DATA AND TEXT MINING TECHNIQUES ON REAL MEDICAL DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY
BURCU DEMİRDÜMEN

IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2008

Title of the Thesis**: Data and Text Mining Techniques on Real Medical Data**

Submitted by **Burcu DEMİRDÜMEN**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University

Prof.Dr. Özhan ULUATAM
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master

M. R. Tolun
Prof. Dr. Mehmet Reşit TOLUN
Head Of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master

Ziya Aktaş
Prof. Dr. Ziya AKTAŞ
Supervisor

Examination Date : 10.09.2008

Examining Committee Members :

Prof. Dr. Ziya AKTAŞ	(Çankaya University)	Ziya Aktaş
Prof. Dr. Mehmet Reşit TOLUN	(Çankaya University)	M. R. Tolun
Doç. Dr. Kemal ARDA	(T. Yüksek İhtisas Hast.)	

STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Burcu DEMİRDÜMEN

Signature :

Date : 10.09.2008

ABSTRACT

DATA AND TEXT MINING TECHNIQUES ON REAL MEDICAL DATA

DEMİRDÜMEN, Burcu

M.Sc., Department of Computer Engineering
Supervisor : Prof. Dr. Ziya AKTAŞ
September 2008, 118 pages

Clinical databases have been accumulating large quantities of data about patients and their medical conditions. Relationships and patterns within that data could provide new medical information and knowledge. Yet, unfortunately, few methodologies have already been developed and applied to discover this hidden knowledge. In this study, the techniques of data mining (also known as Knowledge Discovery in Databases) were used to search for relationships in a clinical database. The study describes the processes involved in mining a clinical database; it also includes the processes of data warehousing, data query and cleaning, and data analysis. The study is extended to demonstrate text mining too.

Keywords: Classification, Decision Tree, Data Mining, Naive Bayes, Pulmonary Embolism, Support Vector Machine, Text Mining.

ÖZ

GERÇEK TIBBİ VERİLER ÜZERİNDE VERİ VE DÖKÜMAN MADENCİLİĞİ TEKNİKLERİNİN UYGULANMASI

DEMİRDÜMEN, Burcu

Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı
Tez Yöneticisi : Prof. Dr. Ziya AKTAŞ
Eylül 2008, 118 sayfa

Klinik veritabanları hastalar ve hastaların sağlık durumlarıyla ilgili olarak çok büyük miktarlarda veri içermektedir. Bu veriler içerisindeki ilişkiler ve benzerlikler yeni ve bilinmeyen medikal bilgi sağlayabilmektedir. Ancak, böylesi gizli bilginin keşfini sağlamak için ne yazık ki henüz çok az yöntem geliştirilmiş ve uygulanmıştır. Bu tez çalışmasında, klinik veritabanlarındaki veri ilişkilerinin araştırılması amacıyla veri madenciliği teknikleri (diğer bir deyişle veritabanlarındaki bilgi keşfi) kullanılmıştır. Bu çalışmada klinik veritabanlarındaki veri madenciliği yanında veri ambarları, veri sorguları, temizleme ve veri analizi süreçleri uygulamaları da gösterilmiştir. Tezde metin madenciliği konusuna da kısaca giriş yapılmıştır.

Anahtar Kelimeler : Destek Vektör Makinaları, Karar Ağaçları, Metin Madenciliği, Navie Bayes Yöntemi, Pulmoner Emboli, Sınıflandırma, Veri Madenciliği.

ACKNOWLEDGEMENT

I would like to express my special thanks to my supervisor Prof. Dr. Ziya AKTAŞ for his supports and suggestions.

Thanks are also extended to Doç. Dr. Kemal ARDA and his colleagues for providing me real medical data and for his guidance in evaluating the results.

Mr. Nuri Cömert and Mr. Sedat Zincirci of ORACLE also deserve special thanks for providing us the needed special ORACLE software for data and text mining.

I would like to express my deepest gratitude to Uğraş, who encouraged and supported me in this demanding study with his never ending patience.

Finally, my deepest thanks are to my parents who supported and motivated me with their never ending patience, tolerance and understanding throughout the study.

TABLE OF CONTENTS

CHAPTERS :

APPENDICES :

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABN	**:**	Adaptive Bayesian Network
AI	**:**	Artificial Intelligence
Ai	**:**	Attribute Importance
HIS	**:**	Hospital Information System
LIS	**:**	Laboratory Information System
MDL	**:**	Minimum Description Length
ODBC	**:**	Open Database Connectivity
ODM	**:**	Oracle Data Mining
OLAP	**:**	Online Analytical Processing
PL/SQL	**:**	Procedural Language/Structured Query Language
RDBMS	**:**	Relational Database Management System
RIS	**:**	Radiology Information System
ROC	**:**	Receiver Operating Characteristic
SQL	**:**	Structured Query Language
SVM	**:**	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1. Statement of the Problem

Information and knowledge are valuable assets and they have been becoming increasingly important to organizations. Society is currently living in an electronic age. With the large amounts of data processing capabilities comes a massive amount of data. Organizations are trying to find ways to strategically use and effectively manage this data.

The collected data and information in the medical area have been increasing rapidly more than ever in the last decades. Since the medical data and information have characteristics of redundancy, multi-attribution, incompletion and closely related with time, medical data mining differs from other applications. As a result of this, it becomes harder to get knowledge extraction within simple database queries or statistical methods using medical data.

Modern medicine generates a great deal of data and information stored in the medical databases. Extracting useful knowledge and providing support for scientific decision-making for the diagnosis and treatment of disease from the databases increasingly becomes necessary. The process of extracting such valuable knowledge from a database is known as Data Mining.

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as some statistical and mathematical techniques. This definition shows that data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis. In this study a database perspective is adopted in the presentation of data mining.

Databases are rich with hidden information and knowledge that can be used for making intelligent business decisions. Data mining involves using different algorithms to extract models describing important data classes or to predict future data trends. Basically, the algorithms try to fit a model closest to the characteristics of data under consideration. Models can be predictive or descriptive. Predictive models are used to make predictions, for example, to make a diagnosis of a particular disease. A patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. Descriptive models are used to identify patterns in data. Classification, regression, and time series analysis are some of the tasks of predictive modeling, whereas clustering, association rules, and visualization are some are the tasks of descriptive modeling.

The purpose of the study is extracting useful knowledge needed in scientific decision-making for the diagnosis of diseases by applying classification algorithms of descriptive models on the patient dataset.

1.2. Previous Work and Objective of the Study

During the literature survey stage of this study, a number of references about medical data mining and knowledge discovery on patient’s dataset have been studied. Some of these references are summarized in the following sections.

KARANIKAS and MAVROUDAKIS [2005] stated that we are living in the “Information Age” and the main characteristic is the amazing growth of data that are being generated and stored making it difficult for humans to understand. In the healthcare sector cost pressure is growing, quality demands are raising and the competitive situation amongst suppliers is mounting. These developments confront hospitals more than ever with the necessities of critically reviewing their own efficiency under both medical and economical aspects. At the same time growing capture of medical data and integration of distributed and heterogeneous databases create a completely new base for medical quality.

STÜHLİNGER [2000] applied intelligent data mining methods to patient data from several clinics and from years 1996 to 1998. They have shown that intelligent data mining in addition to conventional analyses and statistical studies in patient data can deliver further evidence for medical quality management.

IBM company researchers, URAMATO et al. [2004], developed a text-mining system called MedTAKMI for knowledge discovery from biomedical documents. This system is able to parse 11 million MEDLINE citations with a syntactic shallow parser and extract relationships from these parsed entities with hierarchical category identifiers.

Among other references for this study, one may cite COUSEAULT [2004], COULTER [2001], HONIGMAN [2001], MILWARD [2006], PONOMARENKO, [2002], ZHOU [2003].

The objective of this study is predicting useful information about the diagnosis of pulmonary embolism by using supervised (descriptive) data mining techniques and using this knowledge for better clinical decision-making.

The phases of business understanding, data understanding, data preparation, modeling, evaluation and deployment have been prepared by using ORACLE Data Miner tool which holds the process of data mining under a certain discipline.

In data preparation step, a user interface is implemented to take the data which contains information about patient's history, laboratory findings and radiology report details. This data which is taken from the user is saved into the database in XML format. A processor between the database and user interface reads data and save useful information into a new table.

In modeling step, data mining classification algorithms such as Naive Bayes, Support Vector Machine modeling techniques have been compared in order to find the best technique for Pulmonary Embolism data.

In evaluation and deployment step, pulmonary embolism classification model is applied to new data and results are discussed.

1.3. Organization of the Thesis

After introduction in Chapter 1, Chapter 2 provides background for Data Mining. The steps in the evolution of Data Mining are explained. Definitions of Data Mining and Data Warehouses are given. An introduction about the Supervised and Unsupervised Data Mining techniques is given. At the end of the chapter examples of business applications in various sectors and industries that can most benefit from data mining are explained and some Commercial Data Mining Products are covered to understand what software tools can be found on the market and what their features are.

Chapter 3 presents information about the steps of Data Mining Process.

Chapter 4 provides information about the Text Mining and steps of the Text Mining Process.

Chapter 5 presents detailed information about the Algorithms for Classification which are explained shortly in Chapter 2.

Chapter 6 presents information about ORACLE Data Mining. Then explains how Oracle Data Mining supports Data Mining and Text Mining.

Chapter 7 gives brief introduction about the definition, symptoms, test and diagnosis of Pulmonary Embolism. Then the medical data which is used in the case study is explained.

Chapter 8 provides the case study on the Pulmonary Embolism dataset. The application of the data mining process with the help of ORACLE Data Mining is explained step by step. At the end of the chapter results of the applied classification algorithms are compared.

Chapter 9 concludes the study by providing a summary of study and results. Limitations and further improvements of the study are also presented in this chapter.

CHAPTER 2

DATA MINING

2.1. What Motivated Data Mining?

The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [HAN and KAMBER, 2000].

Although data mining is 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 (AI) 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 realised that a number of the techniques that were being used for scientific applications could be applied also 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. Table 1 shows the developments that have taken place over the past 40 years.[IBM, 2001]

Table 2.1. Steps in the Evolution of Data Mining

Evolutionary Step	**Business Question**	**Enabling Technologies**	**Characteristics**
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Prospective, proactive information delivery

2.2. What is Data Mining?

There are a number of definitions for data mining. Becerra-Fernandez, I. et. al [2004] defines as "Discovering new knowledge: Data Mining" and Awad and Ghaziri [2004] states that "Data Mining: Knowing the unknown". According to Buruncuk [2006] data mining is the process of extracting hidden information such as data attributes, trends or patterns from large databases by analyzing data from different perspectives and summarizing it into useful information. The extraction process is achieved usually by finding correlations or patterns among dozens of fields of large databases which are usually constructed as data warehouses. Data mining gains the attention of people as a result of the accumulation of large amounts of data in the databases and the increasing need to analyze and then convert them into meaningful information. In the evolution from business data to business information, each new step has been built upon the previous one. For example, dynamic data access is critical for querying the necessary information and the ability to store large databases is critical to data mining.

Data mining uses the historical accumulated data as a guide, when effective decisions are needed to predict the future. This is achieved by offering a rich capability for modeling historical data and then using this model to predict likely future outcomes. This ability to give advance information about the future is unique to data mining and makes business professionals have a new perspective of factors, which truly contribute to business success or failure.

2.3. Data Warehouses

Data warehouses are closely related to data mining. It is, therefore, proper to have a quick look at the data warehouses.A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart which are analytical data stores designed to focus on specific business functions for a specific community within an organization. Data marts are often derived from subsets of data in a data warehouse, though in the bottom-up data warehouse design methodology the data warehouse is created from the union of organizational data marts. Figure 2.1 shows geographic, analysis, data mining data marts which are derived from a data warehouse [TWO CROWS, 2002].

In order to facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity. The data are stored to provide information from a historical perspective (such as from the past 5-10 years), and are typically summarized. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store, or, summarized to a higher level, for each sales region.

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data [HAN and KAMBER, 2000].

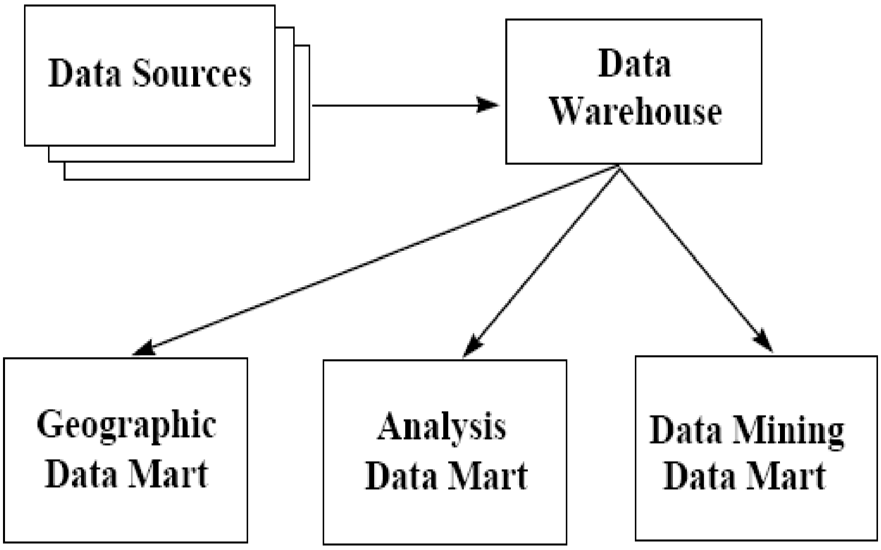


Figure 2.1. Data Mining Data Mart Extracted from a Data Warehouse

A data warehouse is not a necessary requirement for data mining. Yet setting up a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a query database can be an enormous task, sometimes taking years and costing millions of dollars. However, mine data from one or more operational or transactional databases by simply extracting it into a read-only database (Figure 2.2). This new database functions as a type of data mart [TWO CROWS, 2002].

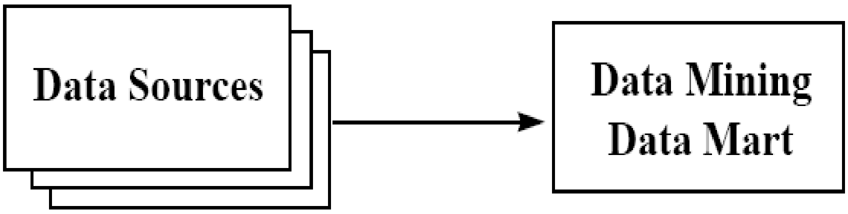


Figure 2.2. Data mining Data Mart Extracted from Operational Databases

2.4. Data Mining and Statistics

Data mining is about discovering new things about business from the data have been already collected. One may think that he already can do these using standard statistical techniques to explore his/her database. In reality what he is normally doing is making a hypothesis about the business issue that he is addressing and then attempting to prove or disprove his/her hypothesis by looking for data to support or contradict the hypothesis. For example, suppose that as a retailer, one believes that customers from “out of town” visit his/her larger inner city stores less often than other customers, but when they do so they make larger purchases. In order to check the validity of such an hypothesis he can simply formulate a database query looking, for example, at his/her branches, their locations, sales figures, customers and then compile the necessary information (average spend per visit for each customer) to prove his/her 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 his/her other customers. In this case, his/her 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 one does not know what patterns of customer behaviors exist. In this case he may simply ask the question, what are their relationships between what my customers spend and where they come from? In this case, he would leave it up to the data mining algorithm to tell him about all of the different types of customers that one had. This should include the out-of-town, weekend shopper. Data mining therefore provides answers, without one having to ask specific questions.

The difference between the statistical and data mining approaches are summarized in Figure 2.3 [IBM, 2001].

In conclusion there are reasons for believing that data mining is nothing new from a statistical viewpoint. But there are also reasons to support the idea that, because of their nature, statistical methods should be able to study and form formalize the methods used in data mining. This means that on one hand we need to look at the problems posed by data mining from a viewpoint of statistics and utility, while on the other hand we need to develop a conceptual paradigm that allows the statisticians to lead the data mining methods back to a scheme of general and coherent analysis.

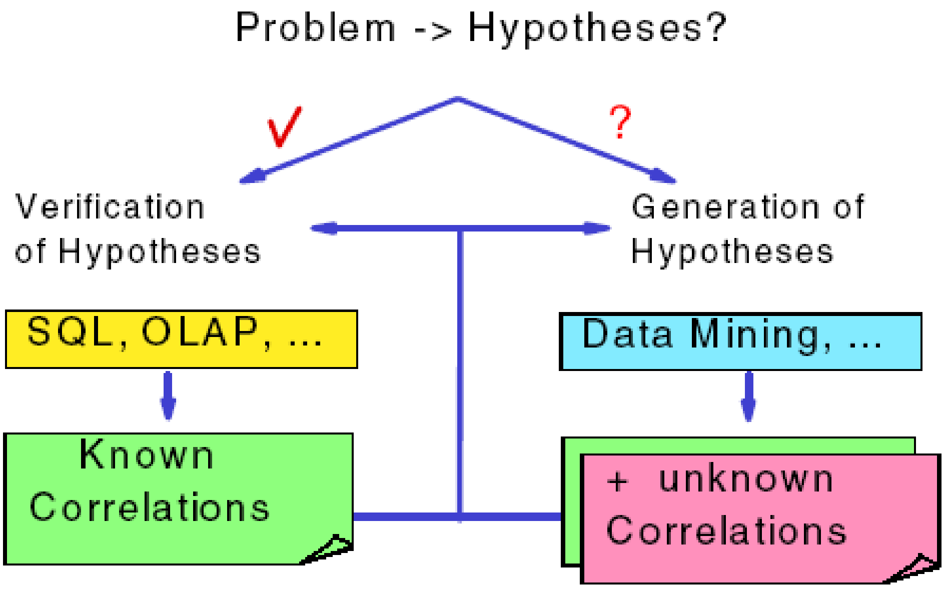


Figure 2.3. Standard and Data Mining Approach on Information Detection

2.5. Data Mining Techniques

There are two types of data mining techniques, namely, supervised data mining and unsupervised data mining.

2.5.1. Supervised Data Mining

Commonly used supervised data mining techniques are classification, regression, attribute importance and anomaly detection and they are summarized below. Some parts are taken from [ORACLE, 2005].

2.5.1.1 Classification

Classification of a data collection consists of dividing the items that make up the collection into categories or classes. In the context of data mining, classification is done using a model that is built on historical data. The goal of predictive classification is to accurately predict the target class for each record in new data, that is, data that is not in the historical data.

A classification task begins with *build data* (also known as *training data*) for which the target values (or class assignments) are known. Different classification algorithms use different techniques for finding relations between the predictor attributes' values and the target attribute's values in the build data. These relations are summarized in a model; the model can then be applied to new cases with unknown target values to predict target values. A classification model can also be applied to data that was held aside from the training data to compare the predictions to the known target values; such data is also known as *test data* or *evaluation data*. The comparison technique is called *testing a model*, which measures the model's predictive accuracy. The application of a classification model to new data is called *applying the model*, and the data is called *apply data* or *scoring data*. Applying a model to data is often called *scoring the data*.

Classification is used in customer segmentation, business modeling, credit analysis, and many other applications. For example, a credit card company may wish to predict which customers are likely to default on their payments. Customers are divided into two classes: those who default and those who do not default. Each customer corresponds to a case; data for each case might consist of a number of attributes that describe the customer's spending habits, income, demographic attributes, etc. These are the predictor attributes. The target attribute indicates whether or not the customer has defaulted. The build data is used to build a model that predicts whether new customers are likely to default.

Classification problems can have either binary or multiclass targets. Binary targets are those that take on only two values, for example, *good credit risk* and *poor credit risk*. Multiclass targets have more than two values, for example, the product purchased (comb or hair brush or hair pin). Multiclass target values are not assumed to exist in an ordered relation to each other, for example, hair brush is not assumed to be greater or less than comb.

2.5.1.2. Regression

Regression models are similar to classification models. The difference between regression and classification is that regression deals with numerical or continuous target attributes, whereas classification deals with discrete or categorical target attributes. In other words, if the target attribute contains continuous (floating-point) values or integer values that have inherent order, a regression technique can be used. If the target attribute contains categorical values, that is, string or integer values where order has no significance, a classification technique is called for. Note that a continuous target can be turned into a discrete target by binning; this turns a regression problem into a problem that can be solved using classification algorithms.

2.5.1.3. Attribute Importance

Attribute Importance **(ai)** provides an automated solution for improving the speed and possibly the accuracy of classification models built on data tables with a large number of attributes. The time required to build ODM classification models increases with the number of attributes. Attribute Importance identifies a proper subset of the attributes that are most relevant to predicting the target. Model building can proceed using the selected attributes only.

Using fewer attributes does not necessarily result in lost predictive accuracy. Using too many attributes (especially those that are "noise") can affect the model and degrade its performance and accuracy. Mining using the smallest number of attributes can save significant computing time and may build better models. The programming interfaces for Attribute Importance permit the user to specify a number or percentage of attributes to use; alternatively the user can specify a cutoff point.

2.5.1.4. Anomaly Detection

Anomaly detection consists of identifying novel or anomalous patterns. Identifying such patterns can be useful in problems of fraud detection (insurance, tax, credit card, etc.) and computer network intrusion detection. An anomaly detection model predicts whether a data point is typical for a given distribution or not. An atypical data point can be either an outlier or an example of a previously unseen class. An anomaly detection model discriminates between the known examples of the positive class and the unknown negative set of counterexamples. An anomaly detection model identifies items that do not fit in the distribution.

2.5.2. Unsupervised Data Mining

Commonly used unsupervised data mining techniques are clustering, association and feature extraction and they are summarized below. Some parts are taken from [ORACLE, 2005].

2.5.2.1. Clustering

Clustering is useful for exploring data. If there are many cases and no obvious natural groupings, clustering data mining algorithms can be used to find natural groupings.

Clustering analysis identifies clusters embedded in the data. A cluster is a collection of data objects that are similar in some sense to one another.

Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models. Clustering models are different from supervised models in that the outcome of the process is not guided by a known result, that is, there is no target attribute. Supervised models predict values for a target attribute, and an error rate between the target and predicted values can be calculated to guide model building. Clustering models, on the other hand, are built using optimization criteria that favor high intra-cluster and low inter-cluster similarity. The model can then be used to assign cluster identifiers to data points.

2.5.2.2. Association

An Association model is often used for market basket analysis, which attempts to discover relationships or correlations in a set of items. Market basket analysis is widely used in data analysis for direct marketing, catalog design, and other business decision-making processes. A typical association rule of this kind asserts that, for example, "70% of the people who buy spaghetti, wine, and sauce also buy garlic bread."

Association models capture the co-occurrence of items or events in large volumes of customer transaction data. Because of progress in bar-code technology, it is now possible for retail organizations to collect and store massive amounts of sales data. Association models were initially defined for such sales data, even though they are applicable in several other applications. Finding association rules is valuable for cross-marketing and mail-order promotions, but there are other applications as well:

catalog design, add-on sales, store layout, customer segmentation, web page personalization, and target marketing.

Traditionally, association models are used to discover business trends by analyzing customer transactions. However, they can also be used effectively to predict Web page accesses for personalization. For example, assume that after mining the Web access log, Company X discovered an association rule "A and B implies C," with 80% confidence, where A, B, and C are Web page accesses. If a user has visited pages A and B, there is an 80% chance that he/she will visit page C in the same session. Page C may or may not have a direct link from A or B. This information can be used to create a dynamic link to page C from pages A or B so that the user can "click-through" to page C directly. This kind of information is particularly valuable for a Web server supporting an e-commerce site to link the different product pages dynamically, based on the customer interaction.

2.5.2.3. Feature Extraction

Feature Extraction creates a set of features based on the original data. A *feature* is a combination of attributes that is of special interest and captures important characteristics of the data. It becomes a new attribute. Typically, there are far fewer features than there are original attributes.

Some applications of feature extraction are latent semantic analysis, data compression, data decomposition and projection, and pattern recognition. Feature extraction can also be used to enhance the speed and effectiveness of supervised learning.

For example, feature extraction can be used to extract the themes of a document collection, where documents are represented by a set of key words and their frequencies. Each theme (feature) is represented by a combination of keywords. The documents in the collection can then be expressed in terms of the discovered themes.

2.6. Data Mining in Practice

Data mining is a broad technology that can potentially benefit any functional areas within a business where there is a major need or opportunity for improved performance and where data is available for analysis that can impact the performance improvement. Table 2 shows examples of business applications in various sectors and industries that can most benefit from data mining [BURUNCUK, 2006].

