing applications.

Time intelligence

Time is an integral component of almost any analytical application. Time is a unique dimension, because it is sequential in character (January always comes before February). True OLAP systems understand the sequential nature of time. Business performance is almost always judged over time, for example, this month versus last month, this month versus the same month last year.

The time hierarchy is not always used in the same manner as other hierarchies. For example, a manager may ask to see the sales for May or the sales for the first five months of 1995. The same manager may also ask to see the sales for blue shirts but would never ask to see the sales for the first five shirts. Concepts such as year-to-date and period over period comparisons must be easily defined in an OLAP system.

In addition, OLAP systems must understand the concept of balances over time. For example, if a company sold 10 shirts in January, five shirts in February, and 10 shirts in March, then the total balance sold for the quarter would be 25 shirts. If, on the other hand, a company had a head count of 10 employees in January, only five employees in February, and 10 employees again in March, what was the company's employee head count for the quarter? Most companies would use an average balance. In the case of cash, most companies use an ending balance.

2.3.4 Statistics

Statistical tools are typically used to address the business problem of generating an overview of the data in your database. This is done by using techniques that summarize information about the data into statistical measures that can be interpreted without requiring every record in the database to be understood in detail (for example, the application of statistical functions like finding the maximum or minimum, the mean, or the variance). The interpretation of the derived measures require a certain level of statistical knowledge.

These are typical business questions addressed by statistics:

- What is a high-level summary of the data that gives me some idea of what is contained in my database?
- Are their apparent dependencies between variables and records in my database?
- What is the probability that an event will occur?
- Which patterns in the data are significant?

To answer these questions the following statistical methods are typically used:

- Correlation analysis
- Factor analysis
- Regression analysis

These functions are detailed in 8.2.2, “Statistical functions” on page 171.

2.3.5 Data mining

However, in contrast with statistical analysis, data mining analyzes all the relevant data in your database and extracts hidden patterns.

Data mining is to some extent based on the techniques and disciplines used in statistical analysis. However, the algorithms used in data mining automate many of the tedious procedures that you would need to go through to obtain the same depth of analysis using traditional statistical analysis.

An introduction to data mining is given in Chapter 3, “A generic data mining method” on page 23.

2.4 Data warehouse, OLAP and data mining summary

If the recommended method when building data warehouses and OLAP datamarts is:

- To build an ODS where you collect and cleanse data from OLTP systems.
- To build a star schema data warehouse with fact table and dimensions tables.
- To use data in the data warehouse to build an OLAP datamart.

Then, the recommended method for building data warehouses and data mining datamarts could be quite the same:

- To build an ODS where you collect and cleanse data from OLTP systems
- To build a star schema data warehouse with fact table and dimensions tables
- To pick the dimension which is of main interest, for example, customers — to use aggregation and pivot on the fact table and maybe one or two other dimensions in order to build a flat record schema or a datamart for the mining techniques.

As a star schema model or multidimensional model, a data warehouse should be a prerequisite for OLAP datamarts, even if it is not a prerequisite for a data mining project, it may help as a design guideline.

OLAP and data mining projects could use the same infrastructure. The construction of the star schema and extracting/transforming/loading steps to build the data warehouse are the responsibilities of the IT department. An IT department should of course take into account the business users' requirements on OLAP as cubes or multidimensional databases, reports, and also data mining models to design the data warehouse.

OLAP and data mining can use the same data, the same concepts, the same metadata and also the same tools, perform in synergy, and benefit from each other by integrating their results in the data warehouse.

3

A generic data mining method

Data mining is one of the main applications that are available to you as part of your overall BI architecture. You may already use a number of analysis and reporting tools to provide you with the day to day information you need. So why is data mining different from the normal types of statistical analysis and other business reporting tools that you use?

In this chapter we describe what data mining is all about and describe some of the things that you can do with the tools and techniques that data mining provides. Gaining an understanding of what data mining can do will help you to see the types of business questions that you can address and how you can take the first steps along the road of mining your own business. To help in this respect we have developed a generic data mining method that you can use as a basic guide. The generic method is explained and in the following chapters we will show how it can be applied to address specific retail business issues.

3.1 What is data mining?

Data mining is treated by many people as more of a philosophy, or a subgroup of mathematics, rather than a practical solution to business problems. You can see this by the variety of definitions that are used, for example:

> *"Data mining is the exploration and analysis of very large data with automatically or semi-automatically procedures for previously unknown, interesting, and comprehensible dependencies"*
>
> Or
>
> *"Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."*

While these definitions have their place, in this book we want to concentrate on the practical issues of data mining and show how to make data mining work for your retail business. Specifically, we want to show you what you have to do to successfully mine your own business and to end up with reliable results that you can put to use.

Although data mining as a subject in its own right, it has only existed for less than 10 years, and its origins can be traced to the early developments in artificial intelligence in the 1950's. During this period, developments in pattern recognition and rule based reasoning were providing the fundamental building blocks on which data mining was to be based. Since this time, although they were not given the title of data mining, many of the techniques that we use today have been in continuous use, primarily for scientific applications.

With the advent of the relational database and the capability for commercial organizations to capture and store larger and larger volumes of data, it was released that a number of the techniques that were being used for scientific applications could be applied in a commercial environment and that business benefits could be derived. The term data mining was coined as a phrase to encompass these different techniques when applied to very large volumes of data. Figure 3-1 shows the developments that have taken place over the past 40 years.

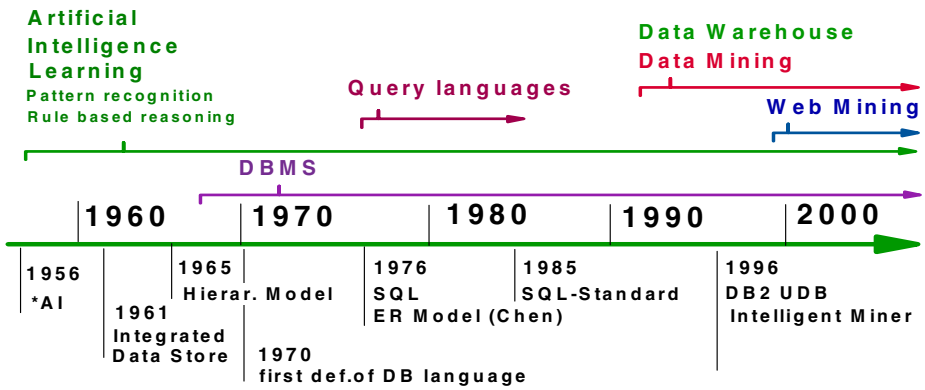


Figure 3-1 A historical view of data mining

Some of the techniques that are used to perform data mining are computationally complex, and in order to discover the patterns existing within large data sets they have to perform a large number of computations. In the last 10 years the growth in the use of large commercial databases (specifically data warehouse) coupled with the need to understand and interpret this data and the availability of relatively inexpensive computers has lead to an explosion in the use of data mining for a wide variety of commercial applications.

3.2 What is new with data mining?

Data mining is about discovering new things about your business from the data you have collected. You may think that you already do this using standard statistical techniques to explore your database. In reality what you are normally doing is making a hypothesis about the business issue that you are addressing and then attempting to prove or disprove your hypothesis by looking for data to support or contradict the hypothesis.

For example, suppose that as a retailer, you believe that customers from “out of town” visit your larger inner city stores less often than other customers, but when they do so they make larger purchases. To answer this type of question you can simply formulate a database query looking, for example, at your branches, their locations, sales figures, customers and then compile the necessary information (average spend per visit for each customer) to prove your hypotheses. However, the answer discovered may only be true for a small highly profitable group of out-of-town shoppers who visited inner-city stores at the weekend. At the same

time, out-of-town customers (perhaps commuters) may visit the store during the week and spend exactly the same way as your other customers. In this case, your initial hypothesis test may indicate that there is no difference between out-of-town and inner-city shoppers.

Data mining uses an alternative approach beginning with the premise that you do not know what patterns of customer behaviors exist. In this case you may simply ask the question, what are their relationships (we sometimes use the term *correlations*) between what my customers spend and where they come from? In this case, you would leave it up to the data mining algorithm to tell you about all of the different types of customers that you had. This should include the out-of-town, weekend shopper. Data mining therefore provides answers, without you having to ask specific questions.

The difference between the two approaches is summarized in Figure 3-2.

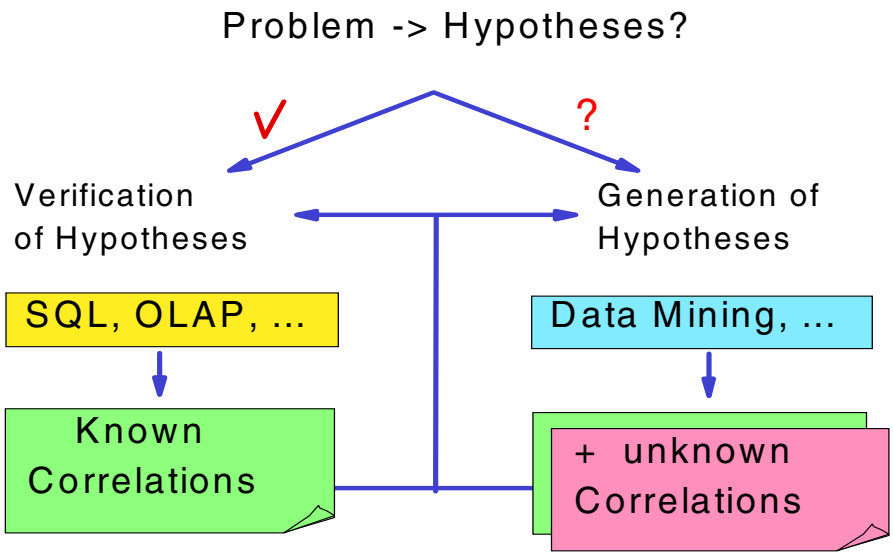


Figure 3-2 Standard and data mining approach on information detection

So how do you set about getting the answers to the sorts of business issues that data mining can address. This is usually a complex issue, but that is why we have written this book. To help in this regard, we follow a **generic method** that can be applied to a wide range of business questions, and in the following chapters we show how it can be applied to solve chosen business issues.

3.3 Data mining techniques

As we explained previously, a variety of techniques have been developed over the years to explore for and extract information from large data sets. When the name data mining was coined, many of these techniques were simply grouped together under this general heading and this has led to some confusion about what data mining is all about. In this section we try to clarify some of the confusion.

3.3.1 Types of techniques

In general, data mining techniques can be divided into two broad categories:

- Discovery data mining
- Predictive data mining

Discovery data mining

Discovery data mining is applied to a range of techniques which find patterns inside your data without any prior knowledge of what patterns exist. The following are examples of discovery mining techniques:

Clustering

Clustering is the term for a range of techniques which attempts to group data records on the basis of how similar they are. A data record may, for example, comprise a description of each of your customers. In this case clustering would group similar customers together, while at the same time maximizing the differences between the different customer groups formed in this way. As we will see in the examples described in this book, there are a number of different clustering techniques, and each technique has its own approach to discovering the clusters that exist in your data.

Link analysis

Link analysis describes a family of techniques that determines associations between data records. The most well known type of link analysis is market basket analysis. In this case the data records are the items purchased by a customer during the same transaction and because the technique is derived from the analysis of supermarket data, these are designated as being in the same basket. Market basket analysis discovers the combinations of items that are purchased by different customers, and by association (or linkage) you can build up a picture of which types of product are purchased together. Link analysis is not restricted to market basket analysis. If you think of the market basket as a grouping of data records then the technique can be used in any situation where there are a large number of groups of data records.

Frequency analysis

Frequency analysis comprises those data mining techniques that are applied to the analysis of time ordered data records or indeed any data set that can be considered to be ordered. These data mining techniques attempt to detect similar sequences or subsequences in the ordered data.

Predictive Mining

Predictive data mining is applied to a range of techniques that find relationships between a specific variable (called the *target variable*) and the other variables in your data. The following are examples of predictive mining techniques.

Classification

Classification is about assigning data records into pre-defined categories. For example, assigning customers to market segments. In this case the target variable is the category and the techniques discover the relationship between the other variables and the category. When a new record is to be classified, the technique determines the category and the probability that the record belongs to the category. Classification techniques include decision trees, neural and Radial Basis Functions (RBF) classifiers.

Value prediction

Value prediction is about predicting the value of a continuous variable from the other variables in a data record. For example, predicting the likely expenditure of a customer from their age, gender and income group. The most familiar value prediction techniques include linear and polynomial regression, and data mining extends these to other techniques, such as neural and RBF value prediction.

3.3.2 Different applications that data mining can be used for

There are many types of applications to which data mining can be applied. In general, other than for the simplest applications, it is usual to combine the different mining techniques to address particular business issues. In Figure 3-3 we illustrate some of the types of applications, drawn from a range of industry sectors, where data mining has been used in this way. These applications range from customer segmentation and market basket analysis in retail, to risk analysis and fraud detection in banking and financial applications.

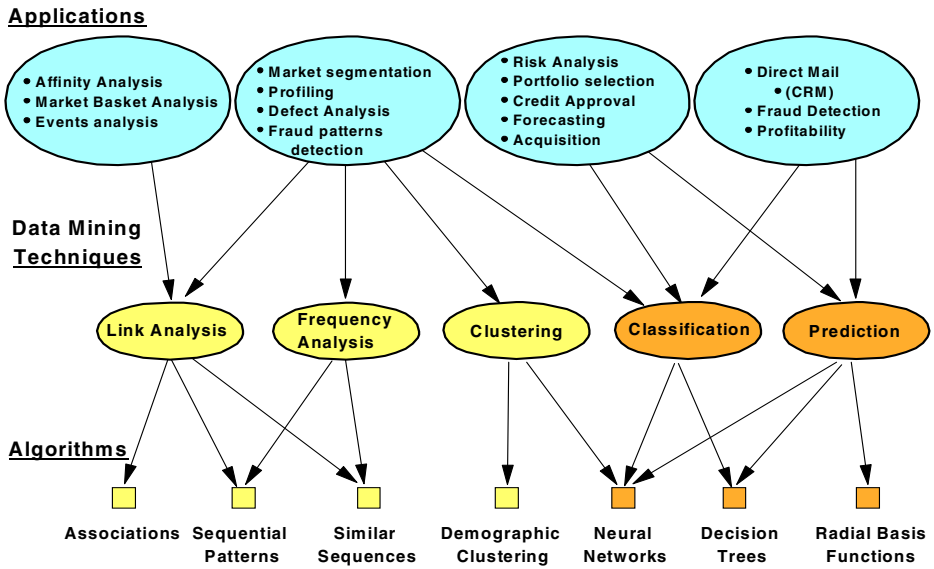


Figure 3-3 From applications to algorithms

Because there is such a bewildering range of things that can be done using data mining, our objective is to concentrate on some of the techniques that apply to your specific industry sector. To assist you in the process of understanding how to decide what techniques are applicable and how to set about the whole process of mining you own business, we have developed a generic data mining method. The objective is to define a sequence of steps that can be followed for all data mining activities, and in subsequent chapters we show how the method is applied to address specific business issues.

3.4 The generic data mining method

This section describes the generic data mining method which comprises seven steps; these are:

- Defining the business issue in a precise statement
- Defining the data model and the data requirements
- Sourcing data from all available repositories and preparing the data (the data could be relational or in flat files, stored in a data warehouse, computed, created on-site or bought from another party. They should be selected and filtered from redundant information)
- Evaluating the data quality

- Choosing the mining function and defining the mining run
- Interpreting the results and detecting new information
- Deploying the results and the new knowledge into your business

These steps are illustrated in Figure 3-4 and the following sections expand on each of the stages.

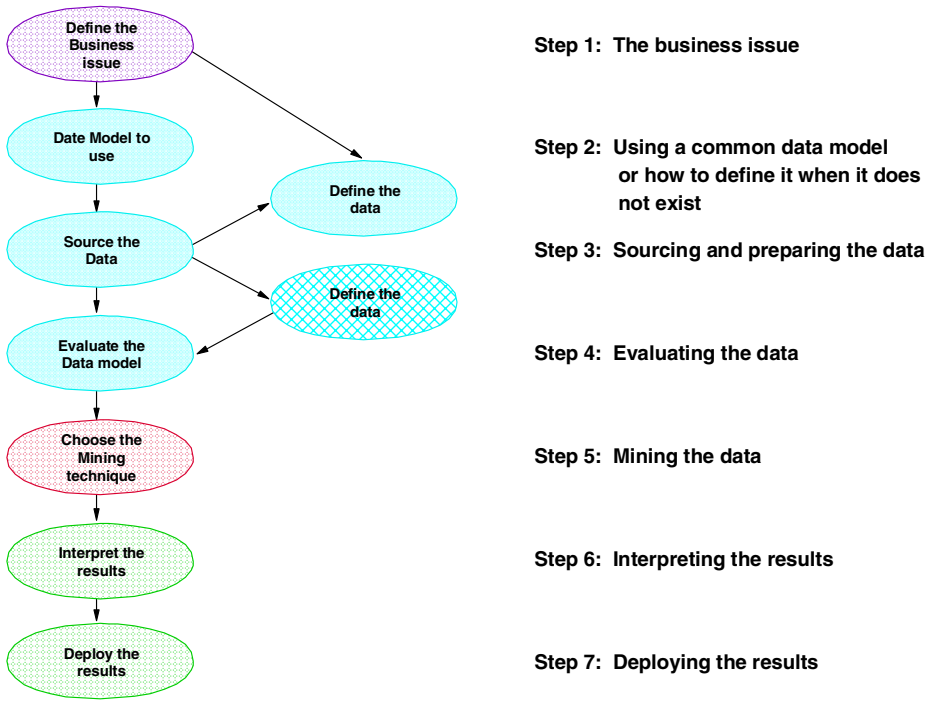


Figure 3-4 The generic method

3.4.1 Step 1 — Defining the business issue

All to often, organizations approach data mining from the perspective that there must be some value in the data we have collected, so we will just use data mining and discover what's there. Using the mining analogy, this is rather like choosing a spot at random and starting to dig for gold. This may be a good strategy if the gold is in large nuggets, and you were lucky enough to choose the right spot to dig, but if the gold could only be extracted by "panning" or some other technique, you may spend a lot of time throwing away the valuable material in a fruitless search, searching for the right thing at the wrong place or with the wrong technique.

Data mining is about choosing the right tools for the job and then using them skillfully to discover the information in your data. We have already seen there are a number of tools that can be used, and that very often we have to use a combination of the tools at our disposal, if we are to make real discoveries and extract the value from our data.

The *first step in our data mining method* is therefore to identify the business issue that you want to address and then determine how the business issue can be translated into a question, or set of questions, that data mining can address.

By *business issue* we mean that there is an identified problem to which you need an answer, where you suspect, or know, that the answer is buried somewhere in the data, but you are not sure where it is.

A business issue should fulfill the requirements of having:

- A clear description of the problem to be addressed
- An understanding of the data that may be relevant
- A vision for how you are going use the mining results in your business

Describing the problem

If you are not sure what questions data mining can address, then the best approach is to look at examples of where it has been successfully used, either in your own industry or in related industries. Many business and research fields have been proven to be excellent candidates for data mining. The major fraction are covered by banking, insurance, retail and telecommunications (telecoms), but there are many others such as manufacturing, pharmaceuticals, biotechnology and so on, where significant benefits have also been derived. Well-known approaches are: customer profiling and cross-selling in retail, loan delinquency and fraud detection in banking and finance, customer retention (attrition and churn) in telecoms, patient profiling and weight rating for Diagnosis Related Groups in health care and so on. Some of these are depicted in Figure 3-5.

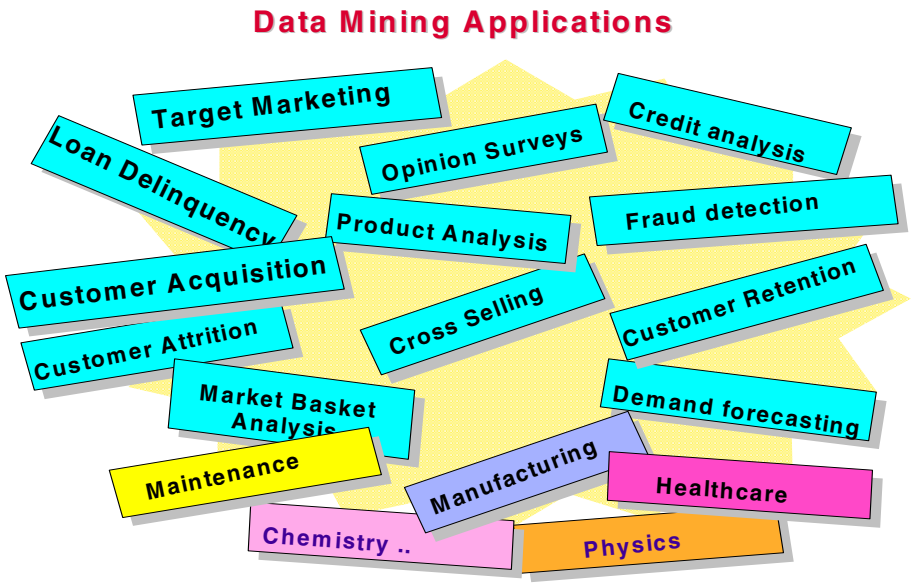


Figure 3-5 Business and research applications

The objective behind this book and others in the series is to describe some of these different issues and show how data mining can be used to address them.

Even where the specific business issue you are trying to address has not been addressed elsewhere, understanding how data mining can be applied will help to define your issue in terms that data mining can answer. You need to remember that data mining is about the discovery of patterns and relationships in your data. All of the different applications are using the same data mining concepts and applying them in subtly different ways.

With this in mind, when you come to define the business issue, you should think about it in terms of patterns and relationships. Take fraud as an example. Rather than ask the question can we detect fraudulent customers, you could ask the question, can we detect a small group of customers who exhibit unusual characteristics that may be indicative of fraud? Alternatively, if you have identified some customers who are behaving fraudulently, the question is, can you identify some unique characteristics of these customers that would enable you to identify other potentially fraudulent customers?

Understanding your data

As you are formulating the business question, you need to also think about whether the data that you have available is going to be sufficient to answer the question. It is important to recognize that the data you hold may not contain the information required to enable you to answer the question you are posing. For example, we suppose you are trying to determine why you are losing customers and the reason is that your competitors are undercutting you on price. If you do not have competitor pricing data in your database, then clearly data mining is not going to provide the answer. Although this is a trivial example, sometimes it is not so obvious that the data cannot provide the answer you are looking for. The amazing thing is how many people still believe that data mining should be able to perform the impossible.

Where the specific business issue has been addressed elsewhere, then knowing what data was used to address the issue will help you to decide which of your own data should be used and how it may need to be transformed before it can be effectively mined. This process is termed the construction of a common data model. The use of common data models is a very powerful aid to performing data mining as we will show when we address specific business issues.

Using the results

When defining the business issue that you want to address with data mining, it is important that you think carefully about how you are going to use the information that you discover. Very often, by considering how you are going to deploy the results of your data mining into your business, will help to clarify the business issue you are trying to address and determine what data you are going to use.

Suppose for example, that you want to use data mining to identify which types of existing customers will respond to new offers or services and then use the results to target new customers. Clearly the variables you use when doing the data mining on your existing customers, must be the same variables that you can derive about your new customers. In this case you cannot use the 6-month aggregated expenditure (*aggregated spend*) on particular products if all you have available for new customers is the expenditure from a single transaction. Thinking about how you are going to use the information you derive places constraints on the selection of data that you can use to perform the data mining and is therefore a key element in the overall process of translating the business issue into a data mining question.

3.4.2 Step 2 — Defining a data model to use

The *second step in the generic data mining method* is to define the data to be used. A business like yours can collect and store vast amounts of data. Usually the data is being stored to support various applications and as we considered in 2.1, “Business Intelligence” on page 8, the best way to do this is to use some form of data warehouse. Although not all data warehouse architectures are the same, one way they can be used efficiently to support your applications is shown in Figure 3-6. In this case each end user application is supported by its own datamart which is updated at regular intervals or when specific data changes, to reflect the needs of the application.

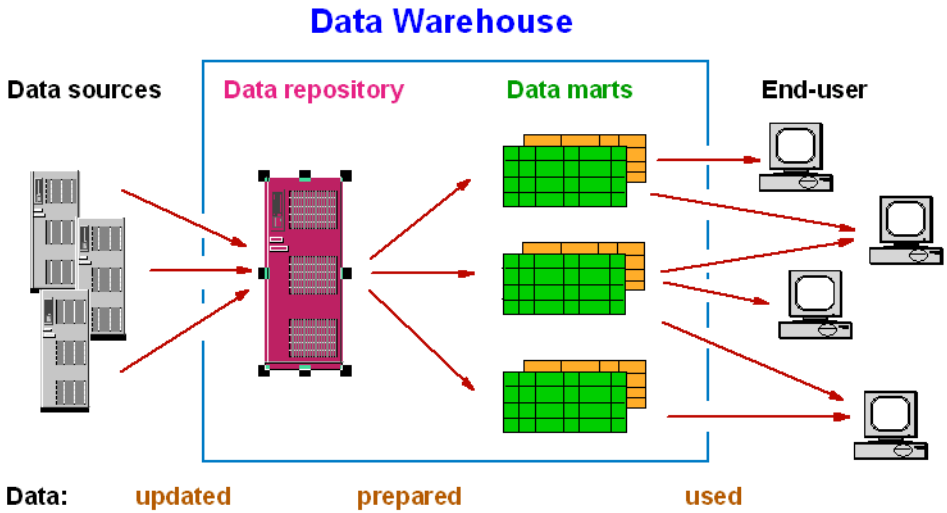


Figure 3-6 Data warehouse architecture

In this structure each datamart has its own specific data and holds knowledge about how the data was derived, the data format used, what aggregations have been performed, what data cleansing has been done and so on. In 2.2.6, “Metadata” on page 12, we described this additional information as metadata. Where the data is being used routinely to support a specific business application, the data and metadata together form what we call a *data model* that supports the application.

Typically the data model will define:

- Data sources used
- Data types
- Data content
- Data description

- Data usage

The *data sources* indicate the physical location from where the data is derived or stored. The *data type* defines how the data is structured (for example, the date time format used). The *data content* lists the tables or data files and the fields which they contain. The *data description* delivers the names and description of these fields. The *data usage* considers the ownership of tables and fields, how users understand their content, and, although often neglected, how the users exploit them. The data model also contains information on when the data is valid, when it must be replicated and so on.

Data mining is just another application and, depending on what it is being used for, requires its own data model. For most data mining applications, the data model required is in the form of a single file or database table, with one record per customer, or department, or whatever is the target of the investigation.

Note: In database terms the required table is called a denormalized table and can be either a physical table in the database or a database 'view' of joined tables each comprising some of the variables consisting of one or more variables

Each record can comprise one or many variables, where each variable may be derived from a number of different data sources but are tied to the same target variable (for example, the customer) in some way. In most business applications the most common data types are:

- Transactional data
- Relationship data
- Demographic data

Transactional data is operational data generated each time some interaction occurs with the target. This data typically contains a timestamp and some transaction identifier together with details of the transaction. This type of data may, for example, relate to point of sales data for a customer in a supermarket, or to the recording of information on a production line in a manufacturing application.

Relationship data is the nonvolatile data containing relatively stable information about customers, products, equipment, items, and working processes.

Demographic data comprises person-specific (customer, patient) data usually coming from external sources. Typically this includes information on age, gender postal code and so on.

Use of common data models

Defining data models for any application is often a complex task and defining data models for data mining is no exception. Where the data model is required to support an application that has specific requirements (for example, some form of business reporting tool) then the data can be defined by asking the end users what types of information they require and then performing the necessary aggregations to support this requirement. In the case of data mining, the challenge is that very often you are not sure at the outset which variables are important and therefore exactly what is required. Generating data models for completely new data mining applications can therefore be a time consuming activity.

The alternative is to use common data models that have been developed to solve similar business issues to the ones you are trying to address. While these types of models may not initially provide you with all of the information you require, they are usually designed to be extendable to include additional variables. The main advantage of using a common data model is that it will provide you with a way of quickly seeing how data mining can be used within your business. In the following chapters we suggest some simple data models that can be used in this way.

3.4.3 Step 3 — Sourcing and preprocessing the data

The *third step in the generic data mining method* is the sourcing and preprocessing of the data that populates the data model. Having a defined data model provides the necessary structure, in terms of the variables that we are going to mine, but we still have to provide the data.

Data sourcing and preprocessing comprises the stages of *identifying, collecting, filtering* and *aggregating* (raw) data into a format required by the data models and the selected mining function. Because sourcing and preparing the data are the most time consuming parts of any data mining project, we describe these crucial steps in broader detail. Where the data is derived from a data warehouse, many of these stages will already have been performed.

The data sources

The data sources can be different by origin and content as shown in Figure 3-7.

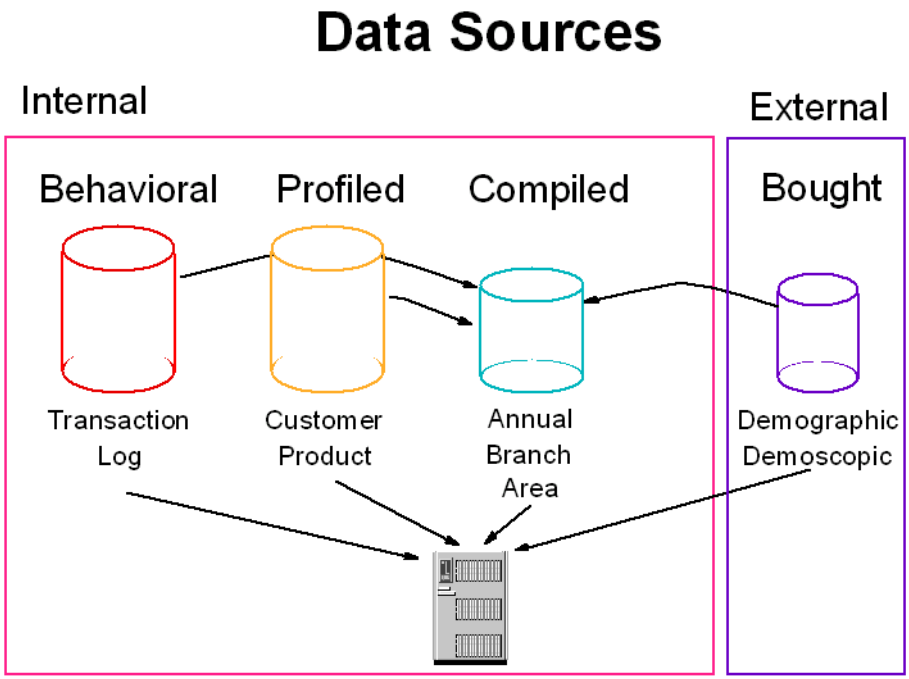


Figure 3-7 Data sources by origin and content

Every business uses standard internal data sources. Many of them are similar from their point of content. Therefore, a *customer database* or a *product database* could be found in nearly any data scenario.

Data mining, in common with many other analysis tools, usually requires the data to be in one consolidated table or file. If the variables required are distributed across a number of sources then this consolidation must be performed such that a consistent set of data records is produced. If the data is stored in relational database tables then the creation of a new tables or a database view is relatively straight forward, although where complex aggregations are required this can have a significant impact on database resources.

Data preprocessing

If the data has not been derived from a data warehouse then the data preprocessing functions of cleansing, aggregated, transforming, and filtering, that we described in 2.2, “Data warehouse” on page 8, must be undertaken. Even when the data is derived from a data warehouse, there may still be some additional transformations of the data that need to be performed before mining can proceed. Structurally the *data preprocessing* can be displayed as in Figure 3-8.

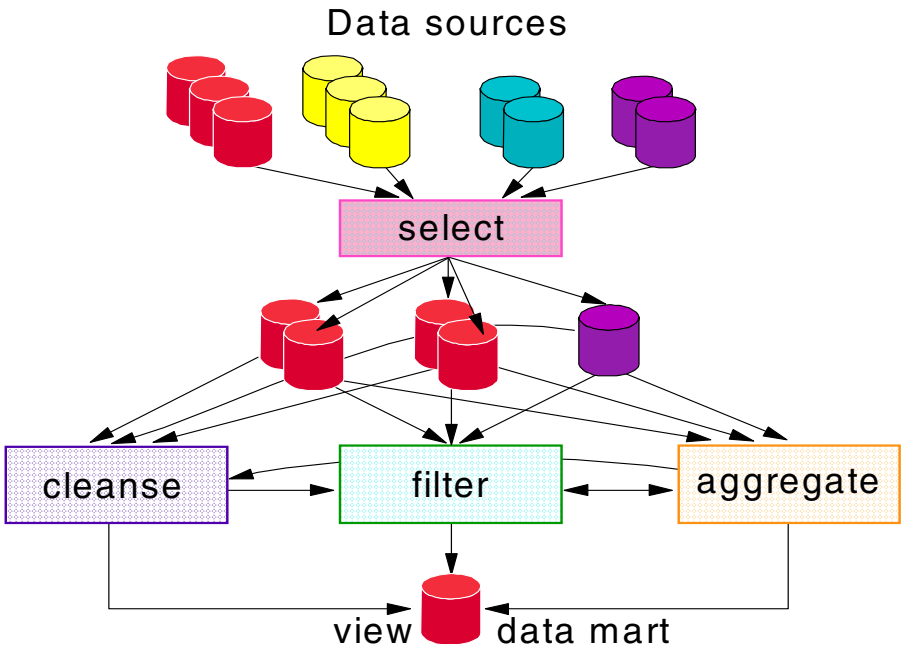


Figure 3-8 Data preprocessing

Data mining tools usually provide limited capability to cleanse the data, because this is a specialized process and there are a number of products that can be used to do this efficiently. Aggregation and filtering can be performed in a number of different ways depending on the precise structure of your data sources. Some of the tools available to do this with the IM for Data product are described in 8.2.1, “Data preparation functions” on page 169.

3.4.4 Step 4 — Evaluating the data model

Having populated the data model with data we still have ensure that the data used to populate our data model fulfills the requirement of completeness, exactness and relevance. To do this we perform the *fourth step in the generic data mining method,* which is to perform an initial evaluation; the steps are described in Figure 3-9.

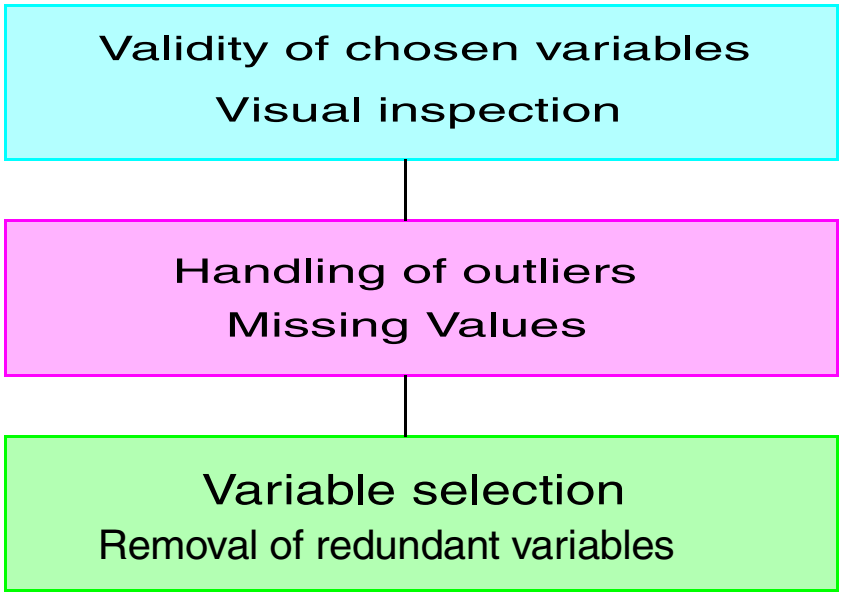


Figure 3-9 Overview: Steps of data evaluation

The first step *visual inspection* comprises browsing the input data with visualizing tools. This may lead to the detection of *implausible distributions*. For example, a wrong join of tables during the data preparation step could result in variables containing values actually belonging to different fields.

The second step deals with the *identification of inconsistencies* and the *resolution of errors*. Unusual distributions found within the first step could be induced by a badly done data collection. Many *outlying and missing values* produce biased results. For example, the interesting information of a correlation between variables indicating the behavior of a customer group and variables describing new offerings could be completely suppressed if too many missing values are accepted. The mitigation of outliers and the transformation of missing values in meaningful information however improves the data quality enormously. Many data mining functions take account of a minor fraction of missing values. But an early treatment of missing values can prevent us from biased mining results and unconsciously progressed errors.

The last step is the final *selection of features/variables* for the mining run. Variables could be superfluous by presenting the same or very similar information as others, but increasing the run time. *Dependent or highly correlated variables* could be found with statistical tests like bivariate statistics, linear and polynomial regression. Dependent variables should be reduced by selecting one variable for all others or by composing a new variable for all correlated ones by factor or component analysis.

Not all variables remaining after the statistical check are nominated as input; only variables with a clear interpretation and variables that make sense for the end user should be selected. A proven data model simplifies this step. The selection of variables in that stage can indeed only be undertaken with practical experience in the respective business or research area.

3.4.5 Step 5 — Choosing the data mining technique

Besides the steps defining business issues, data modeling, and preparation, data mining also comprises the crucial step of the selection of the best suited mining technique for a given business issue. This is the *fifth step in the generic data mining method*. This step not only includes defining the appropriate technique or mix of techniques to use, but also the way in which the techniques should be applied.

Several types of techniques (or algorithms) are available:

- Classification
- Associations
- Clustering
- Value prediction
- Similar patterns
- Similar time sequences

Others have been developed for the detection of different kinds of correlations and patterns inside databases.

The selection of the method to use will often be obvious, for example, market basket analysis in retail will use the associations technique which was originally developed for this purpose. However, the associations technique could be applied to a diverse range of applications, for example, to discover the relationships between faults occurring in production runs and the sources from which the components were derived.

The challenge is usually not which technique to use but the way in which the technique should applied. Because all of the data mining techniques require some form of parameter selection, this then requires experience of how the techniques work and what the various parameters do. In the examples given in the following chapters, we describe some of the parameters that need to be defined and how this can be done.

3.4.6 Step 6 — Interpreting the results

Interpreting the results is *the sixth step in the generic mining method*. The results from performing any type of data mining can provide a wealth of information that can sometimes be difficult to interpret. Our experience is that the interpretation phase requires the input from a business expert who can translate the mining results back into the business context. Because we do not expect the business analyst to necessarily be a data mining specialist, it is important that the results are presented in such a way that they are relatively easy to interpret.

To assist in the interpretation process, it is necessary to have at your disposal a range of tools that enable you to visualize the results and to provide the necessary statistical information that you need to make the interpretation.

In the following chapters, we provide you with a number of examples of the types of visualization techniques that are available and how to understand what the different results are telling you about your business.

3.4.7 Step 7 — Deploying the results

The *seventh and final step in the generic data mining* method is perhaps the most important of all. It deals with the question of how to deploy the results of the data mining into your business. If you only see data mining as an analytical tool, then you are failing to realize the full potential of what data mining has to offer.

As we have already explained, when you perform data mining you can both discover new things about your customers and determine how to classify or how to predict particular characteristics or attributes. In all these cases data mining creates mathematical representations of the data that we call models. These models are very important, because they not only provide you with a deeper insight of your business but can themselves be deployed in or used by other business processes, for example, your CRM systems.

When embarking on any data mining activity you should think carefully about the way in which you intend to use the data mining results and where in your business the results will have the greatest impact. In the following chapters we describe some of the ways in which both data mining results and data mining models can be used.

One particular important development in regard to the deployment of the data mining results is the development of standards for exchanging data mining models and of being able to deploy these models directly into relational databases, such as DB2 Universal Database and ORACLE. The new standard is based on what is called the Predictive Model Markup Language (PMML). This standard provides for the exchange of analytic models like linear regression, clustering, decision tree, and neural network. The most important advantages are:

- Data miner experts on-site are not necessary
- Computational resources are exploited efficiently
- It allows real time (event-triggered) processing and batch processing
- It enables foreign software applications access to the modeling logic
- It allows the generalization of any future model type

Further details on DB2 IM Scoring are presented in 8.3, “DB2 Intelligent Miner Scoring” on page 175.

3.4.8 Skills required

To successfully implement a data mining project using the above method, you will require a mix of skills in the following areas:

- Data manipulation (for example SQL)
- Knowledge of mining techniques
- Domain knowledge or ability to communicate with domain experts
- Creativity

These skills are normally not being incorporated in one person and Figure 3-10 shows the structure of a typical data mining team.

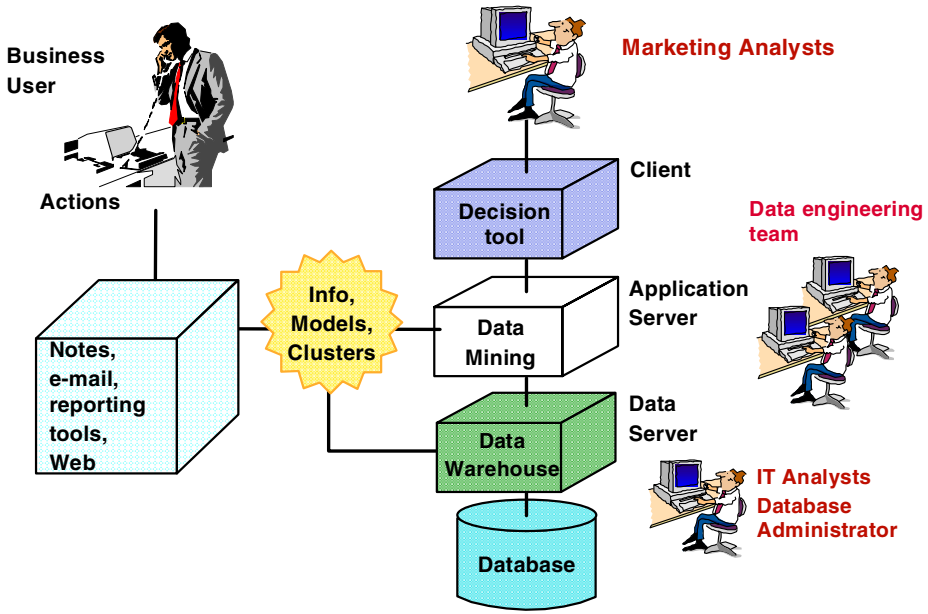


Figure 3-10 The data mining team

Such a team will comprise the following:

- *Marketing analyst* who is informed in the branches of businesses which have been selected for data mining.
- *IT analyst* is needed who is experienced with data management procedures within your business.
- *Data engineering team* who will have the lead and the experience in all data mining topics.
- *Business user* should be selected who can check the usability of the mining result and evaluate the deployment from a solely business perspective. It should be mentioned that not few data mining projects run into problems by underestimating the efforts of searching and delivering raw data in a reasonable quality.
- *Project owner* who is normally the head of a branch inside the company, should support the work and help to resolve problems.

