Bachelor’s Thesis (UAS)

Degree Program in Information Technology

2011

240S07

Xiao Gang

The application of data mining methods

TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR’S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Degree Programme | Information Technology

2011-06-07 | 42

Instructor: Wikström Yngvar M.Eng

Xiaoli Geng

The application of data mining methods

Data mining is becoming more and more important. The aim of this thesis is to study and research data mining, to clarify the background, knowledge and method of data mining, and research some specific areas applications. The aim is also to experiment with an open software by mining some sample data, to prove the advantage and convenience of data mining.

This thesis first introduces the basic concepts of data mining, such as the definition of data mining, its basic function, common methods and basic process, and two common data mining methods, classification and clustering. Then a data mining application in network is discussed in detail, followed by a brief introduction on data mining application in business projects, and some success cases. The last chapter simply introduces a famous open data mining working platform named WEKA, describes the related knowledge of software, characteristics and working process, and takes a simple data mining test based on this software.

This new technology discards unuseful information and obtains useful information, and this can be widely used in various applications. Data mining can extract useful information through many different methods and algorithms, so it can be applied on many different areas or different environments.

KEYWORDS: data mining, useful information, classification, application

FOREWORD

Last summer I got a chance to do my internship at HeBei Technology & Science Management Communication Centre, joining a team working on data mining. That was the first time I touched this filed, I learnt a lot of new knowledge and I am deeply interested in it. We did the project for a seed company, mining some data and statistics, and then reported the results to them.

I was so lucky to have the chance to join this team because this kind project always needs a huge amount of real data. This gave me a practice chance to apply my data mining knowledge into real project which expanded my knowledge area in this field. So when I started to do my thesis, I did some research and study in this filed, and connected the theory background with my practical experience then completed the thesis.

So at first I need to thank to my Background Company, HeBei Technology&Science Management Communication Centre, and also thank you to my project team members. Secondly, I really appreciate my instructor, Mr. Wikström Yngvar; and my language teacher, Mrs. Skarli Poppy; and our thesis guider, Mr. Väänänen Ossi. During my thesis work, they all gave much help and advices. Thank you to all of you!

At last, thanks to my family, to my father and mother. Without you I could not have the chance to study abroad. Thank you!

2011/05/18 Turku

Xiaoli Geng

Contents:

LIST OF FIGURES

1 Introduction

With the development of computer technology, people's ability to collect data and store data has been greatly improved. Scientific research or all areas of social life have accumulated a large amount of data, so, the analysis of these data to discover useful information contained in the data, become a common need in almost all areas. As a result, the role of data mining has been increasingly important. Data mining technology changes these data into useful information and knowledge; the obtained information and knowledge can be widely used in various applications, including business management, production control, marketing analysis, engineering design and science exploration. Therefore, data mining is the natural evaluation result of information technology, which is important.

So I chose this topic as my thesis project. I also used the project test to set up a case study on data mining. This thesis separately introduces the data mining method and different area applications, explains the impacts of data mining nowadays, and the advantages of this technology, and through the test proves and shows this technology iconic ally.

I read some other theses similar with mine, the main things about data mining are the mining methods and the applications. I can find that the main and useful mining methods are similar like I introduced in this thesis, classification and clustering, the mining methods based on decision tree and so on. The final conclusions are expressed the advantages of data mining nowadays and the new applications in many different fields.

The goals of my thesis are to explain the main idea of data mining, and learn some basic common methods, implement an experiment to verify the advantages and the functions of data mining.

2 Background Knowledge

2.1 The background and significance of the project

In the Networked Era, computers and network technologies are changing people's life. Since APARNET was established, the Internet has experienced a rapid development. It has now become a global facility that almost covers every hole and corner on this planet. As a main part of the Internet, network protocols have been well developed to meet a wide range of practical applications. However, with the continuous expansion of its scale to both services and users, the problems that the Internet has to face are also growing.