Table 2.2. Data Mining Business Applications in Various Sectors

Sector / Industry	**Application**
Marketing / Retailing	✓ Market basket analysis ✓ Finding market segments ✓ Identifying loyal customers ✓ Predicting what type customers will respond to mailing ✓ Finding customer purchase behavior patterns ✓ Finding associations among customer characteristics ✓ Determine items for cross selling / up-selling ✓ Detecting seasonal differences in sales patterns ✓ Product placement ✓ Forecasting sales / demand / revenue
Banking / Finance	✓ Predicting customers that are likely to change their credit cards ✓ Identifying loyal customers ✓ Identifying fraudulent behavior ✓ Detecting patterns of fraudulent credit card usage ✓ Credit Scoring ✓ Risk assessment of credit ✓ Determine credit card spending by customer groups ✓ Segmentation of customers ✓ Analysis of customer profitability ✓ Managing portfolios ✓ Forecasting price changes in foreign currency markets ✓ Distribution channel analysis
Telecommunications	✓ Churn analysis
Internet	✓ Text Mining ✓ Web marketing
Manufacturing	✓ Inventory Control ✓ Equipment failure analysis ✓ Resource Management ✓ Process / quality control ✓ Capacity management
Insurance / Healthcare	✓ Identifying fraudulent behavior ✓ Predicting which customers will buy new products ✓ Medical treatment analysis
Transportation	✓ Loading pattern analysis and ✓ Distribution channel analysis

2.7. Commercial Data Mining Products

There are many commercial data mining products currently available, using a variety of technologies to solve classification and estimation problems. Most commonly used ones are CART, Clementine, ORACLE DB Miner, IBM Intelligent Miner, SAS and SPSS. Some of the commercially available data mining products are summarized in APPENDIX A.

CHAPTER 3

THE DATA MINING PROCESS

Referring to ORACLE Data Mining [2005], there are six main steps of data mining (Figure 3.1). Some parts of the process step descriptions are taken from CRISP-DM project's official web page [See http://www.crisp-dm.org].

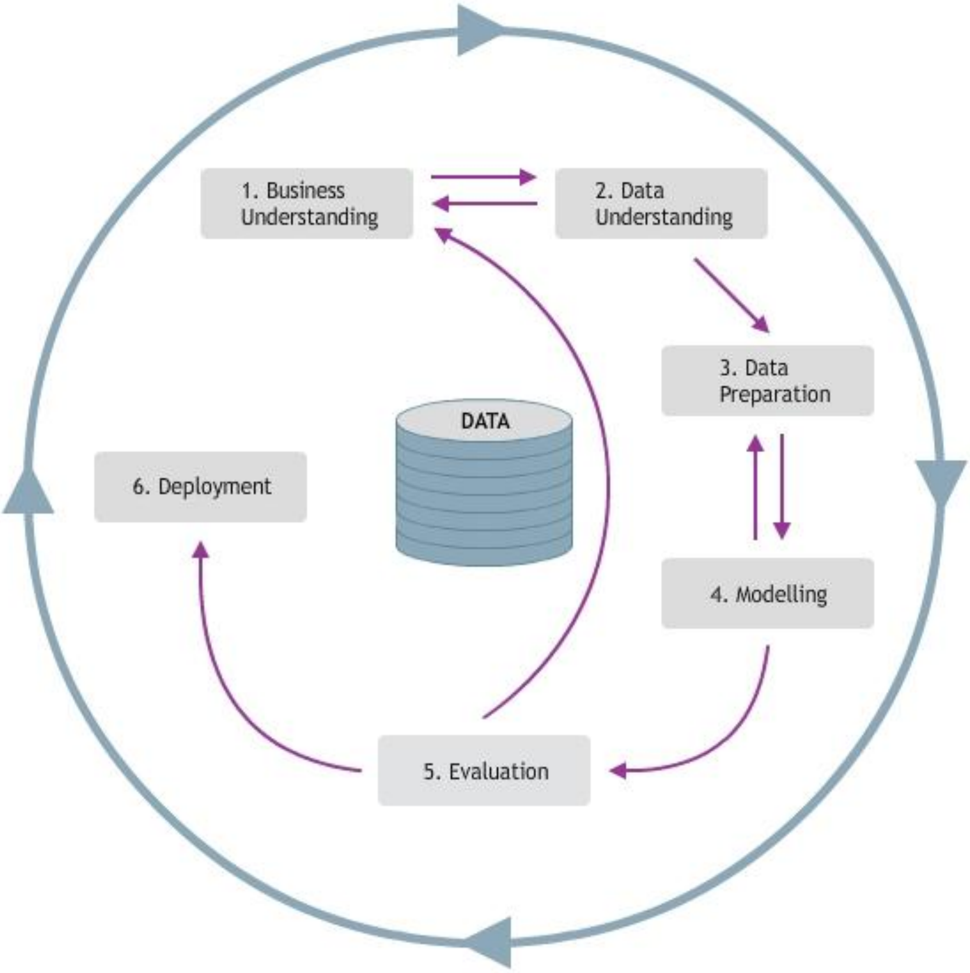


Figure 3.1. Data Mining Processes

3.1. Business Understanding

Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Definition of the objectives involves defining the aims of the analysis. It is not always easy to define the phenomenon we want to analyze. In fact, the company objectives that we are aiming for are usually clear, but the underlying problems can be difficult to translate into detailed objectives that need to be analyzed. A clear statement of the problem and the objectives to be achieved are the prerequisites for setting up the analysis correctly. This is certainly one of the most difficult parts of the process since what is established at this stage determines how the subsequent method is organized. Therefore the objectives must be clear and there must be no room for doubts or uncertainties [GIUDICI, 2003].

3.2. Data Understanding

Start by collecting data, and then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information. Data understanding step involves following sub steps:

A. Collect initial data

Once the objectives of the analysis have been identified, it is necessary to select the data for the analysis. First of all it is necessary to identify the data sources. Usually data is taken from internal sources that are cheaper and more reliable [GIUDICI, 2003]. In general, there are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler). This approach is known as a designed experiment. The second possibility is when the expert cannot influence the data- generation process. This is known as the observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, the sampling distribution is completely unknown after data are collected, or it is partially and implicitly given in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a final application of the results [KANTARDZIC, 2003].

B. Describe data

Describe the data which has been acquired, including the format of the data, the quantity of data, for example number of records and fields in each table, the identities of the fields and any other surface features of the data which have been discovered.

C. Explore data

This task tackles the data mining questions, which can be addressed using querying, visualization and reporting. These include distribution of key attributes, for example the target attribute of a prediction task; relations between pairs or small numbers of attributes; results of simple aggregations; properties of significant sub-populations; simple statistical analyses. These analyses may address directly the data mining goals; they may also contribute to or refine the data description and quality reports and feed into the transformation and other data preparation needed for further analysis.

D. Verify data quality

In this step we examine the quality of the data. Data completeness, correctness etc. is important for the following steps of the data mining process. It is often useful to set up an analysis on a subset or sample of the available data. This is because the quality of the information collected from the complete analysis across the whole available data mart is not always better than the information obtained from an investigation of the samples. In fact, in data mining the analyzed databases are often very large, so using a sample of the data reduces the analysis time.

It also reduces the risk that the statistical method might adapt to irregularities and loses its ability to generalize and forecast.

3.3. Data Preparation

Includes all activities required to construct the final data set (data that will be fed into the modeling tool) from the initial raw data. Tasks include table, case, and attribute selection as well as transformation and cleaning of data for modeling tools. Data preparation step involves following sub steps:

A. Select Data

Decide on the data to be used for analysis.

B. Clean Data

It is often necessary to carry out a preliminary cleaning of the data. It is a formal process used to highlight any variables that exist but which are not suitable for analysis. It is also an important check on the contents of the variables and the possible presence of missing, or incorrect data. If any essential information is missing, it will then be necessary to review the phase that highlights the source [GIUDICI, 2003].

C. Construct Data

This task includes constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes.

D. Integrate Data

These are methods whereby information is combined from multiple tables or records to create new records or values.

E. Format Data

Formatting transformations refer to primarily *syntactic* modifications made to the data that do not change its meaning, but might be required by the modeling tool.

3.4. Modeling

Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

3.5. Evaluation

Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. Determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results is reached.

3.6. Deployment

Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.

CHAPTER 4

TEXT MINING

4.1. What is Text Mining?

Although there are some other references, text mining will be summarized below referring to FELDMAN and SANGER [2007].

Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections.

Certainly, text mining derives much of its inspiration and direction from seminal research on data mining. Therefore, it is not surprising to find that text mining and data mining systems show many high-level architectural similarities. For instance, both types of systems rely on preprocessing routines, pattern-discovery algorithms, and presentation-layer elements such as visualization tools to enhance the browsing of answer sets. Further, text mining adopts many of the specific types of patterns in its core knowledge discovery operations that were first introduced and vetted in data mining research.

Because data mining assumes that data have already been stored in a structured format, much of its preprocessing focus falls on two critical tasks: Scrubbing and normalizing data and creating extensive numbers of table joins. In contrast, for text mining systems, preprocessing operations center on the identification and extraction of representative features for natural language documents. These preprocessing operations are responsible for transforming unstructured data stored in document collections into a more explicitly structured intermediate format, which is a concern that is not relevant for most data mining systems.

Moreover, because of the centrality of natural language text to its mission, text mining also draws on advances made in other computer science disciplines concerned with the handling of natural language. Perhaps most notably, text mining exploits techniques and methodologies from the areas of information retrieval, information extraction, and corpus-based computational linguistics.

4.2. Text Mining Processes

As given by COUSEAULT, [2004] there are approximately five major technique categories in the overall text mining process: a) Document Retrieval, b) Data Extraction, c) Data Cleansing, d) Mining, and e) Visualization. As part of the text mining process, there are a number of technique categories that are subcategories of, or supplements to, these major categories, such as Clustering, Visualization or Summarization (Figure 4.1). This section discusses briefly examples from current research in these core areas: Retrieval, Extraction, Cleansing, Mining and Visualization.

4.2.1. Document Retrieval

“Information Retrieval” is the term often used to describe the act of retrieving information from documents or retrieving documents from a document collection. In this case, what will be described is document retrieval. Much of the research in document retrieval is being done in support of Web search engines. In 2001, 47% of the documents in our information retrieval documents contained either the word “Internet,” “web,” or “www”. The size of the web and subsequently the large number of returns from a search is causing frustration in the user population as they attempt to identify the most relevant pages in a sea of thousands of hits. For example, an "information retrieval" search in 2001 returned 63,600 documents. As a result, researchers are attempting to find new ways of scoring and presenting the search results. The general trend in document retrieval is to incorporate methods that were initially considered post-retrieval techniques and incorporating these techniques into the retrieval process. For example, Google uses a type of link analysis to identify relevant documents, and has been quite successful. Other methods include clustering, both static and dynamic, as a method to present the documents to users and allow them to focus on the area of interest. Clustering may also be used as a type of query expansion. Documents not containing the search term, but clustered with the search term are also returned. Other researchers are looking to improve the document set by returning extracted entities into the search. The results, although they overall have demonstrated some improvement, can also greatly deteriorate the resulting set, depending on the entity that is returned into the search. Further research into the best manner in which the entities can be used may lead to better results. Overall, the trend in retrieval is combining methods and including user interaction.

4.2.2. Data Extraction

Extraction can take two forms; one is to identify the parts of speech and the other is to identify the specific type of entity extracted, such as whether a particular entity is person, organization, phone number, date, address, or geographic location. There are numerous companies working in this domain. Many users require knowing the difference between an organization and a person, or want to be able to associate certain activities with the appropriate proper noun. For these purposes, in a very large corpus, recall is much more important than precision because, in general, it would not be challenging to remove names from the "Proper Name" group, because the number of proper names would not be too large to accomplish this type of revision. However, missing proper names requires that an individual sort through the entire word list to identify proper names missed by the tool. Some of the tools that perform entity extraction boost wonderful results in the 90% area for recall and precision in data, which includes primarily newspaper sources. However, for less predictable

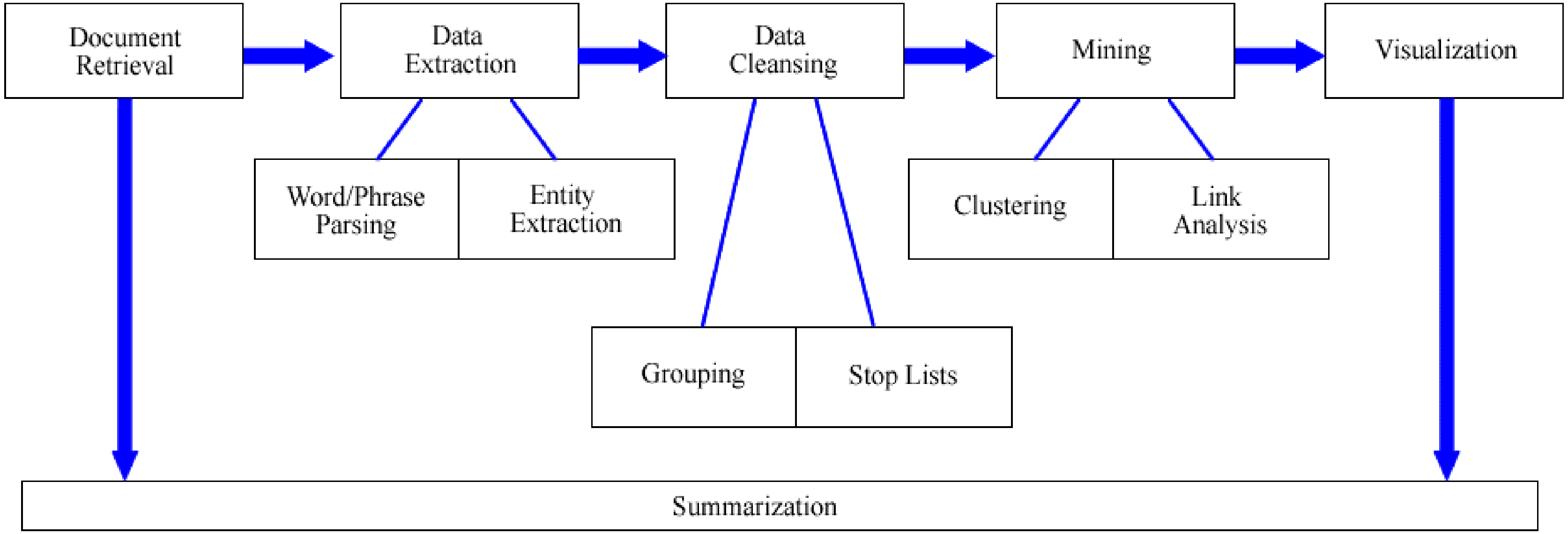


Figure 4.1. Text Mining Processes

formats, such as publication abstracts or web sites, the effectiveness drops considerably. In these areas, the drop in proper noun recall may cause problems. Another type of extraction is “parts of speech,” specifically noun phrases, which are important for capturing domain specific concepts. The problem with parts of speech extraction is that the same word can be used as different parts of speech such as “census” as in “take a census” or “census the population.” A powerful, applicable tool could take advantage of capabilities of both types of extraction to at least be able to identify accurately verbs and complex proper nouns.

4.2.3. Data Cleansing

The Data Cleansing literature is primarily discussed in relation to cluster analysis, but has impacts on all forms of data mining. It consists of the algorithms and methods that determine the final information that feeds the clustering algorithm or link analysis. Data Cleansing impacts the quality of other text mining techniques and determines the quality of the information that is fed into the clustering algorithms as well as how it is structured. For the databases the first decision is what part of the record to use: keywords, title words, abstract phrases/words, for full text words/phrases. In this research, data mining will be applied to abstract phrases.

Other issues are related to the selection and compression of the words that are used. Selection is the way words from text are determined to be candidate keywords for analysis. Selection issues relate to identifying a word as a potential keyword for analysis and determining the significance of that word in the document. The first step in the selection process is the defining of the word. For instance, words can be determined by every space or determined by Natural Language Processing algorithms to identify actual phrases (i.e. “Information Retrieval”). Another approach is simply to use windows of adjacent words. Selection also involves narrowing the number of words for analysis once they have been identified.

The final issue is the determination of strength between keywords based on location of the words. This information is not exactly data cleansing, but the method of capturing this information has an impact on the type of analysis that can be performed on the text data. Some tools may identify that two words are in the same paragraph or the proximity of two words to each other in the document.

4.2.4. Text Mining

In text mining both supervised an unsupervised data mining techniques are used. Some applications of text mining are: classification, clustering, information extraction, and questioning/answering [ORACLE TEXT MINING, 2005].

4.2.4.1. Supervised Classification

The classification task is very well known. The idea is to classify documents into a number of predefined categories. A document can be in one category, multiple categories, or not categories at all. For example you are Head of Information Dissemination at Finance, Inc. You have just arranged with a major newspaper to receive all their newspaper articles (and those of their 200 partners round the world) as an online feed, with new articles dropping in as soon as they're written. You can have everyone list their interests and have an expert scan the incoming stories and send them to everyone on the list.

Assuming you can't justify dedicating one or more people to this task, you can insert all the stories into a database table, and use a free text engine (such as Oracle Text) to index them. Users can then perform their searches - or you could run all the searches and send the results to the interested users.

In order to get the results to users as fast as possible, you are going to have to run these searches very frequently, and you are also going to need to filter out all those hits that you have already sent. This is all pretty inefficient. What you really want to do is to have users tell you what they are interested in, perhaps by registering a collection of queries, and then run each incoming document against the complete set of registered queries.

With the help of rule based classification a set of queries, or “rules”, can be collected in a table - together with additional information such as the name of the user who registered this rule. A special type of index is created on these rules. Then, for each document that is processed (such as the news stories is the scenario above), a search is made for rules that match this particular document.

The process that defines the rules can be manual or automatic. Manual rule generation means that a human expert will create the necessary rules. This is very accurate, but also time consuming and therefore doesn’t scale for large collections and categories. The user must supply a training set consisting of documents already known to belong to one or more categories.

The benefits of automatic rule generation are:

a) No manual creation of rules is required.

b) As the document set grows, new groups of documents can be used for further training.

c) You are making use of existing knowledge to classify future documents.

Let’s assume a routing application where the system receives incoming news articles that should be distributed based on certain categories. Users could spend some time perfecting their ideal queries, and then leave those queries to be run automatically on each incoming document.

However, this does require that users spend a significant amount of time perfecting these queries. For many non-technical users, this may be asking too much.

What might be more useful is for the users to provide a set of example documents from the subject - or “category” - in which they are interested. A useful classification system should then be able to analyze this set of documents, and automatically generate a rule (or set of rules) that would identify future incoming documents in the same subject area. Some text mining tools can do this for you. They use methods to do this, named after their underlying algorithms, which are known as “Decision Trees” and “Support Vector Machine”.

4.2.4.2. Clustering

Clustering is the unsupervised division of patterns into groups. We can distinguish two main clustering techniques: hierarchical and partitioned. The first produces a nested series of partitions and the second produces only one. K-Mean is an example of a partitioned clustering algorithm that produces a single partition of the data set (sometimes called flat output). K-Mean is a trade-off between quality and scalability. The K-Mean algorithm works in “feature space”, where each dimension is represented by one of the features (such as a word or theme) from the document set.

It assigns k centers, or “prototypes”, one for each cluster. It then calculates the distance from each prototype in feature space for each document, and assigns the document to the cluster to which it is closest. It then calculates a new prototype for each cluster based on the mean of each document assigned to it, and repeats the process. The clusters can be used for building features like showing similar documents in the collection.

The benefits are:

a) The automatic discovery of patterns in the collection.
b) It is useful for identifying categories from the collection.
c) It is useful for building abstractions.
d) It provides a statistical snapshot of the collection.

4.2.5. Visualization

Extracting information that no one sees is useless. So a lot of possibilities have been invented of how to visualize the results obtained. The simplest is just to make a table for the user to look up the information he needs. On the other end of the complexity scale are three-dimensional worlds that the user may navigate in. For the visualization of query results hypertext is the classic option. The complexity can be hidden, but if the user is interested in the details, he may just click on the link.

The other issue of visualization is how much data to show to the user. Usually the user is confronted with pure results without the meta-information on how and why the results contain some kind of valuation and the user wants to know what it was exactly that made one result superior to another.

The solution to these problems is transparency. Transparency does not mean that the algorithm that made the decision is explained, but the reason for the decision. An example is the Google highlighting of the search keywords in the results. It says: This result was chosen because it has the keyword in it. This is of course a simplification, because the highlighting is just symbolic for the chain of real events that happened, which is in fact quite complicated.

Not all data mining algorithms support this approach, hiding their reasons in neural networks or complex weighting schemes. Further research may improve the situation [MATHIAK and ECKSTEIN, 2006].

4.3. Text Mining in Practice

Many text mining systems introduced in the late 1990s were developed by computer scientists as part of academic "pure research" projects aimed at exploring the capabilities and performance of the various technical components making up these systems. Most current text mining systems, however – whether developed by academic researchers, commercial software developers, or in-house corporate programmers – are built to focus on specialized applications that answer questions peculiar to a given problem space or industry need. Obviously, such specialized text mining systems are especially well suited to solving problems in academic or commercial activities in which large volumes of textual data must be analyzed in making decisions.

Two areas of analytical inquiry have proven particularly fertile ground for text mining applications. In various areas of corporate finance, bankers, analysts, and consultants have begun leveraging text mining capabilities to sift through vast amounts of textual data with the aims of creating usable forms of business intelligence, noting trends, identifying correlations, and researching references to specific transactions, corporate entities, or persons. In patent research, specialists across industry verticals at some of the world’s largest companies and professional services firms apply text mining approaches to investigating patent development strategies and finding ways to exploit existing corporate patent assets better [FELDMAN and SANGER, 2007].

The most active application area for text mining is in the biosciences. Researchers are exploring enormous collections of biomedical research reports to identify complex patterns of interactivities between proteins. The best known example is Don Swanson's work on hypothesizing causes of rare diseases by looking for indirect links in different subsets of the bioscience literature [HEARST, 2003].

As another example, one of the big current questions in genomics is which proteins interact with which other proteins. There has been notable success in looking at which words co-occur in articles that discuss the proteins in order to predict such interactions. The key is to not look for direct mentions of pairs, but to look for articles that mention individual protein names, keep track of which other words occur in those articles, and then look for other articles containing the same sets of words. This very simple method can yield surprisingly good results, even though the meaning of the texts is not being discerned by the programs. Rather, the text is treated like a "bag of words" [HEARST, 2003].

CHAPTER 5

CLASSIFICATION AND ALGORITHMS

As noted earlier in Chapter 2, supervised data mining techniques in data mining are classification, regression, attribute importance and anomaly detection. In the application of this thesis in the next following chapters, classification will be used. Therefore this chapter is devoted for further information about algorithms for classification.

5.1. Classification and Prediction

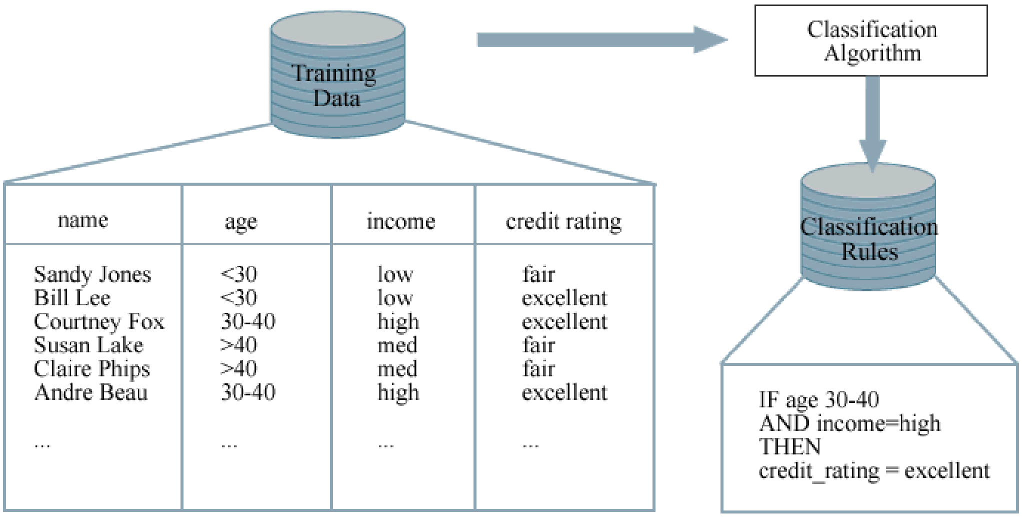
Databases are rich with hidden information that can be used for making intelligent business decisions. Classification and prediction are two forms of data analysis which can be used to extract models describing important data classes or to predict future data trends. Whereas classification predicts categorical labels (or discrete values), prediction models continuous-valued functions. For example, a classification model may be built to categorize bank loan applications as either safe or risky, while a prediction model may be built to predict the expenditures of potential customers on computer equipment given their income and occupation.

Many classification and prediction methods have been proposed by researchers in machine learning, expert systems, statistics, and neurobiology. Most algorithms are memory resident, typically assuming a small data size. Recent database mining research has built on such work, developing scalable classification and prediction techniques capable of handling large, disk resident data. These techniques often consider parallel and distributed processing.

Data classification is a two step process (Figure 5.1.) [HAN and KAMBER, 2000]. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes.

Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. In the context of classification, data tuples are also referred to as *samples, examples, or objects*. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. Since the class label of each training sample is provided, this step is also known as **supervised learning** (i.e., the learning of the model is 'supervised' in that it is told to which class each training sample belongs).