Whether or not these are different individuals clearly depends on the mix of skills that they have, but in general the team must be able to accomplish the following:

- Understanding the data source: There are two aspects of understanding the data source, which are the knowledge about the physical data situation in the company and the usage of the data proposed for data mining. Normally, the data mining expert is not the data administrator, who is responsible for the all data repositories. In this case the interaction with the database owner and the mining expert must be guaranteed.
- Understanding the data preparation: Data preparation needs a lot of expertise in creating new data input (for example, SQL programming skill) and a good understanding of the given raw data and their content. An excellent data miner may not be successful if he/she lacks expertise in the business field under discussion.
- Understanding the algorithms: Using algorithms means to be able to adapt the setting for the various mining runs. Because all data mining functions are highly sophisticated from a implementation point of view, data mining experts are demanded who are well trained with the selected data mining toolkit. Namely, these persons must overview how much effort has to be undertaken to solve single steps of the generic methods (Figure 3-11), and how to solve each task either with the toolkit, or with the database stem, or with additional statistical functions.
- Understanding the business needs: This concerns the interaction with the business end users.

3.4.9 Effort required

It is difficult to be prescriptive about the amount of effort that will be required to perform a typical data mining project. If you were starting from a position of having no data warehouse and with data in disparate files and databases then the type of effort profile required is shown in Figure 3-11.

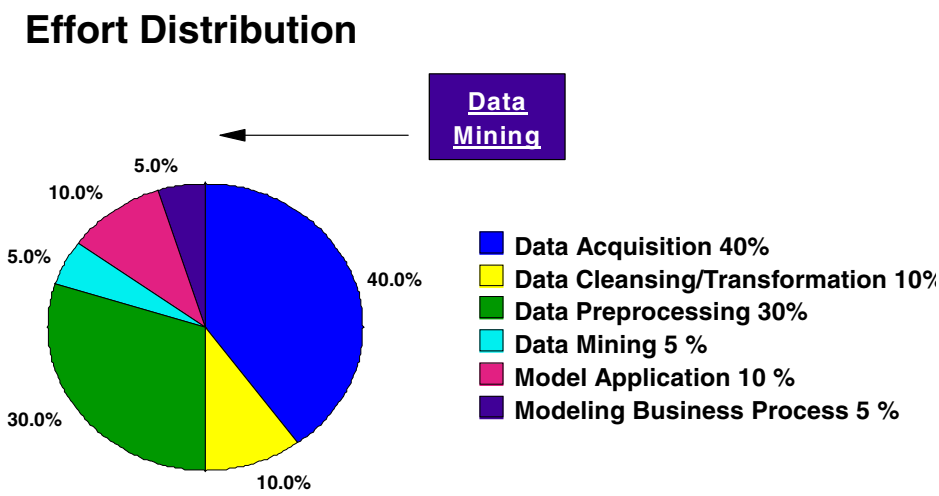


Figure 3-11 Effort distribution

You can see immediately that the vast majority of the effort is spent in the data preparation activities. If you already have the data in a usable form, then the task of mining your own business becomes much easier.

4

How to perform weight rating for Diagnosis Related Groups by using medical diagnoses

Diagnosis Related Groups are a highly discussed topic in the medical area worldwide. The reason for this is that medical services are not based on diagnoses, but on combinations of medical diagnoses (for example, ICD10, International Classification of Medicine) and procedures (for example, ICPM, International Classification of Procedures in Medicine). The weight of Diagnosis Related Groups plays an important role, because the revenue is defined as the product of weight and a fixed amount of money. The weights are predefined and cannot be changed by the physicians.

In this chapter we describe a method to find combinations of medical diagnoses that form a basis for Diagnosis Related Groups (DRG). Using Association Discovery you obtain associative rules that indicate combinations between medical diagnoses. The rules give you a statistically based report about current diagnosis trends and indicate which combinations of rules are more evident than others.

As will be evident in the following chapter the application of Association Discovery for medical diagnoses could become important in detecting higher and lower ranked weights for Diagnosis Related Groups.

4.1 The medical domain and the business issue

Diagnosis Related Groups (see 4.2.1, “Diagnoses data from first quarter 1999” on page 50) represent an international coding system for physicians. They contain a main diagnosis and one or several subdiagnoses. Depending on what the physician indicates as main and subdiagnosis, he or she will get reimbursed more or less money for his or her medical services.

DRGs are based on the ICD10 catalog system (see 4.2.2, “International Classification of Diseases (ICD10)” on page 52), a worldwide standard classification system for medical diagnoses.

Combinations of diagnoses are no longer treated as single medical services but as a group of services. Diagnosis Related Groups form groups of diagnoses that are expressed by a weight in respect to the medical efforts: the more expensive the medical service and the more efforts are investigated, the higher the weight of the Diagnosis Related Groups.

> **Example:** An eighty year old patient is delivered to the hospital with a strong fever and an extreme cold. After examination, the physician’s diagnosis is *Diabetes mellitus* with some form of *pneumonia*. Therefore, the patient receives several types of *antibiotics*, but unfortunately does not recover. He is then sent to intensive care to receive more intensive medical methods that can help to beat the *sepsis*. After a 10-day stay, he gets healthy by special therapies and medicaments; he can leave the hospital without any risks.

To get more money reimbursed for the delivered medical service it is now evident for the physician to choose this diagnosis as the main diagnosis that is more evident than others. For example, if the physician chooses *sepsis* as the main diagnosis and *pneumonia* and *diabetes* as the first and second subdiagnosis, respectively, then this combination leads to a higher revenue than other combinations of the three diagnoses, as described in Table 4-1,Table 4-2 and Table 4-3.

Table 4-1 Calculation of the revenue when SEPSIS is chosen

Main diagnosis	Sub diagnosis	Weight (example)	Revenue (for physician or hospital, in $)
Sepsis	*Pneumonia*, *diabetes mellitus*	1.81	3800

Table 4-2 Calculation of the revenue when PNEUMONIA is chosen

Main diagnosis	Sub diagnosis	Weight (example)	Revenue (for physician or hospital, in $)
Pneumonia	*Sepsis*, *diabetes mellitus*	1.43	3000

Table 4-3 Calculation of the revenue when DIABETES MELLITUS is chosen

Main diagnosis	Sub diagnosis	Weight (example)	Revenue (for physician or hospital, in $)
Diabetes mellitus	*Pneumonia*, *sepsis*	1.24	2600

For each national health care organization that is responsible for the establishment of the DRG catalog, it is not obvious how to find the weights for the Diagnosis Related Groups. Diagnoses like *cold* or *fever* require no big medical services and are therefore not highly to be rated, where *pneumonia* and *diabetes* need more intensive care. On the other side, a ranking like this can change over time and from place to place because combinations of diagnoses and the associated medical service can become cheaper if many of them arise.

From a practical standpoint, a 90% implication of *pneumonia* and *sepsis* is getting cheaper, in respect to cost, to a more seldom implication of *fever* and *cold*, because a medical surgery of diagnoses with high occurrence gets more efficient than more seldom diagnosis combinations; furthermore, appropriate medical surgery can be better established.

4.1.1 Where should we start?

The first stage in the generic method is to translate the business issue you are trying to address, into a question, or set of questions that can be addressed by data mining.

The business problem therefore is: *Can we identify groups of diagnoses that occur more often or more seldom?*

If yes, then this can become a worthwhile contribution to the consistency of the Diagnosis Related Groups.

4.2 The data to be used

You clearly cannot do data mining without having the data about your patients to mine. But what data do you need? *The second stage in our data mining method* is to identify the data required to address the business issue and where we are going to get it from.

For the analysis, data from a German health care organization was used (ZI - Zentralinstitut der kassenärztlichen Vereinigung, Cologne, Germany). The following data sets were available:

- Diagnoses that were recorded by physicians
- ICD catalog system, Version 10

The delivery of the diagnoses are done once a quarter, therefore, we turn our attention to the first quarter in 1999. This data contains more than three million diagnoses for more than one million patients. The quarter is the first one where physicians were definitely be asked to use ICD10.

The ICD Version 10 refers to the worldwide standardized classification system generated by the World Health Organization (WHO). This catalog can be downloaded from `www.who.org`.

4.2.1 Diagnoses data from first quarter 1999

To get an understanding of the data, it is very helpful to know how the data looks. As in many other scenarios, a first step in the exploration of the data is to get the knowledge about the data. Table 4-4 shows the variables that were used for Association Discovery.

Table 4-4 Variables that were used for Association Discovery

Variable name	Variable type	Description
YEAR	Numeric, casted to Categorical	Year when the diagnosis is done
QUARTER	Numeric, casted to Categorical	Quarter when the diagnosis is done
PHY_ID	Categorical	Physician ID (anonymous)
PHY_GROUP	Categorical	Group where the physician belongs to
PAT_ID	Categorical	Patient ID (anonymous)
ICD10	Categorical	ICD10

Variable name	Variable type	Description
COMMENTS	Text, casted to Categorical	Comments that were done by the physician

YEAR and *QUARTER* refer to the date when the diagnosis was done by the physician. However, it is not known exactly when the diagnosis was done, because detailed information, such as Day and/or Time are not available.

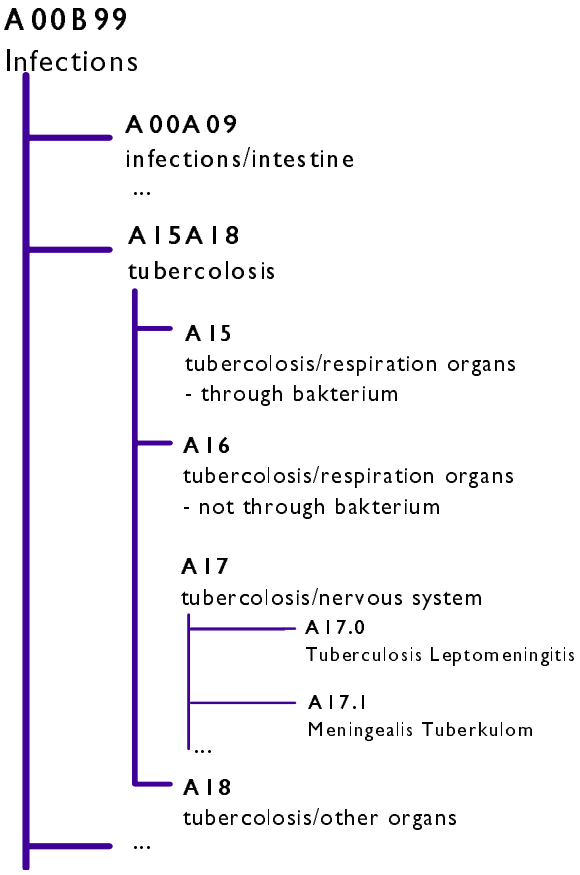


Figure 4-1 ICD10 catalog with different levels

PHY_ID refers to the anonymous physician identification. *PHY_TYPE* indicates the different types of physicians, for example:

- 004 Ophthalmologists
- 007 Surgeons
- 023 Pediatricians
- 038 Neurologists

Because different physician groups intend to diagnose differently, it is necessary to split the whole data in several subsets depending on *PHY_GROUP*.

PAT_ID is an anonymous variable that refers to the patient himself. If a patient visits a physician several times he always get the same patient identification.

ICD10 are diagnoses that are coded by ICD10.

COMMENTS represents textual comments that were additionally done by the corresponding physician. However, *COMMENTS* contains a lot of missing values and therefore is not suited for data mining. In principle, text mining techniques can be used to find some interesting patterns — if *COMMENTS* were filled with sufficient text.

4.2.2 International Classification of Diseases (ICD10)

ICD10 is an abbreviation for *International Classification of Diseases, Version 10*. It is a standardized catalog that was established by the World Health Organization in order to standardize medical diagnoses. In many countries, its usage is required to simplify and structure medical services.

The structure of the ICD10 is a hierarchical system that contains four levels of different degrees (Figure 4-1). For example:

- *A00B99* (Infections) occurs on the first level; A00B99 is an abbreviation for all diagnoses that are between A00 and B99.
- *A15A18* (Tuberculosis) on the second level
- *A17* (Tuberculosis/nervous system) on the third level
- *A17.0* (Tuberculosis Leptomeningitis) on the fourth level

In many cases, physicians tend to use the most specific diagnosis on level 4, but do use codes from level 3, if the disease covers more than one specific code in level 4.

4.3 Sourcing and preprocessing the data

To create our data model we have to take the raw data that we collect and convert it into the format required by the data models. We call this stage in the process sourcing and preprocessing and this is *the third stage in our data mining method*.

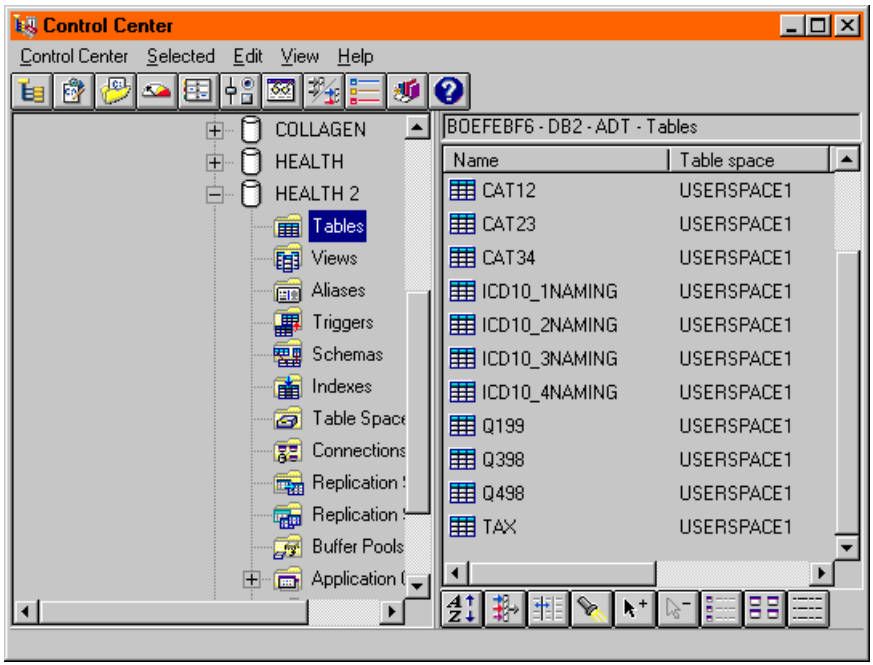


Figure 4-2 Storing the different medical tables Q398, Q498 and Q199 in the database

Figure 4-2 shows the different tables that are stored in the database. In particular, we have the following database tables:

- Q398, Q498 and Q199

  Refers to the diagnoses that were collected for the different quarters

- ICD10_1naming,...,ICD10_4naming

  Refers to the different levels of ICD10 Version 10 catalog

- TAX

  Contains the hierarchical information for the ICD10 catalog

4.4 Evaluating the data

Having created and populated our data models, *the fourth stage in our data mining method* is to perform an initial evaluation of the data itself.

Before starting with data mining as a discovery task we need to check the quality and consistency of the underlying datamarts. This step is necessary, because the quality of computed mining results should be based only on data that is neither missing nor incorrect.

Here we discuss the structure of the diagnoses data and the ICD10 catalog as well. We discuss whether preprocessing is necessary or not and how it will be done.

Furthermore, we turn our attention in the next chapters to ophthalmologists (*PHY_TYPE=004),* because they represent an appropriate group of physicians with a high number of diagnoses and low number of missing ICD entries.

4.4.1 Evaluating diagnoses data

To use the diagnoses data there are some minor processing steps necessary:

- *PAT_ID* and *PHY_ID* are, anonymously, very cryptic. For readability of these variables, it is easier to have numbers instead.
- *YEAR* and *QUARTER* are constant and need no further consideration.
- *PHY_GROUP* is actual and complete.
- *ICD10* is filled and correct in most cases. However, it contains some entries that are not available in the ICD10 catalog: these entries were either removed or, depending on the entry itself, substituted by the most logical alternative.

  As an example, *Z11.1* was recorded, but not defined in the ICD10 catalog. *Z11.1* is therefore substituted by the code on level 3. This is *Z11* which is a special form of screening.
- Furthermore, some entries were empty with only a small amount of information in *COMMENTS*. However, this field is not of interest.
- In some cases, it is possible to correct these diagnoses by the corresponding ICD10 code, and in some cases it is not. *ICD10* was therefore filled with an '*X*' (for example, see Figure 4-5 on page 63).

4.4.2 Evaluating ICD10 catalog

The data from ICD10 catalog are actual, clean and correct, and do not need to be changed.

ICD1	ICD2	ICD3	ICD4
A00B99	A00A09	A00	A00.9
A00B99	A00A09	A01	A01.0
A00B99	A00A09	A01	A01.4
A00B99	A00A09	A01	A01.3
A00B99	A00A09	A01	A01.2
A00B99	A00A09	A01	A01.1
A00B99	A00A09	A02	A02.0
A00B99	A00A09	A02	A02.9
A00B99	A00A09	A02	A02.1
A00B99	A00A09	A03	A03.0
A00B99	A00A09	A03	A03.9
A00B99	A00A09	A03	A03.3
A00B99	A00A09	A03	A03.2
A00B99	A00A09	A03	A03.1
A00B99	A00A09	A04	A04.0
A00B99	A00A09	A04	A04.9
A00B99	A00A09	A04	A04.7
A00B99	A00A09	A04	A04.6
A00B99	A00A09	A04	A04.5
A00B99	A00A09	A04	A04.4

Figure 4-3 Hierarchical system for the ICD10 catalog

However, some questions remain open:

- The description part of each code is often very long. Can we simply cut off the description at a specific position in the text?
- The catalog is divided into several files (first, second, third and fourth levels of ICD10) where each file corresponds to one hierarchy. How can we efficiently establish the hierarchical system?

The first question is more complex than it seems, because a technical truncation of the variable description may lead to an incomplete description of the diagnosis, and therefore to ambiguity. This is because some descriptions differ from others simply by specific medical expressions at the end of the description.

A short-term description is necessary, because associations would be interpretable but would require a lot of additional work. We therefore used only codes with no descriptions. The semantic value was then retrieved manually.

For the second question, the hierarchical system was generated by a script program (Figure 4-3). It contains four columns that correspond to the different levels of ICD10.

4.4.3 Limiting the datamart

We will show how we can find an appropriate solution to the specified question by selecting ophthalmologists. This mainly has two reasons:

- The number of diagnoses that were done using ICD10 is more than 90% in the first quarter 1999. This is the best percentage; all other groups have higher rates of missing diagnoses.
- The average number of diagnoses per visit is almost six. This is the best case; all other groups have lower rates.

To demonstrate how DRGs can be computed, we focus our attention to diagnoses that were done by ophthalmologists. Because the behavior of each physician group is completely different than others groups, a consideration of other physicians — in addition to ophthalmologists — is not acceptable.

4.5 Choosing the mining technique

Choosing the mining techniques to use is *the fifth stage in our generic mining method*.

This section contains three important aspects that need to be considered before we perform data mining on medical datamarts:

- Who should do the mining task and which role plays communication?
- What are the main characteristics of data mining, especially in respect to other analyses strategies?
- Which data mining technique should be chosen to perform the discovery of relevant DRGs?

4.5.1 About the communication between experts

Mining in medicine requires a deep knowledge of the medical scenario. It is not sufficient to simply apply algorithms to the data, because:

- Potentially new information derives only from an existing level of expertise.
- Only communication between physicians and mining experts can lead to appropriate solutions. Some information may seem important to the mining expert, for example, if a mining result has good parameter values, or generally speaking, a good quality measure. For a domain expert, this may be redundant or obvious. Likewise, there could niches in the results that are not obvious to a mining expert but are very valuable to the domain expert.

Therefore, mining of the medical data as well as the interpretation of the results has to be done in collaboration with the domain experts, in this case with physicians.

As shown in Figure 4-4, more than fifty mining runs with different parameter settings were done for the ophthalmologists datamart. Mining the data therefore also means a lot of trying, and it requires some creativity and a feeling of what could and should be done, as well.

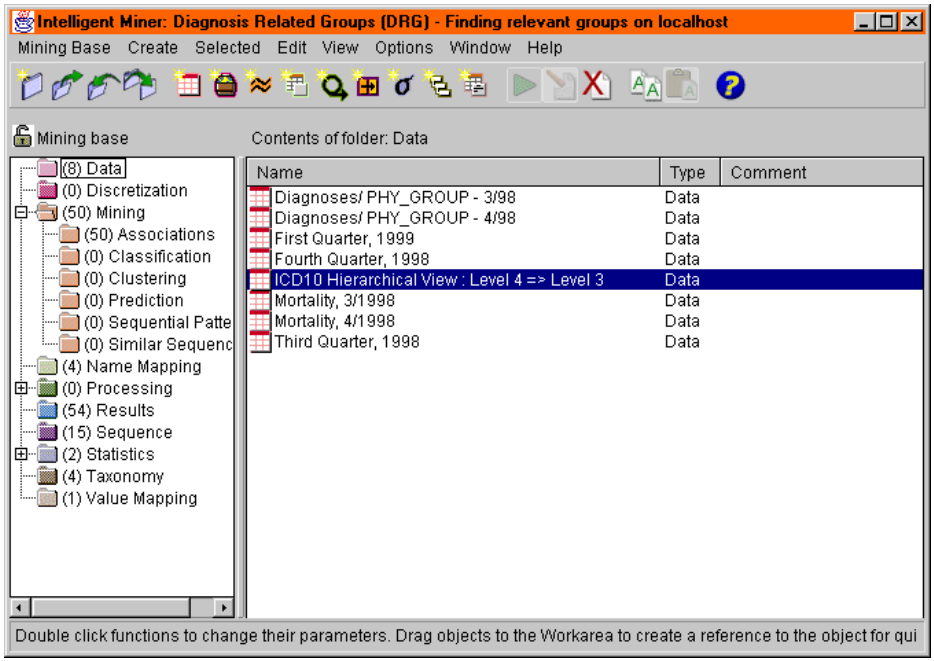


Figure 4-4 Association Discovery through interaction of database and mining technology

4.5.2 About verification and discovery

Although data mining is always said to be a discovery-driven approach, the search for potentially hidden, but useful information by data mining requires a verification of the information, too.

To underline this, the following examples should show which verification steps that were needed to get relevant mining results (see the discussion about the mining results in 4.6, “Interpreting the results” on page 62).

- Verification of the data domain, variables and their values, for example:
  - Using search engines to find explanations to different medical terms like DRG, ICD10, and so on.

- Using medical books or medical Internet sites to get a feeling for medical diseases and the corresponding medical services.
- Preprocessing the data by applying statistical (verification) techniques, for example:
  - Understanding what the ICD10 codes really mean, for example *I10* or *L71.9*
  - Knowing which physician groups are relevant for mining: if the average number of diagnoses per patient is smaller than two, it would not really make sense for this group to be considered.
  - Knowing how many incorrect classifications were done. If the number of missing ICD10 codes is too big, then further concentration on this data is not interesting enough.
  - Checking which patients do share a discovered association rule. Maybe we can find additional diagnoses in their diagnosis history.

In reality, the search for relevant new information by data mining is a game between verification and discovery, and therefore a form of symbiosis.

4.5.3 Let's find associative rules!

Searching combinations of diagnoses is very similar to finding associative rules. For the analysis, we used the Association algorithm of the IM for Data. As we will see in "ATSeries" on page 62, we additionally use ATSeries to observe the behavior of associative diagnoses over time.

Association Discovery

For Association Discovery, we use the following three parameters to identify the relevance and quality of association rules:

- Support
- Confidence
- Lift

Note: *Support* refers to items that are part of an association rule. It defines the probability for each item to occur in the whole population. It is a measure that indicates the quantitative importance.

Note: *Confidence* refers to an associative rule. It is a measure that indicates the qualitative importance of a rule. It is marginal statistics calculated by using the Bayesian probability.

Note: *Lift* refers to an association rule and defines the significance of an association rule. Mathematically, it represents the relative deviation and is defined as the division of confidence and expected confidence.

- If the Lift is exactly 1, then the association rule has no relevance, because it has the same effect as in the random case.
- If the Lift is below 1, then both sides of an association rule inhibit each other with no positive effect.
- If the Lift is greater than 1, then both sides of the association rules have positive impacts on each other: the association rule gets more relevant, the higher the Lift is.

Association Discovery mainly works by two steps that were repeatedly performed by the underlying algorithm:

- Join
- Prune

Join is used to expand the list of associative items where prune is used to filter out those lists that do not satisfy some threshold requirements. To explain the relationship between join and prune in more detail, we demonstrate this by the example in Table 4-5.

Example

Assume we have the following diagnoses available in Table 4-5 for patients 1, 2, 3, and 4.

Table 4-5 First Step: diagnoses that were recorded by the physicians

Patient	Diagnoses
1	J03.9, J06.9, K52.9, Z04
2	H66.9, K52.9, Z04

Patient	Diagnoses
3	J06.9, K52.9
4	J06.9, K52.9, Z04

If we try to generate association on all potential combinations we would get an enormous search space, and, as a side effect, all combinations of rules that are possible at all. In order to restrict the search space, the algorithm allows a pruning of the space by the definition of a minimum threshold (support value) by the user.

Assume we have a minimum support threshold of 40%, which actually means that all combinations of diagnoses that occur less than 40% will be pruned (and therefore neglected). We then get Table 4-6.

Table 4-6 Second Step: Calculate the support of the underlying diagnoses

Diagnosis	Support (= percentage of occurrences)
J03.9	25
J06.9	75
K52.9	100
Z04	75
H66.9	25

Only *J06.9 (accute infection), K52.9 (chronic diarrhoea), and Z04 (special examination)* satisfy the pruning condition; therefore, *J03.9* and *H66.9* are no longer observed.

In the next join step in Table 4-7, we then build combinations out of the remaining three diagnoses.

Table 4-7 Third Step: Generate Combinations and perform pruning

Joined diagnoses	Support (= percentage of occurrences)
J06.9, K52.9	75
J06.9, Z04	50
K52.2, Z04	75

All combination satisfy the support threshold of 40%, therefore no combination is removed by pruning. We then can apply the join process again and get the result in Table 4-8.

Table 4-8 Fourth Step: Generate Combination and perform pruning

Joined diagnoses	Support (= percentage of occurrences)
J06.9, K52.9, Z04	50

J06.9, K52.9 and Z04 define a valid set of diagnoses that satisfy the given threshold. Because they represent the only valid set, no further joining is possible.

Therefore, combinations of these diagnoses define a valid association rule. In fact, *J06.9, K52.9* and *Z04* are nothing else except as an abbreviation to the following association rules:

- If *J06.9* then *K52.9*
- If *J06.9* then *Z04*
- If *Z04* then *K52.9*
- If *J06.9* and *Z04* then *K52.9*
- If *J06.9* and *K52.9* then *Z04*
- If *Z04* and *K52.9* then *J06.9*

However, IM for Data only allows association rules that have one item on the right side of an association rule. Therefore, nine is the number of valid rules.

The IM for Data uses a second threshold (minimum confidence) that reduces the number of rules: by applying a probability function (the Bayesian formula) IM for Data only allows association rules with a minimum quality. For example:

- If *J06.9* AND *Z04* then *K52.9*

We will get a confidence of 100%, because:

- Support(*J06.9* AND *Z04* AND *K52.9*) = 50%
- Support(*J06.9* AND *Z04*) = 50%

Therefore, the division gives 100%.

Assume we define a minimum confidence of 80%. Then the following rules are valid:

- If *J06.9* then *K52.9*
  Confidence = 100%
- If *Z04* then *K52.9*
  Confidence = 100%

If we reduce the Confidence to 75%, then the rules become valid too:

- if *K52.9* then *Z04*
  Confidence = 75%
- if *K52.9* then *J06.9*
  Confidence = 75%

ATSeries

ATSeries stands for Association Time Series and is a tool that was programmed as an extension to the IM for Data for the medical scenario. It acts as a bridge between Associations Discovery and the Similar Sequences algorithm of the IM for Data. ATSeries is not part of IM for Data itself, but it expands the functionality of the IM for Data. In particular:

- ATSeries uses association result files as input. This is interesting if we have association results based on different times or different locations.
- ATSeries offers a pre-selection of interesting items (*frequent itemsets*) by displaying those itemsets items that occur most often.
- ATSeries displays the support, confidence and lift of all association rules that occur at least one time.
- ATSeries transforms support, confidence and lift values of a selected association rule into a format that allows performing time series analysis by IM for Data's Similar Sequence algorithm.

4.6 Interpreting the results

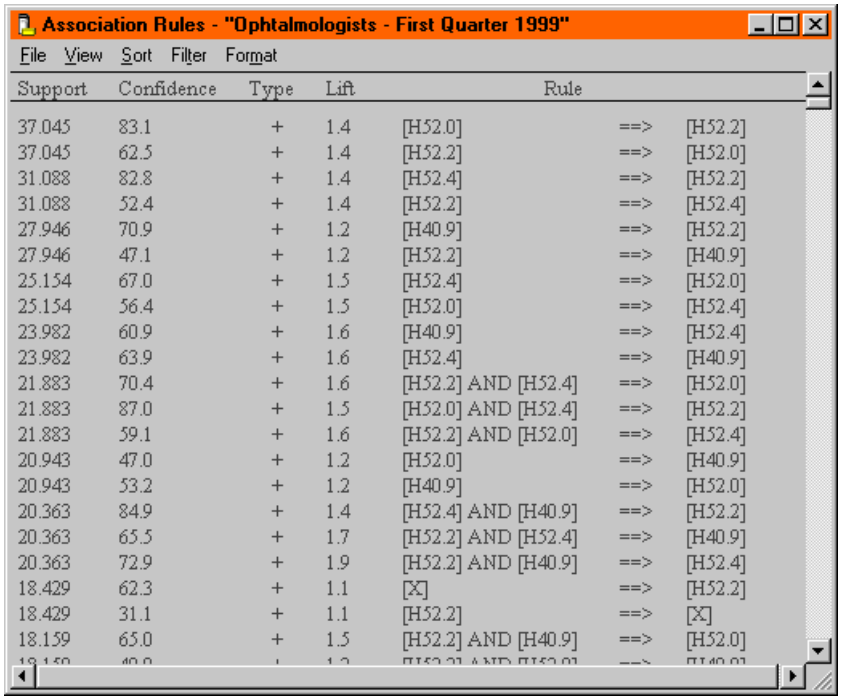
The *sixth stage in our generic mining method* is to interpret the results that we have obtained and determine how we can map them to our business. When you are first confronted with the associations results, the question that you ask is "What does it all mean?".

In this section we describe how to understand and read and interpret the results from the different associations results, but more importantly how to get appropriate association rules from the medical datamart. Then, we argue to observe the discovered rules over time, which leads to a better feeling for the results.

4.6.1 Finding appropriate association rules

Figure 4-5 shows one of the association results that was computed using the Associations Discovery of the IM for Data.

All association rules (for example, [H52.0] ==> [H52.2]) share high supports with high confidences but low lifts. This means that the combination of these diagnoses are mainly obvious: they occur very often (support value) and are strongly correlated (confidence value) but not really relevant (lift value).



Support	Confidence	Type	Lift	Rule		
37.045	83.1	+	1.4	[H52.0]	==>	[H52.2]
37.045	62.5	+	1.4	[H52.2]	==>	[H52.0]
31.088	82.8	+	1.4	[H52.4]	==>	[H52.2]
31.088	52.4	+	1.4	[H52.2]	==>	[H52.4]
27.946	70.9	+	1.2	[H40.9]	==>	[H52.2]
27.946	47.1	+	1.2	[H52.2]	==>	[H40.9]
25.154	67.0	+	1.5	[H52.4]	==>	[H52.0]
25.154	56.4	+	1.5	[H52.0]	==>	[H52.4]
23.982	60.9	+	1.6	[H40.9]	==>	[H52.4]
23.982	63.9	+	1.6	[H52.4]	==>	[H40.9]
21.883	70.4	+	1.6	[H52.2] AND [H52.4]	==>	[H52.0]
21.883	87.0	+	1.5	[H52.0] AND [H52.4]	==>	[H52.2]
21.883	59.1	+	1.6	[H52.2] AND [H52.0]	==>	[H52.4]
20.943	47.0	+	1.2	[H52.0]	==>	[H40.9]
20.943	53.2	+	1.2	[H40.9]	==>	[H52.0]
20.363	84.9	+	1.4	[H52.4] AND [H40.9]	==>	[H52.2]
20.363	65.5	+	1.7	[H52.2] AND [H52.4]	==>	[H40.9]
20.363	72.9	+	1.9	[H52.2] AND [H40.9]	==>	[H52.4]
18.429	62.3	+	1.1	[X]	==>	[H52.2]
18.429	31.1	+	1.1	[H52.2]	==>	[X]
18.159	65.0	+	1.5	[H52.2] AND [H40.9]	==>	[H52.0]

Figure 4-5 Associative rules with high support and confidence but low lift values

- [H52.0] ==> [H52.2]

  This rule says that for 37.045% of all patients with H52.0 (*hypermetropy of the eye axes*) they were also H52.2 (*astigmatism*) diagnosed with a Confidence of 83.1%. Although both Support and Confidence are high values, a Lift of 1.4 indicates that this rule is not really surprising and just on the same expectation level as the random case.

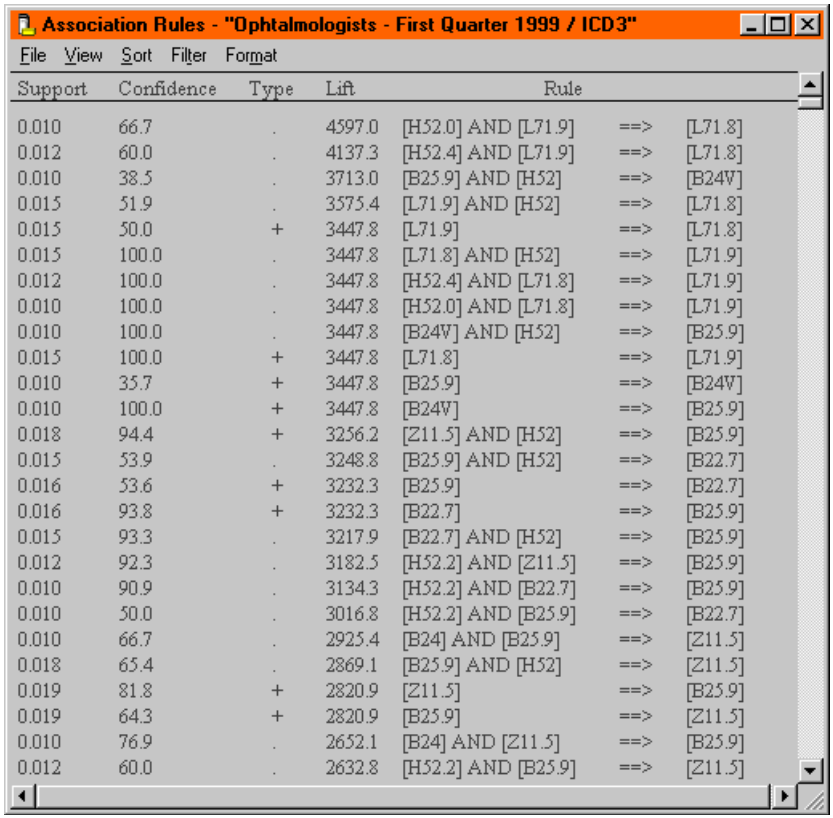
- H52.2 AND H52.4 ==> H52.0

  The first associative rule with three diagnoses contains H52.2 (*astigmatism*), H52.4 (*weakness of the eyes*) and H52.0 (*hypermetropy of the axes*). Here, we also have High support (21.883%) and Confidence (70.4%) but a low Lift value of only 1.6.

  However, all rules have a positive Type (indicated by '+') which is actually an indication that all corresponding diagnoses influence each other.

- H52.2 ==> X (and vice versa)

  These two association rules have an *'X'* in combination with H52.2 (see Figure 4-5) which means that one diagnosis was not coded by ICD10. Because the Confidence of X ==> H52.2 is larger than H52.2 ==> X (62.3% instead of 31.1%) this could lead to the assumption that H52.2 is additionally added by the physician if he delivers an unknown diagnosis.



Support	Confidence	Type	Lift	Rule		
0.010	66.7	.	4597.0	[H52.0] AND [L71.9]	==>	[L71.8]
0.012	60.0	.	4137.3	[H52.4] AND [L71.9]	==>	[L71.8]
0.010	38.5	.	3713.0	[B25.9] AND [H52]	==>	[B24V]
0.015	51.9	.	3575.4	[L71.9] AND [H52]	==>	[L71.8]
0.015	50.0	+	3447.8	[L71.9]	==>	[L71.8]
0.015	100.0	.	3447.8	[L71.8] AND [H52]	==>	[L71.9]
0.012	100.0	.	3447.8	[H52.4] AND [L71.8]	==>	[L71.9]
0.010	100.0	.	3447.8	[H52.0] AND [L71.8]	==>	[L71.9]
0.010	100.0	.	3447.8	[B24V] AND [H52]	==>	[B25.9]
0.015	100.0	+	3447.8	[L71.8]	==>	[L71.9]
0.010	35.7	+	3447.8	[B25.9]	==>	[B24V]
0.010	100.0	+	3447.8	[B24V]	==>	[B25.9]
0.018	94.4	+	3256.2	[Z11.5] AND [H52]	==>	[B25.9]
0.015	53.9	.	3248.8	[B25.9] AND [H52]	==>	[B22.7]
0.016	53.6	+	3232.3	[B25.9]	==>	[B22.7]
0.016	93.8	+	3232.3	[B22.7]	==>	[B25.9]
0.015	93.3	.	3217.9	[B22.7] AND [H52]	==>	[B25.9]
0.012	92.3	.	3182.5	[H52.2] AND [Z11.5]	==>	[B25.9]
0.010	90.9	.	3134.3	[H52.2] AND [B22.7]	==>	[B25.9]
0.010	50.0	.	3016.8	[H52.2] AND [B25.9]	==>	[B22.7]
0.010	66.7	.	2925.4	[B24] AND [B25.9]	==>	[Z11.5]
0.018	65.4	.	2869.1	[B25.9] AND [H52]	==>	[Z11.5]
0.019	81.8	+	2820.9	[Z11.5]	==>	[B25.9]
0.019	64.3	+	2820.9	[B25.9]	==>	[Z11.5]
0.010	76.9	.	2652.1	[B24] AND [Z11.5]	==>	[B25.9]
0.012	60.0	.	2632.8	[H52.2] AND [B25.9]	==>	[Z11.5]

Figure 4-6 Associative rules with low support but high lift and confidence values

Figure 4-6 shows association rules with low support but high lift value.

- H52.0 AND L71.9 (*Acne rosacea*) ==> L71.8 (*Blepharitis*)

  This is the rule with the highest lift (H52.0 denotes a *hypermetropy of the axes*, L71.9 *Acne rosacea* and L71.8 *Blepharitis*). This rule occurs approximately 4000 times higher than the associative rules given above; therefore, for the calculation of the weights (for the DRG's), it causes a higher value than many other combinations of diagnoses probably do.

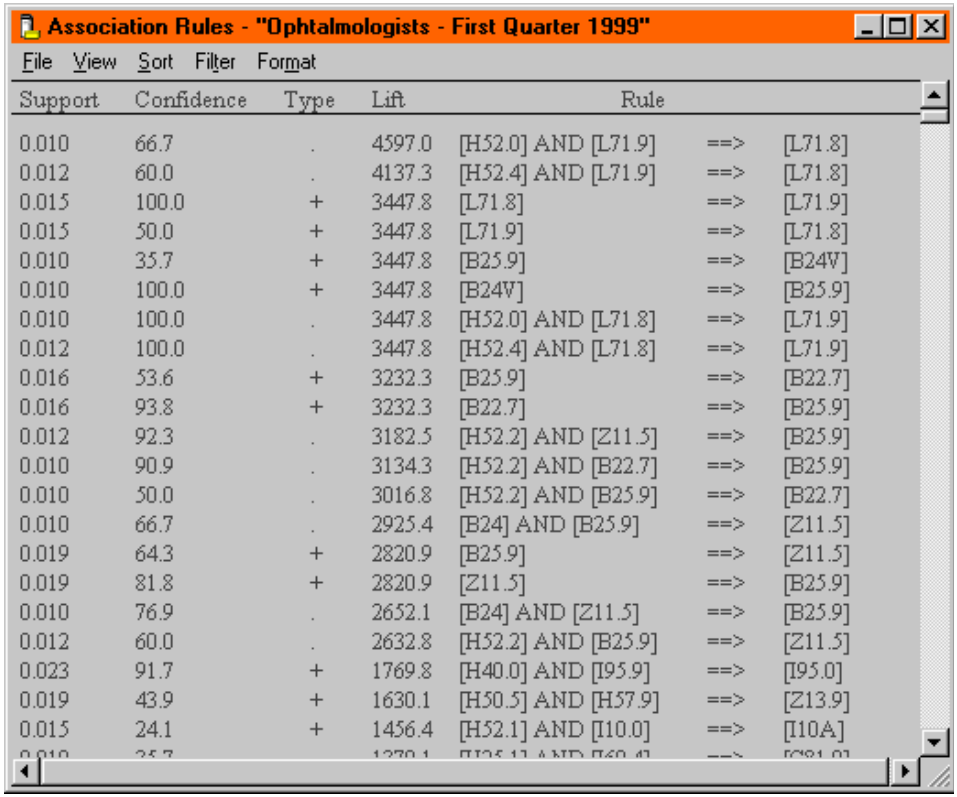


Figure 4-7 Association rules with the integration of ICD10 level 3 for 1/99

Figure 4-7 shows association rules that are calculated using not only Level 4 of the ICD10 code but also Level 3.