Due to the wide use of database management systems, data is piling up as time goes by. People can learn from data, but large bodies of data are unless because people need specific data, not the unassorted one. Over the past few years, the development of knowledge discovery in this field is growing fast because of the large markets and research interests. The progress of computer technology and data collection techniques enable people to collect and store data from a broader range at an unprecedented speed. On the other hand, although modern database technology can help us to store large amounts of data easily, it cannot help us to analyze and understand data, or represent data in an understandable information form. In the past, the common method we used for knowledge acquisition was analysis, filter, comparison, and then we extracted out the knowledge and created rules. However, as the knowledge engineers have limitations on knowledge, so the knowledge we gained will be limited. At present, when the traditional knowledge acquisition faces the huge data warehouse, it cannot do anything, so data mining technology was created to address these challenges.

Data Mining is the process of extracting information and knowledge implicit in from large, incomplete, noisy, fuzzy, random practical application data, people do not know in advance but which is potentially useful [1 , 2].

The reason why data mining has the great importance in the information industry is because large amounts of data need to be changed to useful information that can be understood easily by people, and they also can be widely used in various applications, including business management, production control, marketing analysis, engineering design and science exploration. Therefore, data mining is the natural evaluation result of information technology, which is important.

2.2 The major work and objectives

Data mining algorithms have become a huge technology system after years of development. This involves blending different disciplines and a large number of algorithms and different functions tools. One of the basic objectives of this project is to study data mining techniques, read related data mining materials, understand the basic concepts and general methodology, grasp the common methods and to achieve the preliminary algorithm, especially to master the classification, clustering and feature selection algorithm. Another objective is to study the books and materials related to data mining, read the papers related to network traffic classification based on data mining technology, become familiar with the current network flow, learn the development status and role of data mining in modern society, learn the data mining application technology in network and the application mode in business issues. The last objective is to develop my practical application skills with data mining techniques.

This thesis describes the current network environment now, simply analyze the development and mature status of network technology, discusses the next hot technology spot that can advance the progress of human society, and obtain the current phenomenon “ data explosion but lack of knowledge”. We find that people hope to analyze on a higher level to make better use of these data, this leads to data mining and knowledge discovery techniques, and makes a detailed elaboration and introduction on the data mining method which was proposed in 1980s. Chapter three and Chapter four introduce details on the data mining application in network and business, and several successful cases. These chapters also introduce the data mining method based on statistical features, a typical algorithm based on this method named decision tree algorithm. Finally, the thesis introduces the WEKA software and some relation knowledge, and the test process based on WEKA platform.

2.3Thesis structure

The thesis mainly introduces the basic concepts of data mining, common methods and the application in network and business projects, In addition, it cites some success cases. At last, it describes a simple data mining test based on WEKA software carried out by the author.

The first chapter introduces the research background and significance, the main objective of this project and the main work and arrangement for the overall structure of the thesis.

The second chapter describes the data mining methods. The basic concepts of data mining methods are given. Including the definition of data mining, common methods and the basic processes, the thesis describes the commonly used classification methods and the clustering methods, and then gives the general guidelines of the assessment and classification.

The third chapter describes the data mining application in network, mostly about the network traffic and the data mining for network traffic. Given the concept of network traffic flow, the thesis presents the characteristics of the network features leads network traffic classification methods based on data mining technology, and then introduces a network traffic classification method based on decision tree.

The fourth chapter introduces some applications in business, and some success cases of the data mining application project.

The last chapter is mainly focused the WEKA software, introduces characteristic of the WEKA system, file format, system interface, the mining process, and then describes simple project test on WEKA.

3 Data mining methods

Data mining was developed when artificial intelligence research gradually changed direction into practical applications in the 80 of the 20th century. Data mining is an interdisciplinary project, it makes the application on mining knowledge from the low level to a higher level, and data mining method provides decision support. Many researchers in different areas have devoted themselves to data mining which is an emerging research area, evolving a new technology hot pot.

This chapter describes the general methods of data mining, focusing on the classification and clustering methods and their evaluation criteria in data mining technology.