It contrasts with **unsupervised learning** (or **clustering**), in which the class labels of the training samples are not known, and the number or set of classes to be learned may not be known in advance.



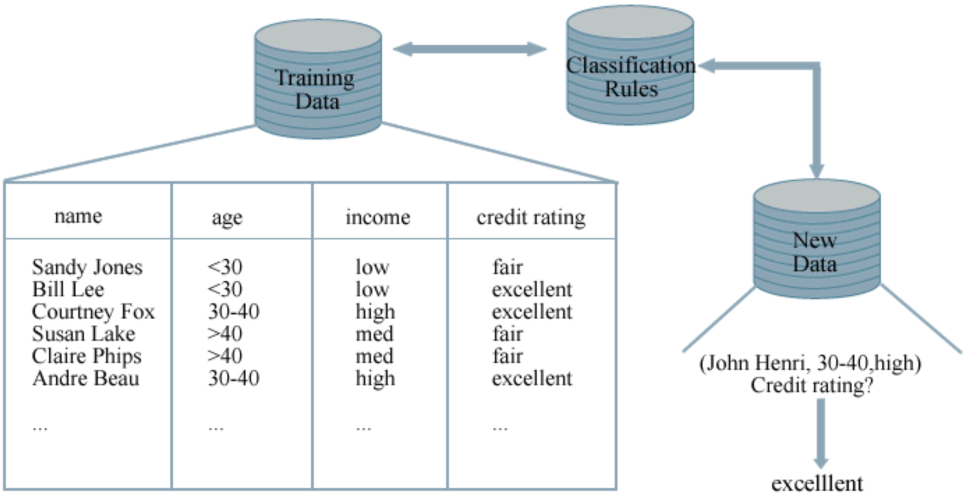
name	age	income	credit rating
Sandy Jones	<30	low	fair
Bill Lee	<30	low	excellent
Courtney Fox	30-40	high	excellent
Susan Lake	>40	med	fair
Claire Phips	>40	med	fair
Andre Beau	30-40	high	excellent
...	...	...	...

IF age 30-40
AND income=high
THEN
credit_rating = excellent

Figure 5.1. The Data Classification Process: Learning

Typically, the learned model is represented in the form of classification rules, decision trees, or mathematical formulae. For example, given a database of customer credit information, classification rules can be learned to identify customers as having either excellent or fair credit ratings (Figure 5.1). The rules can be used to categorize future data samples, as well as provide a better understanding of the database contents.

In the second step (Figure 5.2) [HAN and KAMBER, 2000], the model is used for classification. First, the predictive accuracy of the model (or classifier) is estimated. The holdout method is a simple technique which uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. For each test sample, the known class label is compared with the learned model's class prediction for that sample. Note that if the accuracy of the model were estimated based on the training data set, this estimate could be optimistic since the learned model tends to over fit the data (that is, it may have incorporated some particular anomalies of the training data which are not present in the overall sample population). Therefore, a test set is used.



name	age	income	credit rating
Sandy Jones	<30	low	fair
Bill Lee	<30	low	excellent
Courtney Fox	30-40	high	excellent
Susan Lake	>40	med	fair
Claire Phips	>40	med	fair
Andre Beau	30-40	high	excellent
...	...	...	...

Figure 5.2. The Data Classification Process: Classification

If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known. (Such data are also referred to in the machine learning literature as “unknown" or “previously unseen" data). For example, the classification rules learned in Figure 5.1 from the analysis of data from existing customers can be used to predict the credit rating of new or future (i.e., previously unseen) customers.“How is prediction different from classification?" Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled object, or to assess the value or value ranges of an attribute that a given object is likely to have. In this view, classification and regression are the two major types of prediction problems where classification is used to predict discrete or nominal values, while regression is used to predict continuous or ordered values. In our view, however, we refer to the use of predication to predict class labels as classification and the use of predication to predict continuous values (e.g., using regression techniques) as prediction. This view is commonly accepted in data mining. Classification and prediction have numerous applications including credit approval, medical diagnosis, performance prediction, and selective marketing [HAN and KAMBER, 2000].

5.2. Classification Algorithms

There are various algorithms available for classification. In the following sections, the commonly used algorithms such as decision tree algorithm, Bayesian classification, support vector machine algorithm and adaptive Bayes network algorithm, are briefly explained.

5.2.1. Decision Tree Algorithm

A particularly efficient method for producing classifiers from data is to generate a decision tree. The decision-tree representation is the most widely used logic method. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literature. They are supervised learning methods that construct decision trees from a set of input-output samples. A typical decision-tree learning system adopts a top-down strategy that searches for a solution in a part of the search space. It guarantees that a simple, but not necessarily the simplest, tree will be found. A decision tree consists of *nodes* where attributes are tested. The outgoing *branches* of a node correspond to all the possible outcomes of the test at the node. A simple decision tree for classification of samples with two input attributes X and Y is given in (Figure 5.3) [KANTARDZIC, 2003]. All samples with feature values X > 1 and Y = B belong to Class2, while the samples with values X < 1 belong to Class1, whatever the value for feature Y.

The samples, at a non-leaf node in the tree structure, are thus partitioned along the branches and each child node gets its corresponding subset of samples. Decision trees that use univariate splits have a simple representational form, making it relatively easy for the user to understand the inferred model; at the same time, they represent a restriction on the expressiveness of the model. In general, any restriction on a particular tree representation can significantly restrict the functional form and thus the approximation power of the model. A well-known tree-growing algorithm for generating decision trees based on univariate splits is Quinlan's ID3 with an extended version called C4.5. Greedy search methods, which involve growing and pruning decision-tree structures, are typically employed in these algorithms to explore the exponential space of possible models [KANTARDZIC, 2003].

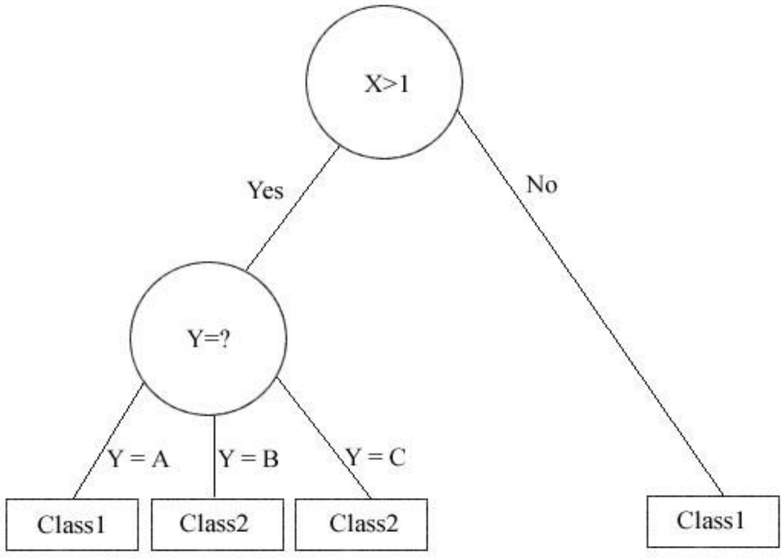


Figure 5.3. A Simple Decision Tree with the Tests on Attributes X and Y

The samples, at a non-leaf node in the tree structure, are thus partitioned along the branches and each child node gets its corresponding subset of samples. Decision trees that use univariate splits have a simple representational form, making it relatively easy for the user to understand the inferred model; at the same time, they represent a restriction on the expressiveness of the model. In general, any restriction on a particular tree representation can significantly restrict the functional form and thus the approximation power of the model. A well-known tree-growing algorithm for generating decision trees based on univariate splits is Quinlan's ID3 with an extended version called C4.5. Greedy search methods, which involve growing and pruning decision-tree structures, are typically employed in these algorithms to explore the exponential space of possible models [KANTARDZIC, 2003].

Decision tree rules provide model transparency so that a business user, marketing analyst, or business analyst can understand the basis of the model's predictions, and therefore, be comfortable acting on them and explaining them to others. In addition to transparency, the Decision Tree algorithm provides speed and scalability. The build algorithm scales linearly with the number of predictor attributes and on the order of nlog(n) with the number of rows, n. Scoring is very fast. Both build and apply are parallelized. The Decision Tree algorithm builds models for binary and multi-class targets. It produces accurate and interpretable models with relatively little user intervention required. The Decision Tree algorithm is implemented in such a way as to handle data in the typical data table formats, to have reasonable defaults for splitting and termination criteria, to perform automatic pruning, and to perform automatic handling of missing values. However, it does not distinguish sparse data from missing data. (See "Sparse Data" for more information.) Users can specify costs and priors. Decision Tree does not support nested tables. Decision Tree Models can be converted to XML.

5.2.2. Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on Bayesian theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved, and in this sense, is considered “naive". Bayesian belief networks are graphical models, which unlike naive Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can also be used for classification [HAN and KAMBER, 2000].

Bayes Theorem

Let X be a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P (H|X), the probability that the hypothesis H holds given the observed data sample X.

P (H|X) is the posterior probability, or a posteriori probability of H conditioned on X. For example, suppose the world of data samples consists of fruits, described by their color and shape. Suppose that X is red and round, and that H is the hypothesis that X is an apple. Then P (H|X) reflects our confidence that X is an apple given that we have seen that X is red and round. In contrast, P (H) is the prior probability, or a priori probability of H.

For our example, this is the probability that any given data sample is an apple, regardless of how the data sample looks. The posterior probability, P(H|X) is based on more information (such as background knowledge) than the prior probability, P(H), which is independent of X.

Similarly, P (X|H) is the posterior probability of X conditioned on H. That is, it is the probability that X is red and round given that we know that it is true that X is an apple. P (X) is the prior probability of X. Using our example; it is the probability that a data sample from our set of fruits is red and round. “How are these probabilities estimated?" P (X), P (H), and P (X|H) may be estimated from the given data, as we shall see below.

Bayes theorem is useful in that it provides a way of calculating the posterior probability, P (H|X) from P (H), P(X), and P (X|H). Bayes theorem is:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

How effective are Bayesian Classifiers?

In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data. However, various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domains. Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers which do not explicitly use Bayes theorem. For example, under certain assumptions, it can be shown that many neural network and curve fitting algorithms output the maximum posteriori hypothesis, as does the naive Bayesian classifier [HAN and KAMBER, 2000].

5.2.3. Support Vector Machine Algorithm

Support Vector Machine (SVM) algorithm is a state-of-the-art classification and regression algorithm. SVM is an algorithm with strong regularization properties, that is, the optimization procedure maximizes predictive accuracy while automatically avoiding over-fitting of the training data. Neural networks and radial basis functions, both popular data mining techniques, have the same functional form as SVM models;

however, neither of these algorithms has the well-founded theoretical approach to regularization that forms the basis of SVM.

SVM projects the input data into a kernel space. Then it builds a linear model in this kernel space. A classification SVM model attempts to separate the target classes with the widest possible margin. A regression SVM model tries to find a continuous function such that maximum number of data points lie within an epsilon-wide tube around it. Different types of kernels and different kernel parameter choices can produce a variety of decision boundaries (classification) or function approximations (regression).Oracle Data Miner SVM implementation which is used in this study supports two types of kernels: linear and Gaussian. Tool also provides automatic parameter estimation on the basis of the characteristics of the data.

SVM performs well with real-world applications such as classifying text, recognizing hand-written characters, classifying images, as well as bioinformatics and biosequence analysis. The introduction of SVM in the early 1990s led to an explosion of applications and deepening theoretical analysis that established SVM along with neural networks as one of the standard tools for machine learning and data mining.

There is no upper limit on the number of attributes and target cardinality for SVMs; the only constraints are those imposed by hardware. SVM is the preferred algorithm for sparse data.

a) Active Learning

SVM models grow as the size of the training data set increases. This property limits SVM models to small and medium size training sets (less than 100,000 cases). Active learning provides a way to deal with large training sets.

The termination criterion for active learning is usually an upper bound on the number of support vectors; when the upper bound is attained, the build stops. Alternatively, stopping criteria are qualitative, such as no significant improvement in model accuracy on a held-aside sample.

Active learning forces the SVM algorithm to restrict learning to the most informative training examples and not to attempt to use the entire body of data. In most cases, the resulting models have predictive accuracy comparable to that of the standard (exact) SVM model. Active learning can be applied to all SVM models (classification, regression, and one-class). Active learning is on by default. It can be turned off.

b) Sampling for Classification

For classification, SVM automatically performs stratified sampling during model build. The algorithm scans the entire build data set and selects a sample that is balanced across target values.

c) Automatic Kernel Selection

SVM automatically determines the appropriate kernel type based on build data characteristics. This selection can be overridden by explicitly specifying a kernel type.

d) Data Preparation and Settings Choice for Support Vector Machines

You can influence both the Support Vector Machine (SVM) model quality (accuracy) and performance (build time) through two basic mechanisms: data preparation and model settings. Significant performance degradation can be caused by a poor choice of settings or inappropriate data preparation. Poor settings choices can also lead to inaccurate models.

SVM has built-in mechanisms that attempt to choose appropriate settings automatically based on the data provided. For some domains the system-determined settings are needed to override [ORACLE, 2005].

5.2.4. Adaptive Bayes Network Algorithm

Adaptive Bayesian Network (ABN) is an Oracle proprietary algorithm that provides a fast, scalable, non-parametric means of extracting predictive information from data with respect to a target attribute. (Non-parametric statistical techniques avoid assuming that the population is characterized by a family of simple distributional models, such as standard linear regression, where different members of the family are differentiated by a small set of parameters.)

ABN, in Single Feature Build mode, can describe the model in the form of human-understandable rules. The rules produced by ABN are one of its main advantages over Naive Bayes. ABN rules provide model transparency so that a business user, marketer, or business analyst can understand the basis of the model's predictions and therefore, be comfortable acting on them and explaining them to others. In addition to rules, ABN provides performance and scalability, which are derived via various user parameters controlling the trade-off of accuracy and build time. ABN predicts binary as well as multiclass targets.

Adaptive Bayesian Network Model Types

An ABN model is an (adaptive conditional independence model that uses the minimum description length principle to construct and prune an array of conditionally independent network features. Each network feature consists of one or more conditional probability expressions. The collection of network features forms a product model that provides estimates of the target class probabilities. There can be one or more network features. The number and depth of the network features in the model determine the model mode.

There are three model modes for ABN:

a) Pruned Naive Bayes (Naive Bayes Build)

b) Simplified decision tree (Single Feature Build)

c) Boosted (Multi Feature Build)

CHAPTER 6

ORACLE DATA MINING

6.1. What is Oracle Data Mining?

Oracle Data Mining (ODM) embeds data mining within the Oracle database. ODM algorithms operate natively on relational tables or views, thus eliminating the need to extract and transfer data into standalone tools or specialized analytic servers. ODM's integrated architecture results in a simpler, more reliable and more efficient data management and analysis environment. Data mining tasks can run asynchronously and independently of any specific user interface as part of standard database processing pipelines and applications. Data analysts can mine the data in the database build models and methodologies, and then turn those results and methodologies into full-fledged application components ready to be deployed in production environments. The benefits of the integration with the database cannot be emphasized enough when it comes to deploying models and scoring data in a production environment. ODM allows a user to take advantage of all aspects of Oracle's technology stack as part of an application. Also, fewer "moving parts" results in a simpler, more reliable, more powerful advanced business intelligence application.ODM provides single-user multi-session access to models. ODM programs can run either asynchronously or synchronously in the Java interface. ODM programs using the PL/SQL interface run synchronously; to run PL/SQL asynchronously requires using the Oracle Scheduler.

Although there are some other references, Oracle data mining will be summarized below referring to ORACLE [2006] and ORACLE DATA MINING [2005].

Oracle Data Mining supports the following data mining functions:

A. Supervised Data Mining:

- **a)** **Classification**: Grouping items into discrete classes and predicting which class an item belongs to
- **b)** **Regression**: Approximating and forecasting continuous values
- **c)** **Attribute Importance:** Identifying the attributes that are most important in predicting results
- **d)** **Anomaly Detection:** Identifying items that do not satisfy the characteristics of "normal" data (outliers)

B. Unsupervised Data Mining:

- **a)** Clustering: Finding natural groupings in the data
- **b)** Association models: Analyzing "market baskets"
- **c)** Feature extraction: Creating new attributes (features) as a combination of the original attributes

6.2. Data for Oracle Data Mining

Data used by ODM consists of tables or views stored in an Oracle database. Both ordinary tables and nested tables can be used as input data. The data used in a data mining operation is often called a data set.

Data has a physical organization and a logical interpretation. Column names refer to physical organization; attribute names, described in the next paragraph, refer to the logical interpretation of the data.

The rows of a data table are often called cases, records, or examples. The columns of the data tables are called attributes or fields; each attribute in a record holds a cell of information. Attribute names are constant from record to record for unnested tables; the values in the attributes can vary from record to record. For example, each record may have an attribute labeled "annual income." The value in the annual income attribute can vary from one record to another.

ODM distinguishes two types of attributes: categorical and numerical. Categorical attributes are those that define their values as belonging to a small number of discrete categories or classes; there is no implicit order associated with the values. If there are only two possible values, for example, yes and no, or male and female, the attribute is said to be binary. If there are more than two possible values, for example, small, medium, large, extra large, the attribute is said to be multiclass.

Numerical attributes are numbers that take on a large number of values that have an order, for example, annual income. For numerical attributes, the differences between values are also ordered. Annual income could theoretically be any value from zero to infinity, though in practice annual income occupies a bounded range and takes on a finite number of values.

Numerical attributes can be transformed to categorical attributes. For example, annual income could be divided into three categories: low, medium, high. Conversely, numerical values can be transformed to categorical values.

Classification and Regression algorithms require a target attribute. A supervised model can predict a single target attribute. The target attribute for all classification algorithms can be numerical or categorical. The ODM regression algorithm supports only numerical target attributes.

Certain ODM algorithms support unstructured text attributes. Although unstructured data includes images, audio, video, geospatial mapping data, and documents or text, ODM supports mining text data only. An input table can contain one or more text columns.

6.2.1. Data Requirements

ODM supports several types of input data, depending on data table format, column data type, and attribute type ODM data must reside in a single table or view in an Oracle database. The table or view must be a standard relational table, where each case is represented by one row in the table, with each attribute represented by a column in the table. The columns must be of one of the types supported by ODM. Column Data Types Supported by ODM:

ODM does not support all the data types that Oracle supports. The supported attribute data types have a default attribute type (categorical or numerical). Each attribute (column) in a data set used by ODM must have one of the following data types:

a) INTEGER
b) NUMBER
c) FLOAT
d) VARCHAR2
e) CHAR
f) DM_NESTED_NUMERICALS (nested column)
g) DM_NESTED_CATEGORICALS (nested column)

Nested table columns can be used for capturing in a single table or view data that is distributed over many tables (for example, a star schema). Nested columns allow you to capture one-to-many relationships (for example, one customer can buy many products). Nested tables are required if the data has more than 1000 attributes; nested tables are useful if the data is sparse, or if the data is already persisted in a transactional format and must be passed to the data mining interface through an object view.

The fixed collection types DM_NESTED_NUMERICALS and DM_NESTED_CATEGORICALS are used to define columns that represent collections of numerical attributes and categorical attributes, respectively.
For a given case identifier, attribute names must be unique across all the collections and individual columns. The fixed collection types enforce this requirement.

Data tables often contain missing values. Certain algorithms assume that a NULL value indicates a missing value; others assume that a NULL value indicates sparse data.

Data is said to be sparse if only a small fraction (no more than 20%, often 3% or less) of the attributes are non-zero or non-null for any given case. Sparse data occurs, for example, in market basket problems. In a grocery store, there might be 10,000 products in the store, and the average size of a basket (the collection of distinct items that a customer purchases in a typical transaction) is on average 50 products. In this example, a transaction (case or record) has on average 50 out of 10,000 attributes that are not null. This implies that the fraction of non-zero attributes in the table (or the density) is approximately 50/10,000, or 0.5%. This density is typical for market basket and text mining problems.

Association models are designed to process sparse data; indeed, if the data is not sparse, the algorithm may require a large amount of temporary space or may not be able to build a model.

Sparse data is represented in a table in such a way that avoids the specification of the most common value to save storage. In such a specification of sparse data, a missing value is implicitly interpreted as the most common value.

Different algorithms make different assumptions about what indicates sparse data. For Support Vector Machine, k-Means, association, and Non-Negative Matrix Factorization, NULL values indicate sparse data; for all other algorithms, NULL values indicate missing values. See the description of each algorithm for information about how it interprets NULL values.

6.2.2. Data Preparation

Data is said to be prepared when certain data transformations required by a data mining algorithm are performed by the user before the algorithm is invoked. For most algorithms, data must be prepared before the algorithm is invoked.

Data preparation can take many forms, such as joining two or more tables so that all required data is in a single table or view, transforming numerical attributes by applying numerical functions to them, recoding attributes, treating missing values, treating outliers, omitting selected columns for a training data set, and so forth. ODM includes transformations that perform the following data-mining-specific transformations:

Certain algorithms are sensitive to outliers. Winsorizing and trimming transformations are used to deal with outliers. Winsorizing involves setting the tail values of an attribute to some specified value. For example, for a 90% Winsorization, the bottom 5% of values are set equal to the minimum value in the 6th percentile, while the upper 5% are set equal to the maximum value in the 95th percentile. Trimming removes the tails in the sense that trimmed values are ignored in further computations. This is achieved by setting the tails to NULL. This process is sometimes called clipping.

Some ODM algorithms may benefit from binning (discretizing) both numeric a categorical data. Naive Bayes, Adaptive Bayesian Network, Clustering, Attribute Importance, and Association Rules algorithms may benefit from binning.Binning means grouping related values together, thus reducing the number of distinct values for an attribute. Having fewer distinct values typically leads to a more compact model and one that builds faster. Binning must be performed carefully. Proper binning can improve model accuracy; improper binning can lead to loss in accuracy.
Normalization converts individual numerical attributes so that each attribute’s values lie in the same range. Values are converted to be in the range 0.0 to 1.0 or the range -1.0 to 1.0. Normalization ensures that attributes do not receive artificial weighting caused by differences in the ranges that they span. Some algorithms, such as k-Means, Support Vector Machine, and Non-Negative Matrix Factorization, benefit from normalization.

6.3. How does Oracle Data Mining Support Data Mining?

a) ODM integrates data mining with the Oracle database and exposes data mining through the following interfaces:

b) Java interface: Java Data Mining (JSR-73) compliant interface that allows users to embed data mining in Java applications.

c) PL/SQL interface: The packages DBMS_DATA_MINING and DBMS_DATA_MINING_TRANSFORM allow users to embed data mining in PL/SQL applications.

d) Automated data mining: The DBMS_PREDICTIVE_ANALYTICS PL/SQL package, automates the entire data mining process from data preprocessing through model building to scoring data.

e) Data mining SQL functions: The SQL Data Mining functions (CLUSTER_ID, CLUSTER_PROBABILITY, CLUSTER_SET, FEATURE_ID, FEATURE_SET, FEATURE_VALUE, PREDICTION, PREDICTION_COST, PREDICTION_DETAILS, PREDICTION_PROBABILITY, and PREDICTION_SET) support deployment of models within the context of existing applications, improve scoring performance, and enable pipelining of results involving data mining predictions.

f) Graphical interfaces: Oracle Data Miner and Oracle Spreadsheet Add-In for Predictive Analytics are graphical interfaces that solve data mining problems.

g) The end result of data mining is a model. Often this model is deployed so that its results can be embedded in an application.

6.4. ODM and Oracle Data Mining Support for Text Mining

ODM provides infrastructure for developing data mining applications suitable for addressing a variety of business problems involving text. Among these, the following specific technologies provide key elements for addressing problems that require text mining:

a) Classification

b) Clustering

c) Feature extraction

d) Association

e) Regression

f) Anomaly Detection

The technologies that are most used in text mining are classification, clustering, and feature extraction.

Table 6.1 summarizes how the ODM (both the Java and PL/SQL interfaces) and Oracle Text support text mining functions.

Table 6.1. ODM and Oracle Text Support Text Mining Functions

Feature	**ODM**	**Oracle Text**
Association	Text data only or text and non-text data	No support
Clustering	k-Means algorithm supports text only or text and non-text data	k-means algorithm supports text only
Attribute importance	No support for text data	No support
Regression	Support Vector Machine(SVM) algorithm supports text data only or text and non-text data	No support
Classification	SVM supports text only or text and non-text data Support for assigning documents to one of many labels	SVM and decision trees support text only Support for assigning documents to one of many labels and also for assigning documents to multiple labels at the same time
One-Class SVM	One-Class SVM supports text only or text and non-text data	No support
Feature extraction (basic features)	The Java API handles the process supports the feature extraction process that transforms a text column to a nested table. The PL/SQL API requires the use of Oracle Text procedures to perform extraction. ODM allows the same degree of control as Oracle Text	Feature extraction is done internally; the results are not exposed
Feature extraction (higher order features)	Non-negative matrix factorization (NMF) supports either text or text and non-text data	No support
Record apply	No support for record apply	Supports record apply for classification
Support for text columns	Features extracted from a column of type CLOB, BLOB, BFILE. LONG, VARCHAR2, XML Type, CHAR, RAW, LONG RAW using an appropriate transformation	Supports table columns of type CLOB, BLOB, BFILE. LONG, VARCHAR2, XML Type, CHAR, RAW, LONG RAW

CHAPTER 7

CASE STUDY: REAL MEDICALDATA - PULMONARY EMBOLISM DATA

Some parts of this chapter which gives information about pulmonary embolism are taken from Mayo Clinic’s and Medicinenet web pages. Diagnosis of pulmonary embolism is explained referring to KEARON [2003] at the end of the chapter.