- B25.9 ==> B24

  This is an interesting rule (B25.9 is *Angiitis,* an uncommon inflammation of the blood vessels and B24 *a HIV disease*), because B24 is additionally marked by a 'V' (*suspicion of*).

  Although only 35.7% of the patients with B25.9 also have B24, it may be that B25.9 could act as trigger for HIV disease. The lift values are very high (>3000); the relevance of these rules are therefore very large.

4.6.2 Association discovery over time

Note: ATSeries is not part of the standard functionality of the IM for Data; It is a tool that was programmed using the IM for Data API.

In Figure 4-8 and Figure 4-9 we show two association rules over time:

- H52.2 ==> M35.0
- B25.9 ==> B24

H52.2 ==> M35.0

Figure 4-8 shows the relationship between H52.2 (*astigmatism*) and M35.0 (*atropic dakryosialoadenopathie*). Astigmatism is a common form of visual impairment in which part of an image is blurred, due to an irregularity in the curvature of the front surface of the eye.

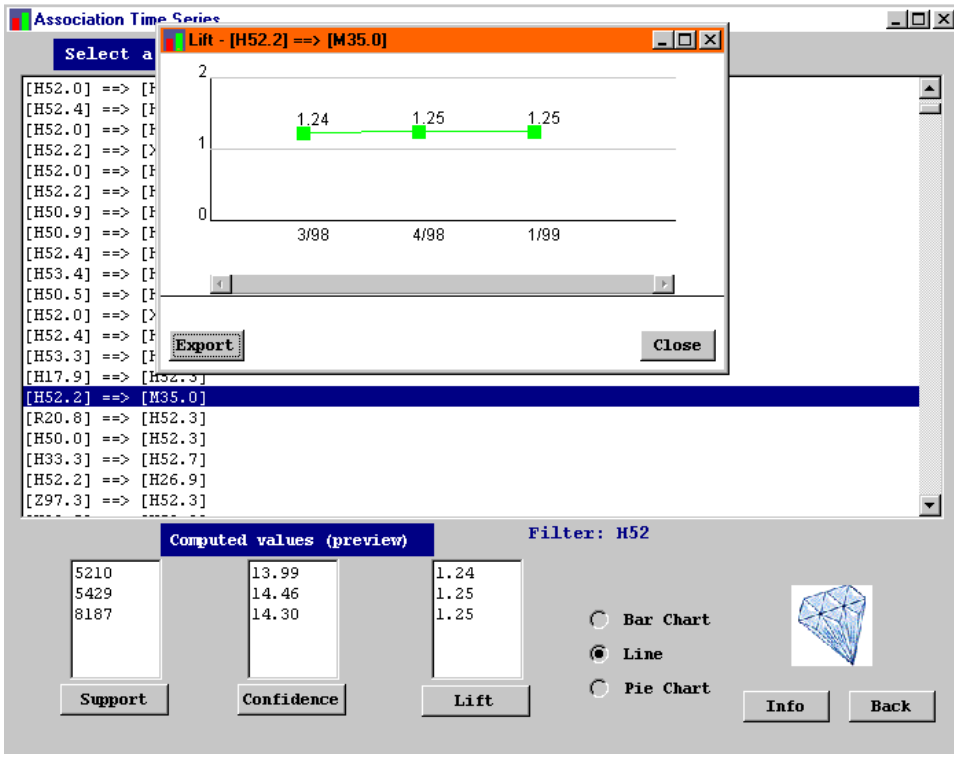


Figure 4-8 Using ATSeries to show low-valued lift over time

We can see that the *lift* values tend to be constant over the three quarters with values smaller than 2.

B25.9 ==> B24

As an example for high *lift* behavior over time, Figure 4-9 shows the selected association rule between B25.9 and B24 for the quarters 3/98, 4/98, and 1/99. The *lift* values are displayed on the right: whereas the lift values are nearly constant for the two quarters in 1998, the lift increases with factor 8 for the first quarter 1999.

For the *confidence*, we can observe a similar behavior (middle window at the bottom): it stays nearly constant (3/98, 4/98) but increases to a value that is five times higher than before (1/99).

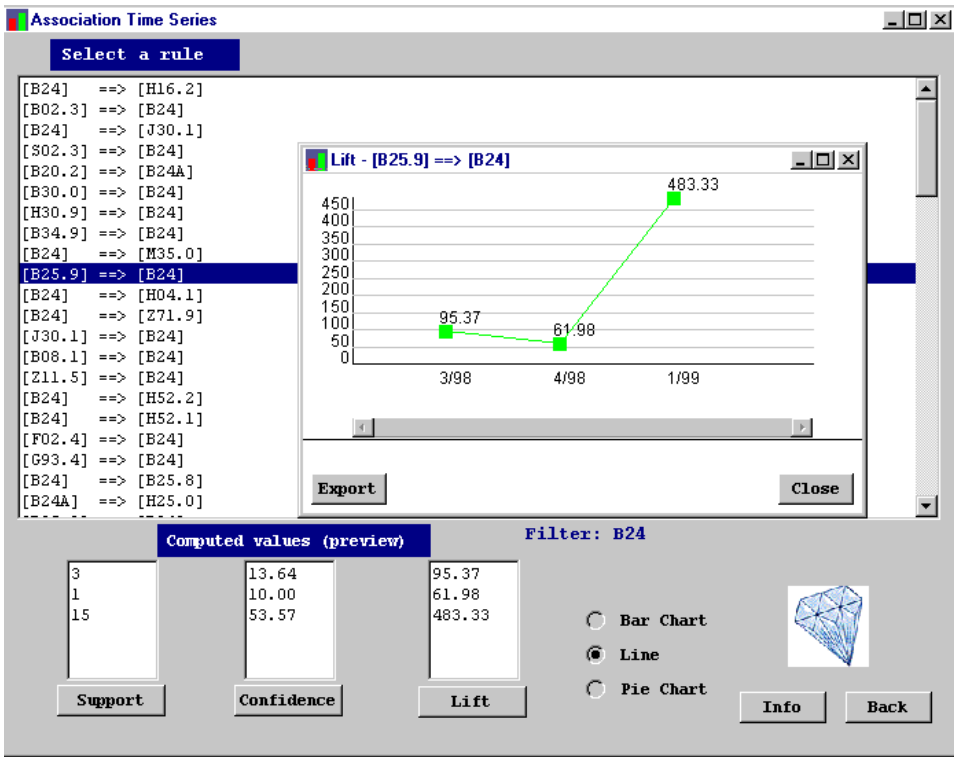


Figure 4-9 Using ATSeries to show high-valued lift over time

Note: ATSeries provides a very special subset of OLAP functionality on a cube with time dimension; it measures then support, confidence and lift overt time.

4.7 Deploying the mining results

The final and *seventh stage in our generic mining method* is perhaps the most important of all. How do you deploy the mining results into your business to derive the business benefits that data mining offers? The reason that this is so important is that all too often data mining is seen as an analytical tool that can be used to gain business insight, but is difficult to integrate into existing systems. In this section, after a short summary about the methods used, we suggest a strategy on how to use results of the Association Discovery to model weights for the Diagnosis Related Groups.

4.7.1 What we did so far

First at all, we tried to find combinations of diagnoses that commonly occur. The diagnoses were based on data that came from ophthalmologists of the first quarter in 1999.

Using Association Discovery by the IM for Data we got *association rules* that were labeled with some parameters: some of the rules had a *high lift* and some of them a *low lift*. We said that the lift corresponds to the relevance of a rule; therefore, all association rules with a high lift were indicated as relevant, and all association rules with lower lift as irrelevant. *Support* and *confidence* had no direct impact.

Then we identified several association rules and showed how they can be discovered over time.

4.7.2 Performing weight rating for Diagnosis Related Groups

Diagnosis Related Groups are defined by the relevance of the diagnosis and the efforts in terms of medical services that need to be investigated. This is expressed by their weights (see 4.1, “The medical domain and the business issue” on page 48): a high relevance and less efforts should have similar weights to those of low relevance and high efforts; but both of these cases are of lower value than high relevance and high efforts.

For a deployment of the model, we define the following two variables as the main components for the rating of the weights:

- Relevance
- Efforts

For ***relevance***, we simply take the lift, because lift represents the relative deviation against the random case: the higher the lift, the more significant the corresponding association rule.

For ***efforts***, however, we need to be more careful: the confidence of an Association Rule represents its strength and, in some senses, its quality. The stronger the implication of the associated diagnoses, the higher the confidence.

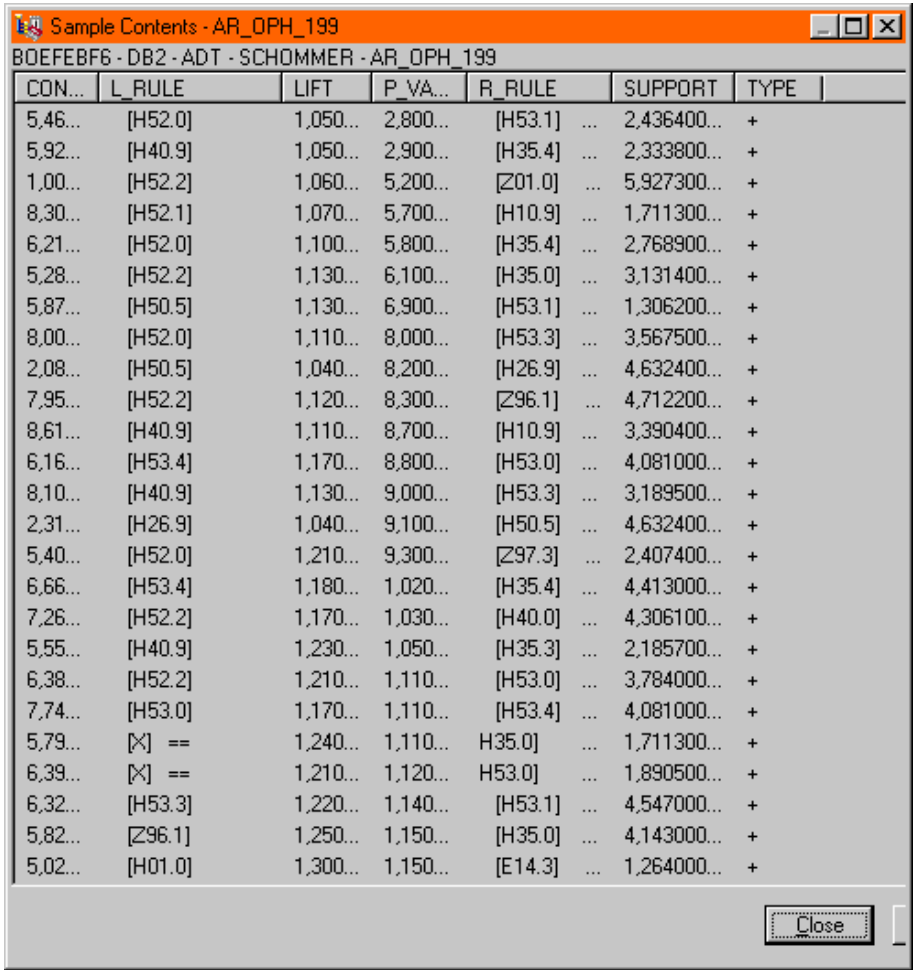
The *support* of the associated diagnoses is the probability of how often the diagnoses occur together. The higher the diagnoses occur, the higher the support is.

In other words:

- *A high support and a high confidence* indicate that the association occurs more often and strong.
- *A low support and a high confidence* indicate that the association occurs more seldom but strong.
- *A high support and a low confidence* indicate that the association occurs more often but weak.
- *A low support and a low confidence* indicate that the association occurs more seldom and weak.

For evaluation reasons, it is clear that ***efforts*** need to be high if *confidence and support are high*. On the other hand, ***efforts*** should be low *if confidence and support are low.* Therefore, we can define ***efforts*** as *the product of confidence and support.*

We can transfer the results of Association Discovery process into the database by using SQL. We then get a datamart that contains all the parameters and the association rules (Figure 4-10). We then can calculate ***efforts*** and ***relevance***.



CON...	L_RULE	LIFT	P_VA...	R_RULE		SUPPORT	TYPE
5,46...	[H52.0]	1,050...	2,800...	[H53.1]	...	2,436400...	+
5,92...	[H40.9]	1,050...	2,900...	[H35.4]	...	2,333800...	+
1,00...	[H52.2]	1,060...	5,200...	[Z01.0]	...	5,927300...	+
8,30...	[H52.1]	1,070...	5,700...	[H10.9]	...	1,711300...	+
6,21...	[H52.0]	1,100...	5,800...	[H35.4]	...	2,768900...	+
5,28...	[H52.2]	1,130...	6,100...	[H35.0]	...	3,131400...	+
5,87...	[H50.5]	1,130...	6,900...	[H53.1]	...	1,306200...	+
8,00...	[H52.0]	1,110...	8,000...	[H53.3]	...	3,567500...	+
2,08...	[H50.5]	1,040...	8,200...	[H26.9]	...	4,632400...	+
7,95...	[H52.2]	1,120...	8,300...	[Z96.1]	...	4,712200...	+
8,61...	[H40.9]	1,110...	8,700...	[H10.9]	...	3,390400...	+
6,16...	[H53.4]	1,170...	8,800...	[H53.0]	...	4,081000...	+
8,10...	[H40.9]	1,130...	9,000...	[H53.3]	...	3,189500...	+
2,31...	[H26.9]	1,040...	9,100...	[H50.5]	...	4,632400...	+
5,40...	[H52.0]	1,210...	9,300...	[Z97.3]	...	2,407400...	+
6,66...	[H53.4]	1,180...	1,020...	[H35.4]	...	4,413000...	+
7,26...	[H52.2]	1,170...	1,030...	[H40.0]	...	4,306100...	+
5,55...	[H40.9]	1,230...	1,050...	[H35.3]	...	2,185700...	+
6,38...	[H52.2]	1,210...	1,110...	[H53.0]	...	3,784000...	+
7,74...	[H53.0]	1,170...	1,110...	[H53.4]	...	4,081000...	+
5,79...	[X] ==	1,240...	1,110...	H35.0]	...	1,711300...	+
6,39...	[X] ==	1,210...	1,120...	H53.0]	...	1,890500...	+
6,32...	[H53.3]	1,220...	1,140...	[H53.1]	...	4,547000...	+
5,82...	[Z96.1]	1,250...	1,150...	[H35.0]	...	4,143000...	+
5,02...	[H01.0]	1,300...	1,150...	[E14.3]	...	1,264000...	+

Figure 4-10 Writing Association Rules into a database table

There are two strategies now to use ***efforts*** and ***relevance***:

- Perform segmentation and try to find clusters that have similar values of *efforts* and *relevance* in common
- Perform bivariate statistics against a target

We only will show the weight ranking by the second strategy. We therefore define a target *weight* that is calculated by the product of ***efforts*** and ***relevance***. Additionally, we use discretization to transform the numerical variable into a categorical data type.

We use the following borders for ***weight***:

- *weight* is **low** if the product of *efforts* and *relevance* is below 30.
- *weight* is **middle-low** if the product of *efforts* and *relevance* is between 30 and 60.
- *weight* is **middle** if the product of *efforts* and *relevance* is between 60 and 120.
- *weight* is **middle-high** if the product of *efforts* and *relevance* is between 120 and 500.
- *weight* is **high** if the product of *efforts* and *relevance* is greater than 300.

Figure 4-11 shows the resulting cluster when bivariate clustering will be applied against *weight*. According to the different values of *weight* we then get five clusters that contain *efforts* and *relevance* as well as the following variables:

- L-RULE: the diagnosis that occurs on the left side of an association rule
- R-RULE: the diagnosis that occurs on the right side of an association rule

LR-RULE: the diagnoses that occurs both on the left and right side of an association rule.

Figure 4-11 Weight rating with 5 clusters that were generated by applying bivariate statistics against weight

Figure 4-12 shows the cluster with low *efforts* and low *relevance*. For example, the cluster contains the following associative diagnoses (*LR_RULE*):

- R0.7 (*pains in breast and heart areas*), H52.0 (*hypermetropy of the eye axes*)
- H00.0 (*abscess*), H51.9 (*defect in the movement of the eyes*)

The associated diagnoses indicated in LR_RULE would then become a lower weighting.

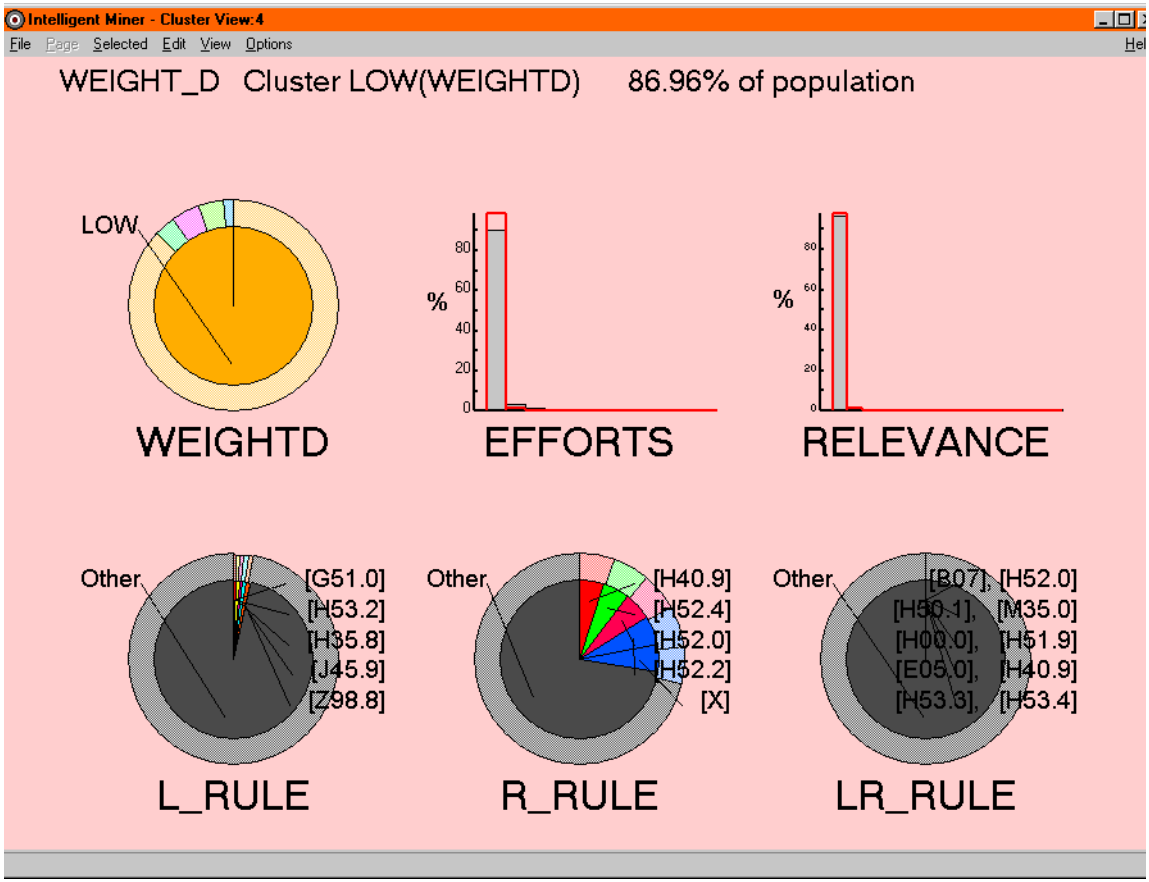


Figure 4-12 Cluster LOW with low EFFORTS and low RELEVANCE

Figure 4-13 shows the cluster with low *efforts* and low *relevance*. For example, the cluster contains the following associative diagnoses (*LR_RULE)*:

- H47.3 (*disease of the pupil*), H53.4 (*anopsy*)
- H02.5 (*Ankyloblepharon*), H33.2 (*ablatio retinae*)

The associated diagnoses indicated in LR_RULE would then become a higher weighting.

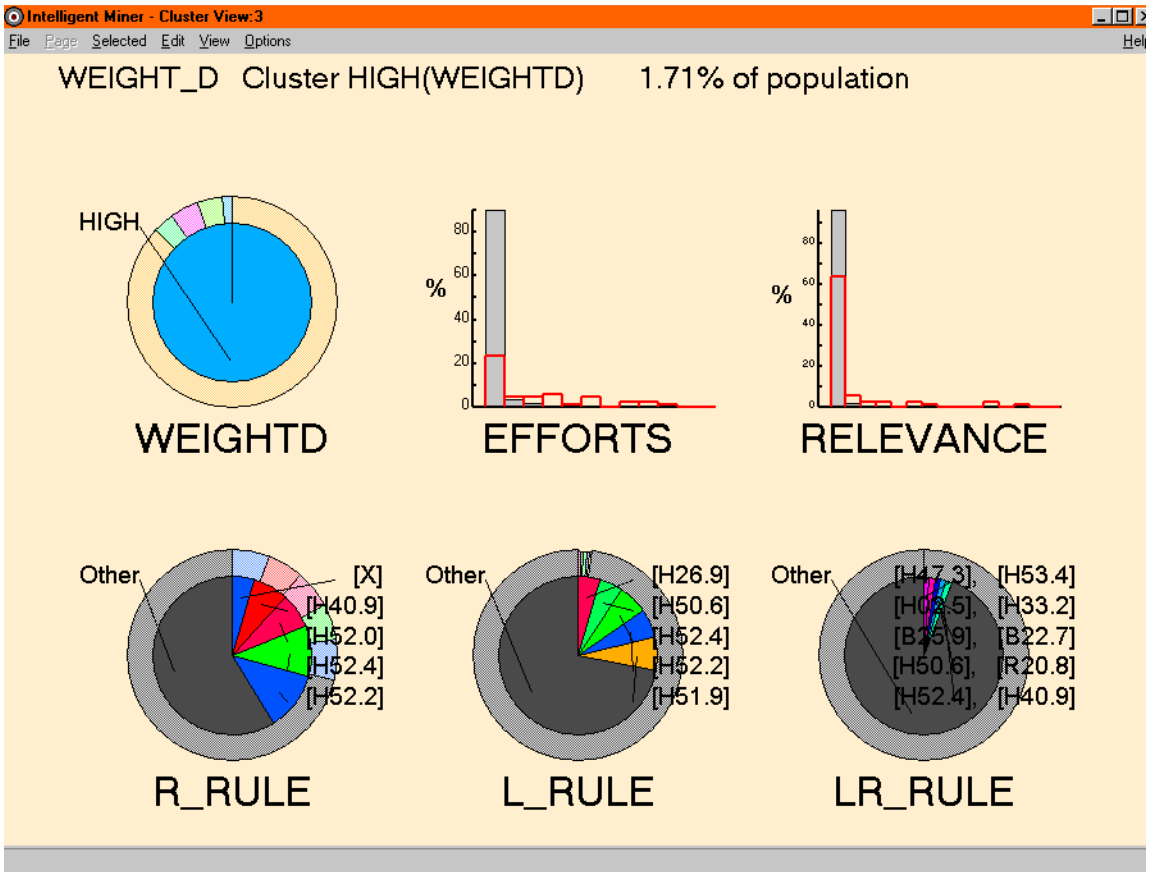


Figure 4-13 Cluster 'HIGH' with high EFFORTS and high RELEVANCE

5

How to perform patient profiling

In this chapter we describe a method to perform patient profiling. This method is performed on patients who were tested for deep vein thrombosis.

Some of patients were diagnosed with thrombosis during an examination test, some of them were not. The challenge in this chapter is to find groups of patients who share a similar behavior. The hope is to detect some new but useful indicators that may be derived from our analysis.

As we show in this chapter, we are not only concerned with a description of techniques of data mining (IM for Data), but also with an analytical way to find relevant patterns.

5.1 The medical domain and the business issue

This section contains a small description of what deep vein thrombosis is. Furthermore, we describe what it causes and when it occurs and we define the business issue that we should start with.

5.1.1 Deep vein thrombosis

A vein is a blood vessel that returns blood from the tissues of the body back to the heart. Here, the body has two distinct systems of veins: a *superficial system* and a *deep system*: the superficial system is made up of veins that are close to the skin. These are the blood vessels that can be seen on hands or arms. The deep system is comprised of veins within the muscles of the body. They are connected by small communicating veins where the body regulates the amount of blood going through both systems, as a way of rigidly controlling the body's central temperature. A Deep Vein Thrombosis (DVT) is therefore a condition where a blood clot forms in a vein of the deep system.

DVT can occur anywhere in the body but is most frequently found in the deep veins of the legs or thighs. Several factors may cause deep vein thrombosis including injury to the vein, slowing of blood flow, and conditions that increase the tendency for the blood to clot. The most common cause of injury to a vein is trauma to the leg, such as occurs with broken bones, severe muscle injury, or surgery. Immobilization is the most common cause of slow blood flow in a vein, because movement of the leg muscles helps keep blood flowing through the deep veins. People who have conditions like one or some of the following are at a higher risk for developing deep vein thrombosis:

- Paralysis from a stroke or spinal cord injury
- A cast on a leg to help heal a fractured bone
- Confinement in bed due to a medical or surgical condition
- Prolonged sitting, especially with crossed legs, in a car, train, or airplane

5.1.2 What does deep vein thrombosis cause?

The most common symptoms of a deep vein thrombosis in the leg are swelling and pain in the affected leg. These symptoms are caused by the accumulation of blood that is unable to get past the clot in the vein and the resulting leakage of fluid from the blood into the muscle. Many other conditions exhibit symptoms similar to those of a deep vein thrombosis, for example, muscle strains, skin infections, and inflammation of superficial veins. A deep vein thrombosis, therefore, is difficult to diagnose without specific tests in which the deep vein system can be examined. Furthermore, many patients with a deep vein thrombosis have no symptoms at all unless the clot dislodges, travels to the lung,

and causes a pulmonary *embolism*. In this case, the patient may develop a rapid heart rate, shortness of breath, sharp chest pain that worsens with deep breathing, or coughs up blood. If the pulmonary emboli are large and block one or both of the major pulmonary arteries sending blood to the lungs, the patient may develop a very low blood pressure, pass out, and possibly die from lung or heart failure. In the case of deep vein thrombosis, however, many other conditions, for example, a heart attack or pneumonia, can mimic a pulmonary embolism. Therefore, specific tests must be done to confirm the diagnosis.

5.1.3 Using venography to diagnose deep vein thrombosis

Several tests, each of which has certain advantages and limitations, can be used to diagnose the presence of deep vein thrombosis. The oldest of these tests is *venography*; this test is performed by injecting a radiopaque fluid into a vein on the top of the foot. The dye flows with the blood and fills the veins of the leg as well as the thigh. An obstructing blood clot in one of these veins can be seen on an X-ray as a dye-free area within the vein. Venography is the most accurate test to identify deep vein thrombosis, but it is painful, expensive, and occasionally can cause painful inflammation of the veins. Furthermore, venography requires a high degree of expertise to perform and interpret correctly.

5.1.4 Deep vein thrombosis and ICD10

The ICD10 catalog contains several entries for Thrombosis. Examples for ICD10 codes are D69, I80, and D69.

5.1.5 Where should we start?

The first stage in the generic method is to translate the business issue you are trying to address, into a question, or set of questions that can be addressed by data mining.

The challenge in this chapter is now to find groups of patients who share a similar behavior during a deep vein thrombosis diagnosis test.

5.2 The data to be used

You clearly cannot do data mining without having the data about your patients to mine. But what data do you need? *The second stage in our data mining method* is to identify the data required to address the business issue and where we are going to get it from.

The underlying data was collected from a hospital. Each patient came to the clinic of the hospital as recommended by his or her home doctor or a general physician in the local hospital. Then he or she was tested for thrombosis using standard tests as described above. As a result the patient got the information whether he or she has thrombosis or not.

The following data was used:

- Demographic data
- Data from special medical tests that were done by the hospital
- Data from historical examination tests

In the following section we describe how this data is used to create a new datamart that contains all the relevant information. Furthermore, we calculate some new variables.

5.3 Sourcing and preprocessing the data

To create our data model we have to take the raw data that we collect and convert it into the format required by the data models. We call this stage in the process sourcing and preprocessing and this is *the third stage in our data mining method*.

This section describes the different data in detail. Whereas, the demographic data is easy to understand, the data from the medical tests and the historical data need a more deeper medical background. The variables and medical expressions, therefore, will be explained in more detail.

5.3.1 Demographic data

The demographic data contains general information about the patient. The data includes all patients and contains about 1000 records (Table 5-1).

Table 5-1 List of the demographic variables

Variable	Variable Type	Description
ID	CATEGORICAL	Identification of the patient (anonymous)
SEX	CATEGORICAL	Gender of the patient
BIRTHDAY	CATEGORICAL	Birthday of the patient
DATE_OF_REC	CATEGORICAL	The first day when the patient was recorded
DATE_OF_DEL	CATEGORICAL	The date when the patient came to the hospital

Variable	Variable Type	Description
ADMISSION	CATEGORICAL	Indicates whether the patient was admitted to the hospital or followed at the hospital
DIA_2	CATEGORICAL	Diagnosis, **not** coded by ICD10

5.3.2 Data from medical tests

This data is special laboratory examinations that was measured by the hospital. The data does not cover all known patients but only a subgroup. This patients were tested on one day for thrombosis. Table 5-2 shows a list of variables that were recorded when the thrombosis test was done in the hospital.

Table 5-2 Variables that were recorded by the examination test

Variable	Variable Type	Description
ID	CATEGORICAL	Identification of the patient
EXAMINATION_ DATE	CATEGORICAL	Date when a test was done
ACL_IGL	NUMERIC	Anti-cardiolipin antibody concentration of Immunoglobulin type G.
ACL_IGM	NUMERIC	Anti-cardiolipin antibody concentration of Immunoglobulin type M.
ACL_IGA	NUMERIC	Anti-cardiolipin antibody concentration of Immunoglobulin type A.
ANA_PATTERN	CATEGORICAL	Antinucleus antibody concentration
DIAGNOSIS	CATEGORICAL	Diagnosis - **not** coded by ICD10
KCT	BINARY	Measure of degree of coalgulation for KCT
RVVT	BINARY	Measure of degree of coalgulation for RVVT
LAC	BINARY	Measure of degree of coalgulation for LAC
SYMPTOMS	CATEGORICAL	Other symptoms observed
THROMBOSIS	CATEGORICAL	Thrombosis

These variables are:

- *Immunoglobin* (Ig) is a protein that is induced by plasma cells and lymphocytes. Immunoglobulins are an essential part of the body's immune system which attach to foreign substances, for example, bacteria and assists

in destroying them. Some classes of immunoglobulins are, for example, A, M, and G.

- *ANA* is an abbreviation of *antinuclear antibodies* that are directed against the structures within the nucleus of the cells. The nucleus is the inner core within each of the body's cells; it contains the genetic material. ANA are found in patients whose immune system can be predisposed to cause inflammation against their own body tissues. Antibodies that are directed against one's own tissues are referred to as auto antibodies. The propensity for the immune system to work against its own body is referred to as autoimmunity.
- Anti nucleus antibodies are "unusual antibodies" that are directed against the structures within the nucleus of the cells. The nucleus is the innermost core within each of the body's cells and it contains the genetic material. ANA are found in patients whose immune system can be predisposed to cause inflammation against their own body tissues. Antibodies that are directed against one's own tissues are referred to as auto antibodies. The propensity for the immune system to work against its own body is referred to as autoimmunity. ANA indicate the possible presence of autoimmunity.
- KCT, LAC, and RVVT are binary variables that indicate either a high (+) or low (-) examination value.
- *SYMTOMPS* refers to other diagnoses; it contains, for example, brain infarct, epilepsy or CVA. CVA is a sudden death of some brain cells due to lack of oxygen when the blood flow to the brain is impaired by blockage or rupture of an artery to the brain.

5.3.3 Historical medical tests

This historical data is based on examination results that were done over a longer period of time. The data does not cover all known patients but only a subgroup. Table 5-3 shows the variables of these tests:

Table 5-3 Data from historical medical tests

Variable	Variable Type	Description
ID	CATEGORICAL	Identification of the patient
DATE	CATEGORICAL	Date when the test was done
GOT	CONTINUOUS	Glutamin oxaloacetic transaminase
GPT	CONTINUOUS	Glutamin pylvic transaminase
LDH	CONTINUOUS	Lactate Dehydrogenase
ALP	CONTINUOUS	Alkaliphosphotase
TP	CONTINUOUS	Total number of proteins

Variable	Variable Type	Description
ALB	CONTINUOUS	Albumin
UA	CONTINUOUS	Urid Acid
UN	CONTINUOUS	Urin Nitrogen
CRE	CONTINUOUS	Creatinine
T-BIL	CONTINUOUS	Bilirubin
T-CHO	CONTINUOUS	Cholesteron
TG	CONTINUOUS	Triglyceride
CPK	CONTINUOUS	Creatinine Phosphokinase
C4	CONTINUOUS	Complement 4
WBC	CONTINUOUS	Number of white blood cells in a volume of blood
RBC	CONTINUOUS	Number of red blood cells
HGB	CONTINUOUS	Hemoglobin
HCT	CONTINUOUS	Hematocrit
PLT	CONTINUOUS	Platelet
IGG	CONTINUOUS	Immunoglobin G

These variables are:

- *LDH* (Lactate Dehydrogenase) is an enzyme that catalyzes the conversion of lactate to pyruvate. This is an important step in energy production in cells. Many different types of cells in the body contain this enzyme. Some of the organs relatively rich in LDH are heart, kidney, liver, and muscle.
- *T-ALB* (Albumin) is the main protein in human blood and the key to the regulation of the osmotic pressure of blood.
- *CREATININE* is a chemical waste molecule that is generated from muscle metabolism. It is produced from creatine, a molecule of major importance for energy production in muscles. Creatinine is transported through the bloodstream to the kidneys. The kidneys filter out most of the creatinine and dispose of it in the urine.
- *T-BIL* (Bilirubin) is a yellow-orange compound produced by the breakdown of hemoglobin from RBC.
- *RBC* (Red blood cells) are the cells that carry oxygen and carbon dioxide through the blood.

- *HGB (*Hemoglobin) is a pigment in the red blood cells. It forms an unstable, reversible bond with oxygen. In its oxygenated state it is called oxyhemoglobin and is bright red. In the reduced state it is called deoxy-hemoglobin and is purple-blue.
- *HCT* (Hematocrit) is the proportion, by volume, of the blood that consists of red blood cells.
- *PLT* (Platelets) are the smallest cell-like structures in the blood and are important for blood clotting and plugging damaged blood vessels.

5.4 Evaluating the data

Having created and populated our data models, *the fourth stage in our data mining method* is to perform an initial evaluation of the data itself.

This section describes the evaluation of the data that acts as a prerequisite for the establishment of a final common datamart. We start with a statistical overview and argue that this should be a first step in getting a feeling for the data.

5.4.1 Demographic data

Figure 5-1 shows the distribution of the demographic data using univariate statistics. The data contains records with approximately 80% females and 20% males. For *ADMISSION*, about 60% of the patients were followed to the hospital; 40% patients were admitted.

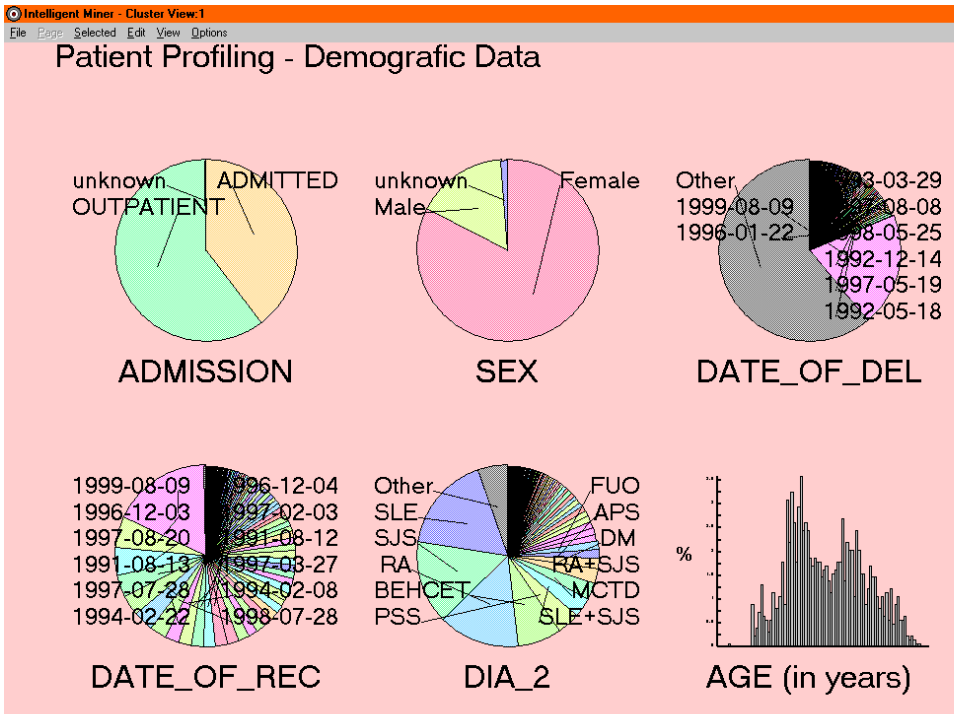


Figure 5-1 Distribution of the demographic data

DATE_OF_REC and DATE_OF_DEL have many distinct values, and it may be therefore suitable to have alternative variable instead (see 5.4.4, “Building a datamart” on page 85).

5.4.2 Data from medical tests

Figure 5-2 describes the distribution of the variables from the laboratory examinations.

- KCT, LAC, and RVVT either have values LOW or HIGH
- ANA is either expressed by a D, S, P or a combination out of these values. The first pattern indicates the main diagnosis, the second one the subdiagnoses.

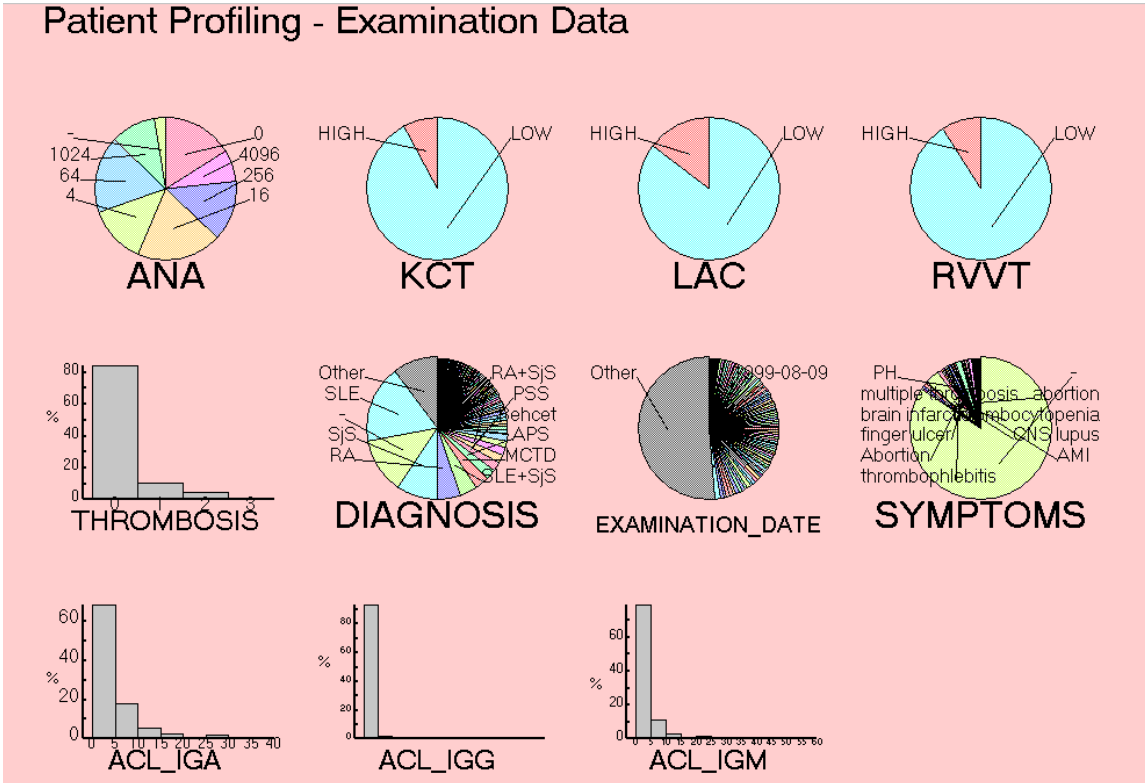


Figure 5-2 Distribution of the data from medical tests

- *THROMBOSIS* has several values:
  - A value of 0 indicates that the patient has no thrombosis
  - A value of 1 indicates that the patient has thrombosis of high degree
  - A value of 2 indicates that the patient has thrombosis of middle degree
  - A value of 3 indicates that the patient has thrombosis of low degree
- *EXAMINATION_DATE* refers to the date when the patient was examined and a test was done.

5.4.3 Historical medical tests

Figure 5-3 shows the distribution of the historical data using univariate statistics.