3.1 The basic concepts of data mining

3.1.1 The origin of data mining

We are now living in the Network Era; computer and network technology is changing people’s social life. The development speed of the Global IP network is doubled every 6 months. In the U.S., it took the radio 38 years to reach 50 million, and it took television 13 years; the users accessing the internet through dial-up reached 50 million in just 4 years.

In retrospect, people would ask: Concerning to the promotion of the progress of human society in history, which technology can be compared with the network technology? What is the next hot point of technology?

Let’s look at some phenomena that can be observed everywhere in our daily life: “The New York Times” printed from 10-20 editions in the 60s; now it prints 100-200 editions, the highest numbers of editions is 1572; In China, “BeiJing Youth Daily ” also prints 16-40 editions, and “Marketing Report” has reached 100 editions already. However, in reality, the reading time per day is usually 30 to 45 minutes, so usually people only can read a 24 edition newspaper. Large amounts of information bring convenience to people and brings a lot of problems at the same time: firstly, information overload is difficult to digest; secondly, it is difficult to identify whether information is true or false; thirdly, it is difficult to ensure information security; fourthly, forms of information are inconsistent and difficult to handle unified. People began to propose a new slogan: “Learn to discard information”. People began to consider, How not to be flooded with the information, but to discover useful knowledge and improve information availability.

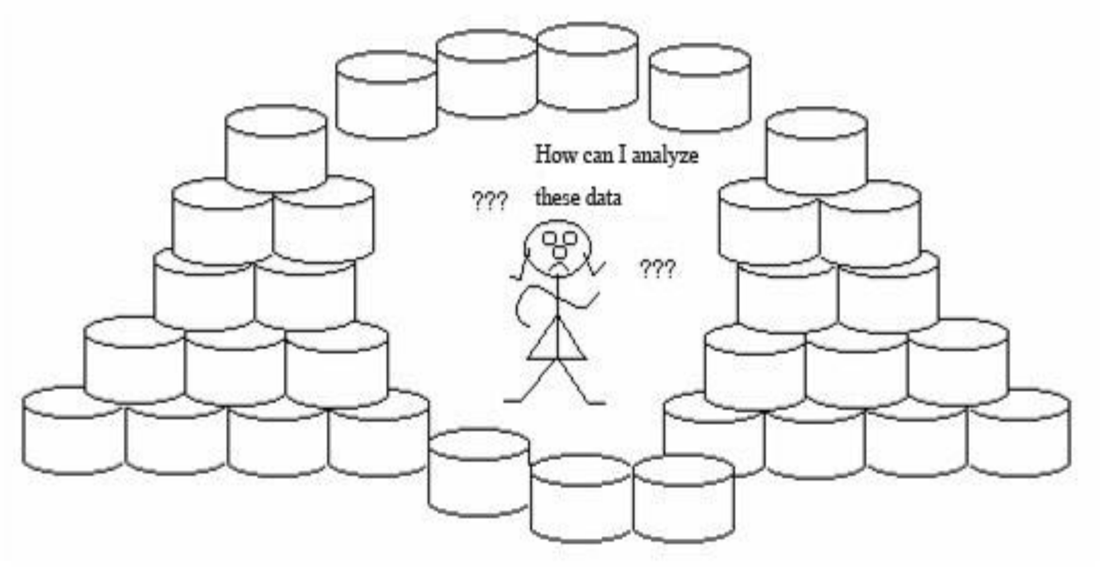


Figure 3.1 How can I analyze these data?

On the other hand, because of the rapid development of database technology, and the widely used database management system, people accumulate more and more data. Behind the data explosion is hidden much important information and people want to make a higher level analysis, in order to make better use of the data. The current database system can be achieved efficiently using data entry, query, statistics and other functions, but cannot find the relations and rules existing inside the data and cannot predict future trends.

The lack of tools mining the knowledge behind the data, lead to “knowledge explosion poor knowledge”.

For this reason, data mining and knowledge discovery technology are being developed, and show strong vitality, thus evolving out of the data mining technology gradually.