7.1. What is Pulmonary Embolism?

The lung is composed of clusters of small air sacs (alveoli) divided by thin, elastic walls (membranes). Capillaries, the tiniest of blood vessels, run within these membranes between the alveoli and allow blood and air to come near each other. The distance between the air in the lungs and the blood in the capillaries is very small, and allows molecules of oxygen and carbon dioxide to transfer across the membranes. The exchange of the air between the lungs and blood are through the arterial and venous system. Arteries and veins both carry and move blood throughout the body, but the process for each is very different. Arteries carry blood from the heart to the body. Veins return blood from the body to the heart. The heart is a two-sided pump. Oxygen-carrying blood travels from the left side of the heart to all the tissues of the body.

The oxygen is extracted by the tissue, and carbon dioxide (a waste product) is delivered back into the blood. The blood now deoxygenated and with higher levels of carbon dioxide, is returned via the veins to the right side of the heart. The blood is then pumped out of the right side of the heart to the lungs, where the carbon dioxide is removed and oxygen is returned to the blood from the air we breathe in, which fills the lungs. Now the blood, high in oxygen and low in carbon dioxide, is returned to the left side of the heart where the process starts all over again. The blood travels in a circle and is therefore referred to as circulation. If a blood clot (thrombus) forms in the one of the body's veins (deep vein thrombosis or DVT), it has the potential to break off and enter the circulatory system and travel through the heart and become lodged in the one of the branches of the pulmonary artery of the lung (Figure 7.1) (See http://www.dangelolaw.com/ortho_evera2.html).

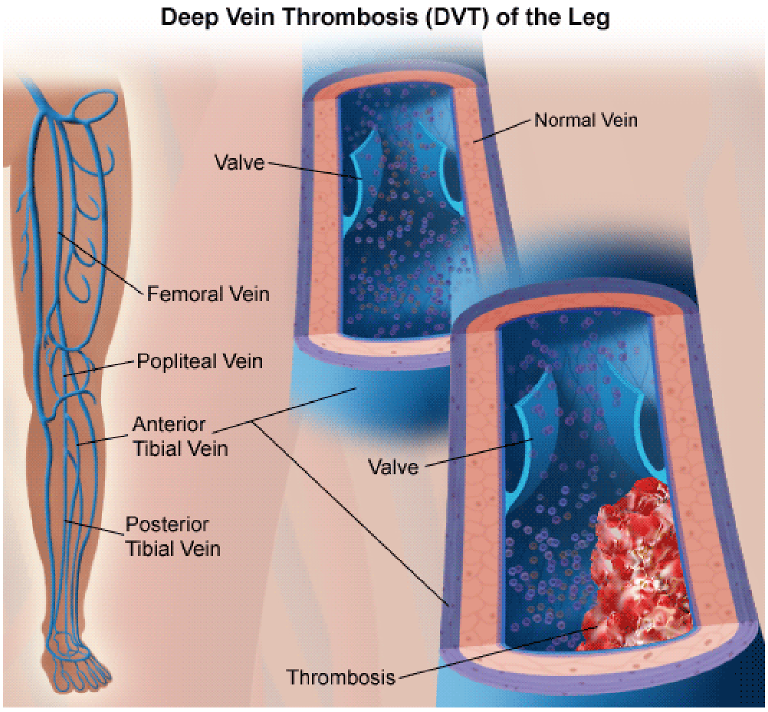


Figure 7.1. Deep Vein Thrombosis (DVT) of the Leg

As given in Figure 7.2 (See http://www.dangelolaw.com/ortho_evera2.html) a clot that travels through the circulatory system to another location is known as an embolus (plural emboli). A pulmonary embolus clogs the artery that provides blood supply to part of the lung. The embolus not only prevents the exchange of oxygen and carbon dioxide, but it also decreases blood supply to the lung tissue itself, potentially causing lung tissue to die (infarct). A pulmonary embolus is one of the life-threatening causes of chest pain and should always be considered when a patient presents to a healthcare provider with complaints of chest pain and shortness of breath.

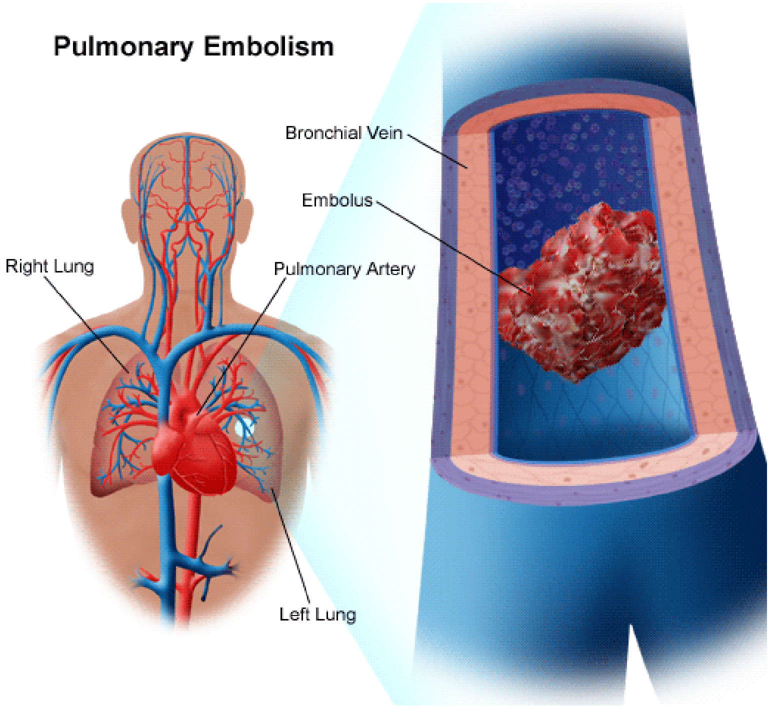


Figure 7.2. Pulmonary Embolism

7.2. Tests and Diagnosis

There are many valuable tests (including clinical assessment) that may be used, singly or in combination, to confirm or exclude the presence of pulmonary embolism with a high degree of confidence. Availability of testing and differences among patient presentations will influence the diagnostic approach used. Some of the diagnosis methods are explained below.

7.2.1. History and Physical Examination

There always needs to be a high a level of suspicion that a pulmonary embolus may be the cause of chest pain or shortness of breath. The healthcare provider will take a history of the type of chest pain, including its onset and associated symptoms that may direct the diagnosis to pulmonary embolism. It may include asking about risk factors for deep vein thrombosis.

Physical examination will concentrate initially on the heart and lungs, since chest pain and shortness of breath may also be the presenting complaints for heart attack, pneumonia, pneumothorax (collapsed lung), and dissection of an aortic aneurysm, among others.

With pulmonary embolism, the chest examination is often normal, but if there is some associated inflammation on the surface of the lung (the pleura), a rub may be heard (pleura inflammation may cause friction which can be heard with a stethoscope).

The surfaces of the lung and the inside of the chest wall are covered by a membrane (the pleura) that is full of nerve endings. When the pleura becomes inflamed, as can occur in pulmonary embolus, a sharp pain can result that is worsened by breathing, so-called pleurisy or pleuritic chest pain. The physical examination may include looking for signs of a deep vein thrombosis in an extremity: warmth, swelling, redness, and tenderness.

It is important to note, however, that the signs associated with deep vein thrombosis may be completely absent even in the presence of a clot.

Again, risk factors for clotting must be taken into consideration when making an assessment.

7.2.2. Basic Testing

Basic testing may include:

- CBC (complete blood count)
- Electrolytes,
- BUN (blood urea nitrogen),
- Creatinine blood test (to assess kidney function),
- Chest X-Ray, and
- Electrocardiogram (EKG or ECG).

The chest X-Ray is often normal in pulmonary embolism. The EKG may be normal, but usually demonstrates a rapid heart rate, so-called sinus tachycardia (heart rate > 100 bpm). If there is significant blockage in a pulmonary artery, it acts like a dam and it is harder for the heart to push blood past the obstructing clot or clots. This can result in a change in the electrical signal passing through the heart by stretching the heart muscle, revealed on an EKG a so-called right heart strain.

Since the cost of missing the diagnosis of pulmonary embolus can be death, the approach to diagnosis is to prove that no pulmonary embolus exists.

7.2.3. Pulmonary Angiogram

The gold standard for the diagnosis of pulmonary embolus is a pulmonary angiogram in which a catheter is threaded into the pulmonary arteries, usually from veins in the leg. Dye is injected and a clot or clots can be identified on imaging studies. This is considered an invasive test and should be performed only by someone with expertise in this procedure.

Fortunately, there are other, less invasive ways to make the diagnosis. The decision as to which test might best make the diagnosis needs to be individualized to the patient and their presentation and situation.

7.2.4. D-dimer Blood Test

If the healthcare provider's suspicion for pulmonary embolism is low, a D-dimer blood test can be used. The D-dimer blood test measures one of the breakdown products of a blood clot. If this test is normal, then the likelihood of a pulmonary embolism is very low. Unfortunately, this test is not specific for blood clots in the lung. It can be positive for a variety of reasons including pregnancy, injury, recent surgery, or infection. Looking at the list of deep vein thrombosis risk factors, one can imagine that a D-dimer blood test may not be helpful in those with significant risk factors for deep vein thrombosis.

7.2.5. CT Scan

If there is greater suspicion, then computerized tomography (CT scan) of the chest with angiography can be done. Contrast dye is injected into an intravenous line in the arm while the CT is being taken, and the pulmonary arteries can be visualized. There are some limitations of the test, especially if a pulmonary embolism involves the smaller arteries in the lung. There are risks with this test since some patients are allergic to the dye, and the contrast dye can be harsh on kidney function especially if the patient's kidney function (as measured by blood tests) is marginal. It may be wise to limit the patient's exposure to radiation, especially in pregnant patients. However, since pulmonary embolus can be fatal, even in pregnancy this test can be performed, preferably after the first trimester.

7.2.6. Ventilation – Perfusion Scans

Ventilation-perfusion scans (V/Q scans) use labeled chemicals to identify inhaled air into the lungs and match it with blood flow in the arteries. If a mismatch occurs, meaning that there is lung tissue that has good air entry but no blood flow, it may be indicative of a pulmonary embolus. These tests are read by a radiologist as having a low, moderate, or high probability of having a pulmonary embolism. There are limitations to the test, since there may be a 5%-10% risk that a pulmonary embolism exists even with a low probability V/Q result.

7.2.7. Venous Doppler Study

Ultrasound of the legs, also known as venous Doppler studies, may be used to look for blood clots in the legs of a patient suspected of having a pulmonary embolus. If a deep vein thrombosis exists, it can be inferred that chest pain and shortness of breath may be due to a pulmonary embolism.

7.2.8. Echocardiography

Echocardiography or ultrasound of the heart may be helpful if it shows that there is strain on the right side of the heart. If non-invasive tests are negative and the healthcare provider still has significant concerns, then the healthcare provider and the patient need to discuss the benefits and risks of treatment versus invasive testing like angiography.

7.3. How is Pulmonary Embolism Diagnosed?

A number of prospectively validated algorithms have been published that emphasize the use of different initial noninvasive tests in conjunction with ventilation–perfusion lung scanning. These include structured clinical assessment and serial venous ultrasonography; sensitive D-dimer assay, empirical clinical assessment and venous ultrasonography at presentation only; and clinical assessment, moderately sensitive D-dimer assay and serial venous ultrasonography. Based on these studies and others that have been discussed, such an algorithm is presented in. Algorithms that incorporate helical CT require further validation.

Figure 7.3 [KEARON, 2003] shows a diagnostic algorithm for pulmonary embolism (estimated frequencies of test results and associated prevalence of pulmonary embolism for a hypothetical cohort of 1000 outpatients). If a very sensitive D-dimer assay is used, it can be the first test performed: a negative result excludes pulmonary embolism regardless of clinical assessment category and a positive test can be followed by a ventilation–perfusion scan [2]. A ventilation–perfusion scan can be performed as the initial test without using clinical assessment of the probability of pulmonary embolism as part of the diagnostic process [3]. Pulmonary angiography or helical CT may be considered if the clinical assessment of pulmonary embolism probability is low, particularly if a D-dimer test has not been done [4]. Additional testing (e.g., helical CT, bilateral venography) may be considered if overall assessment suggests a high probability of pulmonary embolism (e.g., 50%–80%), symptoms are severe or cardiopulmonary reserve is poor [5]. Venography should be considered if there is an increased risk of a false-positive ultrasound result (e.g., previous venous thromboembolism, equivocal ultrasound findings, preceding findings suggest low probability of pulmonary embolism [e.g., 10%]) [6]. It is reasonable not to repeat ultrasound testing, or to do only 1 more ultrasound after 1 week, if preceding findings suggest a low probability of pulmonary embolism (e.g., 10%) [7]. If helical CT is used in place of ventilation–perfusion lung scanning: (i) intraluminal filling defects in segmental or larger pulmonary arteries are generally diagnostic for pulmonary embolism; (ii) all other findings (i.e., a normal CT scan or intraluminal filling defects confined to the subsegmental pulmonary arteries) are nondiagnostic and can be managed as shown for a nondiagnostic lung scan.

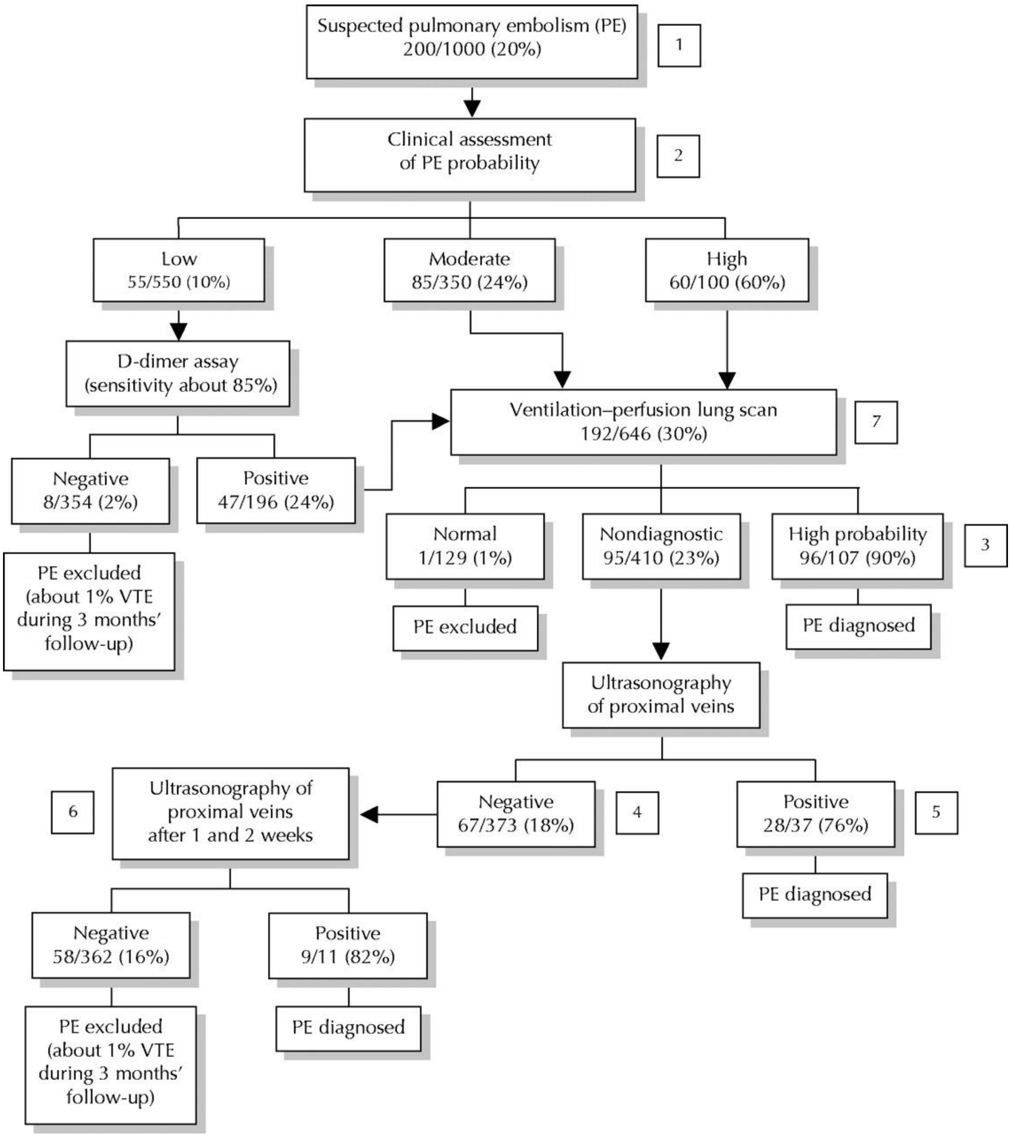
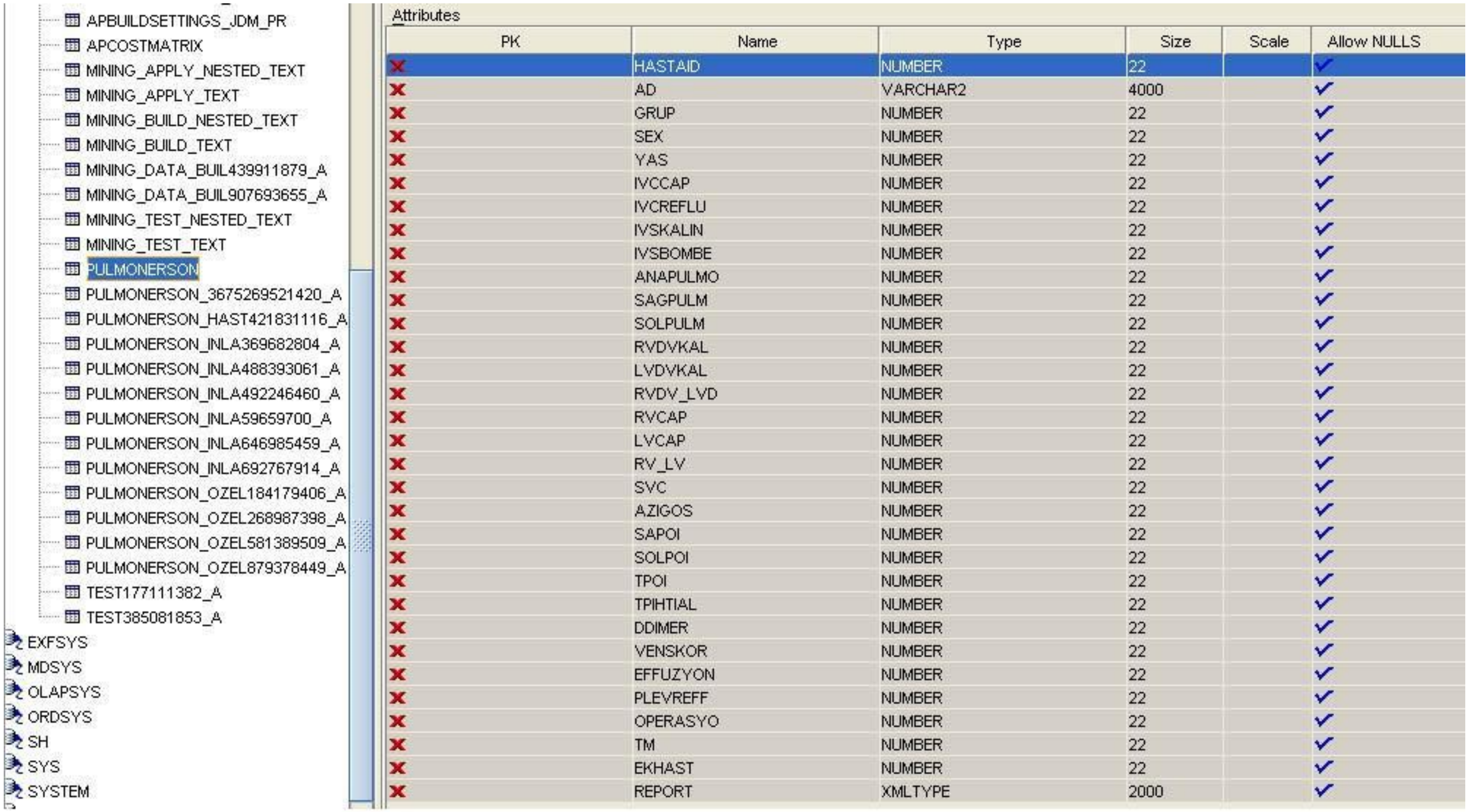

Suspected pulmonary embolism (PE) 200/1000 (20%)
1
Clinical assessment of PE probability
2
Low 55/550 (10%)
Moderate 85/350 (24%)
High 60/100 (60%)
D-dimer assay (sensitivity about 85%)
Ventilation–perfusion lung scan 192/646 (30%)
7
Negative 8/354 (2%)
Positive 47/196 (24%)
PE excluded (about 1% VTE during 3 months' follow-up)
Normal 1/129 (1%)
Nondiagnostic 95/410 (23%)
High probability 96/107 (90%)
3
PE excluded
PE diagnosed
Ultrasonography of proximal veins
6
Ultrasonography of proximal veins after 1 and 2 weeks
Negative 67/373 (18%)
4
Positive 28/37 (76%)
5
PE diagnosed
Negative 58/362 (16%)
Positive 9/11 (82%)
PE excluded (about 1% VTE during 3 months' follow-up)
PE diagnosed

Figure7.3. A Diagnostic Algorithm for Pulmonary

7.4. Medical Data

Clinical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. In this study 75 patients are evaluated for 32 attributes potentially contributing Pulmonary Embolism. Further explanation about attributes of this data set is given in Appendix B.

Because of data used by ODM consists of tables and views stored in an Oracle database Pulmonary Embolism data set is imported into the Oracle database (Figure 7.4 and Figure 7.5).