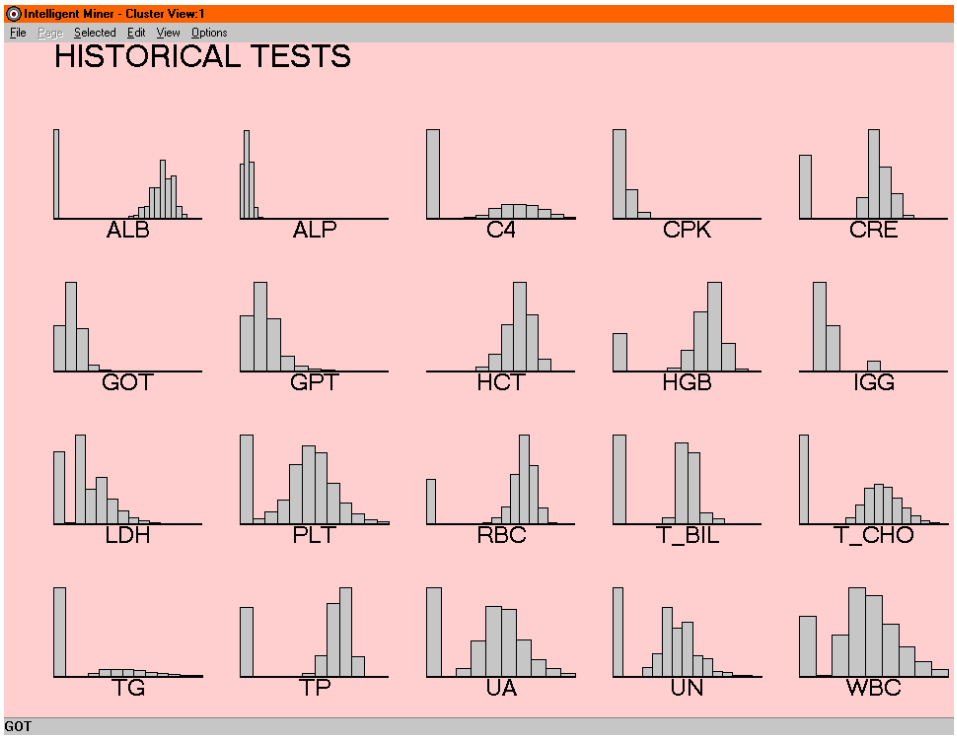


Figure 5-3 Distribution of the data from historical medical tests

Most of the variables contain values that are not taken from an examination but were recorded in order to overcome missing values treatment: these values are indicated on the left side of each histogram. Nevertheless, the distributions are similar in some senses.

5.4.4 Building a datamart

For the analysis, the data sources need to be joined. Because *ID* occurs in all three datamarts (Figure 5-4), we can perform SQL to create a new table. We then get a datamart containing of 39 variables approximately 20,000 records.

However, there is a discussion which variables should be used and how: eventually, new variables have to be calculated. As an example of why this is useful, consider the two variables, *DATE_OF_REC* and *DATE_OF_DEL,* from 5.4.1, “Demographic data” on page 82.

- *DATE_OF_REC* contains about 98 different dates.
- *DATE_OF_DEL* contains about 791 different dates.

Because both variables seem to be very useful, the following new variables are defined (measured in years):

- *DeliveredSinceN* denotes the year when the delivery was done.
- *RecordedSinceN* denotes the year when the first recording was done.

The advantage is to have numerical variables instead; these are easily interpretable.

We have a similar problem for the second data (examination data):

- *EXAMINATION_DATE* has 454 different values.
- *DIAGNOSIS* has 187 different values.

Both are categorical values, and a treatment of categorical values with many different values is always very expensive when we start mining the data.

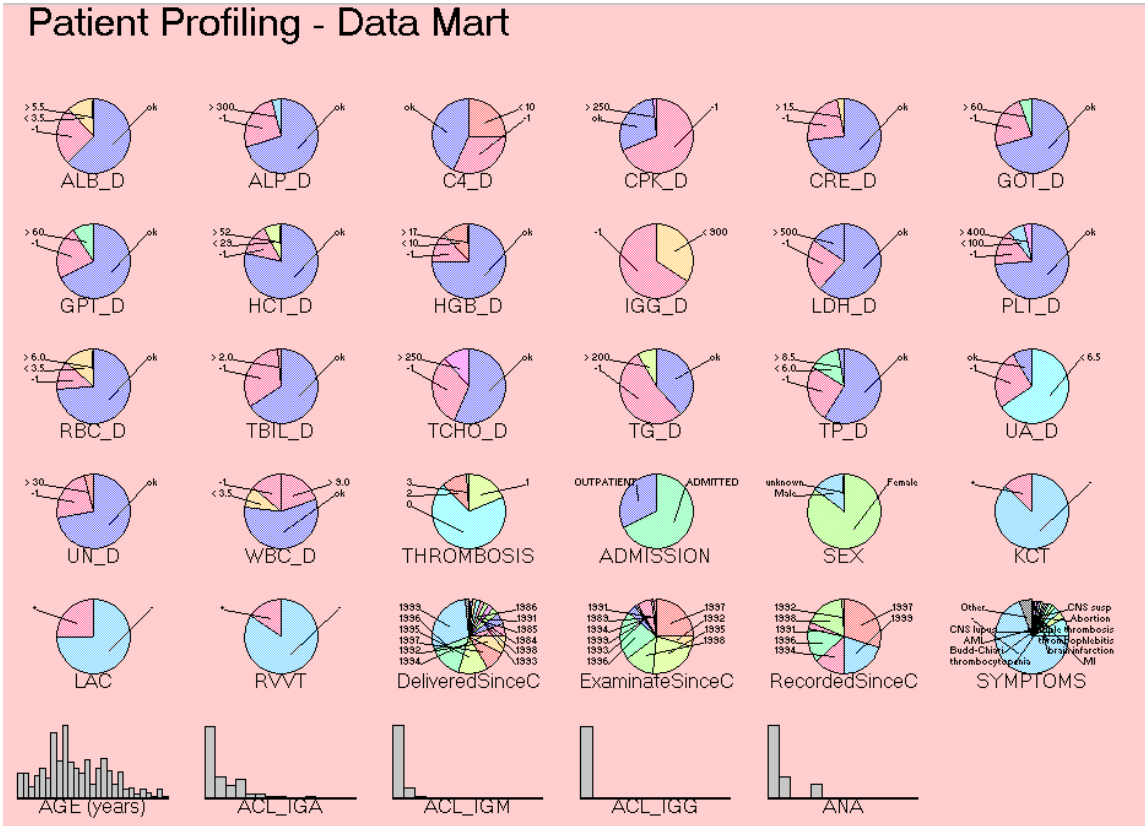


Figure 5-4 Aggregated datamart that contains variables from all three data sources

Therefore, we introduce two new variables that will be used instead:

- *ExaminateSinceN* denotes the year when the examination was done.
- Dia2P contains more generalized medical terms as it is specified in *DIAGNOSIS*.

For the historical data it may be better to have more readability in the variables. Because the normal ranges and abnormal ranges of the variables are known, it is easy to discretize the corresponding variables. In detail, the following discretizations were applied by using corresponding medical sources (Table 5-4).

Table 5-4 Using discretization to transform numerical variables to categorical variables

Variable	Normal Ranges
GOT_D	GOT < 60
GPT_D	GPT < 60
LDH_D	LDH < 500
ALP_D	ALP < 300
TP_D	6.0 < TP < 8.5
ALB_D	3.5 <= ALB < 6.5
UA_D	UA > 6.5
UN_D	UN < 30
CRE_D	CRE < 1.5
TBIL_D	TBIL < 2.0
TCHO_D	TCHO < 250
TG_D	TG < 200
CPK_D	CPK < 250
C4_D	C4 > 10
WBC_D	3.5 <= WBC < 9.0
RBC_D	3.5 <= RBC < 6.0
HGB_D	10 <= HGB < 17
HCT_D	29 <= HCT < 52
PLT_D	100 <= PLT < 400
IGG_D	900 <= IGG < 2000

As a final result, we then get the datamart as shown in Figure 5-4. The number of

records is approximately 20,000.

5.5 Choosing the mining technique

Choosing the mining techniques to use is *the fifth stage in our generic mining method*. This section describes the strategy we use for patient profiling as well as the application of the suggested strategy.

5.5.1 Choosing segmentation technique

Here we perform patient profiling through segmentation. Segmentation is a common term that describes the method to find segments within a datamart. The segments are characterized by the dispersion of the values of these input variables in the input itself. Patients who have a similar behavior will probably be in the same segment.

On the other hand, segmentation is a tedious process, because, from a technical point of view, variables can be correlated or redundant. Therefore, one main goal before applying segmentation is a strategy to find a good data model.

The important point here is quality: quality means both the quality of the datamart as well as quality of the variables itself. But good variables do not automatically define a good data model, and good variables are not given automatically by nature. They have to been carefully examined and, in case of incorrectness to be transformed.

Sometimes we loose some variables because of corruptness or correlation. For correlation, statistical functions like Factor Analysis or Principal Component Analysis (PCA) are suitable, but they are applicable only for numerical, not for categorical variables.

However, we have a lot of categorical values within the aggregated datamart (see 5.3, “Sourcing and preprocessing the data” on page 78). Therefore, numerical variable treatment by Factor Analysis or Principal Component Analysis would not be applicable.

As already discussed in Chapter 6., “Can we optimize medical prophylaxis tests?” on page 111 classification trees define a predictive technique to find characteristics for a given class label. Within the generation process, classification trees always use the variable that splits the current datamart the best. Variables that are of minor importance in this sense are neglected and remain and, in some cases, without any use.

Therefore, classification trees are appropriate in the search for categorical variables. Those variables that are more important to characterize the datamart will occur; those that are less important do probably not occur.

To summarize, we will continue as follows:

- Apply *classification trees* as preprocessing to reduce the number of categorical variables (if necessary).
- Perform segmentation to find segments with patients who share the same behavior for thrombosis. This will be done by *demographic clustering*.

Nevertheless, we will use these variables that will be removed as supplementary variables.

5.5.2 Using classification trees for preprocessing

There are two strategies to apply classification tree for variables reduction:

- **Top Down**
  - Start with all categorical variables and perform *classification*
  - Use only the best variables of the classification tree or remove those variables that do not occur in the tree
  - Continue with the minor variable set
- **Bottom Up**
  - Divide the categorical variables into distinct variable sets and perform Classification for each variable set
  - Use only the best variables in the classification tree or remove those variables that do not occur in the tree
  - Summarize the remaining categorical variables and test the whole model

The advantage of the Bottom Up strategy is that we will get classification trees that describe the class label with only this variable set. This is sometimes necessary if only these variables should included into the model.

In this chapter, we will use Bottom Up strategy by grouping the variables according to their original sets (see 5.4, “Evaluating the data” on page 82) into

- Variable group “*DEMOGRAPHIC*”
- Variable group “*EXAMINATION*”
- Variable group “*HISTORICAL*”

Variable Group DEMOGRAPHIC

For *DEMOGRAPHIC* we use the following variables:

- ADMISSION
- *RecordedSinceC* (from DATE_OF_REC)
- *SEX*
- BIRTHDAY

Dia2P (from DIA_2) cannot be used, because it is obviously correlated with class label *THROMBOSIS*. The same occurs to *DeliveredSinceC* (from DATE_OF_DEL), because all patients who were diagnosed with thrombosis were delivered to the hospital. Therefore, the valid variable group *DEMOGRAPHIC* contains only four variables.

By applying the classification tree technique we get, as the most important variable, *BIRTHDAY*, followed by *ADMISSION* and *RecordedSinceC*. *SEX* occurs one time at level 5.

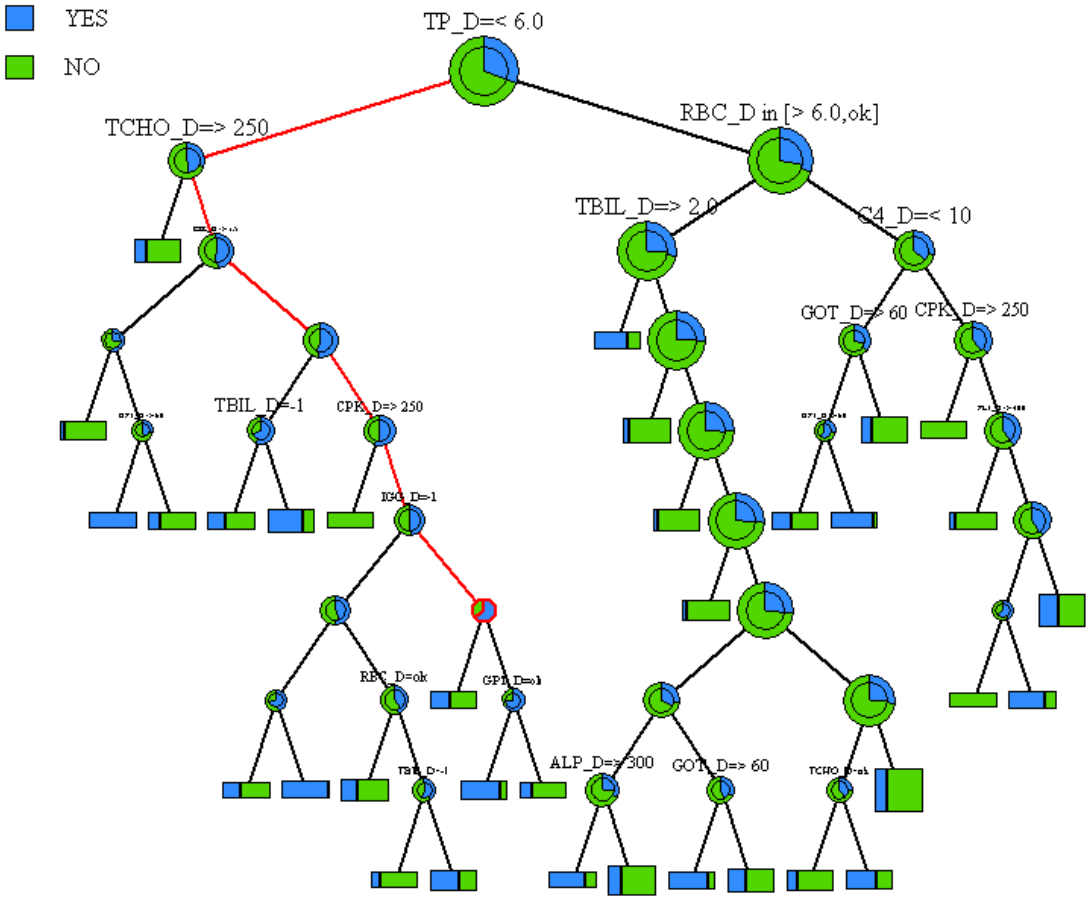


Figure 5-5 Applying classification tree to the variable group EXAMINATION

However, *BIRTHDAY* is very dominant and delivers a quality of more than 90%. It could be that *BIRTHDAY* has some impact to *THROMBOSIS,* because Classification without *BIRTHDAY* delivers a smoother tree with quality around 80%. It is hard to say whether *BIRTHDAY* is influencing *THROMBOSIS,* but it is probable, because this data was recorded during an examination. We can leave the variable in the variable list, but it allows too strong of a domination.

We cannot include the variable BIRTHDAY in the variable list as it has strong correlation, but we will transform it to use the numerical variable *AGE* instead.

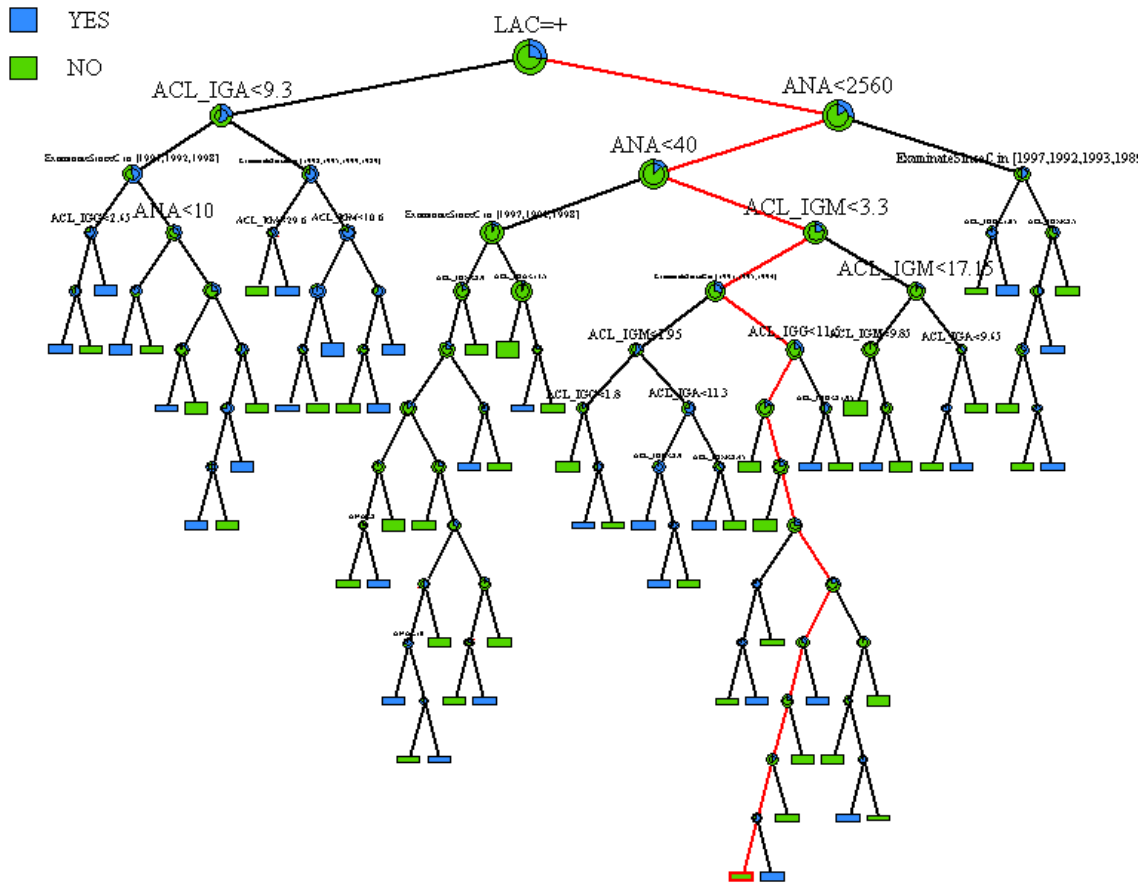


Figure 5-6 Applying Classification Tree to the variable group HISTORICAL

Variable Group *EXAMINATION*

For the variable group *EXAMINATION* we use the following variables:

- ExaminedSinceC (from Examination_Date)

- ACL_IGL
- *ACL*_IGM
- ACL_IGA
- ANA
- KCT
- RVVT
- LAC

SYMPTOMS cannot be used because it contains values that are correlated with class label *THROMBOSIS*. We then get a classification tree as it is shown in Figure 5-5.

The most important variable is *LAC*, followed by *ACL_IGA* and *ANA*. *ExaminateSinceC* and *ACL_IGM* occur on level 3 and 4. However, *RVVT*, KCT and *ACL_IGG* do not occur in the tree. They will be removed from the categorical variable list but further will be considered as supplementary.

However, the resulting Confusion Matrix shows that all records are classified for almost 99%. This is quite nice, but it indicates that some correlations will definitely be present. Therefore, we had better remove at least *LAC*, because this variable may be the reason for this correlation.

However, additional mining runs with a reduced set of variables instead to show a similar behavior: the tree classifies for more than 95%. Another way to identify variables that are highly correlated to each other is to use a correlation matrix which is available in the Principal Component Analysis (PCA); this is called a dimensionality reduction technique, which means to reduce the number of variables; however, this applies to numeric variables only.

To ensure an appropriate datamart we remove all variables from the variable list, but keep them as supplementary.

Variable Group *HISTORICAL*

For the variable group *HISTORICAL* we use all variables (see Table 5-4 on page 87). We then get a classification tree as shown in Figure 5-6.

The most important variable is *TP_D*, followed by *TCHO_D* and *RBC_D. TBIL_D, C4_D, GOT_D, CPK_D, CRE_D* follow on the next levels. Although we have a lot of categorical variables, all appear in the classification tree. We therefore will use all variables in our segmentation model.

A new datamart: DEMOGRAPHIC, EXAMINATION, HISTORICAL

As we discussed for the Bottom Up strategy, we can combine the different variable groups. Because we removed all variables from *EXAMINATION* and most from *DEMOGRAPHIC*, we could think that the variables from *HISTORICAL* may be the important ones. Nevertheless, the resulting classification tree produces a model that has ADMISSION on the top, followed by TP_D and RecordedSinceC. The quality of the model is actually quite acceptable (almost 81%), no variable reduction is therefore necessary.

5.5.3 Applying the model

According to the classification process we use all categorical variables that appear in the variable lists *DEMOGRAPHIC*, *EXAMINATION* and *HISTORICAL* as well as all the numerical variable *AGE* as active. Variables that were removed from the list become supplementary.

We perform segmentation to find segments with patients who share the same behavior for thrombosis. This is done by *by using Condorcet value and demographic clustering.*

Condorcet value: technical definition

Note: The *Condorcet* of the *demographic clustering* is a quality value that measures the goodness of the records belonging to a cluster. The Condorcet is computed as following:

- Normalize all numerical variables.
- Take the first two records and compare them for each variable.
  - If the value of the variable is the same, then increase the number of similar values by one; otherwise, do not.
  - If all values of the common variables are compared, then these records get a similar value that shows the similarity between these two records.
- This procedure is done for all pairs of records. We then get a *similarity matrix* of size n times n (n is the number of variables) where all diagonal entries contain a number of n (because each record is exactly similar to itself) and all other entries a value < n.
- We then have to restructure the similarity matrix depending on which records belong to which cluster.

The Condorcet is then computed by adding all entries of each identified cluster by dividing through the maximum similarity. Because of this normalization step, the Condorcet gets between 0 and 1.

Figure 5-7 shows an example of how the Condorcet is calculated. Then we get a similarity matrix that gives the number of similar (left) and unsimilar matches (right) between two records. For example, the similarity between F and G is 3; between F and B just 1. It is obvious that the similarity values in the diagonal are always the maximum number of similar matches, because, for example, F and F represent the same record.

Similarities and Differences

Example clustering: {FGIJ} {ABC} {DE}{H}

Within Cluster Similarities — **Sum = 74**

	F	G	I	J	A	B	C	D	E	H
F	3	2	3	2	0	0	0	1	1	1
G	2	3	2	3	0	0	0	0	0	2
I	3	2	3	2	0	0	0	1	1	1
J	2	3	2	3	0	0	0	0	0	2
A	0	0	0	0	3	2	1	3	1	0
B	0	0	0	0	2	3	2	2	2	1
C	0	0	0	0	1	2	3	1	1	0
D	1	0	1	0	3	2	1	3	3	1
E	1	0	1	0	1	2	1	3	3	1
H	1	2	1	2	0	1	0	1	1	3

Between Cluster Difference — **Sum = 168**

	F	G	I	J	A	B	C	D	E	H
F	0	1	0	1	3	3	3	2	2	2
G	1	0	1	0	3	3	3	3	3	1
I	0	1	0	1	3	3	3	3	3	2
J	1	0	1	0	3	3	3	3	3	1
A	3	3	3	3	0	1	2	0	2	3
B	3	3	3	3	1	0	1	1	1	2
C	3	3	3	3	2	1	0	2	2	3
D	2	3	2	3	0	1	2	0	0	2
E	2	3	2	3	2	1	2	0	0	2
H	2	1	2	1	3	2	3	2	2	0

Figure 5-7 Demographic clustering: how to calculate the Condorcet value if we have three variables

The colored areas represent the identified clusters. We get the areas by sorting the records in a way that certain records with similarity stand together; whereas, records with an unsimilar relationship do not.

The Condorcet is then calculated either by using both matrices or only the left one (with similar values). As an example, we explain how the Condorcet is calculated in the current version of the IM for Data:

- Summarize all similarity values of each cluster.
- Divide each summarized value by the maximum value that can occur in the cluster.
- Add all normalized values and divide the result by the number of clusters.

Then for Figure 5-8 on page 98, we get the following results:

- Summarizing all similarity values of each cluster yields to 40 (first cluster), 19 (second cluster), 12 (third cluster), 3 (fourth cluster).
- Dividing each summarized value by the maximum gives 0.83 (because the maximum is 48) for the first cluster, 0.70 (because the maximum is 27) for the

second cluster, 1.0 (because the maximum is 12) for the third cluster, and 1.0 (because the maximum is 3) for the fourth cluster.

Adding all normalized values and dividing by the number of clusters leads us then to a Condorcet of 0.88426.

Segmentation by standard demographic clustering

The term “standard demographic clustering” means the application of demographic clustering with standard parameter settings. To be more specific, the following parameter values are important:

- Maximum Cluster
- Similarity

Note: *Maximum Cluster* refers to the number of clusters that are maximally allowed.

Similarity is a threshold that limits the records accepted as best fit for the cluster: if the similarity is high, then only records that satisfy the threshold will be added to the cluster. If the number of Maximum Cluster is low, then records that do not really fit in the cluster (and probably do not satisfy the similarity) must be added to a cluster that allows very deep specification.

A possible strategy is to increase the similarity and the number of the maximum clusters as well (Table 5-5).

Table 5-5 Searching for the best Clustering Model - Part I

No.	No. of Clusters	Similarity	Condorcet	Modify Variables?
1	9	0.50	0.669	NO
2	9	**0.60**	0.711	NO
3	9	**0.70**	0.744	NO
4	9	**0.80**	0.753	NO
5	**9**	0.85	0.749	NO
6	**9**	0.90	0.738	NO
7	7	**0.60**	0.709	NO
8	**20**	**0.80**	0.801	NO
9	**30**	0.80	0.817	NO
10	**50**	0.80	0.831	NO

No.	No. of Clusters	Similarity	Condorcet	Modify Variables?
11	**100**	0.80	0.841	NO
12	**200**	0.80	0.844	NO
13	**500**	0.80	0.846	NO
14	**9**	0.50	0.644	YES, supplementary variables become active
15	**20**	0.80	0.758	YES, supplementary variables become active
16	**50**	0.80	0.799	YES, supplementary variables become active

It is interesting to see that models with all variables (Mining Run 14, 15, and 16) have a worse Condorcet than the corresponding models without these variables (1, 8, 10): preprocessing by tree classification has had a positive influence for the generation of these models.

The best model that can be built is the one done by performing clustering with the parameter setting as it is indicated in Run13. A Condorcet of 0.848 gives us an almost perfect segmentation.

The patient profiling cluster examples are shown in Figure 5-8 and Figure 5-9,

But, are 500 clusters really good to interpret? Should we really use this clustering result as the one that is the best?

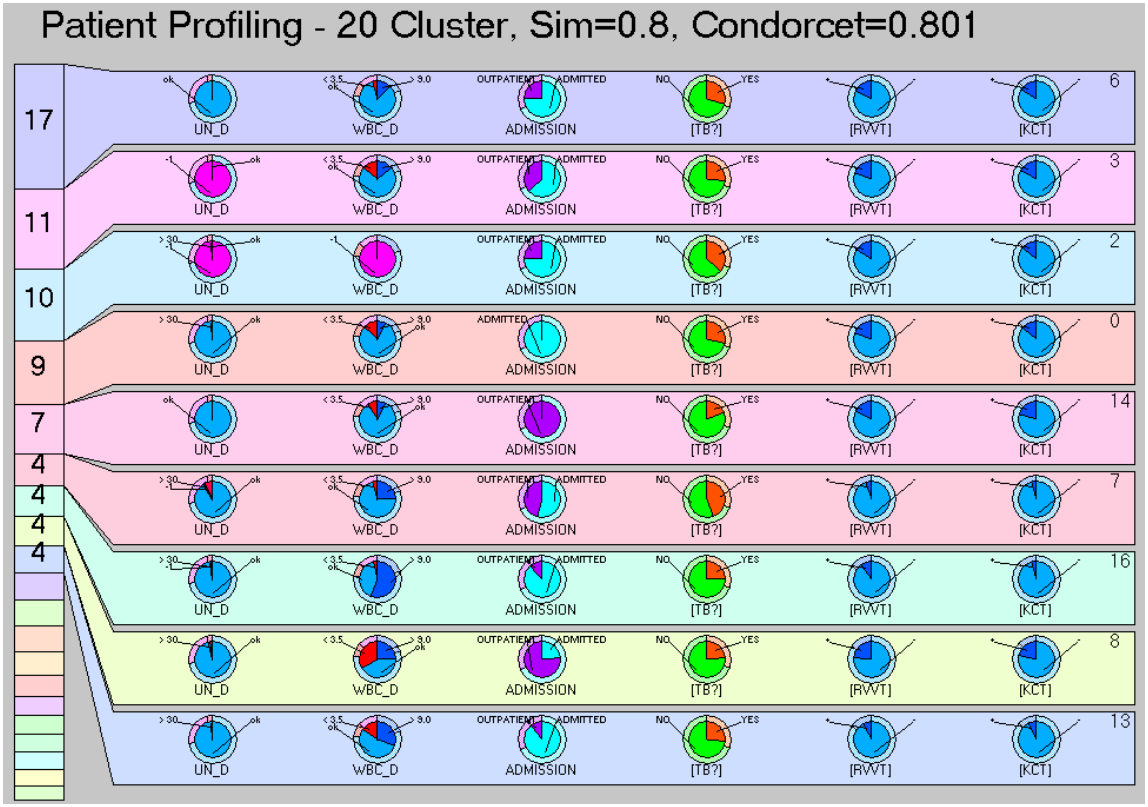


Figure 5-8 Patient profiling: all clusters' view sorted by database view (first part)

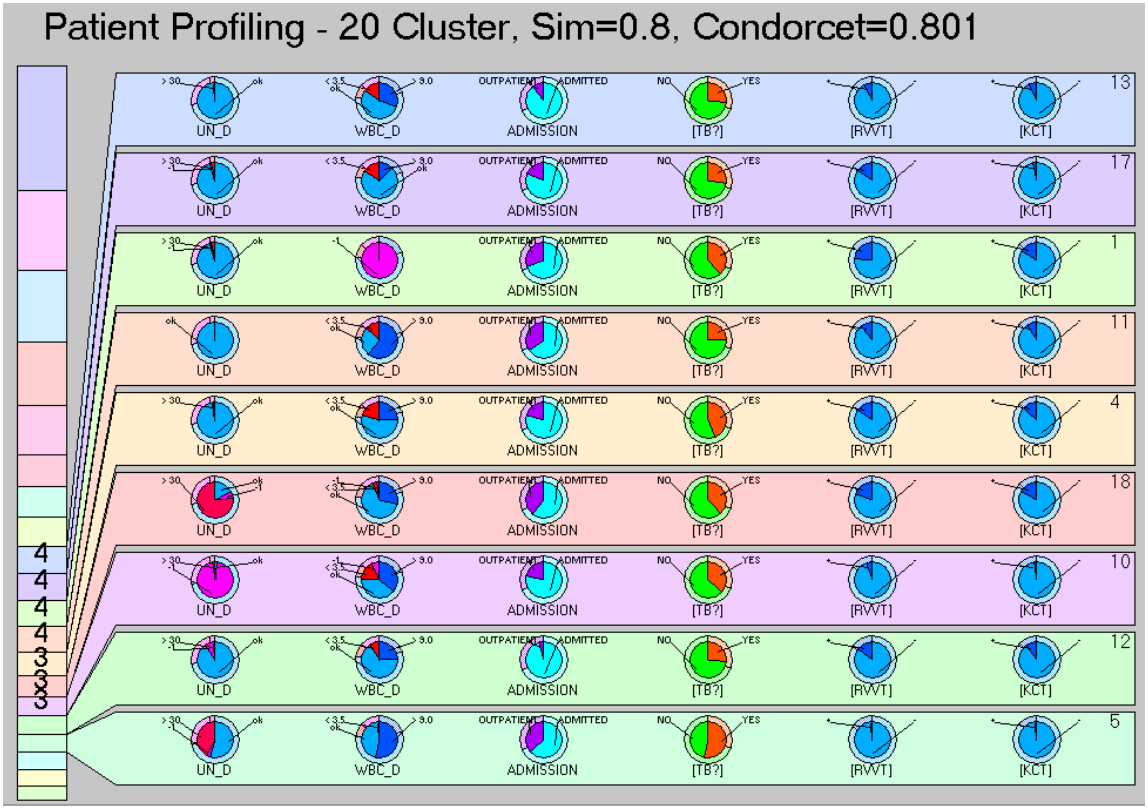


Figure 5-9 Patient profiling: all clusters' view sorted by database view (second part)

The danger here in general is that we get too specific, where all of them probably have a size less than 1%. Furthermore, the readability and interpretation of the clusters become more and more difficult, the more clusters we have: a clustering result with almost 20 clusters is hard to understand, and a clustering result with almost 100 clusters nearly impossible.

Additionally, if we look for the relative improvements of the Condorcet values, it is obvious that the gain of having better results is becoming less after Mining Run 8 (with 20 clusters and a similarity of 0.8). In this respect, it may be more interesting to accept a minor Condorcet with a couple of clusters, rather than having a very high Condorcet with hundreds of small clusters.

On the other hand, niches can be very valuable to be discovered; and, small clusters with high Condorcet may be appreciated, because they can represent unusual or deviant behavior.

Generally, it is the kind of task that finally decides which way we may choose.

In the following section, we use the result of the clustering model that was done with parameter settings of Mining Run 8.

5.6 Interpreting the results

In the previous section we looked at the steps we have to follow to get our mining results using the segmentation technique. The *sixth stage in our generic mining method* is to interpret the results that we have obtained and determine how we can map them to our business. When you are first confronted with the segmentation results, the first question to ask is “What does it all mean?”. In this section we describe how to understand and read and interpret the results.

The resulting cluster can be seen in Figure 5-8 and Figure 5-9.

Cluster 6 contains about 17% of the patients, followed by Cluster 3 (11%) and Cluster 2 (10%). The small clusters represent niches that have a small size but are of highest interest, because they show a relatively high percentage of patients with thrombosis (Cluster 1, 4, 5, 10, and 18 in Figure 5-9).

Now, we will concentrate on Cluster 4 and 5.

5.6.1 Understanding Cluster 4

Cluster 4, shown in Figure 5-10, contains about 3.29% of all patients. The different (active) variables can be interpreted in Table 5-6:

Table 5-6 Cluster 4 - active variables with percentages for this cluster versus all patients

Variable	Semantic interpretation	Relevance
ALB_D	most of the tests were ‘< 3.5’ (65% vs. 34%), and only 34% ‘ok’ (34% vs. 65%).	High
ALP_D	most of the tests were ‘ok’ (90% vs. 70%), 5% of ‘>300’.	Low
C4_D	There are more tests ‘ok’ than in the whole population (55% vs. 42%)	Low
CPK_D	There are more tests ‘ok’ than in the whole population (39% vs. 29%), but the number of missing values are still high (58% vs. 68%)	Low
CRE_D	All tests are ‘ok’; this is very interesting, because about 25% in the whole population are missing.	Middle
GOT_D	Most of the tests are ‘ok’ (94% vs. 70%); almost no test contains missing value (0.1% vs. 24%)	Low
GPT_D	Most of the tests are ‘ok’ (94% vs. 70%); almost no test contains missing value (0.1% vs. 24%)	Low

Variable	Semantic interpretation	Relevance
HCT_D	Most of the tests are '<29' (88% vs. 8%); only 11% of the tests are 'ok' (vs. 78%)	High
HGB_D	All tests show values that are '<10' (100% vs. 12%).	High
IGG_D	There are less tests with '<300' than in general (24% vs. 34%): most of the tests contain missing values (76%)	Middle
LDH_D	There are more tests 'ok' than in general (72% vs. 61%): however, the number of values '>500' is also increased (27% vs. 15%)	Middle
PLT_D	There are a lot of tests with values '>400' (16% vs. 4%); most of the tests are 'ok' (80% vs. 73%), only some of them remain missing (0.3% vs. 16%)	Middle
RBC_D	The number of values '<3.5' has extremely increased (85% vs. 13%); only 14% of the tests are ok (vs. 73%)	High
RecordedSinceC	There are different changes (1998: 25% vs. 17%), but in general no relevant behavior.	Low
SEX	There are more women than men (94% vs. 85%): only 6% are men (vs. 15%)	Middle
TBIL_D	Most of the tests are 'ok' (96% vs. 65%); the number of missing values is extremely low (3% vs. 33%)	Low
TCHO_D	70% of the test are 'ok' (vs. 56%), the number of missing values is quite low (12% vs. 32%)	Low
TG_D	Most of the tests are 'ok' (65% vs. 38%); the number of values '>200' has increased (13% vs. 9%)	Low
TP_D	There is a tremendous increase on tests with values '<6.0' (50% vs. 13%).	High
UA_D	The number of values '<6.5' has increased (88% vs. 65%)	High
UN_D	Almost all tests were 'ok' (98% vs. 71%)	Low
WBC_D	The number of values '<3.5' has doubled (20% vs. 10%); there are more values '>9.0' than for all patients (24% vs. 20%)	Middle
ADMISSION	80% of the patients were admitted to the hospital (vs. 67%).	High
TB?	Almost 43% of the patients have thrombosis (vs. 31%)	High

If we summarize the information, then the variables with high relevance show the following characteristics:

- Albumin (ALB_D) is mainly '<3.5'

- Hematocrit (HCT_D) is mainly ‘<29’
- Hemoglobin (HGB_D) is always ‘<10’
- The red blood cells (RGB_D) is mainly ‘<3.5’
- The total number of proteins (TP_D) is mainly ’<6.0’
- Urin acid (UA_D) is mainly ‘>6.5’

If we take a look at the supplementary variables, we can observe that many of the female patients are much younger, especially between 17 and 20 (8% in this cluster compared to 3% in the whole population), 25 and 27 (15% in this cluster compared to 10% in the whole population) and 40 and 42 (14% in this cluster compared to 6% in the whole population).

The most typical symptom is “CNS lupus” (25% in this cluster compared to 8% in the whole population) that is a chronic inflammatory condition caused by autoimmune disease: an autoimmune disease occurs when the body’s tissues are attacked by its own immune system. Patients with lupus have unusual antibodies in their blood that are targeted against their own body tissues (see `http://www.medicinenet.com`); nearly 58% (compared to 69%) of the symptoms are missing.

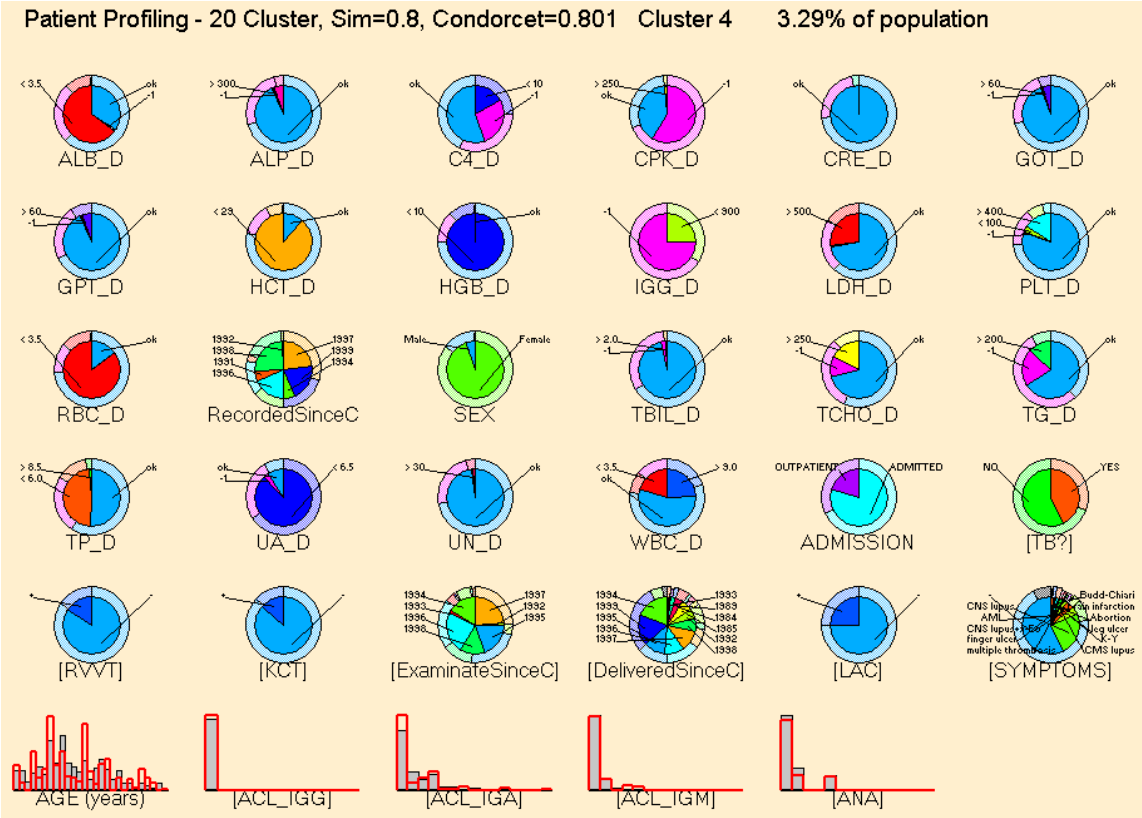


Figure 5-10 Patient Profiling: Cluster 4 with 3.29% of all patients

5.6.2 Understanding Cluster 5

Cluster 5, shown in Figure 5-11 on page 105, contains about 2.42% of all patients. The different (active) variables can be interpreted in Table 5-7.