3.1.2 The definition of data mining

In the 80’ s of 20th century, the Artificial Intelligence (AI) research project was failure. AI turned into practical applications, data mining. Data mining is a new, commercial applications AI research. Nowadays, data mining is receiving more and more business organizations attention. However, what is data mining? In brief, data mining is extracting data or “mining” knowledge from large amounts of data.

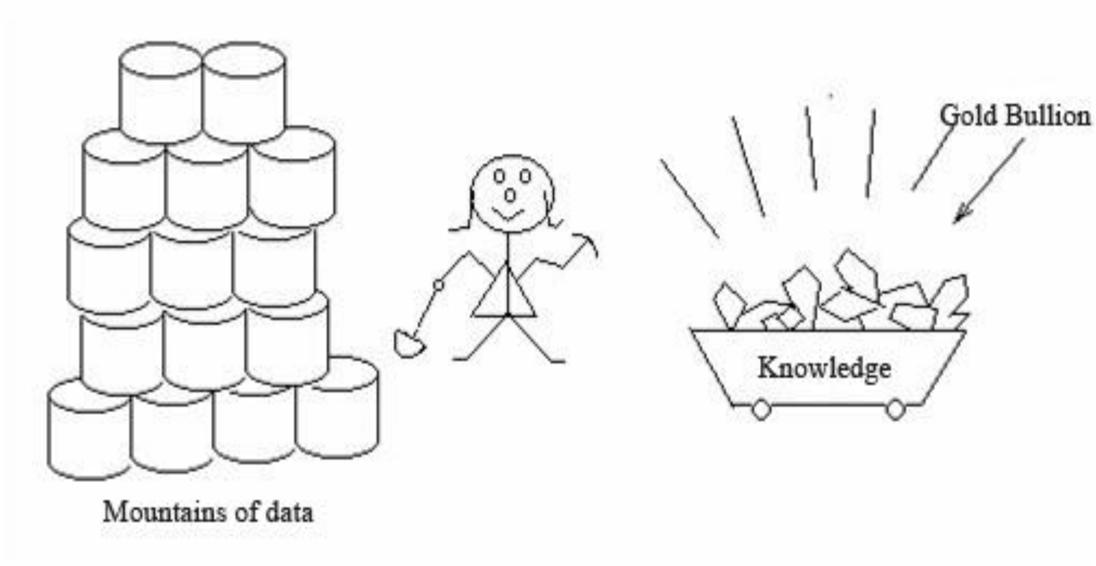


Figure 3.2 Mining data is similar to mining gold

Data mining has been proposed since the 80’ s in the 20th century. In the course of its development, many authorities have raised their own thought on the definition of data mining, the main definition is shown in Table 3-1.

Table 3-1 Different definition of data mining

Researchers	Definitions
SAS	The advanced method of data exploration and establishment of related models based on a large number of data.
Gartner	The process through careful analysis of large amounts of data to reveal
Group	Meaningful new relationships, patterns and trends.
Aaron Zornes	The knowledge mining process from large databases to extract the operational information that we did not know before.
Fayyad	The important process to determine the effective, new and potentially information, and the model can be understood ultimately from the data.
Zekulin	Extracting previously unknown, understandable, actionable information from large databases.
Ferruzza	Used in the knowledge discovery process, some methods to identify unknown relationships and patterns existing in the data .
Jonn	Finding useful patterns while processing the data .
Parsay	A decision to support the process of studying the large data set for those unknown information models.
Bhavani	Finding meaningful new relationships, patterns and trends process in large amounts of data using pattern recognition technology, statistic and mathematical techniques.

Although the definitions in Table 3-1 have some differences, they all highlight data mining as the process from data into an useful model; each useful model provides potential information valuable to users. Its aim is to change the data into knowledge, and increase the intrinsic value of the data.

Therefore data mining can be defined as: data mining is the process which transforms from large, incomplete, noisy, fuzzy, random practical application data into information and knowledge that is implicit and that people do not know in advance but is potentially useful after processing [3]. This definition implies that: data mining sources must be real, substantial and noisy; the found knowledge is the one that users are interested in; the discovered knowledge can be acceptable, understandable and used; it does not require the discovered knowledge to fit all and the goal is to solve specific problems in a specific field.