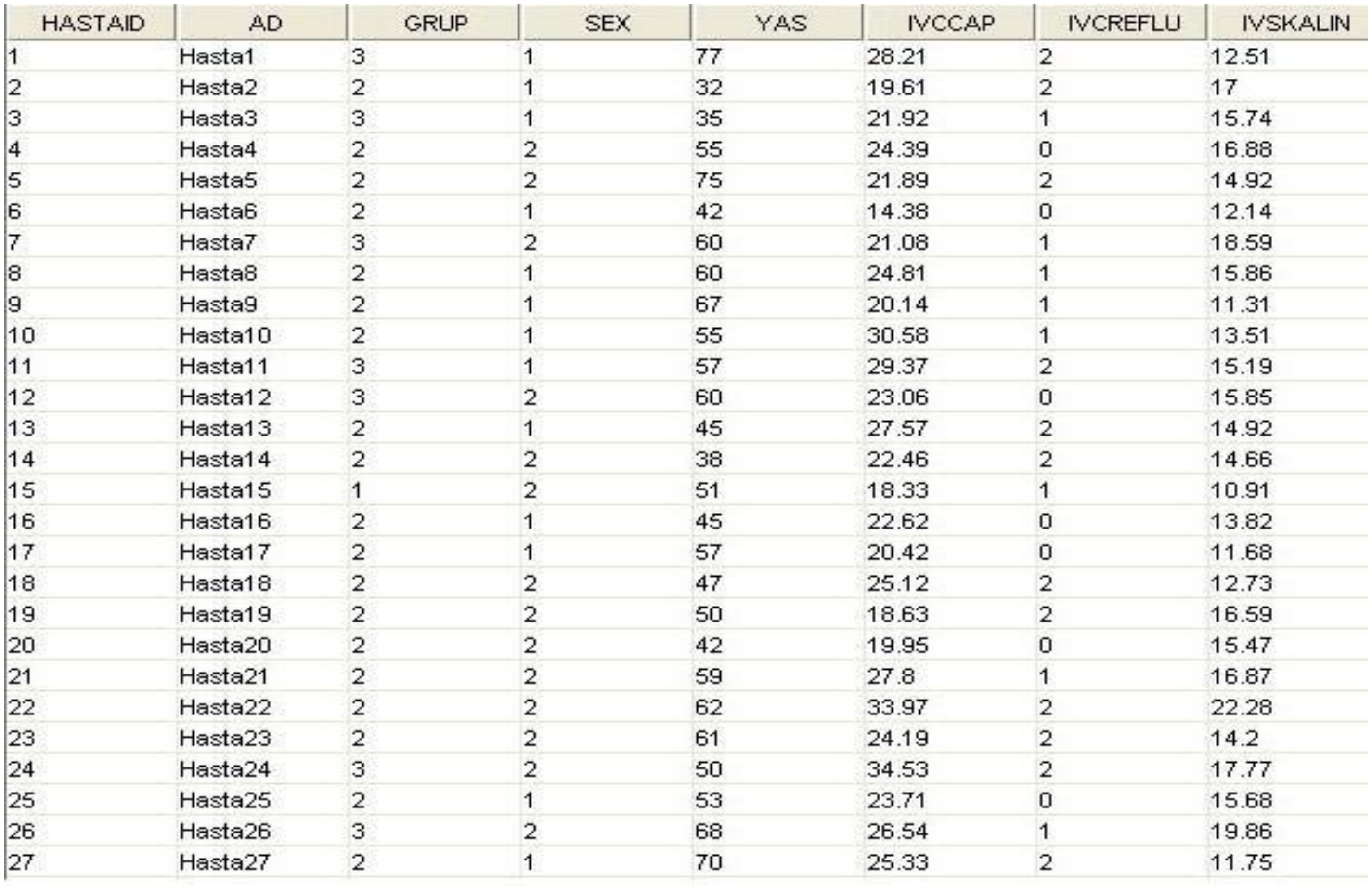


- APBUILDSETTINGS_JDM_PR
- APCOSTMATRIX
- MINING_APPLY_NESTED_TEXT
- MINING_APPLY_TEXT
- MINING_BUILD_NESTED_TEXT
- MINING_BUILD_TEXT
- MINING_DATA_BUIL439911879_A
- MINING_DATA_BUIL907693655_A
- MINING_TEST_NESTED_TEXT
- MINING_TEST_TEXT
- PULMONERSON
- PULMONERSON_3675269521420_A
- PULMONERSON_HAST421831116_A
- PULMONERSON_INLA369682804_A
- PULMONERSON_INLA488393061_A
- PULMONERSON_INLA492246460_A
- PULMONERSON_INLA59659700_A
- PULMONERSON_INLA646985459_A
- PULMONERSON_INLA692767914_A
- PULMONERSON_OZEL184179406_A
- PULMONERSON_OZEL268987398_A
- PULMONERSON_OZEL581389509_A
- PULMONERSON_OZEL879378449_A
- TEST177111382_A
- TEST385081853_A
- EXFSYS
- MDSYS
- OLAPSYS
- ORDSYS
- SH
- SYS
- SYSTEM

Attributes

PK	Name	Type	Size	Scale	Allow NULLS
✗	HASTAID	NUMBER	22		✓
✗	AD	VARCHAR2	4000		✓
✗	GRUP	NUMBER	22		✓
✗	SEX	NUMBER	22		✓
✗	YAS	NUMBER	22		✓
✗	IVCCAP	NUMBER	22		✓
✗	IVCREFLU	NUMBER	22		✓
✗	IVSKALIN	NUMBER	22		✓
✗	IVSBOMBE	NUMBER	22		✓
✗	ANAPULMO	NUMBER	22		✓
✗	SAGPULM	NUMBER	22		✓
✗	SOLPULM	NUMBER	22		✓
✗	RVDVKAL	NUMBER	22		✓
✗	LVDVKAL	NUMBER	22		✓
✗	RVDV_LVD	NUMBER	22		✓
✗	RVCAP	NUMBER	22		✓
✗	LVCAP	NUMBER	22		✓
✗	RV_LV	NUMBER	22		✓
✗	SVC	NUMBER	22		✓
✗	AZIGOS	NUMBER	22		✓
✗	SAPOI	NUMBER	22		✓
✗	SOLPOI	NUMBER	22		✓
✗	TPOI	NUMBER	22		✓
✗	TPIHTIAL	NUMBER	22		✓
✗	DDIMER	NUMBER	22		✓
✗	VENSKOR	NUMBER	22		✓
✗	EFFUZYON	NUMBER	22		✓
✗	PLEVREFF	NUMBER	22		✓
✗	OPERASYO	NUMBER	22		✓
✗	TM	NUMBER	22		✓
✗	EKHAST	NUMBER	22		✓
✗	REPORT	XMLTYPE	2000		✓

Figure 7.4 Pulmonary Embolism Data imported into Oracle Database



HASTAID	AD	GRUP	SEX	YAS	IVCCAP	IVCREFLU	IVSKALIN
1	Hasta1	3	1	77	28.21	2	12.51
2	Hasta2	2	1	32	19.61	2	17
3	Hasta3	3	1	35	21.92	1	15.74
4	Hasta4	2	2	55	24.39	0	16.88
5	Hasta5	2	2	75	21.89	2	14.92
6	Hasta6	2	1	42	14.38	0	12.14
7	Hasta7	3	2	60	21.08	1	18.59
8	Hasta8	2	1	60	24.81	1	15.86
9	Hasta9	2	1	67	20.14	1	11.31
10	Hasta10	2	1	55	30.58	1	13.51
11	Hasta11	3	1	57	29.37	2	15.19
12	Hasta12	3	2	60	23.06	0	15.85
13	Hasta13	2	1	45	27.57	2	14.92
14	Hasta14	2	2	38	22.46	2	14.66
15	Hasta15	1	2	51	18.33	1	10.91
16	Hasta16	2	1	45	22.62	0	13.82
17	Hasta17	2	1	57	20.42	0	11.68
18	Hasta18	2	2	47	25.12	2	12.73
19	Hasta19	2	2	50	18.63	2	16.59
20	Hasta20	2	2	42	19.95	0	15.47
21	Hasta21	2	2	59	27.8	1	16.87
22	Hasta22	2	2	62	33.97	2	22.28
23	Hasta23	2	2	61	24.19	2	14.2
24	Hasta24	3	2	50	34.53	2	17.77
25	Hasta25	2	1	53	23.71	0	15.68
26	Hasta26	3	2	68	26.54	1	19.86
27	Hasta27	2	1	70	25.33	2	11.75

Figure 7.5 Pulmonary Embolism Data Samples

CHAPTER 8

APPLICATION OF ORACLE DATA MINER ON THE CASE STUDY

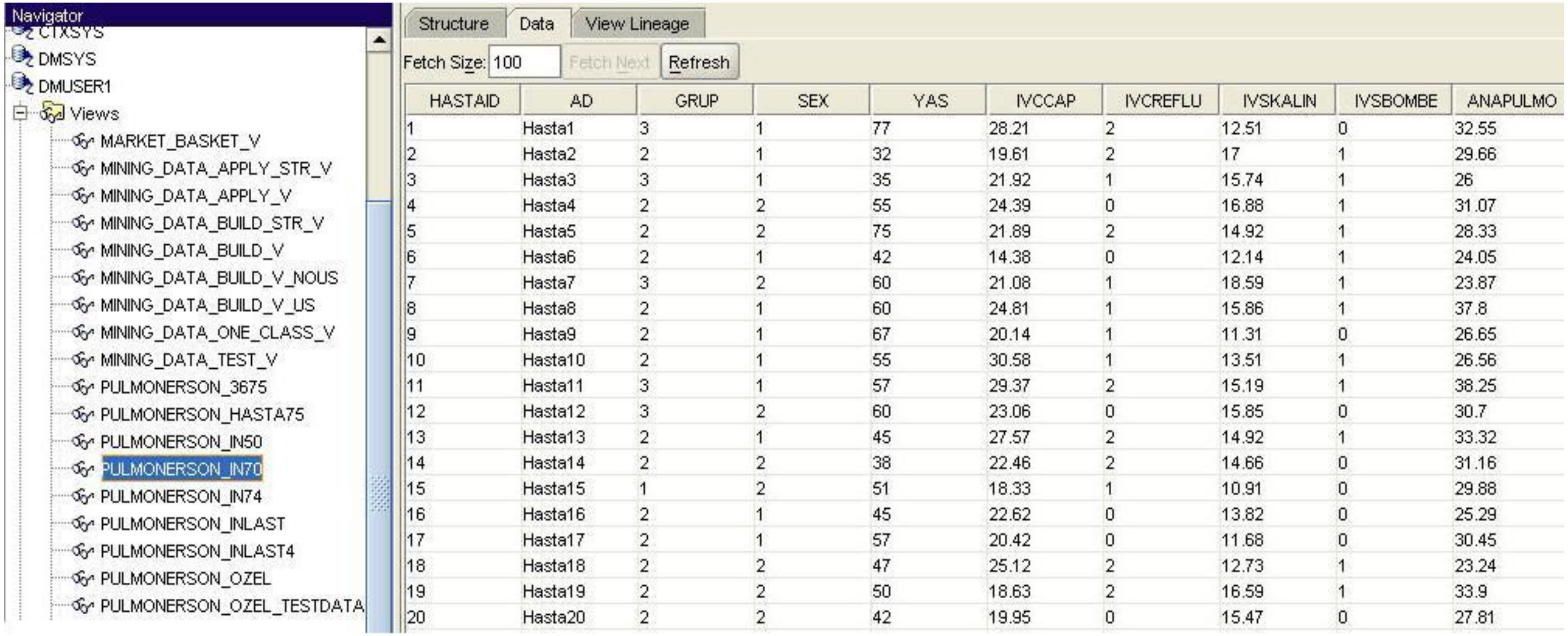
Oracle provides a powerful data mining infrastructure embedded directly into the database. This infrastructure, Oracle Data Mining (ODM), can be accessed by the graphical user interface (GUI) or Oracle Data Miner, Java API, SQL API, Predictive Analytics one-click data mining, and the Clementine data mining interface by SPSS and it automates the process of data mining which is explained in Chapter 2. In this chapter steps which are followed in the case study will be explained. Pulmonary Embolism data set will be used to construct the classification model.

8.1. Data Acquisition and Preparation

Under most circumstances for Supervised Learning problems, data are split into two mutually exclusive subsets, one for building the model, the other for test metrics step of the application. For this reason Pulmonary Embolism dataset is split into two subsets and two views are created in the Oracle database. First view which contains 70 patient’s data is used to build the classification model, second view which contains last 5 patient’s data is used to test model.

Partial view of the data for model building that consists of ten attributes and twenty patient’s data is given as Figure 8.1.

Partial view of the data for testing that consists of ten attributes and five patient’s data is given as Figure 8.2.

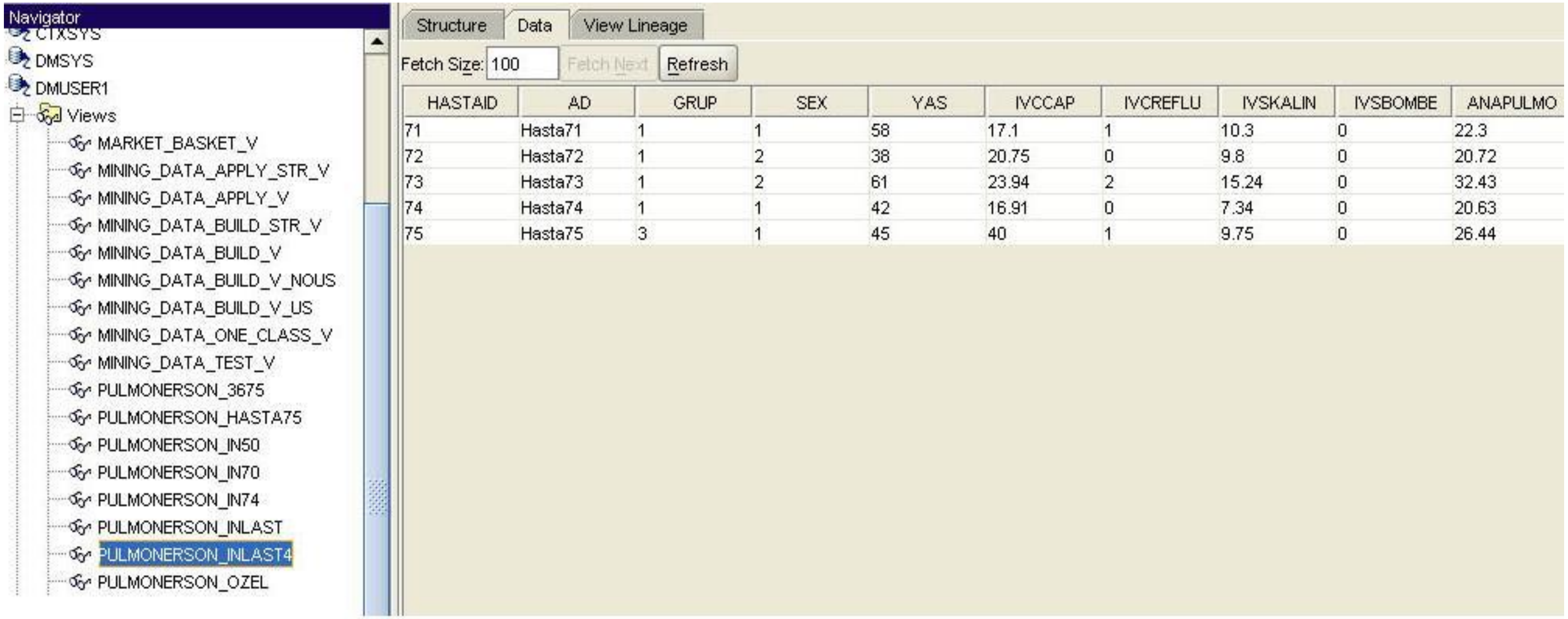


CTXSYS
DMSYS
DMUSER1
Views
MARKET_BASKET_V
MINING_DATA_APPLY_STR_V
MINING_DATA_APPLY_V
MINING_DATA_BUILD_STR_V
MINING_DATA_BUILD_V
MINING_DATA_BUILD_V_NOUS
MINING_DATA_BUILD_V_US
MINING_DATA_ONE_CLASS_V
MINING_DATA_TEST_V
PULMONERSON_3675
PULMONERSON_HASTA75
PULMONERSON_IN50
PULMONERSON_IN70
PULMONERSON_IN74
PULMONERSON_INLAST
PULMONERSON_INLAST4
PULMONERSON_OZEL
PULMONERSON_OZEL_TESTDATA

Fetch Size: 100

HASTAID	AD	GRUP	SEX	YAS	IVCCAP	IVCREFLU	IVSKALIN	IVSBOMBE	ANAPULMO
1	Hasta1	3	1	77	28.21	2	12.51	0	32.55
2	Hasta2	2	1	32	19.61	2	17	1	29.66
3	Hasta3	3	1	35	21.92	1	15.74	1	26
4	Hasta4	2	2	55	24.39	0	16.88	1	31.07
5	Hasta5	2	2	75	21.89	2	14.92	1	28.33
6	Hasta6	2	1	42	14.38	0	12.14	1	24.05
7	Hasta7	3	2	60	21.08	1	18.59	1	23.87
8	Hasta8	2	1	60	24.81	1	15.86	1	37.8
9	Hasta9	2	1	67	20.14	1	11.31	0	26.65
10	Hasta10	2	1	55	30.58	1	13.51	1	26.56
11	Hasta11	3	1	57	29.37	2	15.19	1	38.25
12	Hasta12	3	2	60	23.06	0	15.85	0	30.7
13	Hasta13	2	1	45	27.57	2	14.92	1	33.32
14	Hasta14	2	2	38	22.46	2	14.66	0	31.16
15	Hasta15	1	2	51	18.33	1	10.91	0	29.88
16	Hasta16	2	1	45	22.62	0	13.82	0	25.29
17	Hasta17	2	1	57	20.42	0	11.68	0	30.45
18	Hasta18	2	2	47	25.12	2	12.73	1	23.24
19	Hasta19	2	2	50	18.63	2	16.59	1	33.9
20	Hasta20	2	2	42	19.95	0	15.47	0	27.81

Figure 8.1 Partial View of the Pulmonary Embolism Dataset



Fetch Size: 100

HASTAID	AD	GRUP	SEX	YAS	IVCCAP	IVCREFLU	IVSKALIN	IVSBOMBE	ANAPULMO
71	Hasta71	1	1	58	17.1	1	10.3	0	22.3
72	Hasta72	1	2	38	20.75	0	9.8	0	20.72
73	Hasta73	1	2	61	23.94	2	15.24	0	32.43
74	Hasta74	1	1	42	16.91	0	7.34	0	20.63
75	Hasta75	3	1	45	40	1	9.75	0	26.44

Figure 8.2 Partial View of the Pulmonary Embolism Dataset which Contains 5 Patient's Data

In data acquisition step pulmonary embolism dataset is examined with the help of data mining histogram display and attribute importance. In data preparation step the pulmonary embolism dataset is split into training, or build, and test sets by random selection of cases by using the Split Transformation. Data mining histogram display and attribute importance is explained in the following two subsections.

8.1.1. Data Mining Histogram Display

Histograms are powerful way to visually and statistically explore the data. In this study histogram display is used to see where most of data is concentrated and show if the data is normally distributed. A normal distribution is the traditional bell-shaped curve described in introductory statistics courses. Histograms are very useful for identifying outliers, which are data points lying far outside the normal curve, and can adversely affect model performance.

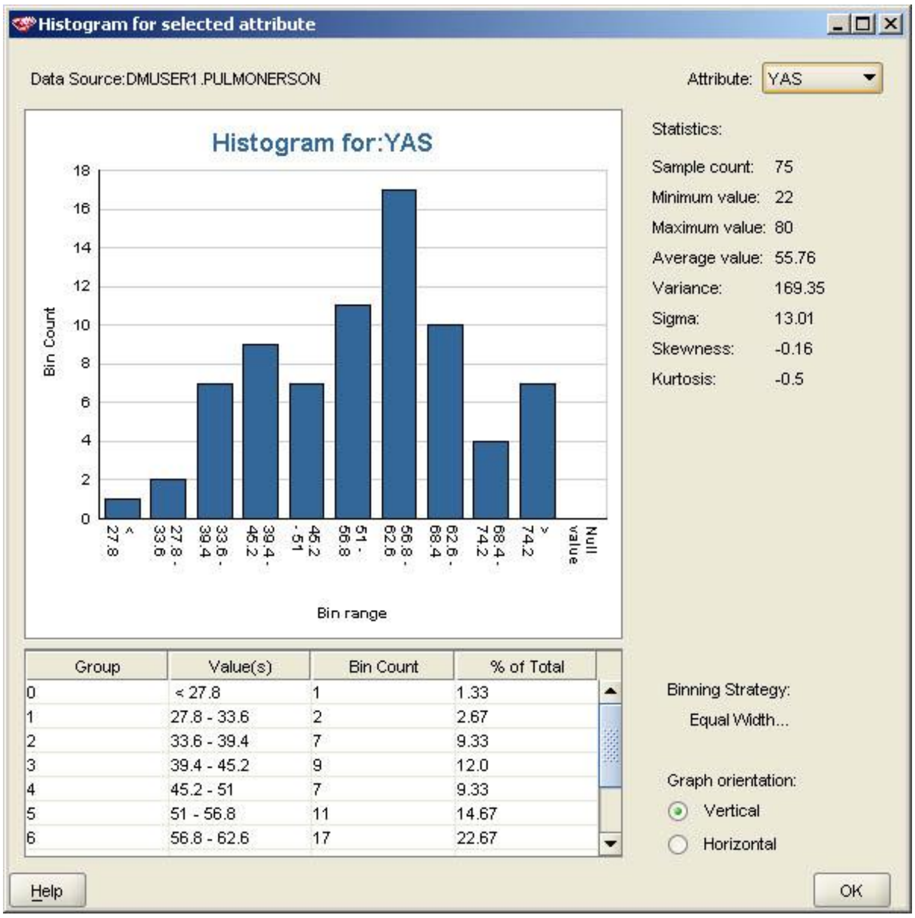
Figure 8.3 shows the histogram of age attribute in Pulmonary Embolism dataset. With number of bins set to 10, the histogram shows 10 bars or groups, age values for each group, number of cases in each bin or bin count, and percent of total.

Numerical attributes like AGE and DDIMER are divided into bins of equal width between the minimum and maximum. It can be seen that in this dataset proportion of patients between 56.8-62.6 ages is large.

8.1.2. Attribute Importance

If a data set has many attributes, it is likely that not all attributes will contribute to a predictive model. Indeed, some attributes may simply add noise, that is, they actually detract from the model's predictive value. ODM provides Attribute Importance (ai) that uses the Minimum Description Length algorithm to rank the attributes by significance in determining the target value.

The time required to build ODM classification models increases with the number of attributes. Attribute Importance identifies a proper subset of the attributes that are most relevant to predicting the target. Model building can proceed using the selected attributes only.



Data Source:DMUSER1.PULMONERSON

Attribute: YAS

Statistics:

Sample count:	75
Minimum value:	22
Maximum value:	80
Average value:	55.76
Variance:	169.35
Sigma:	13.01
Skewness:	-0.16
Kurtosis:	-0.5

Group	Value(s)	Bin Count	% of Total
0	< 27.8	1	1.33
1	27.8 - 33.6	2	2.67
2	33.6 - 39.4	7	9.33
3	39.4 - 45.2	9	12.0
4	45.2 - 51	7	9.33
5	51 - 56.8	11	14.67
6	56.8 - 62.6	17	22.67

Binning Strategy:
Equal Width...

Graph orientation:
● Vertical
○ Horizontal

Figure 8.3 Age Attribute Histogram

Using fewer attributes does not necessarily result in lost predictive accuracy. Using too many attributes (especially those that add noise) can affect the model and degrade its performance and accuracy. Mining using the smallest number of attributes can save significant computing time and may build better models.

Decision Tree and Adaptive Bayes Network algorithms do internal feature reduction, that is, they determine which attributes are important to build the model and use only those attributes. For these kinds of models, it is not necessary to create an ai model to reduce the number of features. Even for these algorithms, reducing the number of features may result in better performance.

An Attribute Importance (ai) model calculates rank and importance for each attribute. The rank of an attribute is an integer. Importance is a real number that may be negative. The rank or importance of an attribute allows you to select the attribute to be used in building models. The correct way to interpret attribute importance is that attributes with a greater numeric value for importance are relatively more important; the most important attribute has rank equal to 1. If the importance of attribute A is 10 times bigger than the importance of attribute B, it doesn’t mean that attribute A is 10 times more important than attribute B; all that it only means that attribute A is more important than attribute B. If the importance of an attribute is a negative number, then that attribute is not correlated with the target.

The problem in this study consists of identifying the high risky patients from among all patients for the purpose of increasing the early diagnosis of pulmonary embolism.

The data mining solution consists of building a predictive model from the results of the test that can be applied to the entire patient base in order to distinguish the high risky patients from the others. In this study the (ai) analysis is used to find the highest ranking attributes and build an effective classification model.

Histogram of pulmonary embolism dataset in attribute importance activity is given as Figure 8.4 where VENSKOR is the most important attribute, followed by TPIHTIAL and so on.

Since the available data is limited the result of (ai) is not used in this study. But, for larger dataset it may be useful.

Rank of pulmonary embolism dataset in attribute importance activity is given as Figure 8.5 where VENSKOR is the most important attribute, followed by TPIHTIAL and so on.