Table 5-7 Cluster 5 - active variables with percentages for this cluster vs. all patients

Variable	Semantic interpretation	Relevance
ALB_D	Most of the tests are ‘ok’ (90% vs. 62%)	Low
ALP_D	Most of the patients have values ‘>300’ (53% vs. 5%)	High
C4_D	Most of the patients have values ‘<10’ (67% vs. 25%); the number of tests that are ‘ok’ is really small (8% vs. 43%)	High
CPK_D	Most of the tests contains missing values (94% vs. 68%)	Low

Variable	Semantic interpretation	Relevance
CRE_D	There is a large number on patients with values '>1.5' (45% vs. 3%); the number of tests with 'ok' has decreased (52% vs. 73%)	High
GOT_D	Most of the tests are 'ok' (93% vs. 70%)	Low
GPT_D	The number of patients with values '>60' has increased (27% vs. 9%); there are almost no missing values (1% vs. 24%)	Middle
HCT_D	Most of the tests are 'ok' (98% vs. 78%)	Low
HGB_D	Most of the tests are 'ok' (90% vs. 74%)	Low
IGG_D	No important behavior; very similar to the whole population	Low
LDH_D	No important behavior; very similar to the whole population	Low
PLT_D	There is lot of patients with values '<100' (20% vs. 7%)	Low
RBC_D	Most of the patients have values 'ok' (95% vs. 73%)	Low
RecordedSinceC	There are many patients who were recorded in 1998 (57% vs. 17%)	Middle
SEX	Most of the patients are male (90% vs. 15%)	High
TBIL_D	There is a large number of patients with values '>2.0' (25% vs. 2%)	Middle
TCHO_D	There is a large number of patients with values '>250' (31% vs. 12%)	Middle
TG_D	There is a large number of patients with values '>200' (22% vs. 8%)	Middle
TP_D	There is a large number of patients with values '<6.0' (32% vs. 13%)	Middle
UA_D	More than half of the patients' values are 'ok' (53% vs. 8%)	Low
UN_D	The number of '>30' values has changed dramatically (42% vs. 4%)	High
WBC_D	The number of '>9.0' values has changed dramatically (54% vs. 19%); almost no patient has a value '<3.5' (0.4% vs. 10%)	Middle
ADMISSION	Less patients were admitted to the hospital (62% vs. 56%)	Low
TB?	54% of the patients have thrombosis (vs. 31%)	High

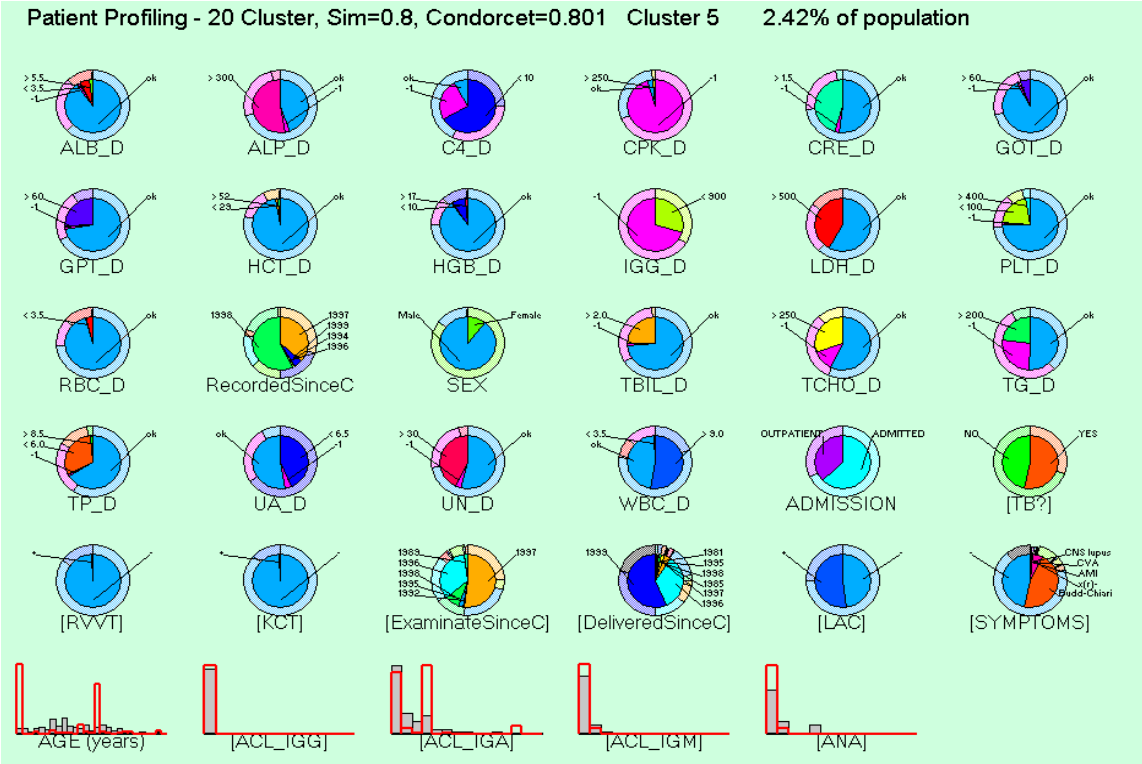


Figure 5-11 Patient Profiling: Cluster 5 with 2.42% of all patients

If we summarize the information, then the variables with high relevance show the following characteristics:

- Alkaliphosphotase (ALP_D) is mainly '>300'
- Complement4 (C4_D) is mainly '<10'
- Creatinine (CRE_D) is mainly '>1.5'
- The gender (SEX) is mainly male
- Urin nitrogen (UN_D) is mainly '>30'

If we look at the supplementary variables, we can observe that many of the male patients are much older, especially between 45 and 47 (33% compared to 4%).

Some of the symptoms were not recorded (46% compared to 69%), the main symptom, however, can be identified as "Budd-Chiari syndrome" (46% compared to 2%).

The “Budd-Chiari” syndrome includes a variety of conditions with obstruction of the hepatic venous outflow, at the level of either the large hepatic veins or the supra hepatic segment of the inferior vena cava. Depending on the cause and pathophysiologic manifestations, distinction is made between primary and secondary forms (see `http://rbrs.org/journal/volume80/page301.html`).

5.7 Deploying the mining results

The final and *seventh stage in our generic mining method* is perhaps the most important of all. How do you deploy the mining results into your business to derive the business benefits that data mining offers? The reason that this is so important is that all to often data mining is seen as an analytical tool that can be used to gain business insight but is difficult to integrate into existing systems.

In this section, after a short summary about the methods we used, we explain how the clustering techniques that we have described can be deployed into your business. We specifically address how recent advances in data mining technology enable you to export the clustering models that you create and import and deploy them to new groups of patients.

5.7.1 What we did so far

In this chapter we proposed a method to perform patient profiling. For this purpose, we performed the following steps:

- Building a datamart out of three different medical datamarts (see 5.4.4, “Building a datamart” on page 85):
  - Datamart that contained the demographics of a patient
  - Datamart that contained the results of an examination test done by physicians
  - Datamart that contained historical data from previous tests
- Performing preprocessing of the variables by using discretization (see 5.5.1, “Choosing segmentation technique” on page 88).
- Testing the categorical variables by using the *classification tree*. The reason was to reduce the number of categorical variables to get a better mining model (see 5.5.2, “Using classification trees for preprocessing” on page 89).
- Performing segmentation by applying demographic clustering. We then could see two clusters with different types of patients with thrombosis (see 5.5.3, “Applying the model” on page 93).

5.7.2 Where can the method be deployed?

Patient profiling is not only a method to explain the characteristics of a group of patients in respect to a specific disease. Moreover, patient profiling can be used to identify new patients who probably have this specific disease.

To be more specific, patient profiling allows the application of an existing profiling system to a new group of patients. It acts as predictive system that is applicable for patients with similar behavior.

The following examples show why patient profiling should be applied in the medical area.

Example 1: Tuberculosis

Tuberculosis (TB) is a communicable disease caused by a bacterium named *mycobacrium tuberculosis* (see `www.medicinenet.com`), often called *tubercle bacillus*. It is spread from person to person through the inhalation of airborne particles containing the bacterium. These particles, also called *droplet nuclei*, are produced when a person with infectious TB of the lung exhales, such as when coughing, sneezing, laughing, speaking, or singing.

These infectious particles can remain suspended in the air and inhaled by someone sharing the same air. TB is transmitted in closed areas where ventilation is poor. The risk of transmission increases when susceptible persons share air for prolonged periods with a person who has untreated pulmonary TB.

If the body's immune system cannot contain the TB bacteria it will continue in producing more and more bacteria. Normally, the infection occurs in the top portion of the lungs, and it may take several months for the symptoms to take affect. Usually, there is a general tiredness or weakness, loss of weight, fever and nightly sweats. If the infection in the lung worsens, further symptoms can include coughing, chest pain, and shortness of breath.

In this respect, there are some arguments that underlay the need of profiling:

- The description of the symptoms refers to characteristics that are already known by physicians. However, it may be the case, that additional, probably unknown symptoms, exist. This may depend on, for example, different countries, different conditions and/or different hygienic situations.
- TB bacteria can mutate from time to time and from place to place, and it is difficult to have the latest bacteria always just in time.
- Sometimes, it is not easy to prove whether a patient has tuberculosis or not. Although TB can be diagnosed in several different ways, for example, by using chest X-rays that reveal evidence of active tuberculosis pneumonia or it may show scarring, suggesting contained inactive TB. It is very important to

have the diagnosis as early as possible. The more advanced the status of the disease, the more difficult it will be to use TB standard tests to recover the patient's health.

Example 2: Narcosis

Another scenario is narcosis that takes place before a patient gets an operation. Normally, the physician decides which kind of narcosis will be used for a patient to be operated on. Sometimes the physician asks the patient to have either a local narcosis or a general one.

However, a narcosis is not so harmless as it seems. If the patient shows some allergy against the medicament this can lead to different forms of pain, for example with lungs, brain or heart. In the worst case, the patient can fall into a coma.

Patient profiling can help to identify patients who have, for example, allergies against a special medicament. Based on this model the physician can use this information as an additional source to ensure the patient's health.

Example 3: Hospital management systems

For the management of a hospital, it is important to know how many patients are admitted. If the capacity of the hospital is full, then only medical service to emergency cases is feasible, and in an extreme case, the patient cannot stay in the hospital for medical care but has to leave, because there are no beds or no rooms available for an operation.

Patient profiling can help to overcome capacity problems, for example:

- Assume, we have records about patients who had the same or similar disease, but stayed different times in the hospital. Reasons for this could either be simple and normal reasons like the different state of health, whether someone smokes or not, and so on. These reasons are already known and often a clear indication of how to manage such cases in respect to the medical services.

  However, there may be some other reasons why a patient's hospital stay could become different.

  As an example, it could be that some patients prefer rooms with windows to the south or the west and therefore have a better regeneration in general; whereas, other patients enjoy to stay in rooms with windows to the north or the east. Of course, at the first look this may be neither logic nor relevant, but it could be a reason for the patients themselves.

- Assume we have records about patients who are delivered to the hospital. Then patient profiling can help, as an additional instrument, estimating which persons probably need to stay in the hospital and which do not.

  As an example, the hospital's computer system could be enriched by rules that were generated by patient profiling. If it remains unclear whether a patient needs to stay locally or really could go home, the application of such rules, and above all the estimation to which the group of patients he or she probably belongs to, could help to overcome capacity problems.

6

Can we optimize medical prophylaxis tests?

In this chapter we will describe how medical patient records can be used to optimize medical prophylaxis tests. By introducing diabetes mellitus — actually one of the most important diseases — we define a method that helps to obtain information about the relevance of different test components.

Because some test components are more important than others, the correct order of these components and/or the right choice of the test components may lead to a faster and more secure strategy to identify diabetes mellitus.

Predictive data mining techniques may be used to perform this task. These techniques allow the generation of prediction models based on historical data that can be used to realize a more effective prophylaxis.

6.1 The medical domain and the business issue

This section contains a small description of what diabetes is. Furthermore, we describe what it causes and the tests to diagnose it when it occurs, and we define the business issue that we should start with.

6.1.1 Diabetes insipidus and diabetes mellitus

Diabetes mainly occurs as one of the following diseases:

- Diabetes insipidus
- Diabetes mellitus

Diabetes insipidus is an endocrine disorder that involves a deficient production or lack of effective action of an antidiuretic hormone (that is *Vasopressin*). An *antidiuretic hormone* (ADH) is made in the hypothalamus, stored in and secreted by the pituitary gland — a small gland that is located below the hypothalamus — and works on the kidney to conserve fluid.

A deficient production of ADH or a lack of effective action of ADH causes a large amount of urine output; it increases thirst, dehydration, and low blood pressure in advanced cases. The average urine volume for a normal adult is about 1.5 liters per day for patients with diabetes insipidus, but can increase to 18 liters daily.

Commonly referred to as 'diabetes', *diabetes mellitus* is a chronic medical condition that is associated with abnormally high levels of glucose in the blood. Elevated levels of blood glucose concentration lead to spillage of glucose into the urine. Normally, blood glucose levels are tightly controlled by insulin, a hormone produced by the pancreas. Insulin lowers the blood glucose level, and when the blood glucose elevates, insulin is released from the pancreas to normalize the glucose level. For patients with diabetes mellitus, the absence or insufficient production of insulin causes hyperglycemia.

Diabetes mellitus is a chronic medical disease that can last a lifetime. Over time, diabetes mellitus can lead to blindness, kidney failure, and nerve damage. It is also an important factor in accelerating the hardening and narrowing of the arteries, leading to strokes, coronary heart diseases, and other blood vessel diseases in the body.

6.1.2 What causes diabetes mellitus?

Diabetes mellitus is caused by an insufficient production of insulin. The early symptoms of untreated diabetes mellitus are related to elevated blood sugar levels, and loss of glucose in the urine. High amounts of glucose in the urine can cause increased urine output and lead to dehydration. Dehydration causes an increased thirst and a consumption of water. The inability to utilize glucose energy eventually leads to weight loss despite an increase in appetite. Some untreated diabetes patients also complain of fatigue, nausea, and vomiting.

Patients with diabetes are prone to developing infections of the bladder, skin, and vaginal areas. Fluctuations in blood glucose levels can lead to blurred vision. Extremely elevated glucose levels can lead to lethargy and coma (diabetic coma).

6.1.3 Tests to diagnose diabetes mellitus

Concerning a prophylactic examination, there is a fast plasma glucose test. It is a preferred test to diagnose diabetes, because it is easy to perform and convenient. After the patient has fasted overnight (at least 8 hours), a single sample of blood is drawn and sent to the laboratory for analysis.

- Normal fasting plasma glucose levels are less than 110 mg/dl.
- Fasting plasma glucose levels of more than 126 mg/dl on two or more tests on different days indicate diabetes.
- If the overnight fasting blood glucose is greater than 126 mg/dl on two different tests on different days, the diagnosis of diabetes mellitus is made.

A random blood glucose test can also be used to diagnose diabetes. Random blood samples (if taken shortly after eating or drinking) may be used to test for diabetes when symptoms are present. A blood glucose level of 200 mg/dl or higher indicates diabetes, but it must be reconfirmed on another day with a fasting plasma glucose or an oral glucose tolerance test.

6.1.4 Where should we start?

The first stage in the generic method is to translate the business issue you are trying to address, into a question, or set of questions that can be addressed by data mining.

The business problem therefore is: Can we correctly order some test components and/or make the right choice of the test components to find a faster, more secure strategy to identify diabetes mellitus?

6.2 The data to be used

You clearly cannot do data mining without having the data about your patients to mine. But what data do you need? *The second stage in our data mining method* is to identify the data required to address the business issue and where we are going to get it from.

6.2.1 Diabetes mellitus and ICD10

The ICD10 catalog contains several entries for diabetes where mostly diabetes mellitus is cataloged. The main groups are E10 through E14 and O24.

6.2.2 Data structure

The data contains examination and test results for patients with suspense of diabetes mellitus. The results were recorded on the basis of a diabetes examination test, based on the criteria of the World Health Organization (WHO). Additionally, a new variable was added that should show whether the patient is diseased or not: a positive examination result was indicated by a '*POSITIVE*' (has diabetes mellitus), a negative result by a '*NEGATIVE*' (does not have diabetes mellitus).

Concerning the patients, we have only female patients who are at least 21 years old of Pima Indian heritage. The number of patients who were examined was in the range of 1000. After finishing the medical tests, two third of the test patients were diseased, one third were healthy (Table 6-1).

Table 6-1 Variables used for the prophylaxis analysis

Variable	Variable type	Description
PREGNANT	CONTINUOUS	Number of times the patient was pregnant.
PLASMA_GLUC_CONC	CONTINUOUS	Plasma Glucose concentration in the blood by a 2hours oral glucose tolerance test
BLOOD_PRESSURE	CONTINUOUS	Diastolic blood pressure
TRICEPS_SKIN	CONTINUOUS	Triceps skin fold thickness [in mm]
INSULIN	CONTINUOUS	Insulin [in ml]
BODY_MASS_INDEX	CONTINUOUS	Body mass index (weight [in kg] divided by squared height [in m]
DIABETES	CONTINUOUS	Diabetes pedigree function

Variable	Variable type	Description
AGE	CONTINUOUS	Age of the patient
INDICATOR	CATEGORICAL	Indicates whether examination test was positive (patient has diabetes mellitus) or negative (patient does not have diabetes mellitus)

6.2.3 Some comments about the quality of the data

There are only approximately 1200 qualitative records available for this analysis. Indeed, this is quite a low number of records, because the application of data mining always requires a large quantity of data.

However, the problem of having qualitative data occurs very often in the medical field and is not potentially new. There are a lot of reasons for this, for example:

- It seems to be irrelevant for the overall diagnosis to perform all tests.
- Medical tests are not done because the test seems to be too expensive.
- Diagnoses are done in an unsatisfactory way, for example by hand-written and unreadable papers and/or unclear dictations.

Additionally, it often occurs that physicians have their own patient records that are, for data security reasons, not available to any other person. However, this also means that physicians do not really know what kind of medical services other physicians do. More expressive patient records are therefore not realistic although this would definitely lead to better health care.

6.3 Sourcing and evaluating data

Having defined our data models, *the third and fourth stage in our data mining method* is to perform an initial evaluation of the data itself and to populate our datamarts.

6.3.1 Statistical overview

All variables are actual and contain no missing values. However, some of them are partially not clean:

- *BLOOD_PRESSURE:* 30 patients share a blood pressure of 0, a sure indication that no examination has taken place.
- *TRICEPS_SKIN*: there is similar situation, because almost 30% of the patients have a skin thickness of 0 cm.

- *INSULIN*: almost 50% of the patients have an insulin concentration of 0.

Nevertheless, all the records that contain these values remain in the data.

As it is shown in Figure 6-1, most of the patients are not diseased (*INDICATOR* = NEGATIVE).

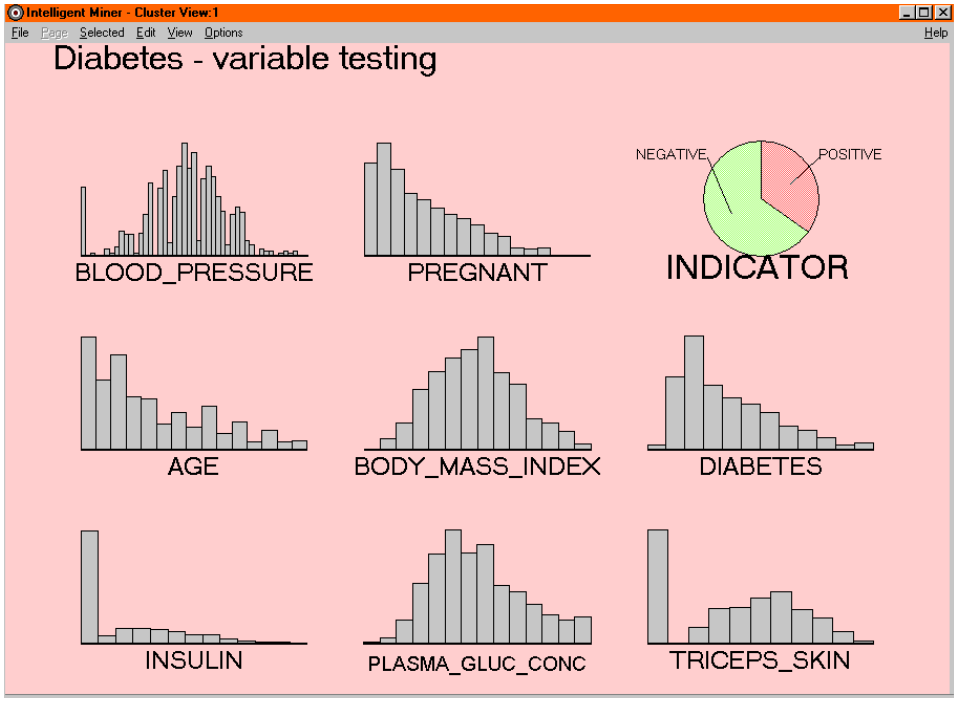


Figure 6-1 Statistical view for the distribution of the variables

Most of the patients were pregnant one or two times (*PREGNANT*), but ranges up to seventeen pregnancies for two patients.

Concerning *DIABETES*, a pedigree function uses a family health history diagrammed as input to indicate the individuals in the family, their relationship to one another or those with disease, and so on. It is not known how this function looks, only that it produces a numeric value.

Concerning *AGE,* about 60% of the patients are younger than forty and approximately 4% older than 60.

INDICATOR is a numerical variable that indicates the result of the examination (negative or positive).

TRICEPS_SKIN is a numeric variable that reflects the thickness of the skin. This is a valid test, because *diabetes mellitus* can be influenced by *Arthritis*.

Figure 6-2 shows the distribution of the variables in comparison to the whole population when only patients were considered who have a negative test result.

The patients are younger (*AGE*) than in general, the thickness of the skin is thinner (*BODY_MASS_INDEX*), and the number of pregnancies is lower than usual.

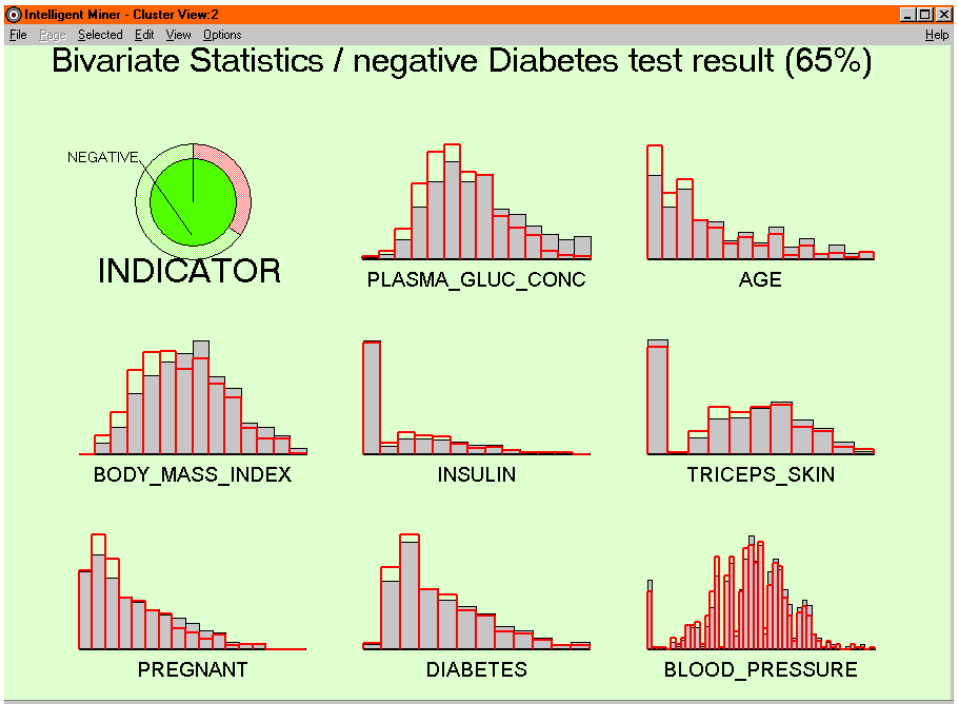


Figure 6-2 Distribution for patients if only negative test results are threatened

Figure 6-3 shows the distribution of the variables in comparison to the whole population when only patients are considered who have positive test results.

The Plasma glucose concentration (*PLASMA_GLUC_CONC*) is higher than in the whole population. In general, these patients are older (*AGE*), with a higher blood pressure (*BLOOD_PRESSURE*) and a higher number of pregnancies (*PREGNANT*). Generally, there is some kind of “right shift” to the data distribution; whereas, most of the variables in Figure 6-2 denote a more concentrated distribution to the left side. Variables of infected patients tend to have higher values in general.

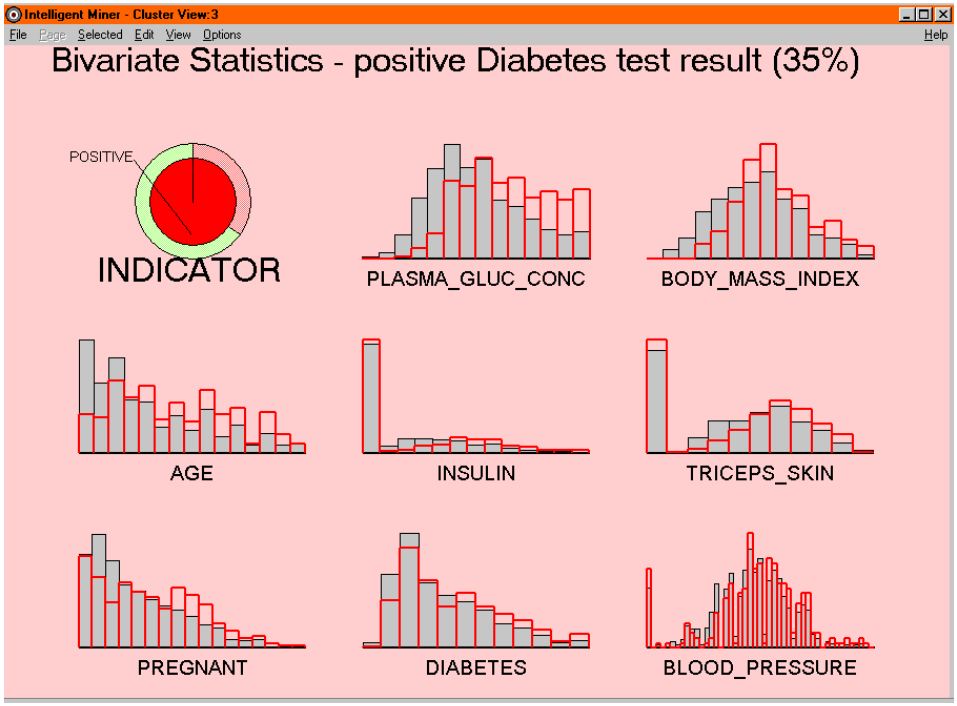


Figure 6-3 Distribution of the variables if only positive tests are threatened

If we compare the two results we come to the following semantic interpretation as described in Table 6-2.

Table 6-2 Comparison between positive and negative tests

Variable	Positive result	Negative result
AGE	Older	Younger
PLASMA_GLUC_CONC	Higher concentration	Lower concentration
BODY_MASS_INDEX	Higher values	Lower values
TRICEPS_SKIN	Higher values	Lower values
PREGNANT	More pregnancies	Lower pregnancies
INSULIN	Higher values	Lower values

In general, patients who are infected with diabetes have higher values in general and are older. However, this is a general trend but is not a secure pattern.

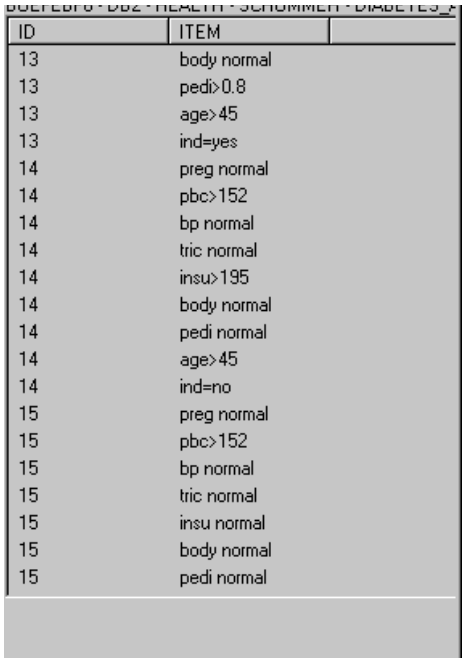
6.3.2 Datamart aggregation for Association Discovery

If we look for medical variables we only have a view from the patient side. These records allow an analysis of patients independently but not an analysis of the test values in a transactional form.

Therefore, it may be interesting to have a transactional view as it occurs in the diagnoses data. Instead of applying techniques to unique patient records, we also could apply Association Discovery to confirm our results or to find potentially new information. The solution to this idea can be realized by aggregating the medical records to a new datamart. Practically speaking, this can be done by using the pivot technique of the IM for Data.

The following steps give one possible way to aggregate the new datamart.

- The new datamart consists of only two columns. The first column contains the patient ID (the Transaction Group). The second column contains the test result for a specific variable (the Transaction Item).
- Because the new datamart will have a transactional view we have to prepare the second column by adding the variable name to the value.
- To avoid too many values for a variable, for example, we probably have more than 100 different values for only one variable, such as for *AGE*. We therefore need to find a way to reduce the number of values.
- The method we use is discretization which means a transfer of numerical to categorical values. This was done as following:
  - We calculate the average value of each variable as well as the standard deviation value.
  - By applying these calculated numbers we get three intervals, namely:
    - All values that are lower than the average minus standard deviation
    - All values that are within the standard deviation
    - All values that are higher than the average plus standard deviation
  - We then have to add new values as shown in Figure 6-4 (where ID represents the patient_id): a value is then “normal” if it occurs within the standard deviation range.



ID	ITEM
13	body normal
13	pedi>0.8
13	age>45
13	ind=yes
14	preg normal
14	pbc>152
14	bp normal
14	tric normal
14	insu>195
14	body normal
14	pedi normal
14	age>45
14	ind=no
15	preg normal
15	pbc>152
15	bp normal
15	tric normal
15	insu normal
15	body normal
15	pedi normal

Figure 6-4 Aggregated datamart calculated by pivot of the medical records

Practically speaking, a small program (written in the programming language *awk*) and a preprocessing function of the IM for Data (Figure 6-5) was used to perform this datamart creation.

- The program operates on flat files. It concatenates the variable name to this value. The concatenated variable then becomes automatically categorical.
- The preprocessing function of the IM for Data can then be used to build the datamart (Figure 6-5).

Figure 6-5 shows the preprocessing function of the IM for Data. All the variables (those with suffix "_D") that are in the selected list box ("add to pivot") are categorical.

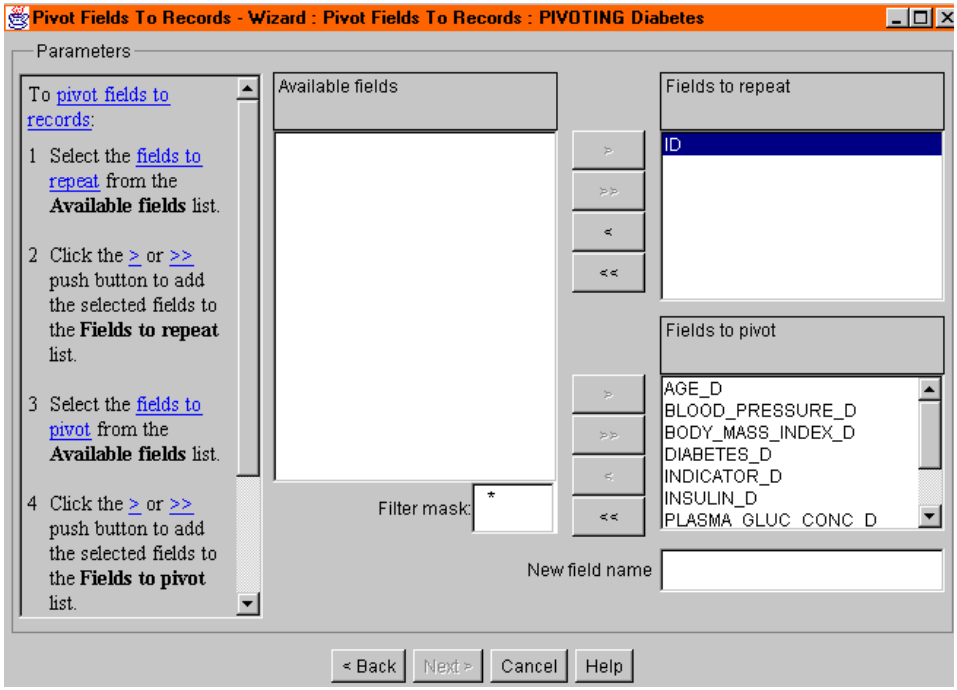


Figure 6-5 Pivoting the data by applying the preprocessing function of the IM for Data

6.4 Choosing the mining technique

Choosing the mining techniques to use is *the fifth stage in our generic mining method*.

The characterization of the underlying medical datamart is done by several algorithms of the IM for Data. As it was already shown in chapter 6.3, “Sourcing and evaluating data” on page 115 the bivariate statistics represent a suited way to describe patients with and without diabetes mellitus. However, they do not deliver predictive rules but only a semantic description of the visualized data distribution.

We therefore turn our attention to predictive modeling methods, for example, *decision trees* and *Radial Basis Functions,* which allow the generation of predictive rules that can be applied to new patients by using the corresponding application mode. Additionally, the models deliver some kind of information about the characteristics of the predicted classes.

It is difficult to calculate good prediction models, because variables may be correlated or medical knowledge is not present. Before you start building predictive modeling, we recommend that you test the underlying variables for their relevance. This can be done using decision trees as a preprocessing technique, because decision trees try to use variables that split the datamart each time. It is important to know which variables occur on top of the tree: those that are not in the tree or are far away from the root of the tree may be negligible. For the generation of a new prophylactic test this may be very interesting to know, because medical tests can then be ordered according to their relevance; and, some time could be saved by getting relevant information at the right time.

Another aspect is quality: depending on the importance of each variable a weighting of a variable may be suited. As described above, the concentration of blood plasma is surely a main contributor in deciding whether a patient has diabetes mellitus or not. From the analytical standpoint it may be interesting to see how the quality of the model behaves if we reduce or increase the weight of this variable.

In general a deeper understanding of the variables' relevance is not really necessary, because only eight variables are present. However, if a larger number of records with a higher variable concentration is available, this may be a good strategy to get an impression of the variables' importance.

An alternative way of getting expressible rules is the computation of associative rules by Association Discovery. Using the aggregated datamart that was based on the medical records, we then get new information that can describe the characteristics of negative and positive test results.

6.5 Interpreting the results

The *sixth stage in our generic mining method* is to interpret the results that we have obtained and determine how we can map them to our business.

This section contains evaluation results by applying two techniques of the IM for Data:

- Classification tree
- Radial Basis Functions

Both techniques are predictive techniques that can build up predictive models based on historical datamarts.

6.5.1 Predictive modeling by decision trees

A first application of the *decision tree* with *INDICATOR* as a class label and all others as active variables delivers a tree model of 83% quality (Figure 6-6).

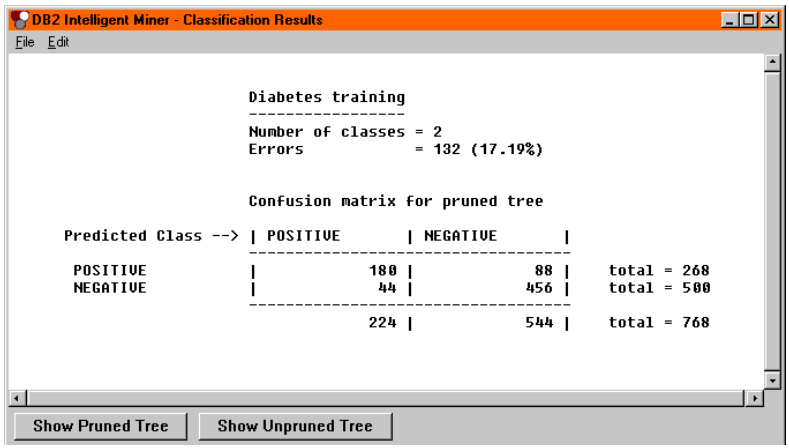


Figure 6-6 First decision tree model where all variables are active

The quality is calculated by dividing the correct classified records by the number of all records. Overall, we have 768 patient records where:

- 180 patients are diseased (POSITIVE) and correctly classified (Predicted class is POSITIVE)
- 44 patients are not diseased (NEGATIVE) and incorrectly classified (Predicted class is NEGATIVE)
- 88 patients are diseased (POSITIVE) and incorrectly classified (Predicted class is POSITIVE)
- 456 patients are not diseased (NEGATIVE) and correctly classified (Predicted class is NEGATIVE)

Therefore, the number of correctly classified records is 636, the number of incorrectly classified records 132. We then have an error rate of 17.19% or a quality of approximately 83%.

The corresponding decision tree is displayed in Figure 6-7 and Figure 6-8 with different classification rules for both the negative and the positive case.

PLASMA_GLUC_CONC represents the most important variable followed by *AGE* and *BODY_MASS_INDEX*, respectively. For example, we get the following decision rules by following the path from the root to the specific node:

- If the plasma glucose concentration is lower than 127.5 and the age of the patient is lower than 28.5, then he or she probably will not be infected by diabetes mellitus. The rule has a quality of 91.5% (Purity, see the bottom right of Figure 6-7).
- If the plasma glucose concentration is greater than 157.5 and the body_mass_index exceeds 29.95, then he or she probably will be infected by diabetes mellitus. The rule has a quality of 87.0% (Purity, see the bottom right of Figure 6-8).

Of course, these are just two rules derived from the tree, but it shows that almost 50% of all patients could be classified by these two rules (363 out of 768).

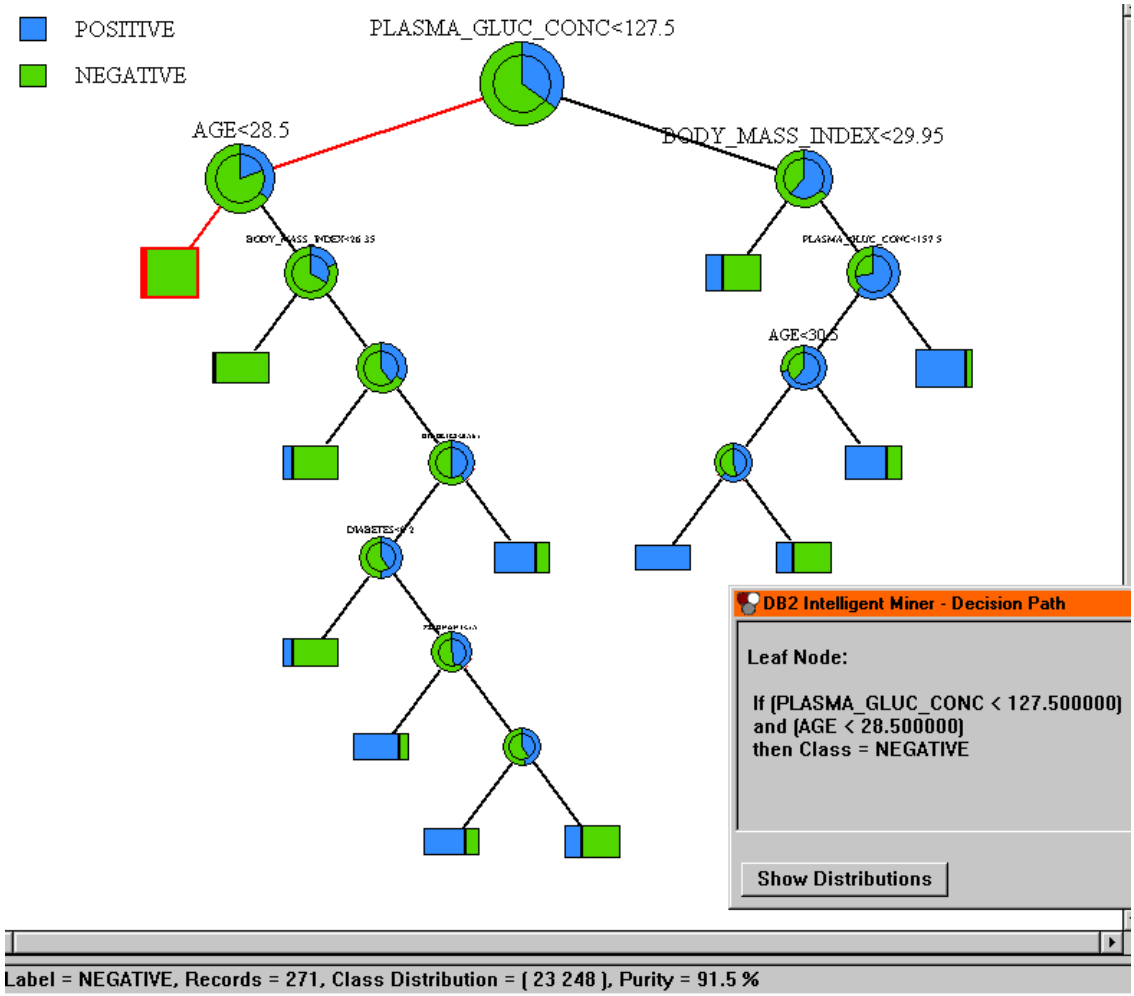


Figure 6-7 Decision tree with rule for patients with negative test result

Additionally, the depth of the tree is 14 but is pruned to depth 8.

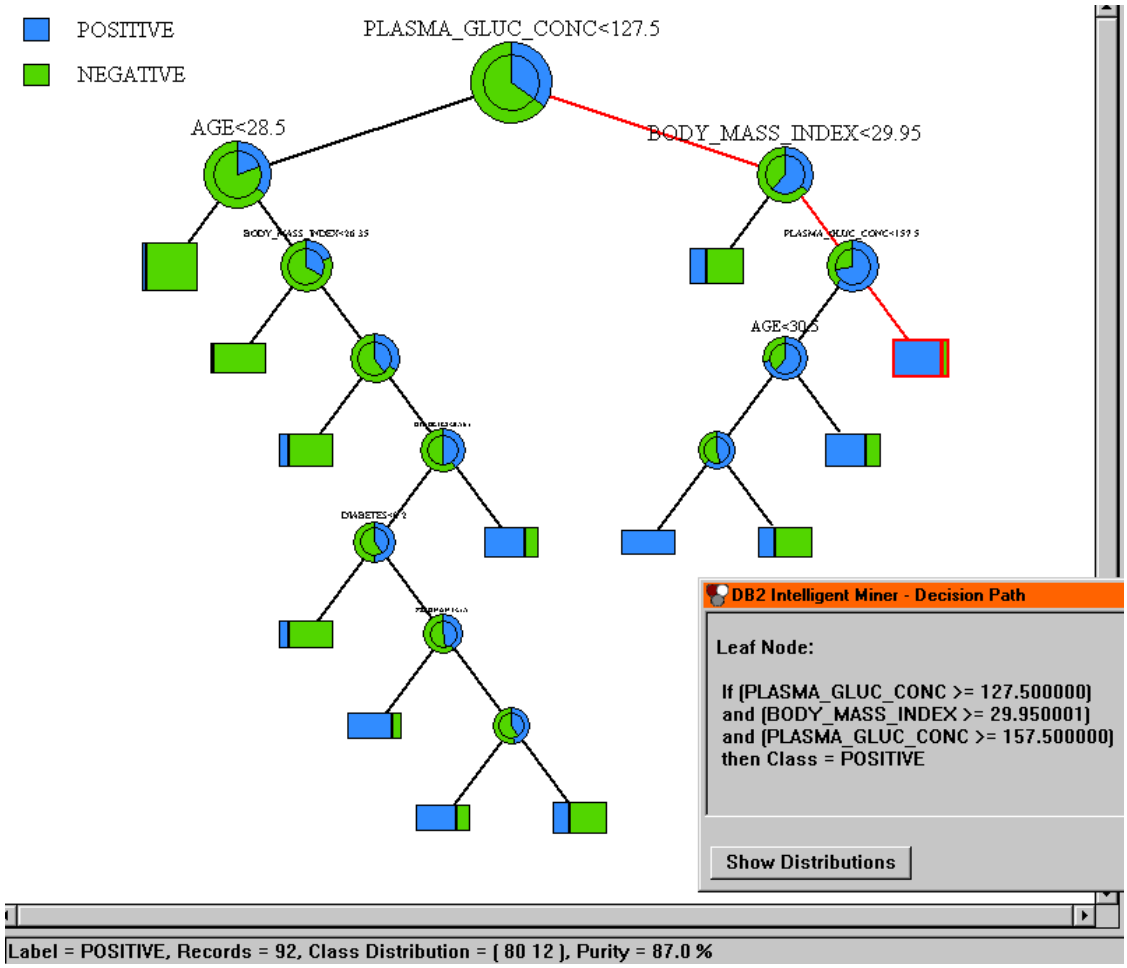


Figure 6-8 Decision tree with rule for patients with positive test result

However, *PLASMA_GLUC_CONC* is a variable that is commonly known as a good test for diabetes mellitus (see above). Therefore, it would make sense to restart the tree classification process without this variable. Hopefully, we than can get more qualitative results.