3.1.3 The basic function of data mining

Generally, the function of data mining can be divided into two classes: (1) on the basis of available data sets generating new, unusual information; (2) Predictive data mining: generating a systematic model described by the known data set. These two types usually contain the following functions:

(1) Concept Description

Concept description describes the meaning of certain objects, and summarizes the relevant characteristics of the objects. The concept description can be achieved through the following methods: data characteristics and data distinguishing. The former describes the common features of certain objects, for example, from the characteristics of quality customers of the bank we can identify potential high quality customers. The latter describes the difference between heterogeneous objects, such as the comparison between credit card fraudsters and non--fraudsters.

(2) Association analysis

Association analysis is finding the interesting connection, correlation or causal structure in items from large amounts of data. If two or more data items' value is repeated and the probability is very high that they have some relationship, then we can structure an association rules for these data. The aim of association analysis is to find the hidden association rule. For example, when the customer buys a computer, he also will buy some software; this is an association rule.

(3) Classification and prediction

Classification is finding a model or function that can describe the typical features of data sets, so that it can identify the ownership or category of the unknown data. For example, credit card applicants will be divided into low, medium and high risk groups.

Prediction is using historical data to identify the principle, structure model and use this model to predict the types and features of the data. An example of which customers will cancel the company's service in the next six months, or predict which customers will apply for more services.

(4) Cluster analysis

Clustering is also called unsupervised learning. The aim of cluster is to divide the data into a series of meaningful subset according to certain rules; in the same cluster, the gap between individuals is smaller; in a different cluster the distance between individuals is much larger. For example, according to the volatility of the stock price, we can divide the stock into different categories; each category contains what, all the information that is very important to the investor.

Cluster analysis is different from classification; clustering analysis is a method that does not give a classification scheme before, but it clusters information according to similarity.

(5) Outlier analysis

If a database contains some data that has inconsistent behavior or models, this kind of data is called outlier. Most data mining methods discard the outlier analysis as noise or unusual, but in some applications, it is necessary to find the usual data, such as finding the buying behaviors of particularly low or particularly high income customers through outlier analysis .

(6) Evolution analysis

Evolution analysis is to structure model based on the change law and evolution trends of data objects. It mainly contains time sequence data analysis and is based on similarity data analysis. For example, 80% of people who have bought laser printers will buy a new toner.

3.1.4 The common methods of data mining

Data mining is developed from artificial intelligence and machine learning methods, combined with the traditional statistical analysis methods, mathematical method and visualization of scientific computing, then formed data mining methods and techniques. Generally, classification according to function normally, commonly used data mining methods is summarized as follows; they are from different angles on data to excavation and finding useful patterns and gaining knowledge.

(1) **Categorization** is finding a set of common characteristics of data objects in the database, and according to the classification model divide data into different classes, in order to map the data elements from the database to a given category, forecasting discrete target variables.

(2) **Cluster analysis** is making a set of data according to the similarities and differences divided into several categories. The basic principle is to make the similarity between the same categories data as large as possible, the similarity between different categories data as small as possible. The purpose is to find a group closely related to the observation group.

(3) **Regression analysis** reflects property values in the time figures in transaction database; creates a function of predictor variables with real data to, finding dependencies between variables or attributes.

(4) **Association rule** describes the relationship rules existing between the data items in database, that is some items appearance in a object, these items will export other items also appear in the same object, the association or correlation hidden in the data, finding the mode has a strong association characteristics in the data.

(5) **Feature analysis** is extracting the characteristic type related to the data from a group of data from the database. These characteristic types show the general characteristics of the data set. In the change and deviation analysis, deviation includes a large class of potentially interesting knowledge, such as abnormal instances in the classification process; the aim is to search the significant differences between the observation result and the reference volume. Unexpected rules mining can be applied to discovery, analysis, identification, assessment and early warning and so on for a variety of abnormal information.