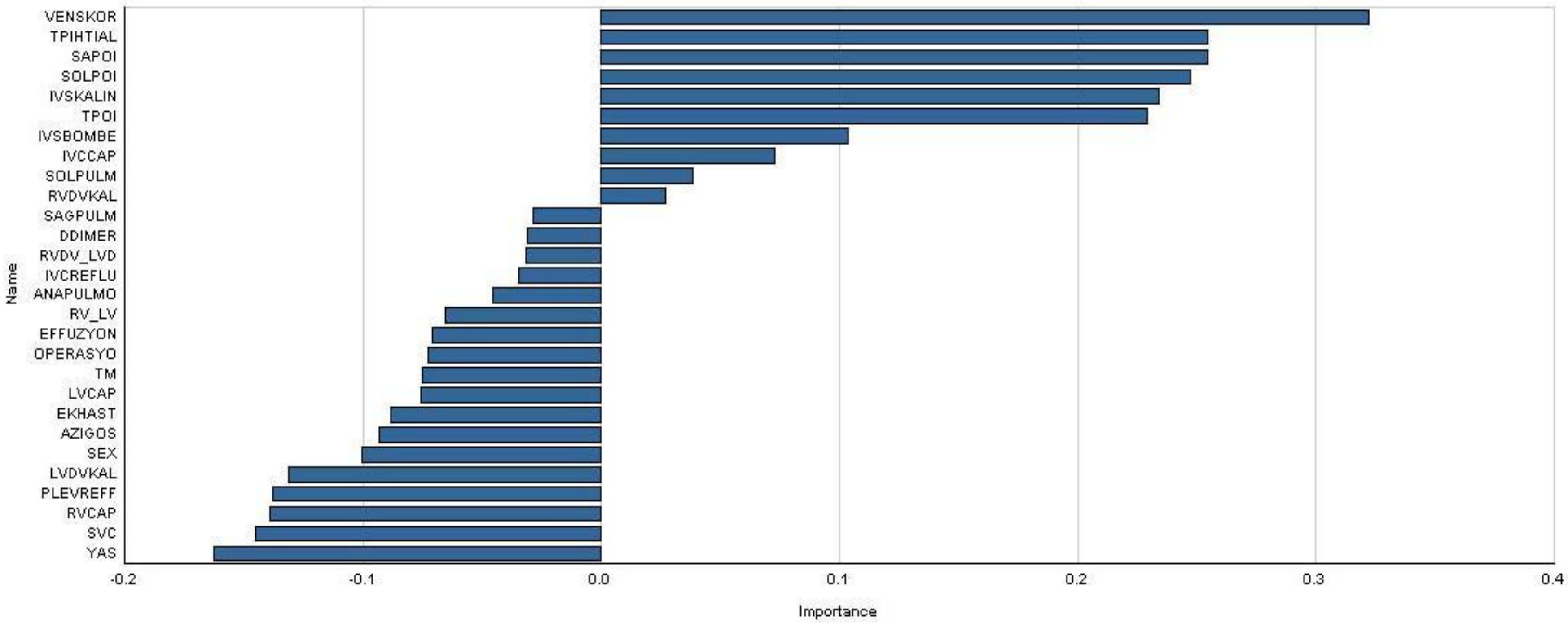
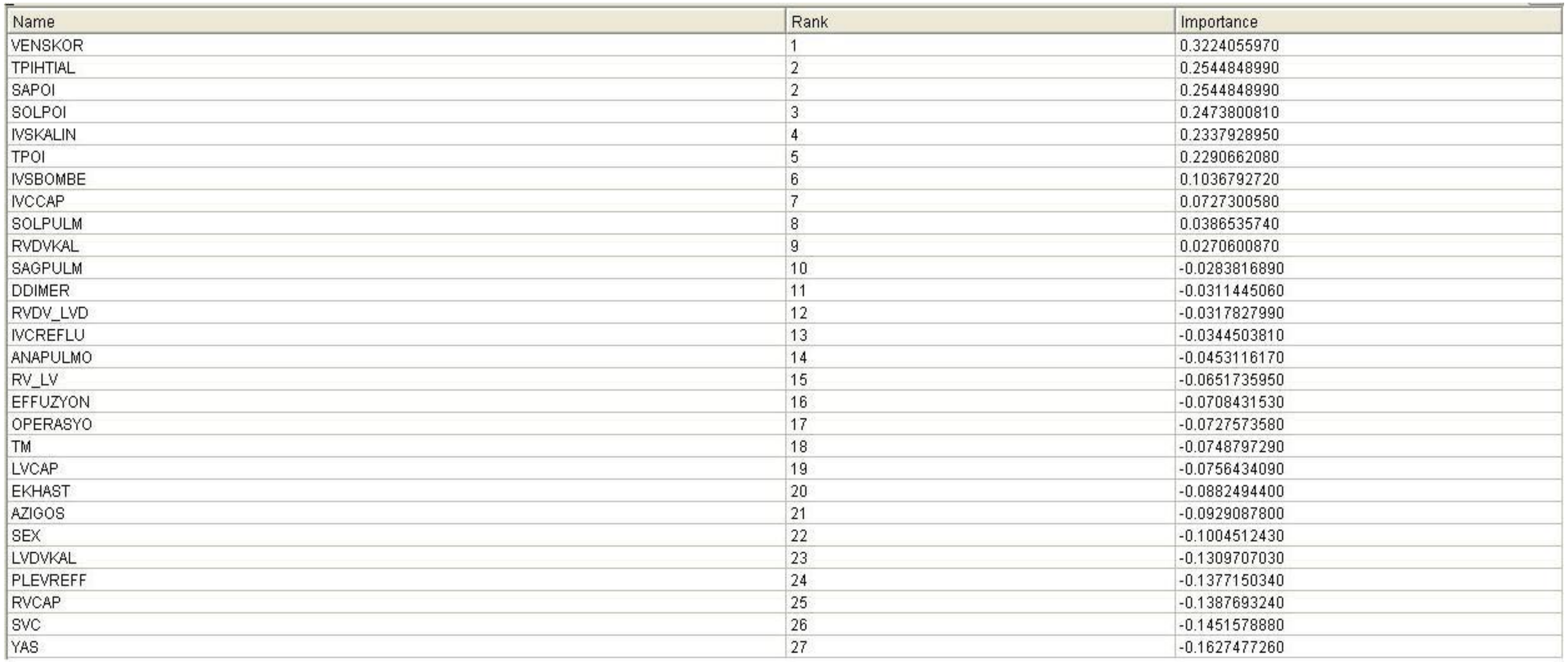


Figure 8.4 Attribute Ranking Activity: Histogram of Pulmonary Embolism Dataset



Name	Rank	Importance
VENSKOR	1	0.3224055970
TPIHTIAL	2	0.2544848990
SAPOI	2	0.2544848990
SOLPOI	3	0.2473800810
IVSKALIN	4	0.2337928950
TPOI	5	0.2290662080
IVSBOMBE	6	0.1036792720
IVCCAP	7	0.0727300580
SOLPULM	8	0.0386535740
RVDVKAL	9	0.0270600870
SAGPULM	10	-0.0283816890
DDIMER	11	-0.0311445060
RVDV_LVD	12	-0.0317827990
IVCREFLU	13	-0.0344503810
ANAPULMO	14	-0.0453116170
RV_LV	15	-0.0651735950
EFFUZYON	16	-0.0708431530
OPERASYO	17	-0.0727573580
TM	18	-0.0748797290
LVCAP	19	-0.0756434090
EKHAST	20	-0.0882494400
AZIGOS	21	-0.0929087800
SEX	22	-0.1004512430
LVDVKAL	23	-0.1309707030
PLEVREFF	24	-0.1377150340
RVCAP	25	-0.1387693240
SVC	26	-0.1451578880
YAS	27	-0.1627477260

Figure 8.5 Attribute Ranking Activity: Ranks of Pulmonary Embolism Dataset

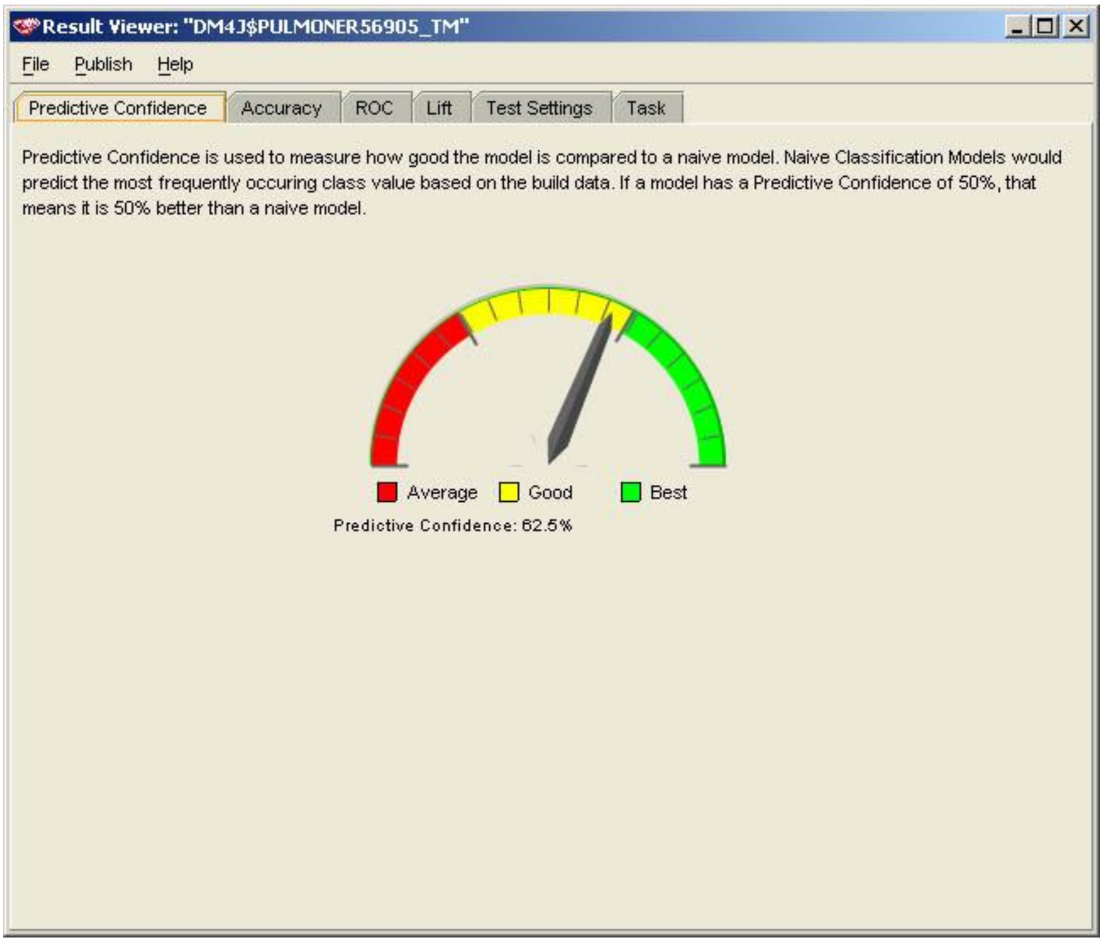
8.2. Building and Evaluation of Models

ODM provides the choice of four different classification models, Naive Bayes, Adaptive Bayes, Decision Tree, and Support Vector Machine. Each classification data mining activity has distinct advantages depending on the data and the problem. In this study several different algorithms are tried on the Pulmonary Embolism dataset and differences between the results are examined. In building mining activity for the classification problem, Naive Bayes model will be used and moved on to Support Vector Machines algorithm.

8.2.1. Building Naive Bayes Model on Pulmonary Embolism Dataset

The Naive Bayes algorithm looks at the historical data and calculates conditional probabilities for the target values by observing the frequency of attribute values and of combinations of attribute values. The further explanation about Naive Bayes algorithm is given in Chapter 5.

Naive Bayes classification model is built and tested with the help of ODM. The view which contains 70 patients' data is selected as input to mining activity. Patients' group attribute is selected as target and patients who have pulmonary embolism but don't have vena thrombosis is selected as the preferred target value. The results and their explanations are given below. Because, the dataset is small and the training data is randomly selected, these results may be slightly different if a larger dataset is used. The naive rule is important as it is commonly used as a baseline for evaluating the performance of classification models and Predictive Confidence is used to measure how good the model is compared to a naive model. In Figure 8.6 the Predictive Confidence is shown. The predictive confidence of 62.5% indicates that the Naive Bayes model which is built for the pulmonary embolism dataset is about 63% better than the naive rule.

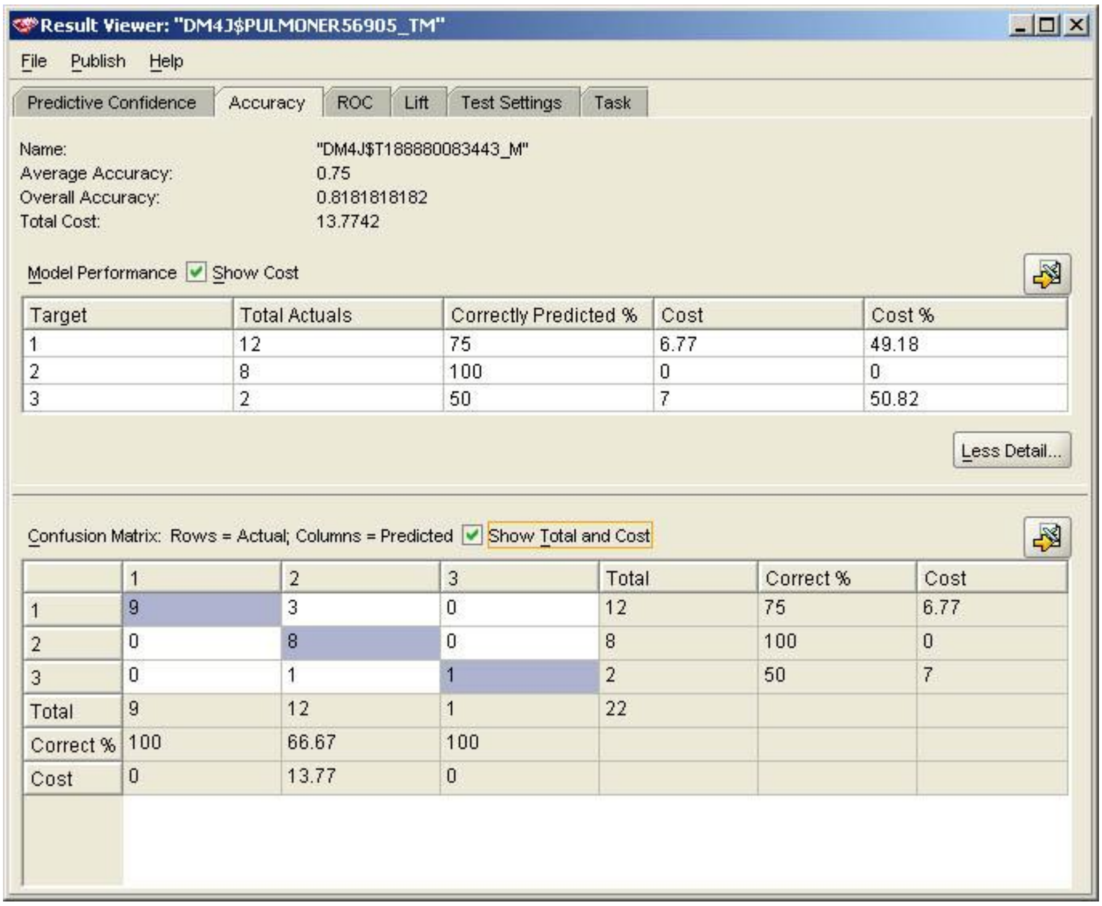


Predictive Confidence is used to measure how good the model is compared to a naive model. Naive Classification Models would predict the most frequently occuring class value based on the build data. If a model has a Predictive Confidence of 50%, that means it is 50% better than a naive model.

Predictive Confidence: 62.5%

Figure 8.6. Predictive Confidence of Naive Bayes Model

In Figure 8.7 the Accuracy tab of ODM's result viewer is shown. This tab contains classification matrix, also called the confusion matrix, where the model is applied to the hold-out test sample. The columns of confusion matrix in the lower half of Figure 8.7 are predictions made by the classification model and the rows are the actual data. Thus, the overall accuracy of model is about 18/22 = 81%, with 9 cases correctly classified as in group 1 which contains patients who don't have pulmonary embolism. 8 cases were accurately classified as in group 2 which contains patients who have pulmonary embolism but don't have vena thrombosis, and 4 were misclassified as in other groups. Similarly, 1 case was accurately classified as in group 3 which contains patients who have both pulmonary embolism and vena thrombosis. The cases that the model misclassified are the false-negative and false positive predictions.



Name: "DM4J$T188880083443_M"
Average Accuracy: 0.75
Overall Accuracy: 0.8181818182
Total Cost: 13.7742

Model Performance ☑ Show Cost

Target	Total Actuals	Correctly Predicted %	Cost	Cost %
1	12	75	6.77	49.18
2	8	100	0	0
3	2	50	7	50.82

Confusion Matrix: Rows = Actual; Columns = Predicted ☑ Show Total and Cost

	1	2	3	Total	Correct %	Cost
1	9	3	0	12	75	6.77
2	0	8	0	8	100	0
3	0	1	1	2	50	7
Total	9	12	1	22		
Correct %	100	66.67	100			
Cost	0	13.77	0			

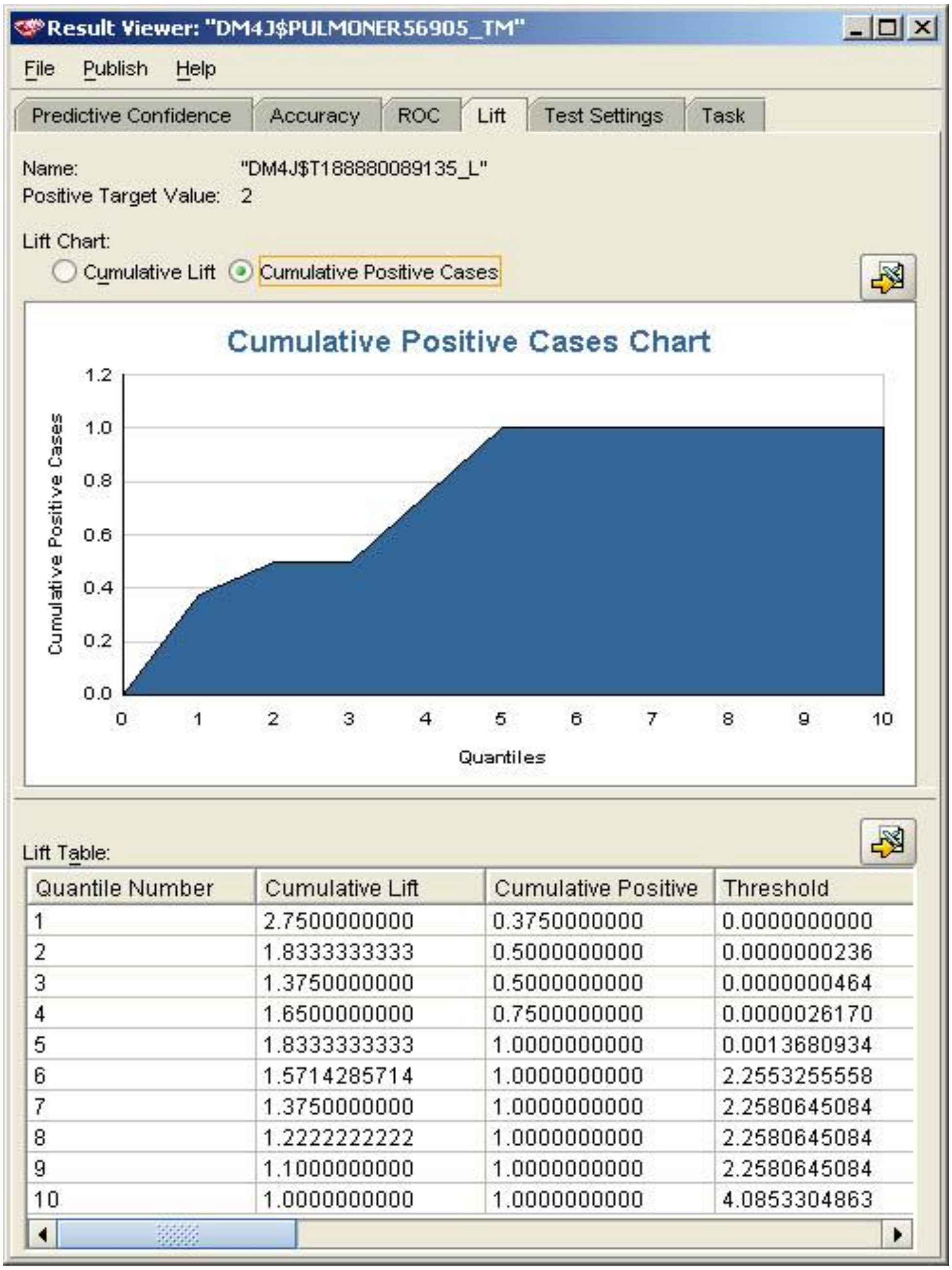
Figure 8.7. Model Accuracy of Naive Bayes Model

In Figure 8.8 the Lift tab of ODM's result viewer is shown. The lift tab is demonstrates two graphical interpretations of the results, the cumulative lift and cumulative positive cases chart. The lift tab is important because it measures the efficiency of the model. The lift curve, also called a gains curve or gains chart, is a popular technique in direct marketing.

In this case the classification model sifts through the records and sort them according to which patients are more likely to belong group 2. The lift curve is used to discover the smallest number of cases with the greatest probability of group 2 patients.

ODM applies the model to the test data, sorts the predicted results by probability, divides the ranked list into 10 equal parts and counts the actual positive values in each part.

The test results indicate that if the quantile number four is taken, the response is at least twice better than expected from random sampling. A good classifier gives high lift to help maximize the number of true diagnosis of pulmonary embolism.



Name: "DM4J$T188880089135_L"
Positive Target Value: 2

Lift Chart:
○ Cumulative Lift ● Cumulative Positive Cases

Lift Table:

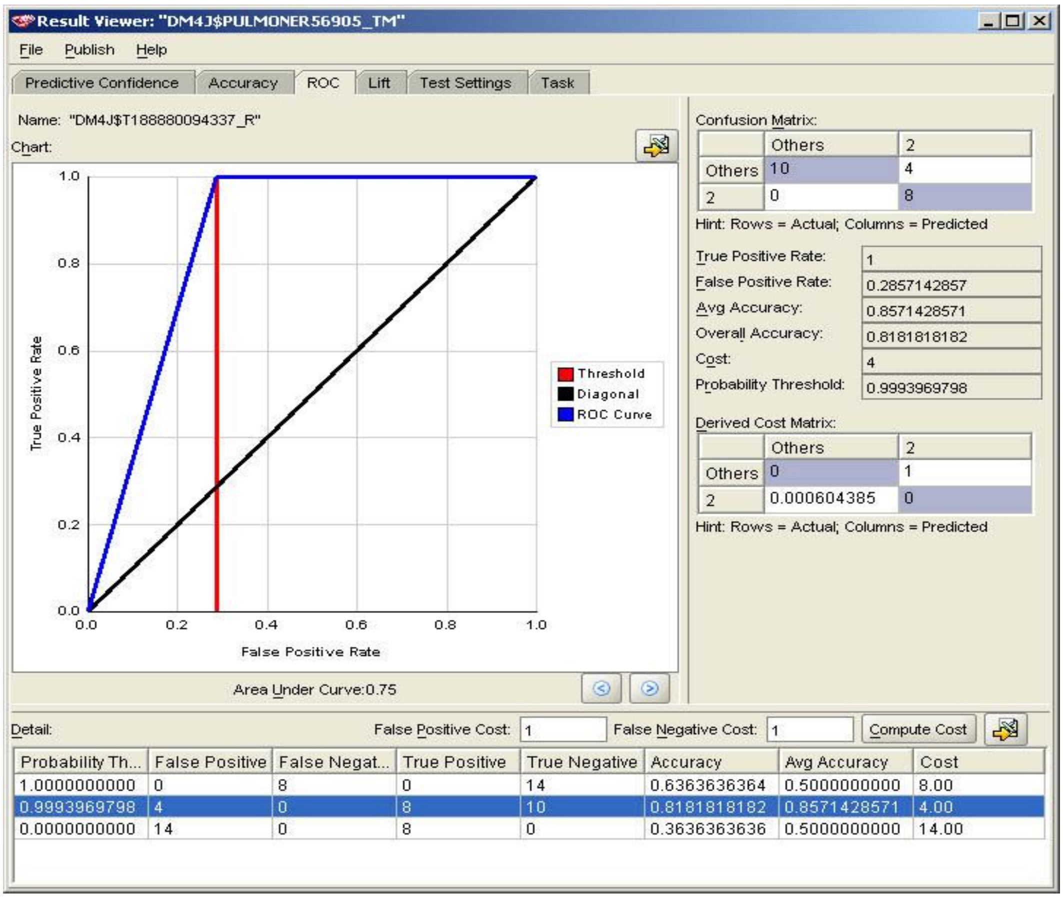
Quantile Number	Cumulative Lift	Cumulative Positive	Threshold
1	2.7500000000	0.3750000000	0.0000000000
2	1.8333333333	0.5000000000	0.0000000236
3	1.3750000000	0.5000000000	0.0000000464
4	1.6500000000	0.7500000000	0.0000026170
5	1.8333333333	1.0000000000	0.0013680934
6	1.5714285714	1.0000000000	2.2553255558
7	1.3750000000	1.0000000000	2.2580645084
8	1.2222222222	1.0000000000	2.2580645084
9	1.1000000000	1.0000000000	2.2580645084
10	1.0000000000	1.0000000000	4.0853304863

Figure 8.8. Model Lift of Naive Bayes Model

In Figure 8.9 the Receiver Operating Characteristic (ROC) tab of ODM’s result viewer is shown. The ROC curve uses the same metric on the y-axis as the lift curve, versus the number of true negatives correctly classified, for different cutoff levels. The default cutoff level is 0.5, meaning that if the probability assigned to a particular case is greater than 0.5, a positive prediction is made.

The false negatives in this model amount to 8 cases. The red vertical line is set to 0.5 probability threshold. The false negative value can be reduced as much as possible with the requirement that this model keep the total number of positive predictions under target number.

By moving the vertical line to the right, the increasing values are changed in the confusion matrix. Changing the probability threshold to 0.886 reduces the false negatives, and keeps the total number of positives.