Figure 6-9 shows the result of a classification run that was done without *PLASMA_GLUC_CONC.* The quality gets worse, because only 79% of the patients can be classified correctly.

There is a similar quality behavior when we try to build classification trees using additional methods like:

- Weighting *PLASMA_GLUC_CONC* with a lower value
- Removing of variables, for example, AGE or BODY_MASS_INDEX
- Changing the maximum allowed depth of the tree

Whereas, the depth of the tree is growing up, the quality decreases up to 72% for the second point. The classification of the negative cases remain nearly stable; whereas, the incorrect classification of the positive cases seem to be responsible for the loss of quality.

If we run a classification by using only *PLASMA_GLUC_CONC* and *AGE*, respectively, then we get a model with only 77% quality. If we run a classification by using only *PLASMA_GLUC_CONC* then the quality decreases to 74%.

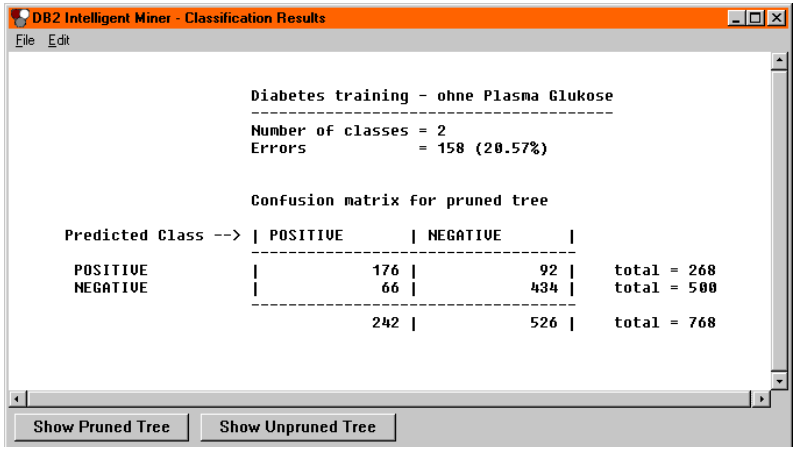


Figure 6-9 Classification Result after removing PLASMA_CONC as active variable

6.5.2 Predictive modeling by Radial Basis Functions

By applying the *Radial Basis Functions* of the IM for Data we get a result as shown in Figure 6-10.

The first region at the top of the visualization shows patients who very probably will have diabetes mellitus; whereas, the last region at the bottom of the visualization shows patients who do not have diabetes mellitus. As an indication whether patients have or do not have diabetes mellitus, you can see the number at the left:

- Region 8 has a high value of 0.66
- Region 15 has a low value of 0.0274

Figure 6-11 shows region 8 in more detail. The region has variables that have the following semantic characteristics:

- Higher concentration of *INSULIN*
- Higher concentration of *PLASMA_GLUC_CONC*
- Higher concentration of *BLOOD_PRESSURE*
- Higher number of pregnancies
- Thicker skins

This is very similar to the information we got by bivariate statistics. And, in respect to predictive modeling by classification trees, it represents a model that is very similar, too.

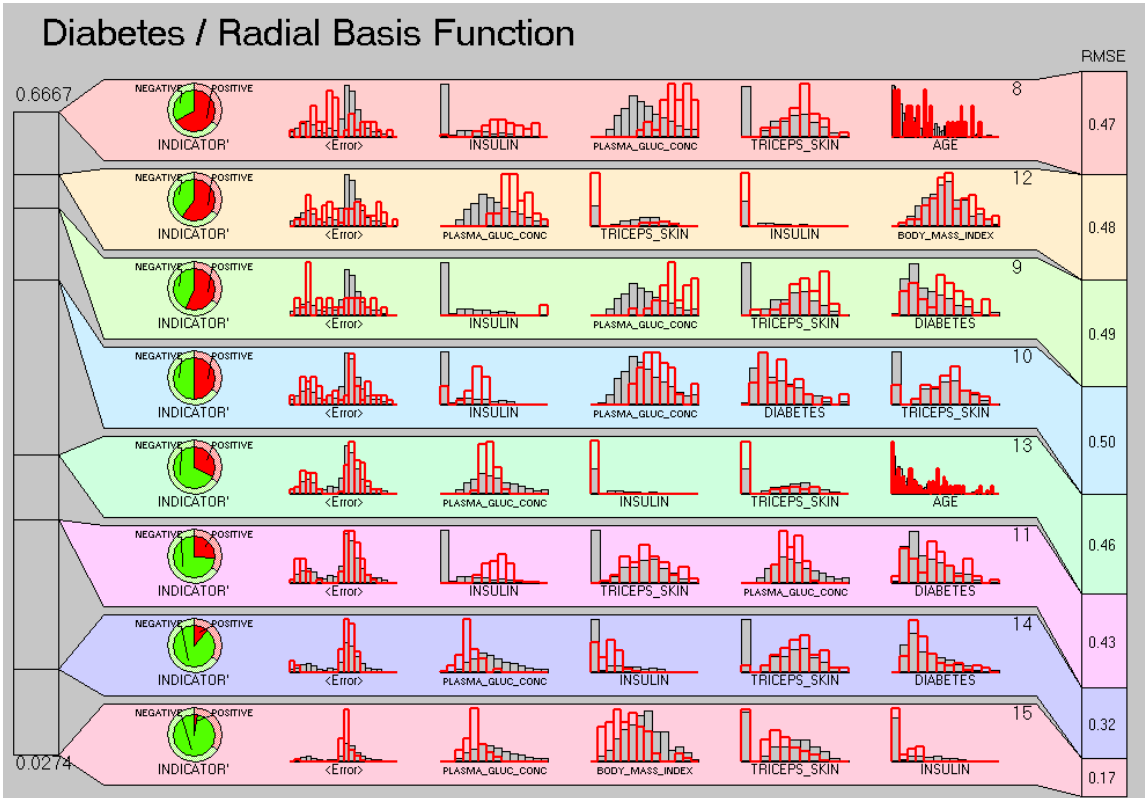


Figure 6-10 Result of the Radial Basis Functions that gives a ranking of patients with diabetes mellitus (region at the bottom) and without (region at the top)

Overall, we can observe that many analysis techniques can be applied; if there is information in the data then this will be shown by all techniques in the same or similar way.

6.5.3 Verification of the predictive models

Before we started with predictive modeling we generated two datamarts for training and test, respectively. Both datamarts contain approximately 65% of patients with diabetes mellitus and 35% without diabetes mellitus. It is discussible whether this relationship is good enough or whether we should take a 50:50 distribution. For the underlying datamart, this discussion was redundant, because the number of the patients with diabetes mellitus is really small (Figure 6-11).

A generation of the datamarts to a 50:50 would decrease the number of records too much. Concerning the data in respect to their size, we chose an approximately 50:50 relation.

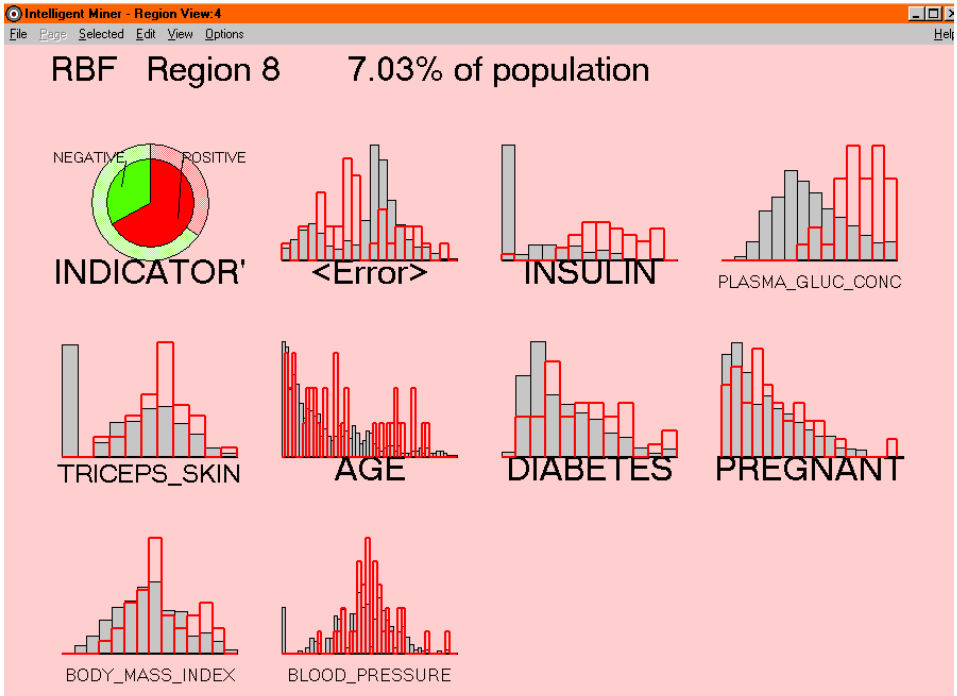


Figure 6-11 Region 8 that shows the patients who probably will have diabetes

The verification of the classification model, that was computed as the best model, produces a quality of almost 75% (Figure 6-12). The value is lower than the training quality value, but secures an appropriate predictive behavior of the classification model. For the Radial Basis Function, we observed a similar behavior: all regions were between 0.65 (patient has positive result) and 0.16 (patient has negative result).

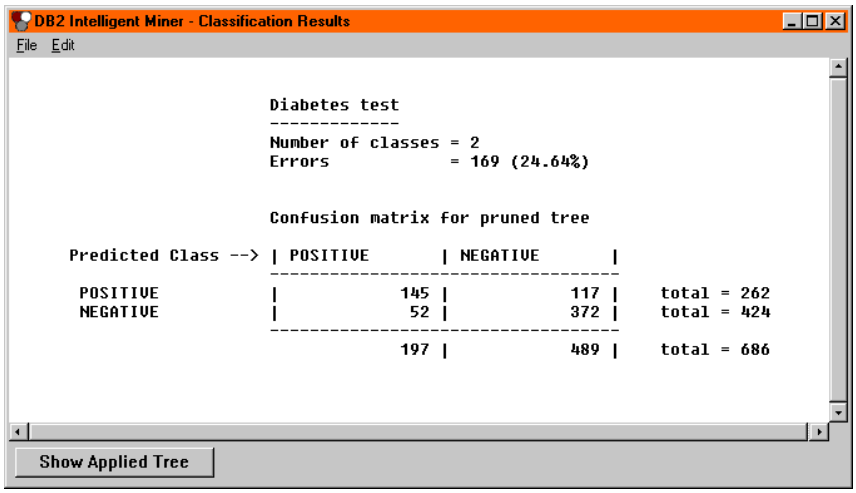
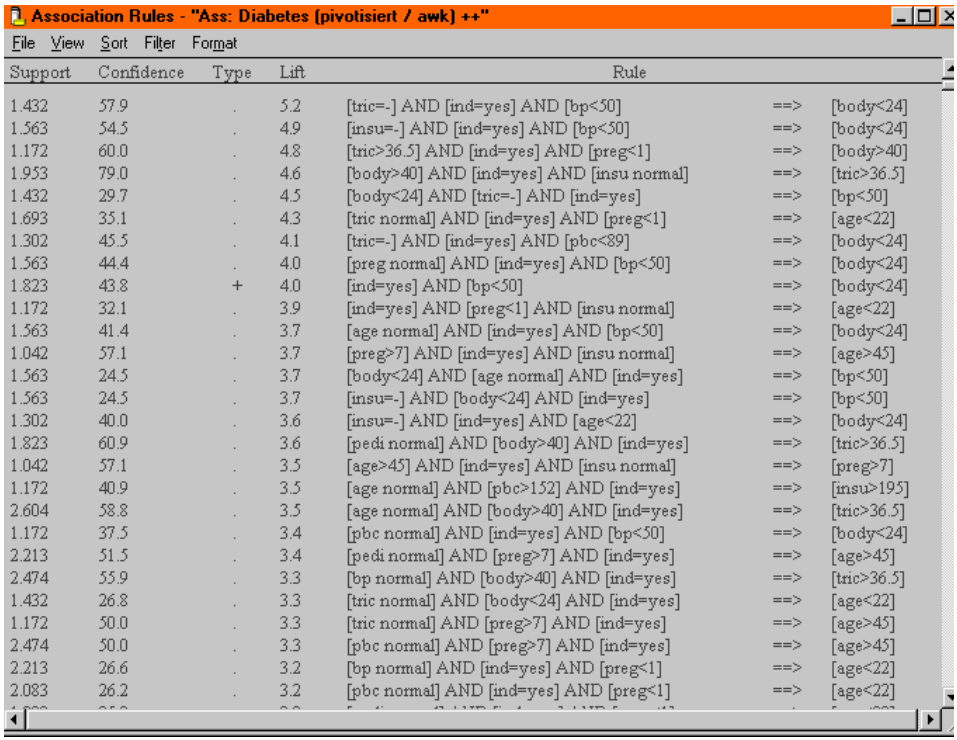


Figure 6-12 Result after testing the classification model

6.5.4 Association Discovery on transactional datamart

The aggregation of the data delivers a new datamart that can be used as a transactional basis to discover associative relationships that may be hidden in the data.

The question is whether we can, in addition to the results we got from the predictive techniques of the IM for Data, discover potentially new information between the variables.



Support	Confidence	Type	Lift	Rule		
1.432	57.9	.	5.2	[tric=-] AND [ind=yes] AND [bp<50]	==>	[body<24]
1.563	54.5	.	4.9	[insu=-] AND [ind=yes] AND [bp<50]	==>	[body<24]
1.172	60.0	.	4.8	[tric>36.5] AND [ind=yes] AND [preg<1]	==>	[body>40]
1.953	79.0	.	4.6	[body>40] AND [ind=yes] AND [insu normal]	==>	[tric>36.5]
1.432	29.7	.	4.5	[body<24] AND [tric=-] AND [ind=yes]	==>	[bp<50]
1.693	35.1	.	4.3	[tric normal] AND [ind=yes] AND [preg<1]	==>	[age<22]
1.302	45.5	.	4.1	[tric=-] AND [ind=yes] AND [pbc<89]	==>	[body<24]
1.563	44.4	.	4.0	[preg normal] AND [ind=yes] AND [bp<50]	==>	[body<24]
1.823	43.8	+	4.0	[ind=yes] AND [bp<50]	==>	[body<24]
1.172	32.1	.	3.9	[ind=yes] AND [preg<1] AND [insu normal]	==>	[age<22]
1.563	41.4	.	3.7	[age normal] AND [ind=yes] AND [bp<50]	==>	[body<24]
1.042	57.1	.	3.7	[preg>7] AND [ind=yes] AND [insu normal]	==>	[age>45]
1.563	24.5	.	3.7	[body<24] AND [age normal] AND [ind=yes]	==>	[bp<50]
1.563	24.5	.	3.7	[insu=-] AND [body<24] AND [ind=yes]	==>	[bp<50]
1.302	40.0	.	3.6	[insu=-] AND [ind=yes] AND [age<22]	==>	[body<24]
1.823	60.9	.	3.6	[pedi normal] AND [body>40] AND [ind=yes]	==>	[tric>36.5]
1.042	57.1	.	3.5	[age>45] AND [ind=yes] AND [insu normal]	==>	[preg>7]
1.172	40.9	.	3.5	[age normal] AND [pbc>152] AND [ind=yes]	==>	[insu>195]
2.604	58.8	.	3.5	[age normal] AND [body>40] AND [ind=yes]	==>	[tric>36.5]
1.172	37.5	.	3.4	[pbc normal] AND [ind=yes] AND [bp<50]	==>	[body<24]
2.213	51.5	.	3.4	[pedi normal] AND [preg>7] AND [ind=yes]	==>	[age>45]
2.474	55.9	.	3.3	[bp normal] AND [body>40] AND [ind=yes]	==>	[tric>36.5]
1.432	26.8	.	3.3	[tric normal] AND [body<24] AND [ind=yes]	==>	[age<22]
1.172	50.0	.	3.3	[tric normal] AND [preg>7] AND [ind=yes]	==>	[age>45]
2.474	50.0	.	3.3	[pbc normal] AND [preg>7] AND [ind=yes]	==>	[age>45]
2.213	26.6	.	3.2	[bp normal] AND [ind=yes] AND [preg<1]	==>	[age<22]
2.083	26.2	.	3.2	[pbc normal] AND [ind=yes] AND [preg<1]	==>	[age<22]

Figure 6-13 Association Discovery for the aggregated datamart

Figure 6-13 shows the result of an Association Discovery sorted by ***lift***. Missing values are marked with an "-". A positive diabetes result is indicated by *ind=yes*, a negative result by *ind=no*. Most of the rules have a neutral relationship beneath each other, an indication that they are not influencing each other very much. Only the rule has a positive type (+) and a Lift that is four times higher than expected:

- If [ind=yes] and [bp<50] ==> [body<24]

If we look at the ***confidence*** we observe a strong quality of almost 43.8% which indicates that a blood pressure lower than 50 together with a positive diabetes result implies *BODY_MASS_INDEX* lower than 24.

Figure 6-14 shows the relationship between *PLASMA_GLUC_CONC* and *INCICATOR.* About 12% of all patients have *PLASMA_GLUC_CONC* that is lower than 89 and a positive test result with a probability of more than 92%.

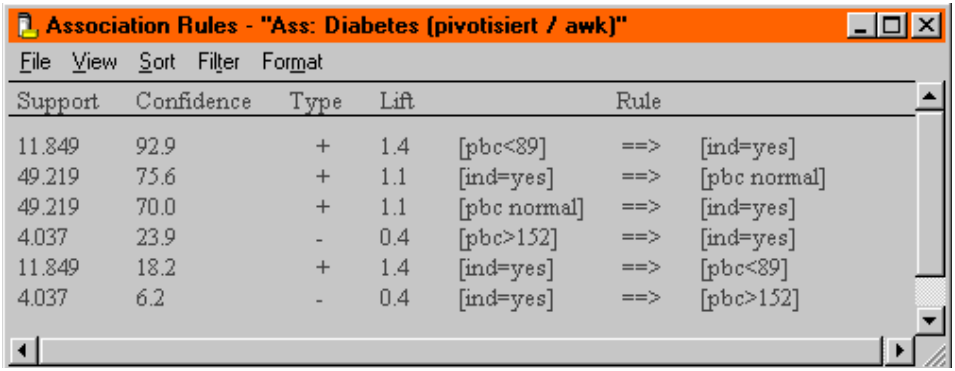


Figure 6-14 Using a filter to identify the relationship between PLASMA_GLUC_CONC (pgc) and INDICATOR (ind)

However, the *lift* is not quite high, therefore the relevance of this rule is not really surprising. But it is interesting that 70% of all patients with a normal value of *PLASMA_GLUC_CONC* also have a positive test result.

6.6 Deploying the mining results

The final and *seventh stage in our generic mining method* is perhaps the most important of all. How do you deploy the mining results into your business to derive the business benefits that data mining offers? The reason that this is so important is that all too often data mining is seen as an analytical tool that can be used to gain business insight but is difficult to integrate into existing systems. In this section we explain how the clustering techniques that we have described can be deployed into your business.

In this section, after a short summary about the methods used, we suggest a strategy on how to optimize medical tests.

6.6.1 What we did so far

In this section we were concerned with the question of whether we can contribute to a more efficient test for diabetes mellitus. Therefore, we performed the following steps:

- Understanding the task (see 6.1.1, “Diabetes insipidus and diabetes mellitus” on page 112)
- Understanding the data (see 6.2.2, “Data structure” on page 114)
- Using bivariate statistics to know what are the characteristics for negative and positive test results (see 6.3.1, “Statistical overview” on page 115)

- Performing predictive modeling by applying classification tree and radial basis functions (see 6.5, “Interpreting the results” on page 122): this modeling was then verified.
- Aggregating and pivoting the data into a transactional view (see 6.3.2, “Datamart aggregation for Association Discovery” on page 119)
- Performing Association Discovery (see 6.5.4, “Association Discovery on transactional datamart” on page 129)

Predictive modeling was done to find characteristic rules that possibly indicate the different groups of patients in respect to diabetes mellitus; additionally, we could use Classification Trees to identify variables within the tree, which are more important than others. As a result, we could show that *PLASMA_GLUC_CONC* is the most important variable within the datamart. If we remove *PLASMA_GLUC_CONC* from the input list, we get worse quality.

Overall, it was quite interesting that, although we used several techniques to find appropriate information, we had the same result, in a different way. By applying the Radial Basis Function, *PLASMA_GLUC_CONC* was also delivered as the main variable responsible for diabetes.

6.6.2 Optimization of medical tests

We can use the information from 6.5, “Interpreting the results” on page 122, to reduce the amount of effort by introducing *PLASMA_GLUC_CONC* as the first test that should be made. If the test results in a value that is greater than 127.5 (Figure 6-8 on page 125) and smaller than 157.5, then only one test must be done additionally to get either a 87% probability that the patient has diabetes or a 68.4% probability that he or she does not.

What is the body/mass index? This question affords the knowledge about the weight and the size of the patient. That is quite easy to ensure and requires only minutes to perform. If the value is greater than 29.95, then the patient has diabetes; otherwise the patient does not.

On the other hand (if *PLASMA_GLUC_CONC* is really lower than 127.5), then the patient has a 91.5% probability of health, if younger than 28.5 years! If not than he or she has at least a 66.8% probability to not be affected by diabetes.

There are a lot of other examples where the application of a method like the proposed model could become deployable. This surely includes tests for:

- Thrombosis (see Section 5.1.4, “Deep vein thrombosis and ICD10” on page 77)
- Tuberculosis (see Section 5.7.2, “Where can the method be deployed?” on page 107)

Overall, the perspective for medical services to be improved by optimizing medical tests makes this model very useful, and very attractive for a hospital; if an optimization of medical tests is possible, then this leads to:

- A reduction of different cost factors in the areas of capacity planning.
- Faster tests; the chance is that diseases are more quickly identified.

6.6.3 Boomerang: improve the collection of data

The application of data mining requires the presence of qualitative data (see 6.2.3, “Some comments about the quality of the data” on page 115). If the data is qualitative, this is suitable for data mining: the discovery of potentially hidden but useful information delivery can then satisfied.

The main task therefore is: How can qualitative data be collected?

Of course, we can ask the physicians to be more careful in collecting the data, in writing clearly and speaking louder. However, we cannot change the physicians' medical responsibility: his or her main task is and remains the application of medical knowledge to ensure the patients health. The physician should not be the one who works with computer all the time.

A possibility to connect medical services and computer-based efforts is the usage of mobile hardware, for example, Palm pilot or PDA. The physician then uses the mobile hardware to save appropriate information, anywhere and anytime. The data is then transferred to a hospital-wide system.

There are different advantages, for example:

- Data will be consistent and clear, because predefined questions with a possible selection of answers could be used.
- Data will be available very soon, because the data transmission can be done every day (automatically).
- Data will be complete, because a patient's record can only be stored if all questions are answered.

There are many other advantages that will come together and the main result is:

- If a test for a disease can be optimized, then it would be possible because the preconditions are satisfied.

And if the data is stored in a way that all physicians in a hospital can access it, this could become a first step in discovering patients' behavior in a new way.

7

Can we detect precauses for a special medical condition?

Diseases are sometimes difficult to identify and often they remain undiscovered. The reasons for this are, for example, ambiguity of the diseases' symptoms, missing medical possibilities, or insufficient experience of the medical staff. New strategies and techniques that help to find precauses for diseases are appreciated.

In this chapter, we present some strategies that describe how we can find precauses for a special disease. We use data that was recorded for patients who were tested for thrombosis. Here, we concentrate on time series data and show what kind of analyses can be done in detail.

By the evaluation of the data we describe how data mining can be used and applied in this medical scenario.

7.1 The medical domain and the business issue

This section contains a small description of what deep vein thrombosis is. Furthermore, we describe what it causes and when it occurs and we define the business issue that we should start with.

7.1.1 Deep Vein Thrombosis

A vein is a blood vessel that returns blood from the tissues of the body back to the heart. Here, the body has two distinct systems of veins: a *superficial system* and a *deep system*: the superficial system is made up of veins that are close to the skin. These are the blood vessels that can be seen on hands or arms. The deep system is comprised of veins within the muscles of the body. They are connected by small communicating veins where the body regulates the amount of blood going through both systems as a way of rigidly controlling the body's central temperature. A Deep Vein Thrombosis (DVT) is therefore a condition where a blood clot forms in a vein of the deep system.

DVT can occur anywhere in the body but is most frequently found in the deep veins of the legs or thighs. Several factors may cause DVT including injury to the vein, slowing of blood flow, and conditions that increase the tendency for the blood to clot. The most common cause of injury to a vein is trauma to the leg, such as occurs with broken bones, severe muscle injury, or surgery. Immobilization is the most common cause of slow blood flow in a vein, because movement of the leg muscles helps keep blood flowing through the deep veins. People who have conditions like one or some of the following are at a higher risk for developing DVT:

- Paralysis from a stroke or spinal cord injury
- A cast on a leg to help heal a fractured bone
- Confinement in bed due to a medical or surgical condition
- Prolonged sitting, especially with crossed legs in cars, trains or airplanes.

7.1.2 What does deep vein thrombosis cause?

The most common symptoms of a DVT in the leg are swelling and pain in the affected leg. These symptoms are caused by the accumulation of blood that is unable to get past the clot in the vein and the resulting leakage of fluid from the blood into the muscle. Many other conditions exhibit symptoms similar to those of a DVT, for example, muscle strains, skin infections, and inflammation of superficial veins. A DVT, therefore, is difficult to diagnose without specific tests in which the deep vein system can be examined. Furthermore, many patients with a Deep Vein Thrombosis have no symptoms at all unless the clot dislodges, travels to the lung, and causes a pulmonary *embolism*. In this case, the patient may

develop a rapid heart rate, shortness of breath, sharp chest pain that worsens with deep breathing, or cough up blood. If the pulmonary emboli are large and block one or both of the major pulmonary arteries sending blood to the lungs, the patient may develop a very low blood pressure, pass out, and possibly die from lung or heart failure. As is the case with DVT, however, many other conditions, for example, a heart attack or pneumonia, can mimic a pulmonary embolism. Therefore, specific tests must be done to confirm the diagnosis.

7.1.3 Can deep vein thrombosis be prevented?

There is good evidence that in certain high risk situations, DVT and the resulting pulmonary emboli can be prevented if certain prophylactic measures are started early. Some issues are:

- Several studies have shown that *graduated compression stockings* can decrease the incidence of DVT in patients who are confined to bed because of medical conditions or surgical procedures. The stockings work by reducing the amount of blood and increasing the flow of blood in the veins of the legs.
- A second effective therapy is *pneumatic compression* of the legs, which is done by applying a plastic stocking to the leg and thigh. The stocking intermittently fills with air and squeezes the leg. This compression stimulates the body to produce factors that help dissolve small blood clots before they can progress to DVT. Here are some preventative medications, such as:
  - Unfractionated heparin: Heparin is an anticoagulant (anti-clotting) medication and useful in preventing thromboembolic complications (clots that travel from their site of origin through the blood stream to clog up another vessel). Heparin is also used in the early treatment of blood clots in the lungs (pulmonary embolisms).
  - Low-molecular-weight heparin.
  - Coumadin: Coumadin is a known teratogen, an agent that can disturb the development of the embryo and fetus and lead to birth defects. Coumadin, when taken by a woman during pregnancy, can cause bleeding into the baby's brain, underdevelopment of the baby's nose and stippling of the ends of the baby's long bones. Each can markedly reduce the incidence of post operative DVT in patients undergoing orthopedic surgery.

7.1.4 Where should we start?

The first stage in the generic method is to translate the business issue you are trying to address, into a question, or set of questions that can be addressed by data mining.

The business problem therefore is: How can we find precauses for a special disease, for example, deep vein thrombosis?

7.2 The data to be used

You clearly cannot do data mining without having the data about your patients to mine. But what data do you need? *The second stage in our data mining method* is to identify the data required to address the business issue and where we are going to get it from.

The underlying datamarts contain data about patients who were recognized as diseased or not by thrombosis. This decision was made on the basis of a thrombosis medical test.

There are three datamarts available:

- Demographic data about the patients
- Examination data that were recorded when the thrombosis test had taken place
- Historical data that contain general medical tests over time

Demographic data is not really interesting for the discovery of time-relevant patterns over time. Demographic data represents some kind of “state of the art” in respect to the patient, and is not really time-specific.

Furthermore, information from the thrombosis test itself is in the same way not interesting. They show a snapshot of the patients’ life that does not indicate behavior over time.

However, there are two variables that need to be used:

- The indication of whether the patient has thrombosis or not (*THROMBOSIS*)
- The date when the thrombosis test had taken place (*EXAMINATION_DATE*)

Both variables are important, because they allow both a comparison between patients with and without thrombosis (*THROMBOSIS*) and a comparison between the patients’ time before and after the thrombosis test.

7.3 Sourcing the data

To create our data model we have to take the raw data that we collect and convert it into the format required by the data models. We call this stage in the process sourcing and preprocessing and this is *the third stage in our data mining method*.

For some of the patients there are several medical tests done over time. For example, the highest number for a patient to be examined is more than 100. In general, this data describes medical tests that were done during several years; they therefore can express very strongly the behavior over time. Table 7-1 shows the variables that were stored in the historical datamart.

Table 7-1 Source the data that come from the historical medical tests

Variable	Variable type	Description
ID	CATEGORICAL	Identification of the patient
DATE	CATEGORICAL	Date when the test was done
GOT	CONTINUOUS	Glutamin oxaloacetic transaminase
GPT	CONTINUOUS	Glutamin pylvic transaminase
LDH	CONTINUOUS	Lactate Dehydrogenase
ALP	CONTINUOUS	Alkaliphosphotase
TP	CONTINUOUS	Total number of proteins
ALB	CONTINUOUS	Albumin
UA	CONTINUOUS	Urin Acid
UN	CONTINUOUS	Urin Nitrogen
CRE	CONTINUOUS	Creatinine
T-BIL	CONTINUOUS	Bilirubin
T-CHO	CONTINUOUS	Cholesteron
TG	CONTINUOUS	Triglyceride
CPK	CONTINUOUS	Creatinine Phosphokinase
C4	CONTINUOUS	Complement 4
WBC	CONTINUOUS	Number of white blood cells in a volume of blood
RBC	CONTINUOUS	Number of red blood cells

Variable	Variable type	Description
HGB	CONTINUOUS	Hemoglobin
HCT	CONTINUOUS	Hematocrit
PLT	CONTINUOUS	Platelet
IGG	CONTINUOUS	Immunoglobin G

7.4 Evaluating the data

Having created and populated our data models, *the fourth stage in our data mining method* is to perform an initial evaluation of the data itself.

The historical thrombosis datamart is represented by a number of records. Each record represents a historical medical test that was done for the patient.

In detail, a record within the datamart contains the patient ID as the identification of the record, followed by the variables that describe the test (Table 7-2).

Table 7-2 Historical medical tests that are represented by records in the datamart

ID	Date	GOT	GPT	LDH	ALP	...
1	860419	24	12	152	63	...
1	860430	25	12	162	76	...
2	960315	22	8	144	68	...
...	...	...	...	...	...	...

For the discovery of patterns over time, however, this is not an appropriate structure. We do not need a representation that is ordered by patient only, but a representation that is ordered by patient and time.

7.4.1 The nondeterministic issue

There are some techniques in the IM for Data portfolio that can be applied for time-analysis (see 7.5, “Choosing the mining technique” on page 148). However, to find which technique to use is not typically determinable. Generally, it depends on several aspects, for example:

- The task we have to fulfill
- The inspiration and/or creativity of the persons who analyzes
- The time that is available

Sometimes, we may think that is necessary to use a specific technique; whereas some time later, we see that other techniques may be more suitable. Overall, many ideas in data mining occur “on the fly”. This means that each time, independent from the current state of the analysis, new data mining tests are possible.

7.4.2 Need for different aggregations

In this chapter, we not only concentrate on a specific technique, but we suggest three techniques that are suited for the discovery of precauses for thrombosis (for more information about the techniques, see 7.5, “Choosing the mining technique” on page 148).

Please note, that these techniques are derived spontaneously, they are not planned; although it could look that way. In reality, only the application of one technique was planned.

To get a basis for all of these techniques we need to preprocess the data in a different way. In fact, there will be two strategies:

Associative aggregation

Associative aggregation means:

- Simplify the entries in the historical datamart (*modification I*) by discretizing all the values. We then get more expressive (categorical) values.
- Expand the entries in the historical datamart (*modification II*) by concatenation with:
  - The time point when the historical medical test had taken place
  - The result of the thrombosis medical test
- Pivot the resulting datamart by ID and Date.

Time Series aggregation

Time Series aggregation means:

- Pivot the historical datamart by ID and Date without any modification.

In general, **Associative aggregation** affords more time than **Time Series aggregation**, because we get a combination of several categorical values instead of more simple numerical values. In the following section, we will describe how to achieve the corresponding new datamarts by performing associative or Time Series aggregation.

7.4.3 Associative aggregation

To get the new datamart by performing associative aggregation involves three steps:

- Discretize the variables in the datamart.
- Concatenate the variables.
- Pivot the variables.

Modification I - Discretization

Table 7-3 shows how the variables of the historical datamart are discretized. The ranges of the different values are based on medical experiences. The values do not represent a standard deviation or some other kind of mathematical value; they are taken as commonly accepted values that are needed to differentiate between normal and abnormal ranges.

Table 7-3 Using discretization to transform numerical variables to categorical ones

Variable	Normal Ranges
GOT_D	GOT < 60
GPT_D	GPT < 60
LDH_D	LDH < 500
ALP_D	ALP < 300
TP_D	6.0 < TP < 8.5
ALB_D	3.5 <= ALB < 6.5
UA_D	UA > 6.5
UN_D	UN < 30
CRE_D	CRE < 1.5
TBIL_D	TBIL < 2.0
TCHO_D	TCHO < 250
TG_D	TG < 200
CPK_D	CPK < 250
C4_D	C4 > 10
WBC_D	3.5 <= WBC < 9.0
RBC_D	3.5 <= RBC < 6.0
HGB_D	10 <= HGB < 17

Variable	Normal Ranges
HCT_D	29 <= HCT < 52
PLT_D	100 <= PLT < 400
IGG_D	900 <= IGG < 2000

Please note that for the discretization itself, each value has to be checked for whether interval it belongs. Discretization will then take place in form of substituting the existing value by the new interval.

Modification II - Concatenation

After finishing discretization the data appears as indicated in Table 7-4. The original values are no longer there since discretization, by substituting numerical by categorical values, leads to new entries.

Table 7-4 Modified historical datamart after discretization has taken place

ID	Date	GOT	GPT	LDH	ALP	...
1	860419	GOT<60	GPT<60	LDH<500	ALP<300	...
1	860430	GOT<60	GPT<60	LDH<500	ALP<300	...
2	960315	GOT<60	GPT<60	LDH<500	ALP<300	...
...	...	...	...	...	...	...

Then, the concatenation with *EXAMINATION_DATE* and *THROMBOSIS* can be done as follows:

- Join the examination datamart and the historical data: only *EXAMINATION_DATE* and *THROMBOSIS* are of interest.
- Compare *EXAMINATION_DATE* with the date when the historical medical test has taken place. Expand the joined datamart by a new variable, *WHEN?,* with the value of:
  - *BEFORE* when the historical medical test had taken place *before* the thrombosis medical test was done
  - *AFTER* when the historical medical test had taken place *after* the thrombosis medical test was done
- Expand the joined datamart by a new column *STAT* that indicates whether the patient was diagnosed with thrombosis or not.
  - *0* indicates that the patient does not have thrombosis.
  - *1* indicates that the patient has thrombosis.
- Concatenate the historical test with all his variables with *WHEN?* and *STAT.*

Finally, we get a new datamart (Table 7-5).

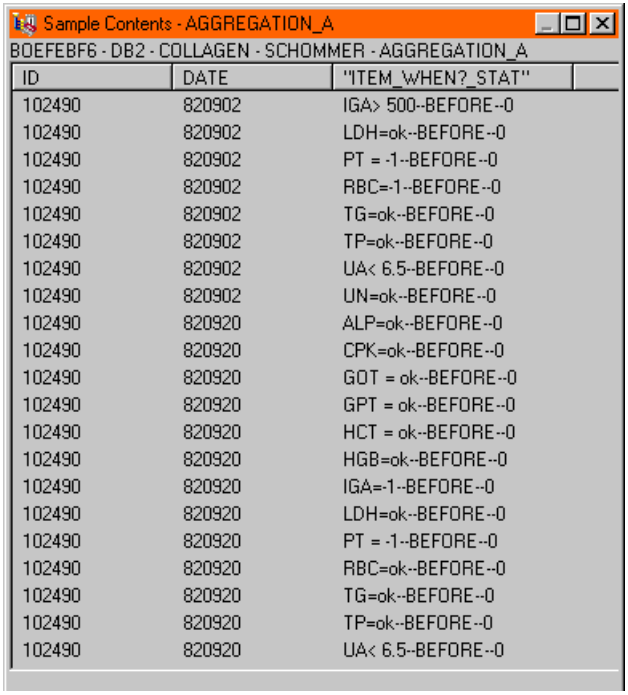
Table 7-5 Modified historical datamart after concatenation has taken place

ID	Date	GOT_D	GPT_D	LDH_D	ALP_D	...
1	860419	GOT<60 -BEFORE-0	GPT<60 -BEFORE-0	LDH<500 -BEFORE -0	ALP<300 -BEFORE -0	...
1	860430	GOT<60 -AFTER -1	GPT<60 -AFTER -1	LDH<500 -AFTER -1	ALP<300 -AFTER -1	...
2	960315	GOT<60 -BEFORE -0	GPT<60 -BEFORE -0	LDH<500 -BEFORE -0	ALP<300 -BEFORE -0	...
...	...	...	...	...	...	...

It is clear that historical tests that were done *before* the thrombosis test had taken place only contain values with suffix'BEFORE--0', and not with suffix'BEFORE--1'. For that point of time, no information in respect to the test result is still available.

Pivot

The last step in getting the new datamart is to pivot the values (Table 7-5) by having ID and Date as fields that need to be repeated (Figure 7-1). The pivoted value is stored in variable *ITEM_WHEN?_STAT*.


Sample Contents - AGGREGATION_A

BOEFEBF6 - DB2 - COLLAGEN - SCHOMMER - AGGREGATION_A

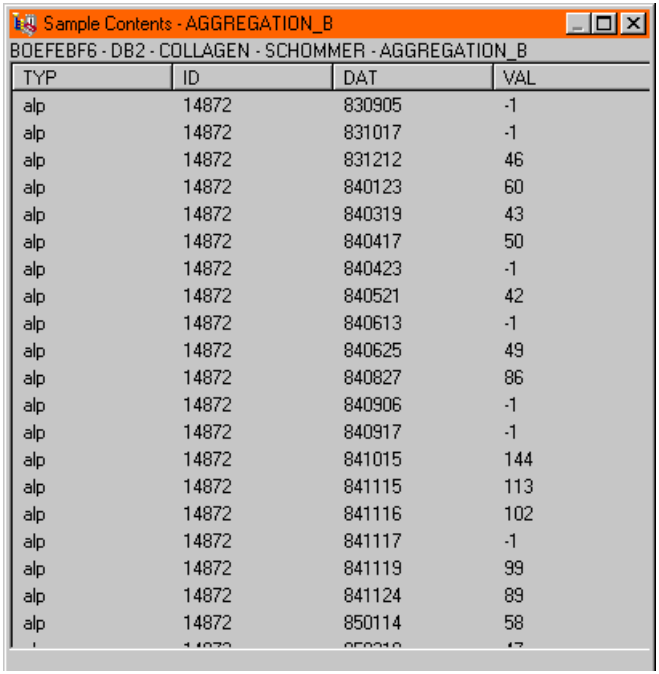
ID	DATE	"ITEM_WHEN?_STAT"
102490	820902	IGA> 500--BEFORE--0
102490	820902	LDH=ok--BEFORE--0
102490	820902	PT = -1--BEFORE--0
102490	820902	RBC=-1--BEFORE--0
102490	820902	TG=ok--BEFORE--0
102490	820902	TP=ok--BEFORE--0
102490	820902	UA< 6.5--BEFORE--0
102490	820902	UN=ok--BEFORE--0
102490	820920	ALP=ok--BEFORE--0
102490	820920	CPK=ok--BEFORE--0
102490	820920	GOT = ok--BEFORE--0
102490	820920	GPT = ok--BEFORE--0
102490	820920	HCT = ok--BEFORE--0
102490	820920	HGB=ok--BEFORE--0
102490	820920	IGA=-1--BEFORE--0
102490	820920	LDH=ok--BEFORE--0
102490	820920	PT = -1--BEFORE--0
102490	820920	RBC=ok--BEFORE--0
102490	820920	TG=ok--BEFORE--0
102490	820920	TP=ok--BEFORE--0
102490	820920	UA< 6.5--BEFORE--0

Figure 7-1 Snapshot of the datamart after performing associative aggregation

7.4.4 Time Series aggregation

The realization of Time Series aggregation is much easier, because it affords no categorical modifications. It only uses pivotization based on numerical values for the creation of the new datamart. In addition to *ID* and *Date* we have to define *TYPE* as the name of the variable that contains the test value. Therefore, the variables *ID, DATE* and *TYPE* have to be repeated during pivot. The pivoted value is stored in *VAL*.