(6) **Neural Nets** imitate the biological neural network. It is a nonlinear prediction model through training learning and it can complete classification, clustering, feature mining, forecasting and other data mining tasks.

(7) **Visualization technology** is a graphics technology. It uses an intuitive graphical presentation of the information model, data association or trend to the decision makers. Visualization improves the efficiency of data mining.

3.1.5 The basic process of data mining

Data mining uses many scientific methods of mathematics, statistics, artificial intelligence and neural network fields, mining the model from large amounts of data is used for decision support, it provides an approach, tool and process for predictive decision support. Many people consider data mining as a fundamental step of Knowledge Discovery Process (KDD) from the data base.

The traditional KDD process is shown as Figure 2-1; the concrete steps are as follows:

(1) Data Cleaning: It eliminates noise or inconsistent or nothing to do with the mining task’s data.

(2) Data Integration: It combines variety of different data sources.

(3) Data Selection: It retrieves and analyzes the data related to task.

(4) Data Transformation: It converts or unifies the data into a form that suitable for mining.

(5) Data Mining: It is the fundamental step of KDD, intending to use the intelligent methods to extract data model.

(6) Pattern Evaluation: According to certain criteria, it identifies the pattern that expresses knowledge.

(7) Knowledge Presentation: It uses visualization and knowledge representation technology to provide the mined knowledge to users.