Name: "DM4J$T188880094337_R"

Area Under Curve:0.75

Confusion Matrix:

	Others	2
Others	10	4
2	0	8

Hint: Rows = Actual; Columns = Predicted

True Positive Rate:	1
False Positive Rate:	0.2857142857
Avg Accuracy:	0.8571428571
Overall Accuracy:	0.8181818182
Cost:	4
Probability Threshold:	0.9993969798

Derived Cost Matrix:

	Others	2
Others	0	1
2	0.000604385	0

Hint: Rows = Actual; Columns = Predicted

Detail: False Positive Cost: 1 False Negative Cost: 1

Probability Th...	False Positive	False Negat...	True Positive	True Negative	Accuracy	Avg Accuracy	Cost
1.0000000000	0	8	0	14	0.6363636364	0.5000000000	8.00
0.9993969798	4	0	8	10	0.8181818182	0.8571428571	4.00
0.0000000000	14	0	8	0	0.3636363636	0.5000000000	14.00

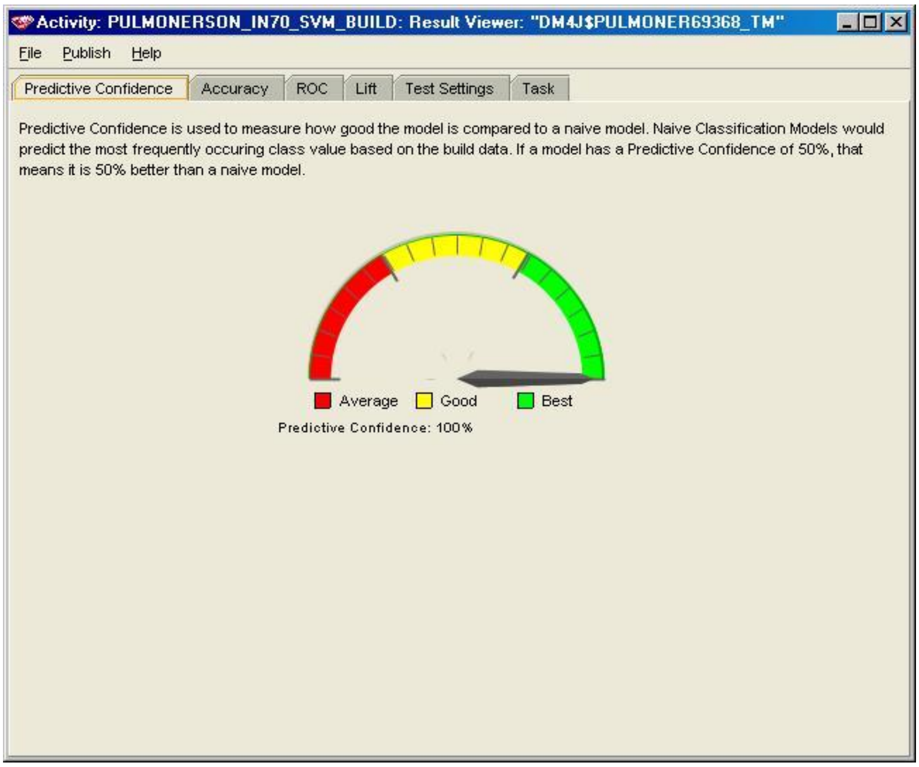
Figure 8.9. ROC of the Naive Bayes Model

8.2.2. Building Support Vector Machine Model on Pulmonary Embolism Dataset

SVM is an algorithm with strong regularization properties, that is, the optimization procedure maximizes predictive accuracy while automatically avoiding over-fitting of the training data. The further explanation about SVM algorithm is given in Chapter 4.

A SVM classification model is built and tested on same views, which are used for building Naive Bayes model, for the comparison of two classification algorithms. The view which contains 70 patients' data is selected as input to mining activity. Patients' group attribute is selected as target and patients who have pulmonary embolism but don't have vena thrombosis is selected as the preferred target value. The results and their explanations are given below. Because, the dataset is small and the training data is randomly selected, these results may be slightly different if a larger dataset is used. The naive rule is important as it is commonly used as a baseline for evaluating the performance of classification models and Predictive Confidence is used to measure how good the model is compared to a naive model. In Figure 8.10 the Predictive Confidence is shown. The predictive confidence of 100% indicates that the SVM model which is built for the pulmonary embolism dataset is about 100% better than the naive rule.

In Figure 8.11 the Accuracy tab of ODM's result viewer is shown. The columns are predictions made by the classification model and the rows are the actual data. The overall accuracy of model is about 100%, with 11 cases correctly classified as in group 1 which contains patients who don't have pulmonary embolism. 9 cases were accurately classified as in group 2 which contains patients who have pulmonary embolism but don't have vena thrombosis. Similarly, 7 cases were accurately classified as in group 3 which contains patients who have both pulmonary embolism and vena thrombosis.

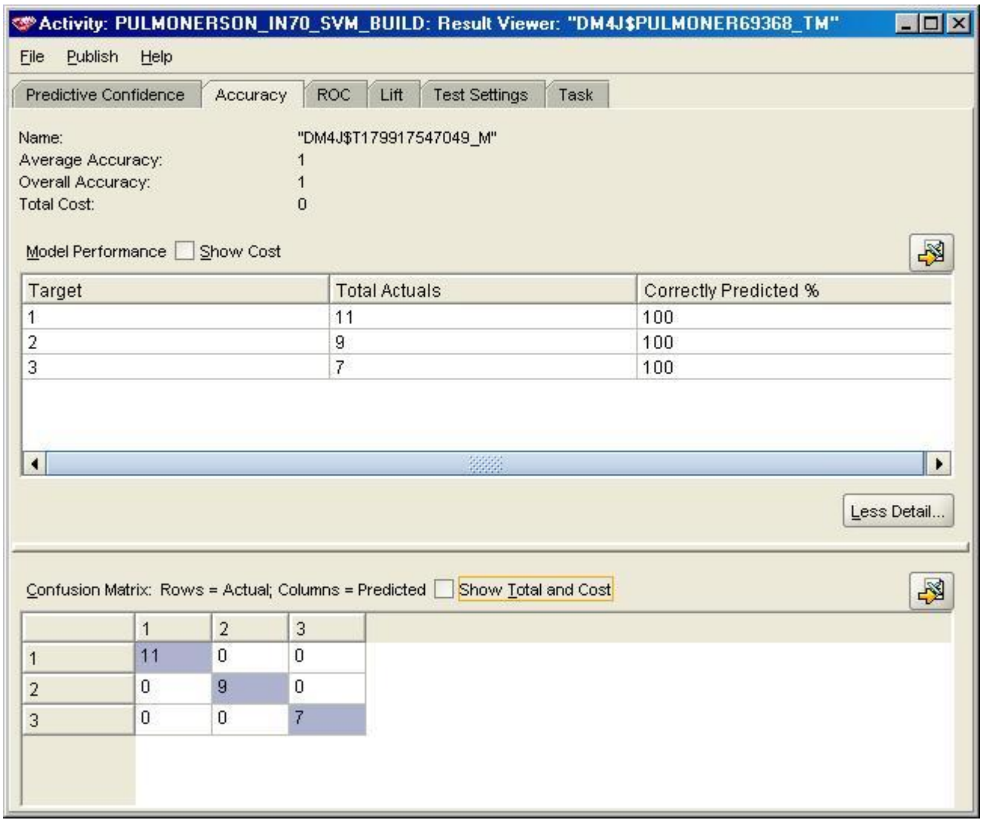

Activity: PULMONERSON_IN70_SVM_BUILD: Result Viewer: "DM4J$PULMONER69368_TM"
File
Publish
Help
Predictive Confidence
Accuracy
ROC
Lift
Test Settings
Task

Predictive Confidence is used to measure how good the model is compared to a naive model. Naive Classification Models would predict the most frequently occuring class value based on the build data. If a model has a Predictive Confidence of 50%, that means it is 50% better than a naive model.

🟥 Average 🟨 Good 🟩 Best

Predictive Confidence: 100%

Figure 8.10. Predictive Confidence of SVM Model



Name:	"DM4J$T179917547049_M"
Average Accuracy:	1
Overall Accuracy:	1
Total Cost:	0

Model Performance □ Show Cost

Target	Total Actuals	Correctly Predicted %
1	11	100
2	9	100
3	7	100

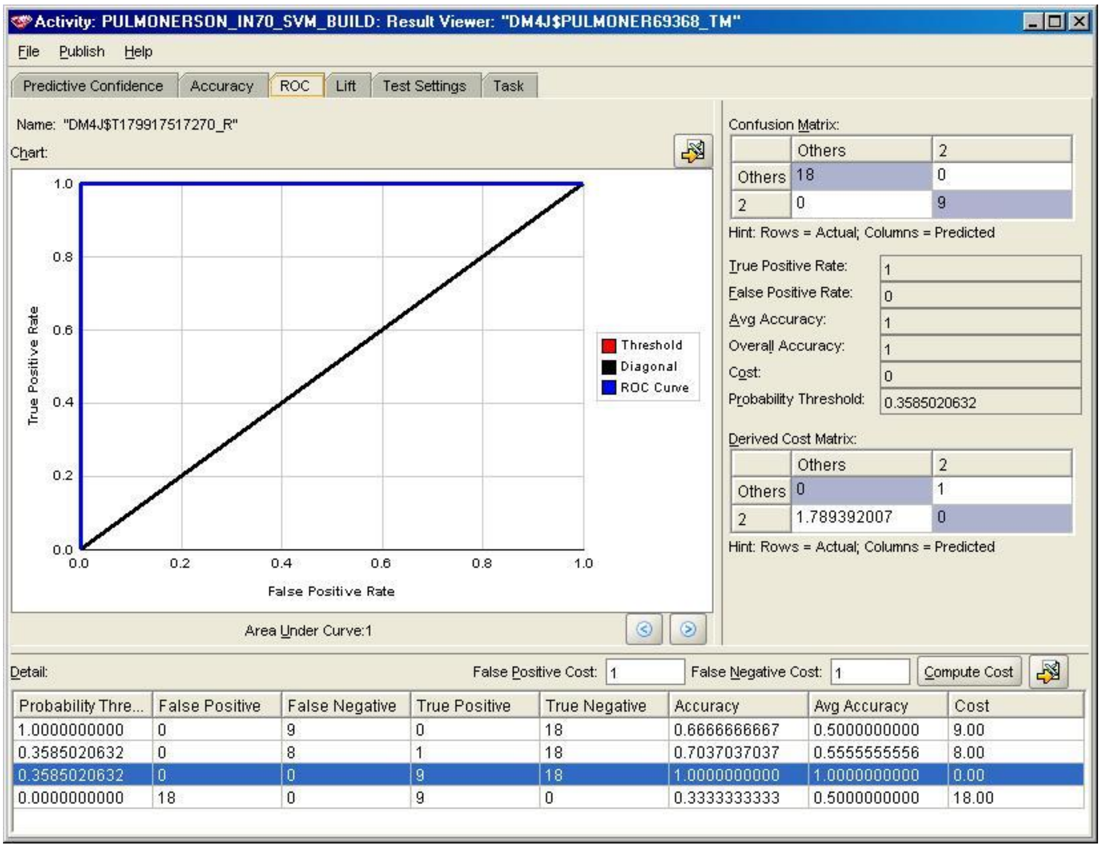
Confusion Matrix: Rows = Actual; Columns = Predicted □ Show Total and Cost

	1	2	3
1	11	0	0
2	0	9	0
3	0	0	7

Figure 8.11. Model Accuracy of SVM Model

In Figure 8.12 the Receiver Operating Characteristic (ROC) tab of ODM's result viewer is shown. The default cutoff level is 0.5, meaning that if the probability assigned to a particular case is greater than0.5, a positive prediction is made.

By increasing the cutoff level, the values are changed in the confusion matrix. Changing the probability threshold to a higher value reduces the false negatives, and keeps the total number of positives.



Name: "DM4J$T179917517270_R"

Area Under Curve:1

Confusion Matrix:

	Others	2
Others	18	0
2	0	9

Hint: Rows = Actual; Columns = Predicted

True Positive Rate: 1

False Positive Rate: 0

Avg Accuracy: 1

Overall Accuracy: 1

Cost: 0

Probability Threshold: 0.3585020632

Derived Cost Matrix:

	Others	2
Others	0	1
2	1.789392007	0

Hint: Rows = Actual; Columns = Predicted

Detail:

False Positive Cost: 1

False Negative Cost: 1

Probability Thre...	False Positive	False Negative	True Positive	True Negative	Accuracy	Avg Accuracy	Cost
1.0000000000	0	9	0	18	0.6666666667	0.5000000000	9.00
0.3585020632	0	8	1	18	0.7037037037	0.5555555556	8.00
0.3585020632	0	0	9	18	1.0000000000	1.0000000000	0.00
0.0000000000	18	0	9	0	0.3333333333	0.5000000000	18.00

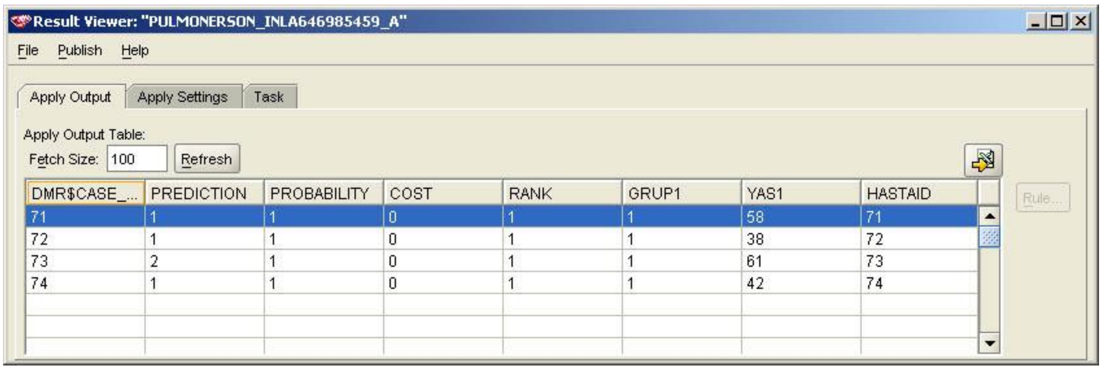
Figure 8.12. ROC of the SVM Model

8.3. Applying the Model

In this step of the data mining processes the pulmonary embolism classification model will be applied to new data. This is also known as scoring the data. When a model is applied to new data, the input data must be prepared and transformed in exactly the same way that the source data for the Build activity was prepared. In this study the view which contains last five patient’s data will be used as input of apply activity. The Apply activity is based on a Build activity, and the Build activity will pass to the Apply activity whatever knowledge is required to prepare the input data appropriately.

When the model is applied to a particular case, a score (normally a probability) is generated for each possible target value, producing a sorted list of values starting with the most likely value and going down to the least likely value. This list has only two entries if the target is binary, but is longer for multi-class problems (for example, group attribute of the pulmonary embolism dataset).

Figure 8.13 shows the scored data after applying Naive Bayes model on test view. The output table contains the HASTAID, the predictor variables; prediction, probability, cost and rank, and finally the additional attributes such as GROUP, YAS.



Apply Output Table:

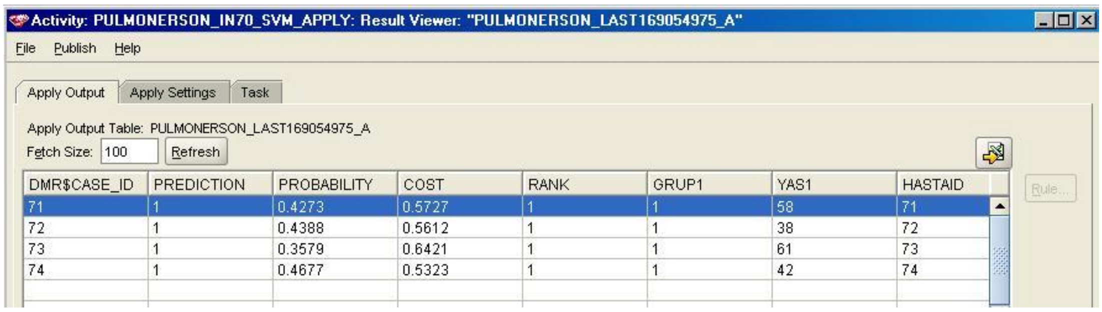
Fetch Size: 100

DMR$CASE_...	PREDICTION	PROBABILITY	COST	RANK	GRUP1	YAS1	HASTAID
71	1	1	0	1	1	58	71
72	1	1	0	1	1	38	72
73	2	1	0	1	1	61	73
74	1	1	0	1	1	42	74

Figure 8.13. Scored Data with Naive Bayes Model

Figure 8.14 shows the scored data after applying SVM model on test view. The output table contains the HASTAID, the predictor variables; prediction, probability, cost and rank, and finally the additional attributes such as GROUP, YAS.

For both figures the prediction is the most likely target value for that case, and the probability is the confidence in that prediction. Cost represents the cost of a wrong prediction, with low cost meaning high probability.


Activity: PULMONERSON_IN70_SVM_APPLY: Result Viewer: "PULMONERSON_LAST169054975_A"
File
Publish
Help
Apply Output
Apply Settings
Task
Refresh
Rule...

Apply Output Table: PULMONERSON_LAST169054975_A

Fetch Size: 100

DMR$CASE_ID	PREDICTION	PROBABILITY	COST	RANK	GRUP1	YAS1	HASTAID
71	1	0.4273	0.5727	1	1	58	71
72	1	0.4388	0.5612	1	1	38	72
73	1	0.3579	0.6421	1	1	61	73
74	1	0.4677	0.5323	1	1	42	74

Figure 8.14. Scored Data with SVM Model

8.4. Comparison of Models

Both Support Vector Machine (SVM) and Naive Bayes algorithms are suitable for classification problems. The SVM algorithms differ from the Naive Bayes by its adaptability for diverse types of data. SVM can be used to solve regression problems, where the predictor values are continuous as opposed to discrete data types.

Another advantage to using SVM is that it can be used to predict outcomes based on text data, so it can be used as part of the classification model for descriptive data such as clinical notes for hospital patients.

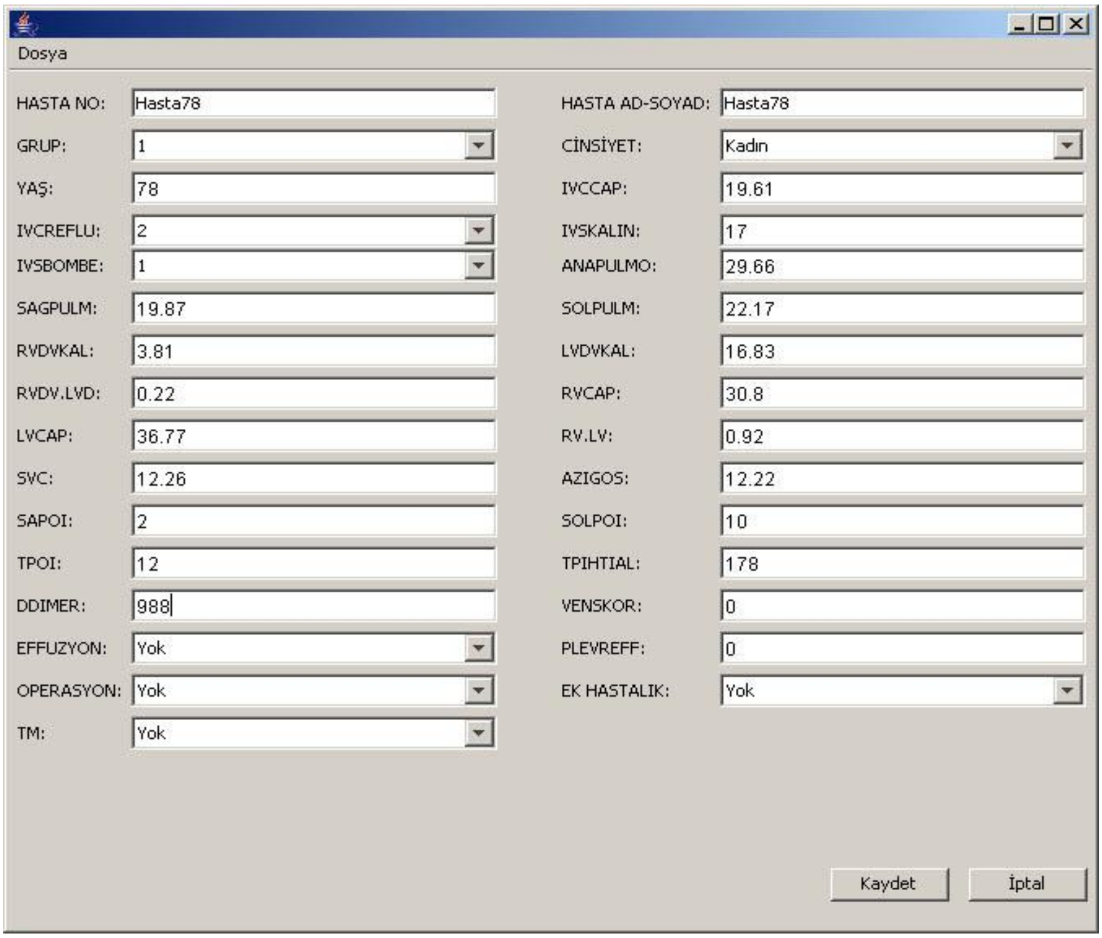
In this study both Naive Bayes and SVM algorithms are applied on the pulmonary embolism dataset and results are explained above. The predictive confidence of the Naive Bayes model which is built for the pulmonary embolism dataset is about 63% better than the naive rule but on the other side SVM model which is built for the pulmonary embolism dataset is about 100% better than the naive rule.

CHAPTER 9

TEXT MINING ON THE MEDICAL DATA

9.1. Application Interface

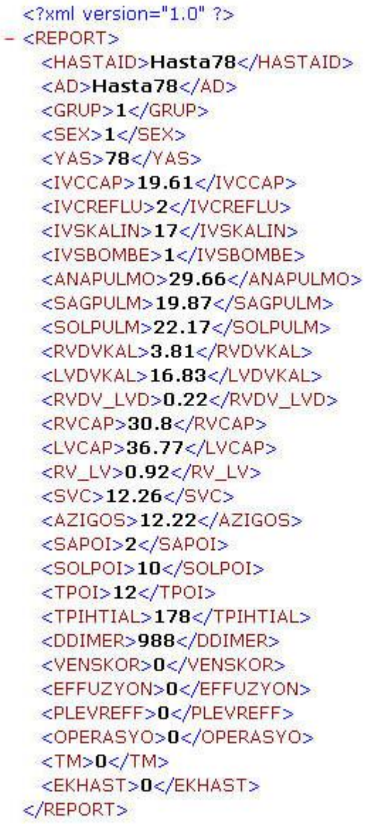
Figure 9.1 shows user interface which is implemented to get information about patient's history, laboratory findings and radiology report details. This data which is taken from the user is saved into the database in XML format.



HASTA NO:	Hasta78	HASTA AD-SOYAD:	Hasta78
GRUP:	1	CİNSİYET:	Kadın
YAŞ:	78	IVCCAP:	19.61
IVCREFLU:	2	IVSKALIN:	17
IVSBOMBE:	1	ANAPULMO:	29.66
SAGPULM:	19.87	SOLPULM:	22.17
RVDVKAL:	3.81	LVDVKAL:	16.83
RVDV.LVD:	0.22	RVCAP:	30.8
LVCAP:	36.77	RV.LV:	0.92
SVC:	12.26	AZIGOS:	12.22
SAPOI:	2	SOLPOI:	10
TPOI:	12	TPIHTIAL:	178
DDIMER:	988	VENSKOR:	0
EFFUZYON:	Yok	PLEVREFF:	0
OPERASYON:	Yok	EK HASTALIK:	Yok
TM:	Yok		

Figure 9.1. A Sample Data for Application Interface

Figure 9.2 shows saved XML format of the report. A processor between the database and user interface reads data and save useful information into a new table.



```
  <?xml version="1.0" ?>
- <REPORT>
    <HASTAID>Hasta78</HASTAID>
    <AD>Hasta78</AD>
    <GRUP>1</GRUP>
    <SEX>1</SEX>
    <YAS>78</YAS>
    <IVCCAP>19.61</IVCCAP>
    <IVCREFLU>2</IVCREFLU>
    <IVSKALIN>17</IVSKALIN>
    <IVSBOMBE>1</IVSBOMBE>
    <ANAPULMO>29.66</ANAPULMO>
    <SAGPULM>19.87</SAGPULM>
    <SOLPULM>22.17</SOLPULM>
    <RVDVKAL>3.81</RVDVKAL>
    <LVDVKAL>16.83</LVDVKAL>
    <RVDV_LVD>0.22</RVDV_LVD>
    <RVCAP>30.8</RVCAP>
    <LVCAP>36.77</LVCAP>
    <RV_LV>0.92</RV_LV>
    <SVC>12.26</SVC>
    <AZIGOS>12.22</AZIGOS>
    <SAPOI>2</SAPOI>
    <SOLPOI>10</SOLPOI>
    <TPOI>12</TPOI>
    <TPIHTIAL>178</TPIHTIAL>
    <DDIMER>988</DDIMER>
    <VENSKOR>0</VENSKOR>
    <EFFUZYON>0</EFFUZYON>
    <PLEVREFF>0</PLEVREFF>
    <OPERASYO>0</OPERASYO>
    <TM>0</TM>
    <EKHAST>0</EKHAST>
  </REPORT>
```

Figure 9.2. Generated XML File for the Sample Data

CHAPTER 10

SUMMARY AND CONCLUSIONS

10.1. Summary

In the study, the supervised (descriptive) data mining techniques using ORACLE Data Mining package is applied on the pulmonary embolism data set and some predictions are gained for better clinical decision-making.