Figure 7-2 shows the datamart after Time Series aggregation had taken place.


Sample Contents - AGGREGATION_B

BOEFEBF6 - DB2 - COLLAGEN - SCHOMMER - AGGREGATION_B

TYP	ID	DAT	VAL
alp	14872	830905	-1
alp	14872	831017	-1
alp	14872	831212	46
alp	14872	840123	60
alp	14872	840319	43
alp	14872	840417	50
alp	14872	840423	-1
alp	14872	840521	42
alp	14872	840613	-1
alp	14872	840625	49
alp	14872	840827	86
alp	14872	840906	-1
alp	14872	840917	-1
alp	14872	841015	144
alp	14872	841115	113
alp	14872	841116	102
alp	14872	841117	-1
alp	14872	841119	99
alp	14872	841124	89
alp	14872	850114	58

Figure 7-2 Snapshot of datamart after performing Time Series aggregation

7.4.5 Invalid values in Time Series aggregation

For Time Series aggregation, there is one point that we were not aware of at this time: invalid values. As seen in Figure 7-2, the variable *VAL* contains a lot of values that are '-1'. In fact, this will lead to a problem, because we have to decide how these values should be treated when using the corresponding time series technique.

There are some methods that may be suited, for example:

- Accept all the values of *VAL* and use them as they are (by '-1').
- Take the last valid value of this variable (for the patient) and substitute the current value by the new one. If there is no valid value before, then delete the current value.
- Delete the entry.

The first point sometimes would result in a time series where gaps occur. Although this may be acceptable, this is not real. Furthermore, the time series would then become very correlated with missing test values.

On the other hand, the substitution of these values by a previous one (point 2) may be acceptable too, but this would definitely lead us to the wrong collection of data. Generally, it would remain unknown whether a test was not done because the physician had no time, had forgotten to perform the test, or assumed the test was not necessary. Some kind of substitution would then lead us to results that would not represent the world as it actually is.

If we delete the corresponding values (point 3), then we reduce the number of values for *VAL* of type *TYP*. The question here is, what percentage we then affect. As an example, consider the scenario in the following example.

Example 7-1 Scenario on medical practice

```
A patient is delivered to the hospital with diagnosis "thrombosis". In order to
control his health some tests are planned that need to be performed. However,
the status of his health is getting better and better, and a lot of tests no
longer remain of interest because:

   From a medical standpoint, these tests generate potentially no new
   information.

   From a management point of view, the cost of these tests could be saved.

Therefore, some tests are not done.
```

The example describes a scenario of how it really is in medical practice. For several reasons, it is possible that a lot of test values for *VAL* of type *TYP* are invalid (for a specific patient, respectively). But, if we delete the values then the question will be, How many values for *VAL* of type *TYP* remain?

Overall, the number of invalid values for Time Series aggregation is about 18.4%. The percentages for the different tests can be seen in Table 7-6.

Table 7-6 All tests (after Time Series aggregation) with the percentage of invalid values

Typ	Relative (in percent)
GOT	20.00
GPT	20.04
LDH	19.52
ALP	21.90
TG	45.54
ALB	22.29
UA	22.31

Typ	Relative (in percent)
UN	20.16
CRE	20.04
TBIL	32.30
TCHO	28.33
TP	21.47
CPK	62.49
C4	27.63
WBC	12.99
RBC	12.99
HGB	12.99
HCT	12.99
PLT	14.12
IGG	33.98

Does it make sense to use all the variables, especially those with a percentage above 25%? The answer is yes and no:

- Yes, because it is generally worthwhile to consider all the different types.
- No, because working with too many invalid values is not efficient.

For analyses, we take the first opportunity, where we define a threshold of maximal invalid values per *TYP* of 25%. Then the following types of *TYP* will be considered: *ALP,* G*OT, GPT, LDH, UN, CRE, WBC,* RBC, HGB, HCT, PLT and TP.

7.5 Choosing the mining technique

Choosing the mining techniques to use is *the fifth stage in our generic mining method*. This section describes the different techniques we used to perform the task.

Concerning associative aggregation, we focus on:

- Association discovery
- Sequence analysis

Concerning time series aggregation, we focus on:

- Detection of similar sequences within the time series

7.5.1 Association discovery

Association discovery is performed by the associations technique of the IM for Data. The idea of this technique is to calculate association rules based on the transactional input data.

Assume, we have associative aggregations the underlying datamart then:

- GPT_D > 60 ==> LDH_D > 500

  Which could become a valid association rule. The association rule says that if *GPT_D* is greater than 60 then also *LDH_D* will be greater than 500.

To characterize an association rule, association discovery uses three parameters:

- *Support*,
- *Confidence*
- Lift

Note: *Support* refers to items that are part of an association rule. It defines the probability for each item to occur in the whole population. It is a measure that indicates the quantitative importance

Note: *Confidence* refers to an associative rule. It is measure that indicates the qualitative importance of a rule. It is a marginal statistic that is calculated using the Bayesian probability.

Note: *Lift* refers to a association rule and defines the significance of a association rule. Mathematically, it represents relative deviation and is defined as the division of confidence and expected confidence.

- If the *lift* is exactly 1, then the association rule has no relevance, because it has the same effect as in the random case.
- If the *lift* is below 1, then both sides of an association rule inhibit each other with no positive effect.
- If the *lift* is greater than 1, then both sides of the association rules have positive impacts on each other: the association rule gets more relevant the higher the *lift* is.

Association discovery works mainly by two steps that were repeatedly performed: *Join* and *Prune*.

In the Join step, items (in case of associative aggregation items are the values of the variable *ITEM_WHEN?_STAT*; they will be joined to *item groups*) were combined together so that each new item group contains exactly one item more than the previous one. For example, the items can be combined to an item group *[GOT>60, LDH>500]*:

- GOT > 60
- LDH > 500

This item group is then checked whether it reaches a given minimum support threshold or not (Join step). If the condition can be satisfied then the item group is accepted; otherwise not.

In fact, the item group *[GOT>60, LDH>500]* is an abbreviation for the following two rules:

- GOT>60 ==> LDH>500
- LDH>500 ==> GOT>60

By the use of the minimum *confidence* threshold it now can be decided whether an association rule is associative or not. If the confidence of an association rule is greater than the threshold, then the association rule is accepted; otherwise not.

7.5.2 Sequence analysis

The main task of sequence analysis is to find categorical sequences over time or for a patient, for example:

- LDH>500 (first event) then
  LDH>500 (second event) then
  LDH=ok (third event)

Sequence analysis will be performed through the sequential patterns technique of the IM for Data. Sequence analysis is similar to association discovery but requires in addition a variable for the time. Furthermore, only a threshold for the support is allowed, because both confidence and lift are calculated on the basis of rules; and not on the basis of sequences.

Assume, we have associative aggregations the underlying datamart. Then, sequence analysis through sequential patterns works as follows:

- As input for the technique we need three variables that play a different role:

  - Transaction group: Transaction group reflects a set of tests that had taken place during a period or for a person. For associative aggregation, *Date* will play this role.
  - Transaction ID: Transaction ID contains the identification of the tests. For associative aggregation, *ID* will be used as the transaction ID.
  - Item: Contains the value of a test, for example, *LDH>500 - BEFORE -0*. For associative aggregation, *ITEM_WHEN?_STAT* will be used as the item.

- The support threshold is defined as the minimum percentage of discovered sequences that occur in associative aggregation for the transaction group. For example, if we use *Date* as to be in months, then a support for the sequence *LDH>500==>GOT>60* is exactly 75%, if this sequence occurs in 75% of all values of *Date*. In detail, if *Date* has 12 different values (Jan - Dec), then support(*LDH>500==>GOT>60*)=75%, if the sequence occurs exactly in nine months.
- The technique works then similar to association discovery and uses the same *join*/*prune* technique to build the sequences. The sequences are then unique and are not a representation of different sequences.
- As in association discovery, sequence analysis allows the implication of a taxonomy. A taxonomy is a hierarchical system that represents different levels of more generalized or more specific information.

7.5.3 Similar sequences

The main idea of similar sequences is to find a pattern of time series that is very similar to a pattern of another time series. For example, it could be that some subsequent values of *LDH* behave very similar to subsequent values of *GOT* over time. If this is true then there is possibly some kind of relation between both tests.

Similar sequences works on numerical values; it uses the following parameters to calculate the sequences:

- *Epsilon*: The maximal envelope drawn around a sequence. Another sequence is similar in respect to epsilon of this sequence if it fits into this envelope.
- *Gap*: The maximum length of subsequences that can be ignored, although it is not epsilon-similar (it does not fit into the envelope that was specified).
- *Window Size*: The length (number of successive values) of a subsequence (window) specifying the atomic unit for matching. In this window no outliers are allowed.
- *Matching length*: The minimal length of similar subsequence (number of successive values) that should be considered. The matching length is given

as the fraction of the minimum length of a similar subsequence and the total length of the whole sequence.

whether a curve has more values than another one does not play a role. For example, if we have 500 tests for *LDH* and 60 tests for *GOT*, then nevertheless there could be some similar sequences in there. The reason is that the sequences will be normalized.

Valid sequences are calculated by moving a window with size *Window_Size* through both time series. Using *Epsilon* and *Gap*, two sequences become similar if the matching between them contains at most *Gap* non similar values. If the number of non similar values is greater then *Gap*, then the window is continuously shifted to the right.

7.6 Interpreting the results

The *sixth stage in our generic mining method* is to interpret the results that we have obtained and determine how we can map them to our business. When you are first confronted with different mining techniques results, the first question to ask is “What does it all mean?”. In this section we describe how to understand and interpret the results from the different mining techniques.

We present some of the results we got from the analysis for associative aggregation and time series aggregation. Association discovery and sequence analysis are used for associative aggregation and similar sequences for time series aggregation.

Please note that for both associative aggregation and time series aggregation the following times are important:

- Time *before* the thrombosis medical test had taken place
- Time *when* the thrombosis medical test had taken place
- Time *after* the thrombosis medical test had taken place

For associative aggregation, this is indicated by ‘AFTER/BEFORE’, but the second point (when) does not exist; for time series aggregation, however, we need to search all of them individually.

7.6.1 Results for associative aggregation

The results for associative aggregation are widespread and allow a large bundle of different strategies. We will present some of them, beginning with simple statistics.

Simple statistics

Table 7-7 shows the relative behavior of some values in *ITEM_WHEN?_STAT*. The values represent ranges that are not normal in the medical sense.

The table contains the status before and after the thrombosis medical test had taken place. Additionally, it is indicated whether the test was positive (*AFTER-1*) or negative (*AFTER-0*). The comments on the right side describes whether there is a trend or not.

Table 7-7 Distribution snapshot of some of the values for type TYP before and after medical test (in percent)

ITEM_ WHEN?_ STAT	BEFORE-0	AFTER-0	AFTER-1	TREND
GPT>60	**8.7**	10.0	**15.2**	Increases high if test is positive (74%), normal if test is negative (14.9%)
LDH>500	**12.5**	21.0	**26.6**	Increases very high if test is positive (113%) and high, if test is negative (68%)
ALP>300	**3.0**	5.0	**14.0**	Increases very high if test is positive (366%) and high, if test is negative (66%)
TP<6.0	**11.0**	13.0	**30.0**	Increases very high if test is positive (173%) and normal, if test is negative (18%)
PLT<100	**6.4**	6.5	**11.8**	Increases high if test is positive (84%) and remains stable, if test is negative
PLT>400	**2.8**	**7.7**	1.3	Decreases high if test is positive (-54%) and increases very high, if test is negative (175%)
RBC<3.5	**12.0**	12.0	**28.0**	Increases very high if test is positive (133%) and remains stable if test is negative
HGB<10	**12.7**	10.0	**20.0**	Increases high if test is positive (57%) and decreases, if test is negative (-21%)
HCT<29	**8.0**	7.0	**15.0**	Increases high if test is positive (87%) and decreases, if test is negative (-12.5%)

The general behavior of all of these tests is that the corresponding concentration increases strongly if the thrombosis test was positive (except for *Platelet (PLT>400*)).

The behavior of *Alkaliphosphotase* (*ALP*) is the most interesting, because the concentration increases 366% if the test was positive. *ALP* could therefore be a very suitable test to indicate the occurrence of thrombosis.

Association discovery

Figure 7-3 shows association rules that show the status of the patients with and without thrombosis after the medical test had taken place.

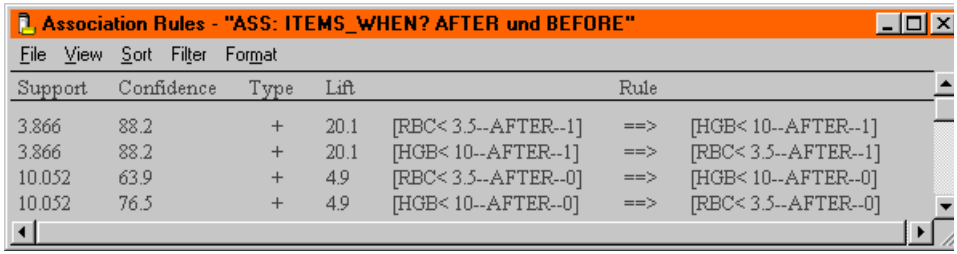


Figure 7-3 RBC<3.5 --AFTER--1 ==> HCT<10 --AFTER--1

A strong association is between *Red Blood Cells* (*RBC<3.5*) and *Hemoglobin* (*HCT<10*): if the thrombosis test is:

- *Positive,* then patients with result *RBC<3.5* implies result *HCT<10* with a Confidence of 88%. The Lift is 20.1 which indicates a very significant relationship between both test results.
- *Negative,* then patients with result *RBC<3.5* implies result *HCT<10* but with a minor Confidence of almost 64% and a Lift that is 5 times less than in the positive case. However, the support is more than 10% which indicates a high number of tests.

Figure 7-4 shows similar results for *Lactate Dehydrogenase* and *Urin nitrogen*.

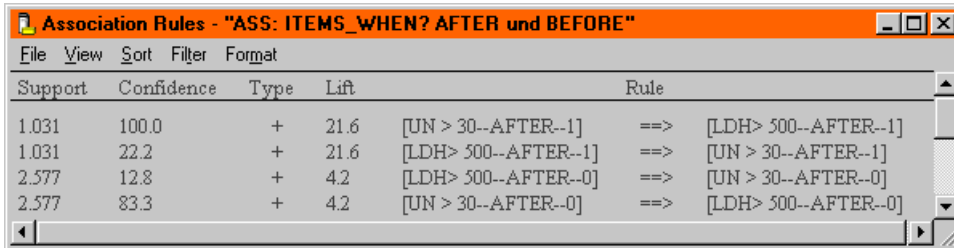


Figure 7-4 LDH>500--AFTER--1 ==> UN>30--AFTER--1

Here is a strong association between *LDH>500* and *UN>30.* If the thrombosis test is:

- *Positive,* then each patient with *UN>30* also has *LDH>500.* The Lift is 21.6 which indicates a very high significance. Almost 1% of all the patients share this behavior.

- *Negative,* then almost 83% with *UN>30* also have *LDH>500*. The Lift is quite low (4.2).

Another interesting relationship is between *Alkaliphosphotase* (*ALP*) and *Glutamin oxaloacetic transaminase (GOT)* as seen in Figure 7-5.

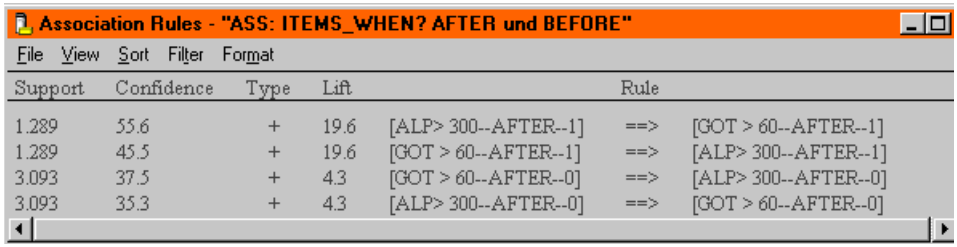


Figure 7-5 ALP>300--AFTER--1 ==> GOT>60--AFTER--1

If the thrombosis test was:

- *Positive,* then *ALP>300* and *GOT>60* have a Lift of 19.6 and a Confidence of almost 56%.
- *Negative,* then *ALP>300* and *GOT>60* have a Lift of only 4.3 and a Confidence of almost 35%

The last relationship can be seen in Figure 7-6 between *Glutamin pylvic transaminase (GPT)* and *Hematocrit (HCT)*.

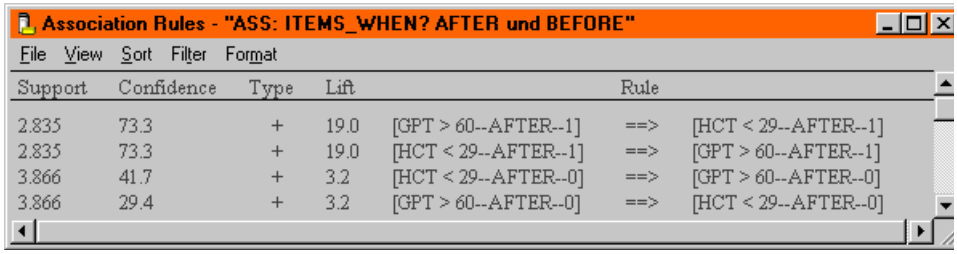


Figure 7-6 GPT>60--AFTER--1 ==> HCT<29--AFTER--1

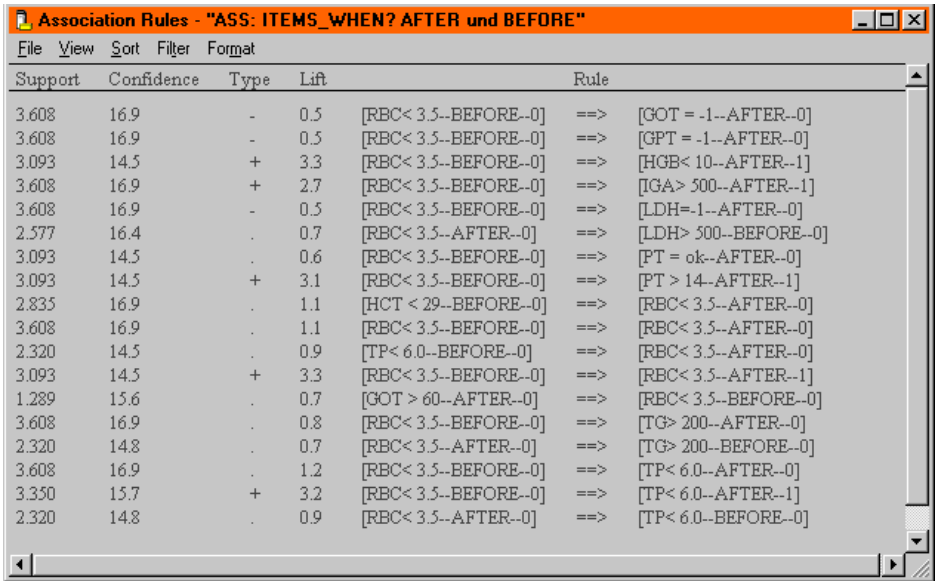
If the thrombosis test was:

- *Positive,* then almost 74% of patients with *GPT>60* also have *HCT<29*. The Lift is 19.0; it indicates a high significance for this rule.
- *Negative,* then only 29.4% of patients with *GPT>60* also have *HCT<29.* The Lift is 3.2; it indicates a less significance that is close to the random case.

Figure 7-7 gives a trend comparison between patients with a positive and a negative thrombosis test result.

For example, look at the following rows:

- 10: *RBC<3.5--BEFORE--0 ==> RBC<3.5--AFTER--0*
- 12: *RBC<3.5--BEFORE--0 ==> RBC<3.5--AFTER--1*



Support	Confidence	Type	Lift			Rule
3.608	16.9	-	0.5	[RBC< 3.5--BEFORE--0]	==>	[GOT = -1--AFTER--0]
3.608	16.9	-	0.5	[RBC< 3.5--BEFORE--0]	==>	[GPT = -1--AFTER--0]
3.093	14.5	+	3.3	[RBC< 3.5--BEFORE--0]	==>	[HGB< 10--AFTER--1]
3.608	16.9	+	2.7	[RBC< 3.5--BEFORE--0]	==>	[IGA> 500--AFTER--1]
3.608	16.9	-	0.5	[RBC< 3.5--BEFORE--0]	==>	[LDH=-1--AFTER--0]
2.577	16.4	.	0.7	[RBC< 3.5--AFTER--0]	==>	[LDH> 500--BEFORE--0]
3.093	14.5	.	0.6	[RBC< 3.5--BEFORE--0]	==>	[PT = ok--AFTER--0]
3.093	14.5	+	3.1	[RBC< 3.5--BEFORE--0]	==>	[PT > 14--AFTER--1]
2.835	16.9	.	1.1	[HCT < 29--BEFORE--0]	==>	[RBC< 3.5--AFTER--0]
3.608	16.9	.	1.1	[RBC< 3.5--BEFORE--0]	==>	[RBC< 3.5--AFTER--0]
2.320	14.5	.	0.9	[TP< 6.0--BEFORE--0]	==>	[RBC< 3.5--AFTER--0]
3.093	14.5	+	3.3	[RBC< 3.5--BEFORE--0]	==>	[RBC< 3.5--AFTER--1]
1.289	15.6	.	0.7	[GOT > 60--AFTER--0]	==>	[RBC< 3.5--BEFORE--0]
3.608	16.9	.	0.8	[RBC< 3.5--BEFORE--0]	==>	[TG> 200--AFTER--0]
2.320	14.8	.	0.7	[RBC< 3.5--AFTER--0]	==>	[TG> 200--BEFORE--0]
3.608	16.9	.	1.2	[RBC< 3.5--BEFORE--0]	==>	[TP< 6.0--AFTER--0]
3.350	15.7	+	3.2	[RBC< 3.5--BEFORE--0]	==>	[TP< 6.0--AFTER--1]
2.320	14.8	.	0.9	[RBC< 3.5--AFTER--0]	==>	[TP< 6.0--BEFORE--0]

Figure 7-7 Some association rules that indicate the behavior before and after the thrombosis test

Although patients with a negative thrombosis test results imply more *RBC<3.5* than patients with positive thrombosis test result (16.9% vs. 14.5%), behavior of *RBC* for patients with a positive test result is more significant, because the Lift in *RBC<3.5--BEFORE--0 ==> RBC<3.5--AFTER--1* row 12 is three times higher than for *RBC<3.5--BEFORE--0 ==> RBC<3.5--AFTER--0* in row 10.

A similar behavior can be found for the following two association rules:

- Row 16: *RBC<3.5--BEFORE--0 ==> TP<6.0--AFTER--0*
- Row 17:*RBC<3.5--BEFORE--0 ==> TP<6.0--AFTER--1*

Although patients with a negative thrombosis test result show a higher confidence for the association rule than patients with positive thrombosis test (16.9% vs. 15.75), the association rule (*RBC<3.5--BEFORE--0 ==> TP<6.0--AFTER--1*) in row 17 is more significant, because the Lift is almost three times higher than for *RBC<3.5--BEFORE--0 ==> TP<6.0--AFTER--0* in row 16.

Sequence analysis

Figure 7-8 shows a snapshot of sequences that were done associative aggregation.

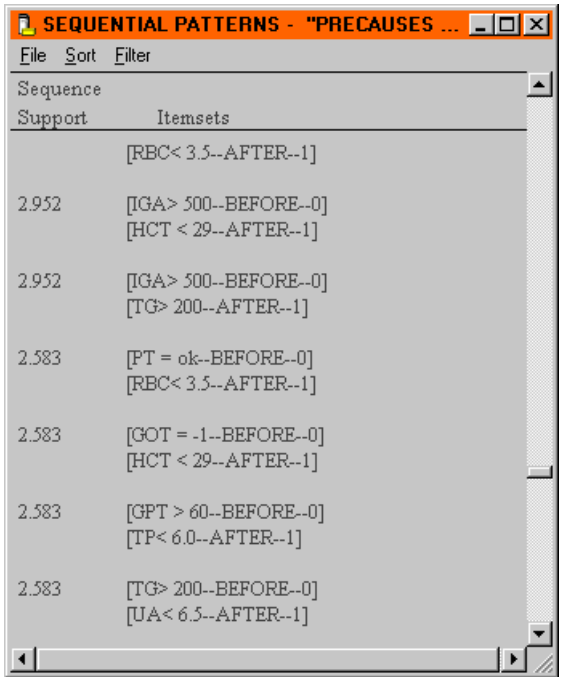


Figure 7-8 Sequence analysis by sequential patterns of length 2 for associative aggregation

The first sequence shows a combination of *Immunoglobulin* (*IGA*) and *Triglyceride* (*TG*) where *IGA* was measured before (*BEFORE*) and *TG* after the thrombosis test had taken place. Almost 3% of all patient had first an abnormal value of more than 500 for *IGA* and then an abnormal value of more than 200 for *TG*. The result of thrombosis test was positive (AFTER--1), the corresponding patients are therefore all diseased.

The last sequence shows a combination of *Glutamin pylvic transaminase* (*GPT*) and the *total number of proteins* (*TP*). Before the test was performed, 2.583% of all patients had an abnormal value of GPT (>60) and an abnormal value of TP (< 6.0) after the test. The result of thrombosis test was positive (AFTER--1), and the corresponding patients are therefore all diseased.

7.6.2 Results for Time Series aggregation

Some results for Time Series aggregation can be seen in Figure 7-9, Figure 7-10 and Figure 7-11. The figures contain a snapshot of all sequences that were discovered. Overall, there are two scenarios that were done for Time Series aggregation:

- Scenario 1: Comparisons between different time series for one patient are observed.
- Scenario 2: Comparisons between one time series for different patients are observed.

Figure 7-9 shows a result for Scenario 1. The displayed time series belongs to patient 5512586, a nine year old boy who was admitted to the hospital by his physician on September 30, 1997. He was diagnosed with thrombosis.

The first time series shows the behavior of *Hematocrit* (*HCT*), the second time series the behavior of *Urin nitrogen* (*UN*).

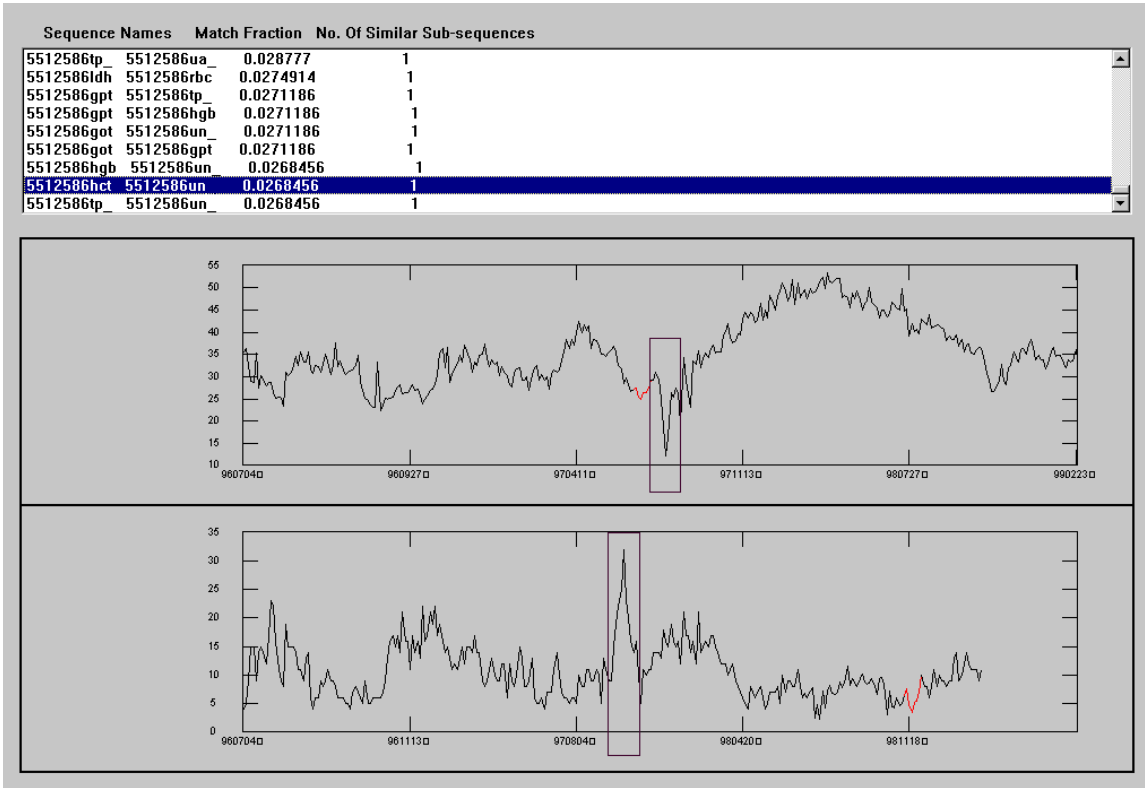


Figure 7-9 Time series between HCT and UN for patient 5512586

The thrombosis test for patient 5512586 had taken place on September 30, 1997. It is interesting that at this time the two time series show a very strange behavior: *Hematocrit (HCT)* gets a deep decrease of less than 15, and the value for *Urin nitrogen* (*UN*) increases almost 35. This is very abnormal since during the whole time such behavior could not be observed.

However, it is obvious that both time series absolutely are not similar. But it is interesting that just at the date where the thrombosis test was taken such a behavior occurs. It is very probable, that patient 5512586 got the thrombosis at this point of time.

Figure 7-10 shows a result for Scenario 2. There is one time series (*Urin nitrogen* (*UN*)) that belongs to patient 5512586 (first time series) and patient 5762587 (second time series).

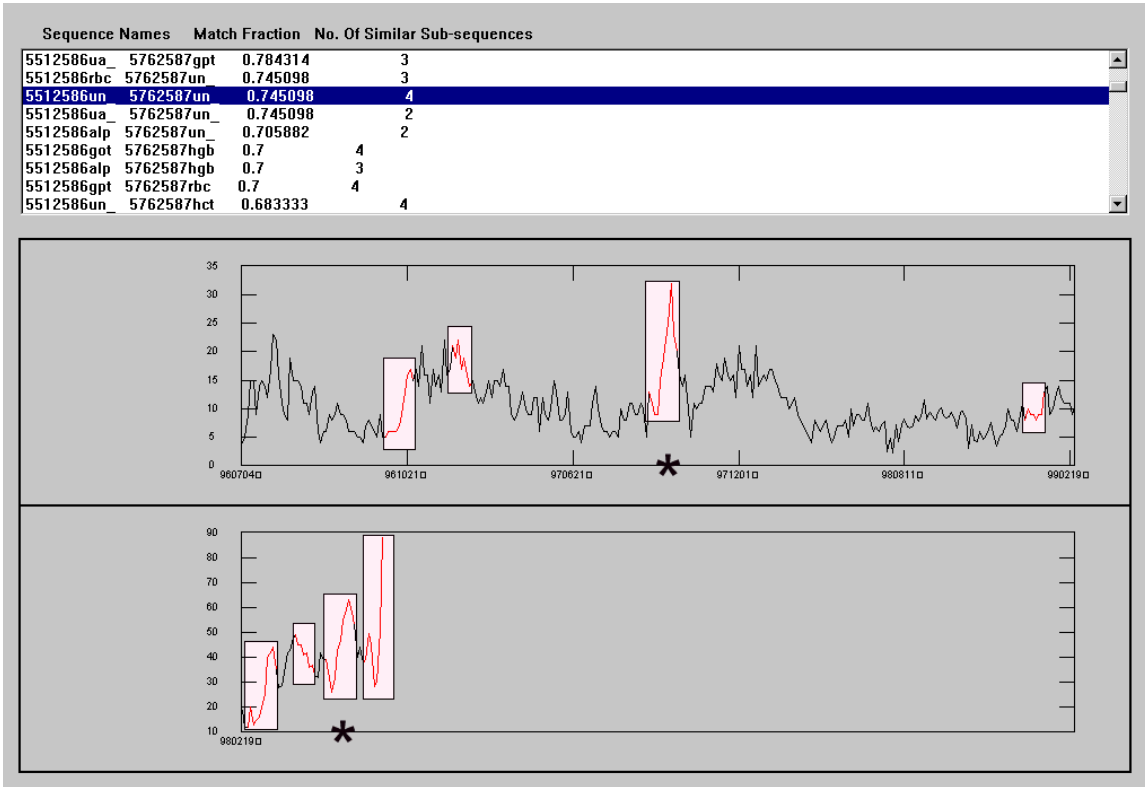


Figure 7-10 Similar sequences between UN for patient 5512586 and 5762587

Patient 5762587 is a 28 year old girl who was admitted to the hospital by her physician on March 26, 1998. She was diagnosed with thrombosis.

The surrounded sequences show the patterns that were found as to be similar for patients 5512586 and 5762587. The star indicates the date when the thrombosis test had taken place.

Generally, there is a common behavior of these curves over time. The vertical range is different because the values for Urin nitrogen are different for women and men in general. The first three patterns can be characterized as follows:

- First of all we register an increase of UN from 5 to 16 (5512586) and 15 to 42 (5762587), respectively.
- Then there comes a small decrease from 20 to 15 (5512586) and 50 to 35 (5762587), respectively.
- When the thrombosis test is performed we register a very high pike of over 30 (5512586) and 60 (5762587), respectively.

However, there were no further examinations done for the young girl after pattern 4. The answer to the specific question is yes: she died because of thrombosis.

Figure 7-11 shows another result for Scenario 2. There is one time series (*Lactate Dehydrogenase (LDH)*) that belongs to patient 5512586 and patient 4904003.

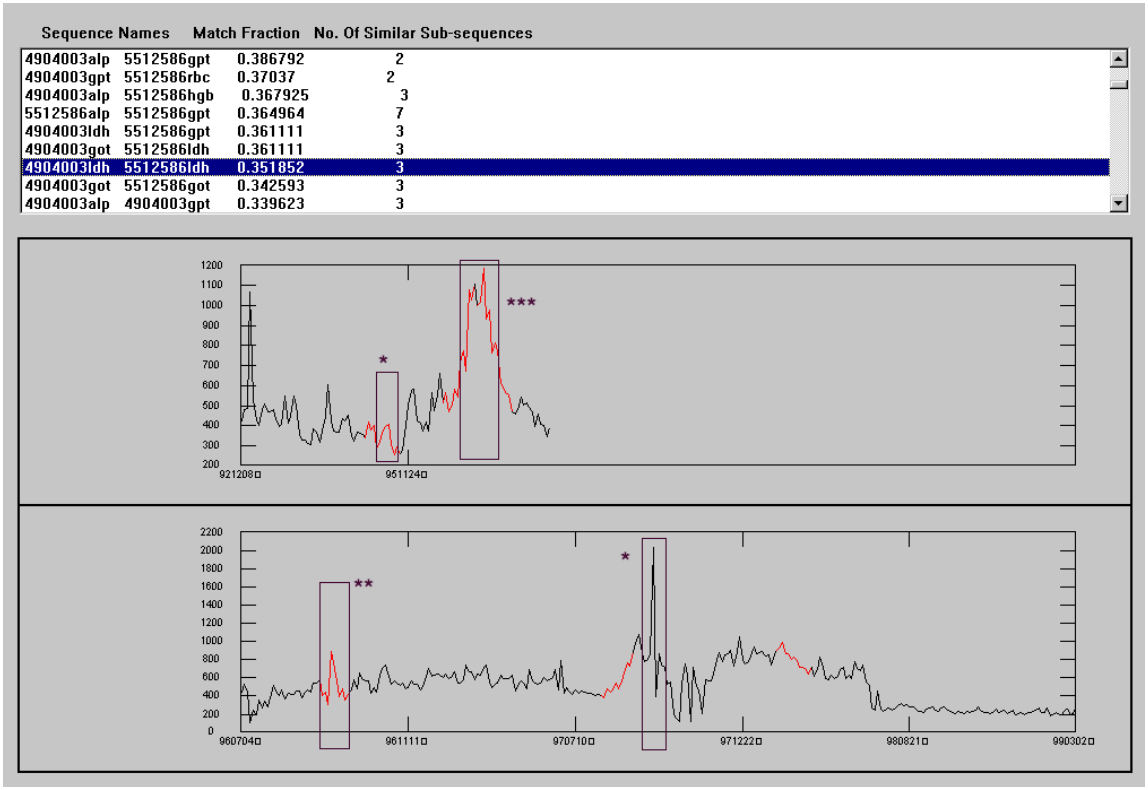


Figure 7-11 Similar sequences between LDH for patients 5512586 and 4904003

Patient 4904003 is 23 year old girl who was admitted to the hospital by her physician on November 9, 1995. She was diagnosed without thrombosis.

As we can see in Figure 7-12, the sequences of the time series that are indicated are very important. In detail, they have the following meanings:

- * indicates the date when the thrombosis test had taken place.

  For patient 5512586 we have September 30, 1997, for patient 4904003 November 9, 1995.

- ** indicates a pike that occurs approximately 2 years before the medical test was done for patient 5512586.

- *** indicates a pike that was measured approximately 2 years after the examination was done for patient 4904003.

Although patient 4904003 was not indicated to have thrombosis, there is some similar sequences to patient 5512586. But the important thing is that the sequences that were measured for patient 4904003 at thrombosis test time are very similar to the sequence that was indicated 2 years before for 5512586 (and where patient 5512586 was tested and diagnosed with thrombosis).

Therefore, it could be that *Lactate dehydrogenase* (*LDH*) could be an indicator for thrombosis.

7.7 Deploying the mining results

The final and *seventh stage in our generic mining method* is perhaps the most important of all. How do you deploy the mining results into your business to derive the business benefits that data mining offers? The reason this is so important is that all too often data mining is seen as an analytical tool that can be used to gain business insight but is difficult to integrate into existing systems.

In this section, after a short summary about the strategies we performed, we will propose some ideas about how to use this strategy in other medical scenarios.

7.7.1 What we did so far

In this chapter we were concerned with the question whether we can discover some precauses for disease or not. In detail, we performed the following steps:

- We used three datamarts that were recorded for a large number of patients:
  - The first datamart contained demographic data.
  - The second datamart contained data that was recorded because of a thrombosis medical test.
  - The third datamart contained historical data consisting of different tests that were recorded over years.
- Generating two aggregations out of the thrombosis data by:
  - Discretization, concatenation and pivoting time series data (associative aggregation): this aggregation contained records with two markers. The first marker indicated whether the test was done before or after the thrombosis test. The second marker indicated whether the thrombosis test was positive or negative.
  - Pivoting time series data (time series aggregation)
- For associative aggregation, we used association discovery (by associations) and sequence analysis (by sequential patterns) to search for precauses for

thrombosis. By the use of both markers we could detect association rules and sequential patterns that described the behavior of several tests over time.

- For time series aggregation, we used similar sequences to search for similar time patterns that occur in two time series. We found out that *LDH* could be an appropriate test to identify thrombosis in advance.

7.7.2 How can the model be deployed?

There is a lot of medical scenarios where time series can be analyzed by the usage of the strategy described above. This section picks up two scenarios that will be shortly discussed.

Application in medical tests

As we used thrombosis data in this chapter, the strategy we discussed is also applicable in many other scenarios of this kind, for example:

- Specific medical tests that occur from time to time: cancer examinations or vaccinations
- Standard examinations in hospitals: blood pressure or heart frequency

In each of these scenarios, the benefit for the physician is to get new information that can act as a trend or alarm indicator.

Assume, we have patients in the hospital who need to be continuously tested, for example, blood pressure or heart rate. The question now is: Can we say something about the patient's health state in general if we know that the an increase of the heart rate for 20% implies automatically an increase of LDH of 10%? If yes, then the use of our suggested strategy could become an alarm system that can be introduced in hospitals.

Furthermore, we could use this strategy as a prophylaxis test for diverse diseases, for example, cancer or tuberculosis. The question here is: Can we find some trends over time that are sure indications for cancer or tuberculosis? If yes, then our strategy could become a milestone in the research of trend analyses.

Bioinformatics

Another important topic in the application of time series analysis is *bioinformatics*. Bioinformatics is the analysis of biological information using computers and statistical techniques, the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. It is used in analyzing genomes, proteomes (protein sequences), three-dimensional modeling of biomolecules and biologic systems, and so on.

In its simplest form, a DNA (*Desoxyribonucleic acid*) record can be represented as a string of nucleotides with some tag or identifier, for example:

>eIF4E [org=Drosophila melanogaster] [strain=Oregon R]

...CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGT TCA...