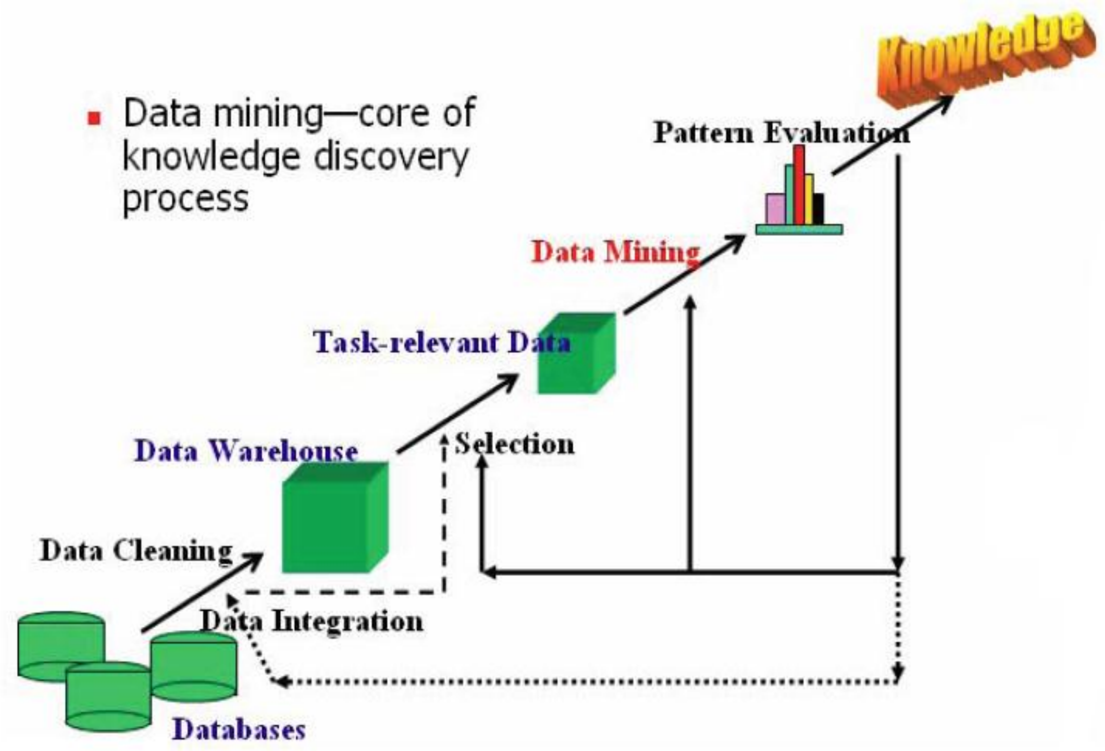


Figure 3-1-5 the basic process and major steps of data mining

Figure 3-1-5 describes a typical data mining basic process and its major steps, includes the related data selection from database; cleaning and integrating the selected data, data transformation, using data mining algorithm for pattern mining, interpretation and evaluation of the model obtained. Input data can be stored in various forms, can intensively reside in the database, or be distributed across multiple sites, after the data selection operation has formed the target data. Data cleaning fills in the missing values, smoothens noise data to eliminate noise and repeated observations, selects records and features related to the current data mining tasks. Data integration combines the data from multiple data sources. A data transformation changes the data type into a type suitable for mining and is a success-oriented pre-treatment process for data mining. Analysis and processing on the transformed data use data mining algorithms, mining the contains model, and reveals the discipline in the data. Then, follows the interpretation and evaluation on the obtained model, if the patterns obtained from mining have no practical significance, or cannot get through the statistical measure or hypothesis testing, then they are regarded as mendacious data mining results and are deleted. This process requires repeated, this repetition process will approach the essence of things, continuous processing has a solution to optimization problem.

The data mining process is not automatic. Most of the work needs to be done manually. Data mining has strict requirements on the data and data pre-processing is the most arduous and time-consuming step in the whole process, generally accounting for 60% of the time throughout the process, and the mining work accounts for only 10% of the total workload.

3.2 Classification and Clustering

This section will highlight these two important data mining methods.

3.2.1 Classification

In data mining, classification is a systematic method based on the input data to establish a classification model. Classification task [8] is learning to get a target prediction function f. This function is also called as the classification model, in the prediction or identification process, f makes each attribute set x map to a predefined class label y. The examples of classification include the decision tree classification method, rule-based classification, Naïve Bayesian classification method, support vector classification method, Neural Networks classification method, etc.. All these technologies use a learning algorithm to determine the classification model, are expected to fit the input data very well, and correctly predict the unknown samples class label.

Classification is generally divided into two steps: (1) the learning process which creates the classification model to describe or identify the data type or data concepts. (2) the prediction or identification process which uses the classification model to predict the unknown object.

The learning process constructs a model by analyzing the data tuple described by the property, describes the intended data set , the class label in the figure is played, classification model is provided by the decision tree. We assume that each data tuple has an attribute called class label, then this attribute marks this data tuple as an intended class. Multi-data tuples with class label are combined together to form the training data set. A single tuple is called training sample in training data set; a training sample is randomly selected by the sample groups.

The classification model can be expressed in a variety of shapes, such as decision tree, IF-THEN rule, mathematics formula or Neural Networks. A decision tree is a structure similar to the flow chart, every node represents an attribute value test, every branch represents a test output, and leaves represent class or its distribution. A decision tree is easily converted into the classification rule form which is easy to understand.

The predict process is shown as follows: classifying the data tuple with the unknown class label by using the classification model obtained from the previous step. Test data is a set of data tuple with a class label, but it does not need the test label in the testing process. Before we apply the classification model to the prediction, we first assess the evaluation index on test data sets from the classification model. If this model's evaluation index on these data sets is acceptable, then we can use it for that data tuple with unknown class label to predict classification.

3.2.2 Clustering method

The process of making the congregation of abstract objects group as multiple clusters formed by similar objects is called Clustering [8]. In the clustering process, one basic principle is maximizing the similarity in each cluster and minimizing the similarity between the various clusters. After clustering, the data objects in one cluster can be treated as a whole and have the common class label. Clustering is different from classification, the cluster' s class attribute and the number of clusters are unknown before clustering on the data, or do not consider the data tuple with class label during study, instead that use clustering analysis to obtain the clustering class label based on the clustering result .

Because of since the requirements of society, clustering analysis has become a very active research topic in data mining, but the huge, complex data sets also present special challenges to cluster analysis. The typical requirements are mainly the following aspects: (1) scalability (2) the ability to handle different types of property (3) the ability discovery arbitrary shape cluster (4) the ability be used to determine the input parameters minimum domain knowledge and the sensitive of input record order (5) the ability to handle noisy data (6) the ability to handle high dimensional data (7) Based on constraints clustering (8) Interpretability and usability.
Generally, the main clustering algorithms can be divided into the following categories:

Partitioning method: This method first creates an initial division, then interactive through moving the object in the division interval to improve the partitioning. But this method can only find spherical clusters.