On a set of real medical data, the following steps are performed:

a) In data preparation step, a user interface is implemented to take the data which contains information about patient’s history, laboratory findings and radiology report details. This data which is taken from the user is saved into the database in XML format. A processor between the database and user interface reads data and save useful information into a new table.

b) In modeling step, data mining classification algorithms such as Naive Bayes, Support Vector Machine modeling techniques have been compared in order to find the best technique for Pulmonary Embolism data.

c) In evaluation and deployment step, pulmonary embolism classification model is applied to new data and results are discussed.

10.2. Conclusions

The pulmonary embolism dataset, that is available, has 32 attributes that contains information about patient’s history, laboratory findings and radiology report details.

By applying classification techniques on this dataset some useful predictions can be gained about pulmonary embolism patients.

The predictive confidence of the classification algorithms that are used is over 85% for the pulmonary embolism dataset. This value is acceptable for better clinical decision-making. For optimal efficiency, choice of the initial diagnostic test should be guided by clinical assessment of the probability of pulmonary embolism and by patient characteristics that may influence test accuracy. The predictions which is gained in this study enables pulmonary embolism to be diagnosed or excluded in a minimum number of steps. However, even with the appropriate use of combinations of noninvasive tests, it is often not possible to definitively diagnose or exclude pulmonary embolism at initial presentation. Most of these patients can be managed safely without treatment or pulmonary angiography by repeating ultrasound testing of the proximal veins after one and 2 weeks to detect evolving deep vein thrombosis. A diagnostic algorithm for pulmonary embolism was already given as Figure 7.3. As it is in that figure, process is quite complicated. By using the approach and the tool of the thesis study, the process may be simplified.

10.3. Extension of the Study

A further extension of this work is to integrate this study into a Hospital Information System (HIS), Radiology information system (RIS), laboratory information system (LIS) and make available classification to provide real time assistance to practitioners.

The key objective of the study was to demonstrate how ORACLE Data Miner software package is applied on a set of actual medical data. That is performed successfully.

Unfortunately the available dataset was very limited. Trial with larger sets will give a better insight about the application.

REFERENCES

[1] **AWAD, E.M.** and **H.M.GAHAZIRI** (2004), *Knowledge Management, Pearson, Prentice Hall.*

[2] **BECERRA-FERNANDEZ, I** et. al. (2004), *Knowledge Management, Pearson, Prentice Hall.*

[3] **BERRY, W.B.** (2004), *Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity, Springer-Verlag New York.*

[4] **BURUNCUK, G.** (2006) *Data Mining for Customer Segmentation and Profiling: A Case Study for a Fast Moving Consumer Goods (FMCG) Company, MS Thesis Study, Istanbul, Bogazici Press.*

[5] **CRISP-DM:** http://www.crisp-dm.org

[6] **COUSEAULT, C.R.** (2004) A *Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions Thesis Study, PhD Thesis Study, Georgia Institute of Technology.*

[7] **ELDER, J.F. and D.W, ABOOTT** (1998), *A Comparison of Leading Data Mining Tools, Fourth International Conference on Knowledge Discovery & Data Mining, New York.*

[8] **FELDMAN, R. and J. SANGER** (2007), *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press.*

[9] **FERNEDA, E.** (2008), *Emerging technologies of text mining: techniques and applications, IGI Global.*

[10] **GANZERT, S**. et. al. (2002), *Analysis of Respiratory Pressure-Volume Curves in Intensive Care Medicine Using Inductive Machine Learning, Artificial Intelligence in Medicine, 26(2002), s.69-86.*

[11] **GIUDICI, P.** (2003), *Applied Data Mining*, Wiley.

[12] **HAMM, C.K.** (2007), *Oracle Data Mining – Mining Gold from Your Data Warehouse, Rampant TechPress.*

[13] **HAN, J. and M. KAMBER** (2000), *Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.*

[14] **HAND, D.** et. al. (2001), *Principles of Data Mining, Prentice Hall.*

[15] **HEARST, M.** (2003), *What is Text Mining?*, SIMS, UC Berkeley.

[16] **HONIGMAN, B**., et. al. (2001), *A computerized method for identifying incidents associated with adverse drug events in outpatients, International Journal of Medical Informatics, 61(2001), s. 21-32.*

[17] **COULTER, D.** et. al. (2001), *Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study, BMJ, 322, 19 MAY 2001, p:1207-1209.*

[18] http://www.dangelolaw.com/ortho_evera2.html

[19] http://en.wikipedia.org/wiki/Pulmonary_embolism

[20] http://www.mayoclinic.com/health/pulmonary-embolism/DS00429

[21] **WEDRO, B (2008)**, http://www.medicinenet.com/pulmonary_embolism/article.htm

[22] **IBM (2001)**, *Mining Your Own Business in Health Care Using DB2 Intelligent Miner for Data, IBM CORPORATION.*

[23] **IBM (2002)**, *Simple Integration of Advanced Data Mining Functions, IBM CORPORATION.*

[24] **KARANIKAS, H. and T. MAVROUDAKIS** (2005)**,** *Text Mining Software Survey, RANLP Text Mining Workshop.*

[25] **KANTARDZIC, M.** (2003)*, Data Mining: Concepts, Models, Methods, and Algorithms,* John Wiley & Sons, New York.

[26] **KEARON, C.** (2003)**,** *Diagnosis of pulmonary embolism, CMAJ, January 21, 2003; 168 (2).*

[27] **KING, M.A. and J.F. ELDER** (1998), *Evaluation of Fourteen Desktop Data Mining Tools, Department of Systems Engineering University of Virginia.*

[28] **MATHIAK, B. and S. ECKSTEIN (2006),** *Five Steps to Text Mining in Biomedical Literature. Second International Symposium on Semantic Mining in Biomedicine 2006.*

[29] **MILWARD,D.,** et. al. (2006), *Flexible Text Mining Strategies for Drug Discovery, Second International Symposium on Semantic Mining in Biomedicine 2006.*

[30] **ORACLE DATA MINING (2005)**, *Oracle Data Mining Concepts, Oracle.*

[31] **ORACLE TEXT MINING (2005)**, *Text Mining with Oracle Text, Oracle.*

[32] **ORACLE (2006)**, *Oracle 10g Release 2 Data Mining Tutorial, Oracle.*

[33] **PONOMARENKO, J**., et. al. (2002), *Mining DNA sequences to predict sites which mutations cause genetic diseases, Knowledge-Based Systems, 15(2002) 225-233*

[34] **RUD, O.** (2001), *Data Mining Cookbook*, John Wiley & Sons.

[35] **STÜHLİNGER, W.,** et. al. (2000), *Intelligent Data Mining for Medical Quality Management, CiteSeer Scientific Literature Digital Library.*

[36] **THEARLING, K.** (2000), *An Introduction to Data Mining: Discovering hidden value in your data warehouse.*
http://www.thearling.com/text/dmwhite/dmwhite.htm

[37] **TWO CROWS (2002)**, *Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation.*

[38] **URAMATO, N.,** et. al. (2004), *A text-mining system for knowledge discovery from biomedical documents, IBM Systems Journal, VOL 43, N0 3.*

[39] **WASAN, K.,** et. al. (2006), *The impact of data mining techniques on medical diagnostics, Data Science Journal, Vol. 5 (2006) pp.119-126.*

[40] **WING, V. and S.CHO (**1999), *PhD Thesis Study, Hong Kong University of Science and Technology, Hong Kong.*

[41] **Yİ, S.W.** (2003), *Introduction to medical data mining, Science Direct, Journal of Biomedical Informatics, Volume 39, Issue 3, June 2006, Pages 249-251.*

[42] **ZAHID, S. and H., ZAIDI** (2002), *Distributed Data Mining From Heterogeneous Healthcare Data Repositories: Towards an Intelligent Agent-Based Framework.*

[43] **ZHOU, Z. (2003)**, *Three Perspectives of Data Mining, Artificial Intelligence, 143(2003), p:139-146.*

APPENDIX A

COMMERCIAL DATA MINING PRODUCTS

Some of the commercial data mining products available are summarized below [Kantardzic, 2003]

- **AgentBase/Marketeer**

AgentBase/Marketeer is, according to its designers, the industry's first second-generation data-mining product. It is based on emerging intelligent-agent technology. The system comes with a group of wizards to guide a user through different stages of data mining. This makes it easy to use. AgentBase/Marketeer is primarily aimed at marketing applications. It uses several data mining methodologies whose results are combined by intelligent agents. It can access data from all major sources, and it runs on Windows95, Windows NT, and the Solaris operating system.

- **ANGOSS Knowledge Miner**

ANGOSS Knowledge Miner combines ANGOSS Knowledge Studio with proprietary algorithms for click stream analysis; it interfaces to Web log reporting tools.

- **Autoclass III**

Autoclass is an unsupervised Bayesian classification system for independent data. It seeks a maximum posterior probability to provide a simple approach to problems such as classification, clustering, and general mixture separation. It works on UNIX platforms.

- **BusinesMiner**

BusinessMiner is a single-strategy, easy-to-use tool based on decision trees. It can access data from multiple sources including Oracle, Sybase, SQL Server, and Teradata. BusinessMiner runs on all Windows platforms, and it can be used stand-alone or in conjunction with OLAP tools.

- **CART**

CART is a robust data-mining tool that automatically searches for important patterns and relationships in large data sets and quickly uncovers hidden structures even in highly complex data sets. It works on the Windows, Mac, and Unix platforms.

- **Clementine**

Clementine is a comprehensive toolkit for data mining. It uses neural networks and rule-induction methodologies. The toolkit includes data manipulation and visualization capabilities. It runs on Windows and UNIX platforms and accepts the data from Oracle, Ingres, Sybase, and Informix databases. A recent version offers sequence association and clustering for Web-data analyses.

- **Darwin (now part of Oracle)**

Darwin is an integrated, multiple-strategy tool that uses neural networks, classification and regression trees, nearest-neighbor rule, and genetic algorithms. These techniques are implemented in open client/server architecture with a scalable, parallel-computing implementation. The client-side unit can work on Windows and the server on Unix. Darwin can access data from a variety of networked data sources including all major relational databases. It is optimized for parallel servers.

- **DataEngine**

DataEngine is a multiple-strategy data-mining tool for data modeling, combining conventional data-analysis methods with fuzzy technology, neural networks, and advanced statistical techniques. It works on the Windows platform.

- **Data Mining Suite**

Data Mining Suite is a comprehensive and integrated set of data-mining tools. The main tools are IDIS (Information Discovery System) for finding classification rules, IDIS-PM (Predictive Modeler) for prediction and forecasting, and IDIS-Map for finding geographical patterns. Data Mining Suite supports client/server architecture and runs on all major platforms with different database-management systems. It also discovers patterns of users' activities on Web sites.

- **Data Surveyor**

Data Surveyor is a single-strategy (classification) tool. It consists of two components: a front-end a back-end. The front-end is responsible for data mining using the tree-generation methodology. The back-end consists of a fast, parallel, database server where the data are loaded from a user's databases. The back-end runs on parallel Unix servers and the front-end works with Unix and Windows platforms.

- **DataMind**

DataMind's architecture consists of two components: DataCruncher for serverside data mining and DataMind Professional for client-side specification and viewing results. It can implement classification, clustering, and association-rule technologies. DataMind can be set up to mine data locally or on a remote server, where data are organized using any of the major relational databases.

- **Datasage**

Datasage is a comprehensive data-mining product whose architecture incorporates a data mart in its data-mining server. The user accesses Datasage through an interface operating as a thin client, using either a Windows client or a Java-enabled browser client.

- **DBMiner**

DBMiner is a publicly available tool for data mining. It is a multiple-strategy tool and it supports methodologies such as clustering, association rules, summarization, and visualization. DBMiner uses Microsoft SQL Server 7.0 Plato and runs on different Windows platforms.

- **Decision Series**

Decision Series is a multiple-strategy tool that uses artificial neural networks, clustering algorithms, and genetic algorithms to perform data mining. It can operate on scalable, parallel platforms to provide speedy solutions. It runs on standard industry platforms such as HP, SUN, and DEC, and it supports most of the commercial, relational database-management systems.

- **Decisionhouse**

Decisionhouse is a suite of tightly integrated tools that primarily support classification and visualization processes. Various aspects of data preparation and reporting are included. It works on the Unix platform.

- **Delta Miner**

Delta Miner is a multiple-strategy tool supporting clustering, summarization, deviation-detection, and visualization processes.

A common application is the analysis of financial controlling data. It runs on Windows platforms and it integrates new search techniques and "business intelligence" methodologies into an OLAP front-end.

- **Emerald**

Emerald is a publicly available tool still used as a research system. It consists of five different machine-learning programs supporting clustering, classification, and summarization tasks.

- **Evolver**

Evolver is a single-strategy tool. It uses genetic-algorithm technology to solve complex optimization problems. This tool runs on all Windows platforms and it is based on data stored in Microsoft Excel tables.

- **GainSmarts**

GainSmarts uses predictive-modeling technology that can analyze past purchases and demographic and lifestyle data to predict the likelihood of response and other characteristics of customers.

- **IBM Datajoiner**

Datajoiner allows the user to view multivendor-relational and nonrelational, local and remote-geographically distributed databases as local databases to access and join tables without knowing the source locations.

- **IBM Intelligent Miner**

Intelligent Miner is an integrated and comprehensive set of data-mining tools. It uses decision trees, neural networks, and clustering. The latest version includes a wide range of text-mining tools. Most of its algorithms have been parallelized for scalability. A user can build models using either a GUI or an API. It works only with DB2 databases.

- **KATE**

KATE is a single, rule-based strategy tool consisting of four components: KATE-editor, KATE-CBR, KATE-Datamining, and KATE-Runtime. It runs on Windows and Unix platforms, and it is applicable to several databases.

- **Kensington 2000**

Kensington 2000 is an internet-based knowledge-discovery and-management platform for the analyses of large and distributed data sets.

- **Kepler**

Kepler is an extensible, multiple-strategy data-mining system. The key element of its architecture is extensibility through a "plug-in" interface for external tools without redeveloping the system core. The tool supports data mining tasks such as classification, clustering, regression, and visualization. It runs on Windows and Unix platforms.

- **Knowledge Seeker**

Knowledge Seeker is a single-strategy desktop or client/server tool relying on a tree-based methodology for data mining. It provides a nice GUI for model building and letting the user explore data. It also allows users to export the discovered data model as text, SQL query, or Prolog program. It runs on Windows and Unix platforms, and accepts data from a variety of sources.

- **MATLAB NN Toolbox**

A MATLAB extension implements an engineering environment (i.e. a computer-based environment for engineers to help them solve their common tasks) for research in neural networks and its design, simulation, and application. It offers various network architectures and different learning strategies. Classification and function approximations are typical data-mining problems that can be solved using this tool. It runs on Windows, Mac, and Unix platforms.

- **Marksman**

Marksman is a single-methodology tool based on artificial neural networks. It provides a number of useful data-manipulation features, which are very important in preprocessing. Its design is optimized for the database-analysis needs of direct-marketing professionals, and it runs on PC/Windows platforms.

- **MARS**

MARS is a logistic-regression tool for binary classification. It automatically handles missing values, detection of interaction between input variables, and transformation of variables.

- **MineSet**

MineSet is comprehensive tool for data mining. Its features include extensive data manipulation and transformation capabilities, varius data-mining approaches, and powerful visualization capabilities. MineSet supports client/server architecture and runs on Silicon Graphics platforms.

- **NETMAP**

NETMAP is a general purpose, information-visualization tool. It is most effective for large, qualitative, text-based data sets. It runs on Unix workstations.

- **Neuro Net**

NeuroNet is a publicly available software for experimentation with different artificial neural-network architectures and types.

- **NeuroSolutions V3.0**

NeuroSolutions V3.0 combines a modular, icon-based artificial neuralnetwork design, and it solves data-mining problems such as classification, prediction, and function approximation. Its implementations are based on advanced learning techniques such as recurrent backpropagation and backpropagation through time. The tool runs on all Windows platforms.

- **OCI**

OCI is publicly available software for data mining. It is specially designed as a decision tree induction system for applications where the samples have continuous feature values.

- **OMEGA**

OMEGA is a system for developing, evaluating, and implementing predictive models using the genetic-programming approach. It is suitable for the classification and visualization of data. It runs on all Windows platforms.

- **Partek**

Partek is a multiple-strategy data-mining product. It is based on several methodologies including statistical techniques, neural networks, fuzzy logic, genetic algorithms, and data visualization. It runs on UNIX platforms.

- **Pattern Recognition Workbench (PRW)**

Pattern Recognition Workbench (PRW) is a comprehensive multiple-strategy tool. It uses neural networks, statistical pattern recognition, and machinelearning methodologies. It runs on all Windows platforms using a spreadsheet-style GUI. PRW automatically generates alternative models and searches for the best solution. It also provides a variety of visualization tools to monitor model building and interpret results.

- **Readware Information Processor**

Readware Information Processor is an integrated text-mining environment for intranets and the Internet. It classifies documents by content, providing a literal and conceptual search. The tool includes a ConceptBase with English, French, and German lexicons.

- **SAS (Enterprise Miner)**

SAS (Enterprise Miner) represents one of the most comprehensive sets of integrated tools for data mining. It also offers a variety of data manipulation and transformation features. In addition to statistical methods, the SAS Data Mining Solution employs neural networks, decision trees, and SAS Webhound that analyzes Web-site traffic. It runs on Windows and Unix platforms and it provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess).

- **Scenario**

Scenario is a single-strategy tool that uses the tree-based approach to data mining. The GUI relies on wizards to guide a user through different tasks, and it is easy to use. It runs on Windows platforms.

- **Sipina-W**

Sipina-W is publicly available software that includes different traditional data-mining techniques such as CART, Elisee, ID3, C4.5, and some new methods for generating decision trees.

- **SNNS**

SNNS is a publicly available software. It is a simulation environment for research on and application of artificial neural networks. The environment is available on Unix and Windows platforms.

- **SPIRIT**

SPIRIT is a tool for exploration and modeling using Bayesian techniques. The system allows communication with the user in the rich language of conditional events. It works on Windows platforms.

- **SPSS**

SPSS is one of the most comprehensive integrated tools for data mining. It has data-management and data-summarization capabilities and includes tools for both discovery and verification. The complete suite includes statistical methods, neural networks, and visualization techniques. It is available on a variety of commercial platforms.

- **S-Plus**

S-Plus is an interactive, object-oriented programming language for data mining. Its commercial version supports clustering, classification, summarization, visualization, and regression techniques. It works on Windows and Unix platforms.

- **STATlab**

STATlab is a single-strategy tool that relies on interactive visualization to help a user perform exploratory data analysis. It can import data from common relational databases and it runs on Windows, Mac, and Unix platforms.

- **STATISTICA-Neural Networks**

STATISTICA-Neural Networks is a single-strategy tool includes a standard back propagation-learning algorithm and iterative procedures such as Conjugate Gradient Descent and Levenberg-Marquardt. It runs on all Windows platforms.

- **Strategist**

Strategist is a tool based on Bayesian-network methodology to support different dependency analyses. It provides the methodology for integration of expert judgments and data-mining results, which are based on modeling of uncertainties and decision-making processes. It runs on all Windows platforms.

- **Syllogic**

Syllogic Data Mining Tool is a toolbox that combines many data-mining methodologies and offers a variety of approaches to uncover hidden information. It includes several data-preprocessing and -transformation functions. It is available on

Windows NT and Unix platforms and it supports most of the commercial relational databases.

- **Teradata Warehouse Miner**

Teradata Warehouse Miner provides different statistical analyses, decisiontree methods, and regression methodologies for in-place mining on a Teradata database-management system.

- **TiMBL**

TiMBL is a publicly available software. It includes several memory-based learning techniques for discrete data. A representation of the training set is explicitly stored in memory, and new cases are classified by extrapolation from the most similar cases.

- **TOOLDIAG**

TOOLDIAG is a publicly available tool for data mining. It consists of several programs in C for statistical pattern recognition of multivariate numeric data. The tool is primary oriented toward classification problems.

- **WINROSA**

WINROSA is a software tool that complements many other tools available for building fuzzy logic systems. It automatically generates fuzzy rules from the available data set. It works on Windows platforms.

- **Viscovery©SOMine:**

This single-strategy data-mining tool is based on self-organizing maps and is uniquely capable of visualizing multidimensional data. Viscovery©SOMine supports clustering, classification, and visualization processes. It works on all Windows platforms.

- **Weka (2.2):**

Weka is a software environment that integrates several machine-learning tools within a common framework and a uniform GUI. Classification and summarization are the main data-mining tasks supported by the Weka system.

- **WUM**

WUM 6.0 is a publicly available integrated environment for Web-log preparation, querying, and visualization of summarized activities on a Web site.

APPENDIX B

ATTRIBUTES of DATASET

74 patients are diagnosed for the Pulmonary Embolism and the following attributes are obtained:

1. **Name**: Patient’s name
2. **Group**: 74 patients are divided into three groups:
   a. Patients who don’t have pulmonary embolism
   b. Patients who have pulmonary embolism but don’t have vena thrombosis
   c. Patients who have both pulmonary embolism and vena thrombosis
3. **Sex**: 1 for female, 2 for male
4. **Age:** Patient’s age
5. **IVC Diameter:** Inferior vena cava diameter (millimeter)
6. **IVC Reflux:** Patients are dived into three categories, 0 for who don’t have inferior vena cava reflux, 1 for who have inferior vena cava reflux in proximal kardiyak neighborhood, 2 for who have reflux at hepatic level
7. **IVS Thickness:** Interventricular septum thickness (millimeter)
8. **IVS Camber**: Patients are divided into two categories, 0 whose interventricular septum is flat, 1 whose interventricular septum is camber
9. **Main Pulmonary Arterial Diameter:** Main pulmonary arterial diameter (millimeter)

10. **Right Pulmonary Arterial Diameter:** Right pulmonary arterial diameter (millimeter)

11. **Left Pulmonary Arterial Diameter:** Left pulmonary arterial diameter (millimeter)

12. **Rvw Thickness:** Right ventricle wall thickness (millimeter)
13. **Lvw Thickness:** Left ventricle wall thickness (millimeter)
14. **Rvw.Lvw:** The ratio of right ventricle wall thickness to left ventricle wall thickness (millimeter)

15. **Rv Diameter:** Right ventricle diameter (millimeter)
16. **Lv Diameter:** Left ventricle diameter (millimeter)
17. **Rv.Lv:** The ratio of right ventricle diameter to left ventricle diameter
18. **Svc Diameter:** Superior vena cava diameter
19. **Azygos:** Azygos vein diameter at the expansion of superior vena cava location
20. **Index of Right Pulmonary Arterial Obstruction:** Index of right pulmonary arterial obstruction (sağ üstloba giden 3 arterin, orta loba giden 2 arterin her birine, alt loba giden 5 dalın her birine eğer tam obstrüksiyon oluşturan trombüs varsa 2, parsiyel obstrüksiyon yapıyorsa 1 ve lümen açıksa 0 değeri verilerek, max sağ akciğer obstrüksiyon indeksi 20 olacak şekilde skorlandı)[*]

21. **Index of Left Pulmonary Arterial Obstruction:** Index of left pulmonary arterial obstruction (sol üstloba giden 3 arterin, lingulaya giden 2 arterin her birine, alt loba giden 5 dalın her birine eğer tam obstrüksiyon oluşturan trombüs varsa 2, parsiyel obstrüksiyon yapıyorsa 1 ve lümen açıksa 0 değeri verilerek, max sol akciğer obstrüksiyon indeksi 20 olacak şekilde skorlandı)

22. **Total of Obstruction Index for Pulmonary:** Total of obstruction index which is calculated for pulmonary

[*] Some of the original comments on data are preserved.

23. **Determined Total Clot Area of Each Pulmonary Arterial :** Determined Total Clot Area of Each Pulmonary Arterial (square millimeter)

24. **D-Dimer:** D-Dimer degree in the serum

25. **Venskor:** : Alt ekstremite venöz doppler us incelemesi sonrası her bir ekstremite için cfv, sfv, dfv, vsm, popliteal ven, vena safena parva ve vena cava inferiorda lümei tamamen tıkayan trombüs için 2, parsiyel trombüs için 2 değeri verilerek toplam venöz pıhtı yükü skoru max 14 olacak şekilde hesaplandı

26. **Plevral Effüzyon:** : Tek taraflı plevral effüzyon1, iki tarafli ise 2, her bir fissürde saptanan sıvı için artı 1'er puan daha hesaplanarak, ayrıca perikardial sivayada ek 1 puan vererek max 6 puan olacak sekilde skorladim(ama böyle bir skorlama aslında yok isterseniz bunu değerlendirmeyebiliriz ama ben pulmoner arterlerdeki pıhtı ile ilişkisi varmı diye böyle puanlayarak değerlendirmeye çalıştım)

27. **Operation**: 1 for patients who have tracheotomy, abdominal operation, heart operation etc., 0 for others

28. **Tumor:** 1 for Pulmonary, stomach etc. tumor, 0 for others

29. **Other Disease:** 1 for patients who have tracheotomy etc., 0 for others