Where:

- A indicates Adenine
- G indicates Guanine
- C indicates Cytosine
- T indicates Thymine

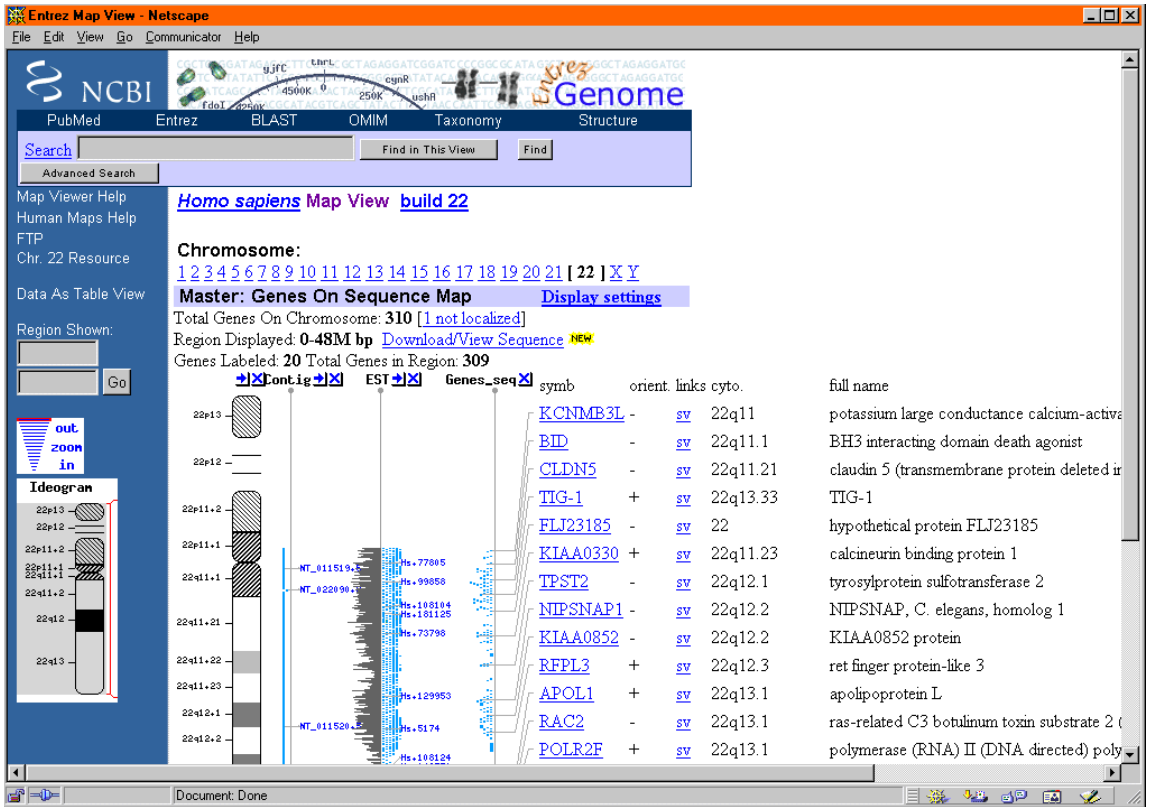


Figure 7-12 Homo sapiens: X-Chromosome 22

The DNA sequence in Figure 7-12, that can be found at `http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/maps.cgi?org=hum&chr=22` is represented in *FASTA* format that is used in a variety of molecular biology software suites:

- ">" indicates the beginning of a new file.
- eIF4% represents an identifier, is followed by the DNA sequence in lowercase or uppercase letters, usually with 60 characters per line.
- [org = Drosophila melanogaster] an additional comment

Nucleotides are a subunit of DNA consisting of a nitrogenous base, a phosphate molecule, and a sugar molecule (deoxyribose in DNA). Thousands of nucleotides are linked to form a *DNA* molecule.

As an example for a *DNA*, Figure 7-12 shows the X-chromosome number 22 with all genes that are already known and in the right order.

Some questions that arise from this scenario are:

- Can we perform sequence analysis in order to detect sequence patterns that occur very often in the chromosome?
- If a mutation takes place in a chromosome, does there exist any relationships between nucleotides? If yes, does a mutation of the one nucleotid also influence the other ones and can we use one of the techniques described above to find such relationships?
- If we translate the activities of the nucleotides into a frequency, can we then detect similar sequences that occur over time? Can we then find indicators that are probably responsible for mutation?
- Sequence tagged site (STS) are a short (200 to 500 base pairs) DNA sequence that has a single occurrence in the human genome. Can we detect such STSs using our suggested strategies?
- Genetic disorders resulting from the combined action of alleles of more than one gene (for example, heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles; therefore, the hereditary patterns are usually more complex than those of single gene disorders. Can we detect such polygenic disorders using our suggested strategies?
- A problem in bioinformatics is the determination of the order of the nucleotides in a DNA molecule or the order of amino acids in a protein. This is referenced as *sequencing*. Can we detect such poly genic disorders using our suggested strategies?

8

The value of DB2 Intelligent Miner for Data

Throughout this book we have been concentrating on how data mining can be used to address specific business issues, rather than concentrating on the detailed capability of the data mining tool that we use. It should come as no surprise to you that we have been using the IBM DB2 Intelligent Miner for Data product (IM for Data) to perform all of the data mining functions described in the previous chapters. Although we have used some of the data mining techniques that are available to us, IM for Data offers a number of other tools and techniques that we have not been able to describe how to use.

In this chapter we take the opportunity to explain the advantages and benefits of using IM for Data by presenting an overview of where the product is placed in relation to its competitors and also a summary of all of its functions and capabilities.

The recent introduction of DB2 Intelligent Miner Scoring (IM Scoring) opens up a whole range of new possibilities for deploying the results of data mining into your business process. You may be concerned that IM Scoring is only applicable to IM for Data and DB2 databases. This is not the case, and we provide some further details on the background to IM Scoring.

8.1 What benefits does IM for Data offer?

The IBM mining products offer a full range of data mining techniques and capabilities that can be used to address a wide range of business problems. The techniques included support the identification of unknown patterns and trends within your stored data. If you use DB2 as your database, then IM for Data is the only mining product that is fully integrated with the DB2 database. If you do not use DB2, then you can still use IM for Data through the Relational Connect feature in DB2 for read only access or through a companion product called DataJoiner. DataJoiner provides additional capabilities that give federated access to a number of different data sources in read and write mode. So even if your data is distributed across a number of databases and file systems you can still perform data mining.

Some straightforward benefits are:

- High performance and scalability of mining functions. Therefore, there is no need for sampling.
- Customized front-ends and visualizers.
- Excellent clustering algorithm for demographic data. Best suited for detection of niche clusters.
- Highly efficient association algorithm and visualizer. Only association algorithm which supports taxonomies.
- Mines any known database format via Relational Connect feature (part of IBM DB2 Universal Database) or IBM DataJoiner.
- Coupling with SAP Business Information Warehouse.
- Real time model deployment through IM Scoring.
- Running on parallel machines.

IM for Data is one of the industry leaders in integration of data mining and database technology.

8.2 Overview of IM for Data

IM for Data provides the user with the complete spectrum of state-of-the-art data mining algorithms, together with a range of tools for data preparation and statistical analysis. The data mining algorithms are categorized as follows:

- Clustering
- Associations discovery
- Sequential patterns discovery
- Classification

- Prediction
- Similar time sequence discovery

Further details of each of these algorithms are given in the following sections.

Each data mining algorithm can be considered as a tool that is used to perform a specific discovery task (that means using the mining analogy — discovering seams of information). Continuing the mining analogy, the tools can also be used in concert to perform more specific complex operations and to extract diamonds or nuggets of valuable information. Next, we describe the different components and functions of the product and how this is achieved.

8.2.1 Data preparation functions

The data to be mined can be in the form of flat files or held in tables within a relational database. If the database option is used, then the preferred database is the IBM DB2 Universal Database (DB2 UDB). Connection to other databases is possible using the DB2 Relational Connect feature to accede in read-only mode to ORACLE, SYBASE, SQLServer or using IBM DataJoiner product to accede in read and write mode to INFORMIX, ORACLE, SYBASE, TERADATA, SQLServer.

Once the desired input data has been selected, it is usually necessary to perform certain transformations on the data. IM for Data provides a wide range of data preparation functions that help to quickly transform the data to be analyzed. The data preparation functions of IM for Data are:

- Aggregate values: To aggregate values of existing fields, for example, monthly salary to annual salary.
- Calculate values: To create new fields with the result of a calculation of existing fields. The Calculate values preprocessing function creates new fields using SQL expressions. These new fields are appended to the input data to create the output data.
- Clean up data sources: To delete database tables or views in the database, usually those no longer used as input or output data.
- Convert to lower-case or upper-case: To convert one or more fields in the output data.
- Copy records to file: To copy records from a database table or view to a flat file (you can also sort by the field you specify).
- Discard records with missing values: To remove input data records containing a missing (NULL) value in any of the fields you specify.
- Discretize into quantiles: To assign input data records to the number of quantiles you specify.

- Discretize using ranges: To assign input data records by splitting the value range of a continuous field into intervals, and then mapping each interval to a discrete value.
- Encode missing values: To encode missing values in the input data by specifying one or more fields to search for missing values and further specifying which value to use as a replacement for any missing values in these fields.
- Encode non-valid values: To encode values found in the first input data if they do not match valid values from the second input data and further specifying which value to use as a replacement for any such value.
- Filter fields: To filter the input data fields to get an output table or view containing only the fields you specify or the ones you didn't specify.
- Filter records: To filter the input data records to get only the records you specify for which a given condition is true.
- Filter records using a value set: To compare field values in a first input data with values in a value set specified for a second input data; then filter the records whose input field contains a value present in the value set.
- Get random sample: To reduce input data to a smaller sample by specifying the size of the sample as a percentage of the input data.
- Group records: To summarize groups of records into a single record that contains aggregated values of the group.
- Join data sources: To join two database tables or views based on one or more pairs of join fields from the input data.
- Map values: To map values found in the first input data to values found in the second input data.
- Pivot fields to records: To split each record of the input data into multiple records.
- Run SQL statements: To submit SQL statements.

Data preparation functions are performed through the GUI, reducing the time and complexity of data mining operations. The user can transform variables, input missing values, and create new fields through the touch of a button. This automation of the most typical data preparation tasks is aimed at improving productivity by eliminating the need for programming specialized routines.

8.2.2 Statistical functions

Although the data mining tools are designed to discover information from the data, understanding the data structure in terms of outlying values or highly correlated features is often necessary if the full power of the mining techniques is to be realized. Therefore after transforming the data, the next stage is usually to analyze it. IM for Data provides a range of statistical functions to facilitate the analysis and the preparation of data, as well as providing forecasting capabilities. For example, you can apply statistical functions like regression to understand hidden relationships in the data, or use factor analysis to reduce the number of input variables. The statistical functions included are:

- Factor analysis: Discovers the relationships among many variables in terms of a few underlying, but unobservable, quantities called factors.
- Linear regression: Used to determine the best linear relationship between the dependent variable and one or more independent variables.
- Polynomial regression: Used to determine the best polynomial relationship between the dependent variable and one or more independent variables.
- Principal component analysis: Used to rotate a coordinate system so that the axes better match the data distribution. The data can be now described with fewer dimensions (axes) than before.
- Univariate curve fitting: Finds a mathematical function that closely describes the distribution of your data.
- Univariate and bivariate statistics: Descriptive statistics, especially means, variances, medians, quantiles, and so on.

8.2.3 Mining functions

All of the mining functions can be customized using two levels of expertise. Users who are not experts can accept the defaults and suppress advanced settings. However, experienced users who want to fine tune their application are provided with the capability to customize all settings according to their requirements. It is also possible to define the mode in which the data mining model will be performed. The possible modes are:

- Training mode: In which a mining function builds a model based on the selected input data.
- Test mode: In which a mining function uses new data with known results to verify that the model created in training mode produces adequate results.
- Application mode: In which a mining function uses a model created in training mode to predict the specified field for every record in the new input data.

The user can also use data mining functions to analyze or prepare the data for a further mining run. The following sections describe each mining algorithm in more detail, using typical commercial examples to illustrate the functionality.

Clustering

Clustering is used to segment a database into subsets, the clusters, with the members of each cluster having similar properties. IM for Data can perform clustering by using either a statistical clustering algorithm (Demographic Clustering) or a neural network algorithm (Kohonen Clustering), depending on the type of the input data set. The neural clustering algorithm requires the user to specify the number of clusters required; the statistical clustering algorithm automatically determines the “natural” number of clusters.

When clustering is performed there are no preconceived notions of what patterns exist within the data; it is a discovery process. The results of the clustering process can be visualized (see 4.6, “Interpreting the results” on page 62) to determine the composition of each cluster. Visualization graphically presents the statistical distributions of the characteristics of those records that compose the cluster in comparison with the data set as a whole. Tabular output is also provided to enable further analysis.

In addition to producing graphical and tabular output, a “cluster model” is also generated (Training Mode). It is also possible to generate a user-defined table, which can include selected information from the input records, together with the cluster number of the segment to which the record has been assigned. The output table can also include details on the next nearest cluster and a measure of the confidence in the degree of matching to the nearest and next nearest clusters for each record (Test Mode). An Application Mode is also provided, in which new data records are assigned to clusters and an output table generated.

In the commercial environment clustering is used in the areas of cross-marketing, cross-selling, customizing marketing plans for different customer types, deciding on media approach, understanding shopping goals, and so forth.

Associations

The association algorithm, developed at the IBM Almaden Research Center in San Jose, California, compares lists of records to determine if common patterns occur across the different lists. In a typical commercial application the algorithm looks for patterns such as whether, when a customer buys paint, they also buy paintbrushes. More specifically, it assigns probabilities; for example, if a

customer buys paint, there is a 20% chance that they will buy a paintbrush. The advantage of this approach is that it compares all possible associations. It also finds multiple associations, for example, if a customer buys paint and paint brushes, there is a 40% chance they will also buy paint thinner.

When the algorithm runs, it potentially creates hundreds or thousands of such rules. The user can however select a subset of rules that have either higher confidence levels (a high likelihood of B given A) or support levels (the percent of transactions in the database that follow the rule) or high lift (the ratio of measured to expected confidence for a rule). It is up to the user to read the rules and decide if the rules are:

- Chance correlations (for example, paint and hair rollers were on sale the same day and therefore were correlated by chance).
- Known correlations (for example, the paint and paint brush correlation is something that would have been known).
- Unknown but trivial correlations (for example, red gloss paint and red non gloss paint correlation may be something unknown, and is unimportant to know).
- Unknown and important correlations (for example, paint and basketballs, which may be something previously unknown and very useful in both organization of advertising and product placement within the store).

Association discovery is used in market basket analysis, item placement planning, promotional sales planning, and so forth.

The association algorithm also includes the capability to include a taxonomy for the items in the lists (for example, paint and a paintbrush are hardware) and the algorithm will discover associations across the taxonomy (for example, there is a 50% confidence that customers who buy hardware also buy soft furnishing).

Sequential patterns

The purpose of discovering sequential patterns is to find predictable patterns of behavior over a period of time. This means that a certain behavior at a given time is likely to produce another behavior or a sequence of behaviors within a certain time frame.

The rule generation method is a variation of the association technique. It analyzes the shopping behavior of customers, for example, over time. Instead of looking at 10,000 purchases, the algorithm looks at 10,000 sets of purchases. These sets are, for example, lists of purchases from a sequence of shopping trips by a single customer. As a typical commercial example, one set of lists may be the purchases of computer:

- Computer in December

- Computer games and joy stick in January
- Additional computer memory and larger hard drive in March

If this sequence, possibly with different time scales but the same order, were repeated across a number of customers, then the sequential association algorithm would typically return a rule, such as:

> If following the purchase of a computer, the customer purchases computer games, then there is a 30% chance that extra computer memory will be purchased in a subsequent visit to the store.

The algorithm also includes the capability to define minimum and maximum time periods between the items in the lists. This would, for example, enable the above rule to include the statement that computer memory will be purchased no earlier than one month and within three months of the purchase of the computer games.

Sequential pattern detection can therefore be used to discover associations over time. This is especially useful in commercial applications, such as direct marketing, or the design special advertising supplements, and so on.

Classification

Classification is the process of automatically creating a model of classes from a set of records that contain class labels. The induced model consists of patterns, essentially generalizations over the records that are useful for distinguishing the classes. Once a model is induced, it can be used to automatically predict the class of other unclassified records. IM for Data has two classification algorithms, a tree induction algorithm (modified CART regression tree) and a neural network algorithm (back propagation), to compute the classes.

The tree and neural network algorithms develop arbitrary accuracy. While neural networks often produce the most accurate classifications, trees are easy to understand and modify and the model developed can be expressed as a set of decision rules.

Commercial applications of classification include credit card scoring, ranking of customers for directed mailing, and attrition prediction. One of the main uses of the tree algorithm is to determine the rules that describe the differences between the clusters generated by the clustering algorithm. This is achieved by taking the output table from the clustering algorithm and constructing the decision tree using the cluster label as the class.

Value prediction

Value prediction is similar to classification; the goal is to build a data model as a generalization of the records. However, the difference is that the target is not a class membership but a continuous value, or ranking. IM for Data has two prediction algorithms: a neural network algorithm and a Radial Basis Functions (RBF) algorithm. The radial basis function is particularly efficient and is appropriate for value prediction with very large data sets.

Similar time sequences

The purpose of this process is to discover all occurrences of similar subsequences in a database of time sequences. Given a database of time sequences, the goal is to find sequences similar to a given one, or find all occurrences of similar sequences. The powerful alternatives afforded by multiple methods are enhanced by the fact that several of the methods are supported by more than one mining technique. Multiple techniques are often used in combination to address a specific business problem.

8.2.4 Creating and visualizing the results

Information that has been created using statistical or mining functions can be saved for further analysis in the form of result objects. The result objects can be visualized using a variety of graphical displays or the results exported to spreadsheets (for example, EXCEL, LOTUS 123), or to browsers (for example, Netscape, Explorer), or to specific statistical packages (for example, SPSS).

Result objects can be used in several ways:

- To visualize or access the results of a mining or statistical function
- To determine what resulting information you want to write to an output data object
- To be used as input data, when running a mining function in test mode to validate the predictive model representation by the result
- To be used as input data, when running a mining function in application mode to apply the model to new data

8.3 DB2 Intelligent Miner Scoring

DB2 Intelligent Miner Scoring (IM Scoring) is an economical and easy-to-use mining deployment capability. It enables users to incorporate analytic mining into Business Intelligence, eCommerce and OLTP applications. Applications score records (segment, classify or rank the subject of those records) based on a set of predetermined criteria expressed in a data mining model.

These applications can better serve business and consumer users alike — to provide more informed recommendations, to alter a process based on past behavior, to build more efficiencies into the online experience; to, in general, be more responsive to the specific situation at hand. All scoring functions offered by the DB2 Intelligent Miner for Data are supported.

The IM Scoring is an add-on service to DB2, consisting of a set of User Defined Types (UDTs) and User Defined Functions (UDFs), which extends the capabilities of DB2 to include some data mining functions. Mining models continue to be built using the IM for Data, but the mining application mode functions are integrated into DB2. Using the IM Scoring UDFs, you can import certain types of mining models into a DB2 table and apply the models to data within DB2. The results of applying the model are referred to as scoring results and differ in content according to the type of model applied. The IM Scoring includes UDFs to retrieve the values of scoring results.

The results of applying the model are referred to as scoring results and differ in content according to the type of model applied. The IM Scoring includes functions to retrieve the values of scoring results.

The IM Scoring is available on the following operating systems:

- AIX
- Solaris
- Windows NT, Windows 2000
- Linux, Linux/390

Summary of functionality

The application mode for the following IM for Data mining and statistical functions are supported by the IM Scoring:

- Demographic and neural clustering
- Tree and neural classification
- RBF and neural prediction
- Polynomial regression

Scoring functions are provided to work with each of these types. Each scoring function includes different algorithms to deal with the different mining functions included within a type, for example, the clustering type includes demographic and neural clustering and so, scoring functions for clustering include algorithms for demographic and neural clustering. For all the supported mining functions, you build and store the model using the IM for Data. Models must then be exported to an external file.

Exchanging models

In support of facilitating the exchange of mining models between applications, IM Scoring makes full use of the Predictive Model Markup Language (PMML) published by Data Mining Group.

PMML is a standard format. Based on the Extensible Markup Language (XML), it provides a standard by which data mining models can be shared between the applications of different vendors. It provides a vendor-independent method of defining models so that proprietary issues and incompatibilities are no longer a barrier to the exchange of models between applications. You can find more information about PMML on the Web site of the Data Mining Group, at:

`http://www.dmg.org`

The IM Scoring includes a facility for converting models, built using IM for Data, to the PMML format. Using this facility, you can select the PMML format when you export the model from the IM for Data GUI. Conversion to PMML is not necessary when importing models into DB2 using the IM Scoring functions. The model import functions read models in either PMML or IM for Data format.

Using the PMML standard allows the models created by IM for Data to be used in databases other than DB2, and IM Scoring also supports ORACLE Cartridge Extenders.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 181.

- *Intelligent Miner For Data Applications Guide,* SG24-5252
- *Intelligent Miner For Data: Enhance Your Business Intelligence,* SG24-5422
- *Getting Started with Data Warehouse and Business Intelligence,* SG24-5415
- *Mining Relational and NonRelational Data with IBM Intelligent Miner For Data Using Oracle, SPSS, and SAS As Sample Data Sources,* SG24-5278
- *Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data*, SG24-6271

Other resources

These medical research publications are relevant to understanding the medical research sector:

- *Intelligent Data Analysis in Medicine and Pharmacology.* Nada Lavrac, et al. Kluwer Academic Publishing, June 1997. ISBN: 0792380002
- *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* Andreas D. Baxevanis and B.F. Francis Ouellette. Wiley-Liss, April 2001. ISBN: 0471383910

These external publications are also relevant as further information sources on data mining:

- *Data Preparation For Data Mining.* Dorian Pyle. Morgan Kaufmann Publishers, March 1999. ISBN: 1558605290
- *Data Mining Your Website*. Jesus Mena. Digital Press, July 1999. ISBN: 1555582222
- *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation.* Ian H. Witten and Eibe Frank. Morgan Kaufmann Publishers, October 1999. ISBN: 1558605525

- *Mastering Data Mining: The Art and Science of Customer Relationship Management.* Michael J. A. Berry and Gordon Linoff. John Wiley & Sons, December 1999. ISBN: 0471331236
- *Data Mining Techniques: For Marketing, Sales, and Customer Support.* Michael J. A. Berry and Gordon Linoff. John Wiley & Sons, May 1997. ISBN: 0471179809
- *Advances in Knowledge Discovery and Data Mining.* Usama M. Fayyad, et al. MIT Press, March 1996. ISBN: 0262560976

These IBM publications are also relevant when using IBM DB2 Intelligent Miner for Data:

- *Using the Intelligent Miner for Data V6.1,* SH12-6394
- *Intelligent Miner for Data V6.1 Using the Associations Visualizer,* SH12-6396
- *Intelligent Miner Scoring, Administration, and Programming for DB2,* SH12-6719

Referenced Web sites

These Web sites are also relevant as further information sources:

- `http://www.medicinenet.com/`
  MedicineNet.com Web site
- `http://www.rbrs.org/journal/volume80/page301.html`
  Budd-Chiari syndrome (hepatic veins, thrombosis) Web site
- `http://www.ibm.com/software/`
  IBM Software home page
- `http://www.ibm.com/software/data/iminer/fordata/`
  IBM DB2 Intelligent Miner For Data Web site
- `http://www.ibm.com/software/data/`
  IBM Database and Data Management home page
- `http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/maps.cgi?org=hum&chr=22`
  NCBI: Homo sapiens map view of genes on sequence
- `http://www.dmg.org`
  The Data Mining Group Web site

How to get IBM Redbooks

Search for additional Redbooks or Redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redbooks become Redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Special notices

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere.,TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

Other company, product, and service names may be trademarks or service marks of others.

Glossary

A

adaptive connection. A numeric weight used to describe the strength of the connection between two processing units in a neural network. Values typically range from zero to one, or -0.5 to +0.5.

aggregate. To summarize data in a field.

albumin (ALB). Albumin is the main protein in human blood and the key to the regulation of the osmotic pressure of blood.

antinuclear antibodies (ANA). ANA are directed against the structures within the nucleus of the cells. They are found in patients whose immune system can be predisposed to cause inflammation against their own body tissues. Anti nucleus antibodies are unusual antibodies that are directed against the structures within the nucleus of the cells. The nucleus is the innermost core within each of the body's cells and it contains the genetic material.

application programming interface (API). A functional interface supplied by the operating system or a separate licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.

architecture. The number of processing units in the input, output, and hidden layer of a neural network. The number of units in the input and output layers is calculated from the mining data and input parameters. An intelligent data mining agent calculates the number of hidden layers and the number of processing units in those hidden layers.

association time series (ATSeries). ATSeries stands for Association Time Series and is a tool that was programmed as an extension to the IM for Data for the medical scenario. It acts as a bridge between Associations Discovery and Similar Sequences. ATSeries is not part of the product itself, but it expands the functionality of the IM for Data through several functionalities.

associations. The relationship of items in a transaction in such a way that items imply the presence of other items in the same transaction.

attributes. Or variable or field.Characteristics or properties that can be controlled, usually to obtain a required appearance. For example, color is an attribute of a line. In object-oriented programming, a data element defined within a class

auto-antibodies. Antibodies that are directed against one's own tissues are referred to as auto-antibodies.

autoimmunity. The propensity for the immune system to work against its own body is referred to as autoimmunity.

B

back-propagation. A general-purpose neural network named for the method used to adjust weights while learning data patterns. The classification -neural function uses such a network.

bilirubin (BIL). Bilirubin is a yellow-orange compound produced by the breakdown of hemoglobin from red blood cells.

bioinformatics. Bioinformatics is the analysis of biological information using computers and statistical techniques, the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. It is used in analyzing genomes, proteomes (protein sequences), three-dimensional modeling of biomolecules and biologic systems, and so on.

blood pressure (BP). Blood pressure indicates the diastolic blood pressure.

body mass Index. The Body mass index is defined as weight [in kg] divided by squared height [in m]

bucket. One of the bars in a bar chart showing the frequency of a specific value.

C

categorical values. Nonnumeric data represented by character strings: for example colors.

causes for diabetes mellitus. Diabetes mellitus is caused e.g. by an insufficient production of insulin.
The early symptoms of untreated diabetes mellitus are related to elevated blood sugar levels, and loss of glucose in the urine. High amounts of glucose in the urine can cause increased urine output and lead to dehydration. Dehydration causes an increased thirst and a consumption of water.
The inability to utilize glucose energy eventually leads to weight loss despite an increase in appetite. Some untreated diabetes patients also complain of fatigue, nausea, and vomiting.
Patients with diabetes are prone to developing infections of the bladder, skin, and vaginal areas. Fluctuations in blood glucose levels can lead to blurred vision. Extremely elevated glucose levels can lead to lethargy and coma (diabetic coma).

characteristics of deep vein thrombosis. Deep vein thrombosis causes a pulmonary embolism, the patient may develop a rapid heart rate, shortness of breath, sharp chest pain that worsens with deep breathing, or cough up blood. If the pulmonary emboli are large and block one or both of the major pulmonary arteries sending blood to the lungs, the patient may develop a very low blood pressure, pass out, and possibly die from lung or heart failure. As is the case with deep vein thrombosis, however, many other conditions, for example, a heart attack or pneumonia, can mimic a pulmonary embolism.

chi-square test. A test to check whether 2 variables are statistically dependent or not. Chi-square is calculated by subtracting the expected frequencies (imaginary values) from the observed frequencies (actual values). The expected frequencies represent the values that were be expected if the variable question was statistically independent.

classification. Assignment of objects into groups based on their characteristics.

cluster. A group of records with similar characteristics.

clustering. A mining function that creates groups of data records within the input data on the basis of similar characteristics. Each group is called a cluster.Assignment of objects into groups based on their characteristics.

confidence, confidence factor. Indicates the strength or the reliability of the associations detected.

continuous, continuous field. A field that can have any floating point number as its value.

Coumadin. Coumadin is a known teratogen, an agent that can disturb the development of the embryo and fetus and lead to birth defects. Coumadin - taken by a woman during pregnancy -

can cause bleeding into the baby's brain, underdevelopment of the baby's nose and stippling of the ends of the baby's long bones. Each can markedly reduce the incidence of post operative deep vein thrombosis in patients undergoing orthopedic surgery.

creatinine (CRE). Creatinine is a chemical waste molecule that is generated from muscle metabolism. It is produced from creatine, a molecule of major importance for energy production in muscles. Creatinine is transported through the bloodstream to the kidneys. The kidneys filter out most of the creatinine and dispose of it in the urine.

D

database view. An alternative representation of data from one or more database tables. A view can include all or some of the columns contained in the database table or tables on which it is defined.

data format. There are different kinds of data formats, for example, database tables, database views, pipes, or flat files.

datamart. A datamart is a set of a data that can either be extracted from a data warehouse or computed by the mining analyst.

data type. There are different kinds of IM For Data data types, for example, discrete numeric, discrete nonnumeric, binary, or continuous.

deep vein system. The deep system is comprised of veins within the muscles of the body. They are connected by small communicating veins where the body regulates the amount of blood going through both systems as a way of rigidly controlling the body's central temperature.

deep vein thrombosis (DVT). A deep vein thrombosis is a condition where a blood clot forms in a vein of the deep system.

diabetes. Diabetes occurs mainly as one of the following diseases:
- Diabetes Insipidius
- Diabetes Mellitus.

diabetes mellitus. Commonly referred to as 'Diabetes' is Diabetes mellitus. It is a chronic medical condition that is associated with abnormally high levels of glucose in the blood. Elevated levels of blood glucose concentration lead to spillage of glucose into the urine. Normally, blood glucose levels are tightly controlled by insulin - a hormone produced by the pancreas. Insulin lowers the blood glucose level, and - when the blood glucose elevates - insulin is released from the pancreas to normalize the glucose level. For patients with diabetes mellitus, the absence or insufficient production of insulin causes then hyperglycemia.
Diabetes mellitus is a chronic medical disease that can last a lifetime. Over time, diabetes mellitus can lead to blindness, kidney failure, and nerve damage. It is also an important factor in accelerating the hardening and narrowing of the arteries, leading to strokes, coronary heart diseases, and other blood vessel diseases in the body.

diabetes insipidus. Diabetes insipidus is an endocrine disorder that involves a deficient production or lack of effective action of an antidiuretic hormone (that is Vasopressin). An antidiuretic hormone (ADH) is made in the hypothalamus, stored in and secreted by the pituitary gland - a small gland that is located below the hypothalamus - and works on the kidney to conserve fluid.
A deficient production of ADH or a lack of effective action of ADH causes a large amount of urine output; it increases thirst, dehydration, and low

blood pressure in advanced cases. The average urine volume for a normal adult is about 1,5 liters per day, for patients with diabetes insipidus, but can increase to 18 liters daily.

diagnosing deep vein thrombosis. A Deep vein thrombosis is difficult to diagnose without specific tests in which the deep vein system can be examined. Furthermore, many patients with deep vein thrombosis have no symptoms at all unless the clot dislodges, travels to the lung, and causes a pulmonary embolism.

diagnosis Related Groups (DRG). DRGs represent an international coding system. They contain a main diagnosis and one or several subdiagnoses. Depending on what the physician indicates as main and subdiagnoses he will get more or less money reimbursed for his medical service.

discrete. Pertaining to data that consists of distinct elements such as character or to physical quantities having a finite number of distinctly recognizable values.

discretization. The act of transforming a set of continuous values in a set of discrete values.

DNA. DNA indicates Desoxyribonucleic acid. DNA consists of four bases: Adenin and Thymin, Guanin and Cytosin.

F

FASTA. The FASTA format is used in a variety of molecular biology software suites. It defines a string that describes a DNA sequence, for example: >eIF4E [org=Drosophila melanogaster] [strain=Oregon R] where:

- '>' indicates the beginning of a new file.
- eIF4% represents an identifier, is followed by the DNA sequence in lowercase or uppercase letters, usually with 60 characters per line.
- [org = Drosophila melanogaster] an additional comment.

field. Or variable or attribute. A set of one or more related data items grouped for processing. In this document, with regard to database tables and views, field is synonymous with column in a database table.

G

GOT. GOT indicates Glutamin Oxaloacetic Transaminase.

GPT. GPT indicates Glutamin Pylvic Transaminase.

graduated stockings. Several studies have shown that *graduated compression stockings* can decrease the incidence of DVT in patients who are confined to bed because of medical conditions or surgical procedures. The stockings work by reducing the amount of blood and increasing the flow of blood in the veins of the legs.

H

hematocrit (HCR). Hematocrit is the proportion, by volume, of the blood that consists of red blood cells.

hemoglobin (HGB). Hemoglobin is a pigment in the red blood cells. It forms an unstable, reversible bond with oxygen. In its oxygenated state it is called oxyhemoglobin and is bright red. In the reduced state it is called deoxy-hemoglobin and is purple-blue.

heparin. Heparin is an anticoagulant (anti-clotting) medication and useful in preventing thromboembolic complications (clots that travel from their site of origin through the blood stream to clog up another vessel). Heparin is also used in the early treatment of blood clots in the lungs (pulmonary embolisms).

I

immunoglobin (IG). Immunoglobin is a protein that is induced by plasma cells and lymphocytes. Immunoglobulins are an essential part of the body's immune system which attach to foreign substances, for example bacteria, and assist in destroying them. Some classes of immunoglobulins are for example A, M, G.

International Classification of Diseases (ICD) Version 10. ICD refer to the worldwide standard classification system for diagnoses that was generated by the World Health Organization (WHO). The complete catalog can be downloaded at www.who.org.
Examples are:

- A00B99 Infections
- A15A18 Tuberculosis
- A17 Tuberculosis/nervous system
- A17.0 Tuberculosis Leptomeningitis
- B24 HIV disease
- B25.9 Angiitis, an uncommon inflammation of the blood vessels
- H00.0 Abscess
- H02.5 Ankyloblepharon
- H33.2 Ablatio retinae
- H47.3 Disease of the pupil
- H51.9 Defect in the movement of the eyes
- H52.0 Hypermetropy of the axes
- H52.2 Astigmatism
- H52.4 Weakness of the eyes
- H53.4 Anopsy
- J06.9 Accute infection
- K52.9 Chronic diarrhoe
- L71.8 Blepharitis
- L71.9 Acne rosacea
- M35.0 Atropic dakryosialoadenopathie
- R0.7Pains in breast and heart areas
- Z04 Special Examination
- Z11 Screening

K

Kohonen Feature Map. A neural network model comprised of processing units arranged in an input layer and output layer. All processors in the input layer are connected to each processor in the output layer by an adaptive connection. The learning algorithm used involves competition between units for each input pattern and the declaration of a winning unit. Used in neural clustering to partition data into similar record groups.

L

large item sets. The total volume of items above the specified support factor returned by the Associations mining function.

lactate dehydrogenase (LDH). Lactate dehydrogenase is an enzyme that catalyzes the conversion of lactate to pyruvate. This is an important step in energy production in cells. Many different types of cells in the body contain this enzyme. Some of the organs relatively rich in Lactate dehydrogenase are heart, kidney, liver, and muscle.

learning algorithm. The set of well-defined rules used during the training process to adjust the connection weights of a neural network. The criteria and methods used to adjust the weights define the different learning algorithms.

M

metadata. Data that describes data objects.

mining. Synonym for discovery-driven analyzing or searching patterns in data.

model. A specific type of neural network and its associated learning algorithm. Examples include the Kohonen Feature Map and back propagation.

N

neural network. A collection of processing units and adaptive connections that is designed to perform a specific processing function.

nucleus. The nucleus is the inner core within each of the body's cells; it contains the genetic material.

nucleotides. Nucleotides are a subunit of DNA consisting of a nitrogenous base, a phosphate molecule, and a sugar molecule (deoxyribose in DNA). Thousands of nucleotides are linked to form a *DNA* molecule.

P

plasma glucose concentration. Plasma Glucose concentration in the blood by a 2 hours oral glucose tolerance test.

platelets (PLT). Platelets are the smallest cell-like structures in the blood and are important for blood clotting and plugging damaged blood vessels.

pneumatic compression of the legs. The pneumatic compression of the legs is a therapy against deep vein thrombosis, which is done by applying a plastic stocking to the leg and thigh. The stocking intermittently fills with air and squeezes the leg. This compression stimulates the body to produce factors that help dissolve small blood clots before they can progress to deep vein thrombosis. Medications are for example:

- unfractionated heparin
- low-molecular-weight heparin

prediction. The dependency and the variation of one field's value within a record on the other fields within the same record. A profile is then generated that can predict a value for the particular field in a new record of the same form, based on its other field values.

R

radial basis function (RBF). In data mining functions, radial basis functions are used to predict values. They represent functions of the distance or the radius from a particular point. They are used to build up approximations to more complicated functions.

red blood cells (RBC). Red blood cells are the cells that carry oxygen and carbon dioxide through the blood.

record. A set of one or more related data items grouped for processing. In reference to a database table, record is synonymous with row.

region. (Sub)set of records with similar characteristics in their active fields. Regions are used to visualize a prediction result.

rule. A clause in the form head <== body. It specifies that the head is true if the body is true.

rule body. Represents the specified input data for a mining function.

rule group. Covers all rules containing the same items in different variations.

rule head. Represents the derived items detected by the Associations mining function.

S

scaling. To adjust the representation of a quantity by a factor in order to bring its range within prescribed limits.

self-organizing feature map (SOM). See *Kohonen Feature Map*.

sensitivity analysis report. An output from the Classification - Neural mining function that shows which input fields are relevant to the classification decision.

sequence tagged site (STS). Sequence tagged site (STS) are short (200 to 500 base pairs) DNA sequence that has a single occurrence in the human genome.

sequencing. A problem in bioinformatics is the determination of the order of the nucleotides in a DNA molecule or the order of amino acids in a protein. This is referenced as sequencing.

sequential patterns. Inter transaction patterns such that the presence of one set of items is followed by another set of items in a database of transactions over a period of time.

similar (time) sequences. Occurrences of similar sequences in a database of time sequences.

structured query language (SQL). An established set of statements used to manage information stored in a database. By using these statements, users can add, delete, or update information in a table, request information through a query, and display results in a report.

support, support factor. Indicates the occurrence of the detected association rules and sequential patterns based on the input data.

superficial vein system. The superficial vein system is made up of veins that are close to the skin. These are the blood vessels that can be seen on hands or on arms.

symptoms of deep vein thrombosis. The most common symptoms of deep vein thrombosis are in the leg: they are swelling and affect pain in the leg. These symptoms are caused by the accumulation of blood that is unable to get past the clot in the vein and the resulting leakage of fluid from the blood into the muscle. Many other conditions exhibit symptoms similar to those of a deep vein thrombosis, for example, muscle strains, skin infections, and inflammation of superficial veins.

T

taxonomy. Represents a hierarchy or a lattice of associations between the item categories of an item. These associations are called taxonomy relations.

taxonomy relation. The hierarchical associations between the item categories you defined for an item. A taxonomy relation consists of a child item category and a parent item category.

tests to diagnose diabetes mellitus. In concern of prophylactic examination, there exist a fast plasma glucose test. It is a preferred test to diagnose diabetes since it is easy to perform and convenient. After the patient has fasted overnight (at least 8 hours), a single sample of blood is drawn and sent to the laboratory for analysis. Normal fasting plasma glucose levels are less than 110 mg/dl, fasting plasma glucose levels of more than 126 mg/dl on two or more tests on different days indicate diabetes. If the overnight fasting blood glucose is greater than 126 mg/dl on two different tests on different days, the diagnosis of diabetes mellitus is made. A random blood glucose test can

also be used to diagnose diabetes. Random blood samples (if taken shortly after eating or drinking) may be used to test for diabetes when symptoms are present. A blood glucose level of 200 mg/dl or higher indicates diabetes, but it must be reconfirmed on another day with a fasting plasma glucose or an oral glucose tolerance test.

triceps skin. Triceps skin fold thickness [in mm].

tuberculosis (TB). TB is a communicable disease caused by a bacterium named mycobacrium tuberculosis. It is spread from person to person through the inhalation of airborne particles containing the bacterium. These particles, also called droplet nuclei, are produced when a person with infectious TB of the lung exhales, such as when coughing, sneezing, laughing, speaking, or singing. These infectious particles can remain suspended in the air and inhaled by someone sharing the same air. TB is transmitted in closed areas where ventilation is poor. The risk of transmission increases when susceptible persons share air for prolonged periods with a person who has untreated pulmonary TB. If the body's immune system can not contain the TB bacteria it will continue in producing more and more bacteria. Normally, the infection occurs in the top portion of the lungs, and it may take several months that the symptoms will affect. Usually, there occurs a general tiredness or weakness, loss of weight, fever and nightly sweats. If the infection in the lung worsens, then further symptoms can include coughing, chest pain, and shortness of breath.

trained network. A neural network containing connection weights that have been adjusted by a learning algorithm. A trained network can be considered a virtual processor: it transforms inputs to outputs.

transaction. A set of items or events that are linked by a common key value, for example, the articles (items) bought by a customer (customer number) on a particular date (transaction identifier). In this example, the customer number represents the key value.

transaction id. The identifier for a transaction, for example, the date of a transaction.

U

UA. UA indicates Urin Acid.

UN. UN indicates Urin Nitrogen.

V

vector. A quantity usually characterized by an ordered set of numbers.

venography to identify deep vein thrombosis. Venography is the oldest of tests that can be used to identify deep vein thrombosis: this test is performed by injecting a radiopaque fluid into a vein on the top of the foot. The dye flows with the blood and fills the veins of the leg as well as the thigh. An obstructing blood clot in one of these veins can be seen on an X-ray as a dye-free area within the vein. Venography is the most accurate test to identify deep vein thrombosis, but it is painful, expensive, and occasionally can cause painful inflammation of the veins. Furthermore, venography requires a high degree of expertise to perform and interpret correctly.

Vein. A vein is a blood vessel that returns blood from the tissues of the body back to the heart. The body has two distinct systems of veins: a superficial system and a deep system.

W

WBC. WBC indicates the White Blood Cells of the human body.

weight. The numeric value of an adaptive connection representing the strength of the connection between two processing units in a neural network.

where is ANA found? ANA is found in patients whose immune system can be predisposed to cause inflammation against their own body tissues. Antibodies that are directed against one's own tissues are referred to as auto antibodies.

Index

Q

R

S

T

U

V

W

Mining Your Own Business in Healthcare Using DB2 Intelligent Miner for Data

(0.2”spine)
0.17”<->0.473”
90<->249 pages

Mining Your Own Business in Health Care

Using DB2 Intelligent Miner for Data

Redbooks

Exploring the health care business issues

Addressing the issues through mining algorithms

Interpreting and deploying the results

The new challenge of integrated solutions is to get more knowledge from data in order to build the most valuable solutions. This IBM Redbook is a solution guide to address the business issues in health care by real usage experience and to position the value of DB2 Intelligent Miner for Data in a Business Intelligence architecture as an integrated solution.

Typical health care issues are addressed in this redbook, such as: How to calculate weight for diagnoses related groups? Are there characteristic groups of patients in my population? Can we optimize medical tests for a specific disease? Can we detect precauses for a special medical condition?

This book also describes a data mining method to ensure that the optimum results are obtained. It details for each business issue:

- What common data model to use
- How to source the data
- How to evaluate the model
- What data mining technique to use
- How to interpret the results
- How to deploy the model

Business users who want to know the payback on their organization when using the DB2 Intelligent Miner for Data solution should read the sections about the business issues, how to interpret the results, and how to deploy the model in the enterprise.

Implementers who want to start using mining techniques should read the sections about how to define the common data model to use, how to source the data, and how to choose the data mining techniques.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6274-00 ISBN 0738423025