Density-based method: If the density area only surrounds a threshold, it continues to cluster. This method can be used to filter “noise data” and to find arbitrary shape clusters.

Grid-based methods: This method makes object be spaced into limits units. This method has a fast processing speed.

After years of research, now there is a great number of clustering algorithm, the comparison between main clustering algorithms [9] is shown in Table 3-2.

Table 3-2 Main clustering method comparison

Algorithm	**Algorithm Efficiency**	**Appropriate Data Type**	**Cluster Type Found**	**Sensitivity on dirty or abnormal data**	**Sensitivity on the order of the input data**
k-means	high	Value	Convex or Spherical	Sensitive	Insensitive
k-medoids	Low	Value	Convex or Spherical	Less Sensitive	Insensitive
K-pototypes	General	Mixed	Convex or Spherical	Sensitive	Less Sensitive
CLARA	Lower	Value	Convex or Spherical	Sensitive	Less Sensitive
CLARANS	Lower	Value	Convex or Spherical	Insensitive	Very Sensitive
BIRCH	High	Value	Convex or Spherical	Insensitive	Sensitive
CURE	Higher	Value	Arbitrary Shape	Insensitive	Less Sensitive
DBSCAN	General	Value	Arbitrary Shape	Sensitive	Sensitive
STING	High	Value	Horizontal or Vertical	Sensitive	Insensitive
Wave Cluster	High	Value	Arbitrary Shape	Insensitive	Insensitive

3.3 Chapter Summary

This chapter introduces the data mining methods. It first describes the basic concept of data mining, the common data mining methods and the basic flow of data mining. And then it highlights the classification and clustering methods and finally gives a common evaluation criterion. It makes a macro and micro average evaluation index. These indicators will serve as the evaluation criteria for the feature selection and sub-clustering.

4 Data mining application in the network

With the rapid development of network technology, applications based on network are becoming more and more complex. Various legal or illegal applications not only waste more and more network resources, but also bring great threat to network security. Because of the advantage of using resources, traffic engineering has become important. Network traffic identification and classification is an important basis for network management, traffic monitoring, service analysis, network accounting and many other aspects.

The major task of traffic classification is according to the TCP or UDP streams of information or property that can be measured or obtained through observation points. Theses streams go through the network link or device, such as ports, message content, connection information, traffic statistics, etc., to speculate that the upper network application or layer protocol (such as WWW, FTP, P2P, etc.) belongs to the category with similar characteristics which are the present service in the current data. The key to traffic classification is to use the information as the classification's basis and then to use that kind of method as the classification method. At present, the machine learning classification based on statistical characteristics of traffic is an important research issue. This chapter gives some basic concepts of the traffic classification. It is mainly concerned with how to apply the Machine Learning classification method on the network traffic classification.

4.1 The definition of Network Traffic

Network protocol is usually developed for different levels; each layer is responsible for different communication functions. A protocol family, such as TCP/IP, is a set of multiple protocol combinations on different levels. The TCP/IP is generally considered as a four protocol system:

(1) The Link Layer usually includes the device drivers in the operating system and the corresponding interface card in computers.

(2) The Network Layer treats groups in the network activities. For example, the routing group. In the TCP/IP protocol family, network protocols include IP (Internet Protocol), ICMP (Internet Control Massage Protocol), and IGMP (Internet Group Management Protocol).

(3) The Transport Layer provides the port-to-port communications for the applications on two hosts. In the TCP/IP protocol, there are two different transport protocols:

TCP (Transmission Control Protocol) and UDP (User Datagram Protocol).

(4) The Application Layer is, responsible for dealing with specific application details. Common application protocols are Telnet, FTP, SMTP, SNMP, WWW and so on.

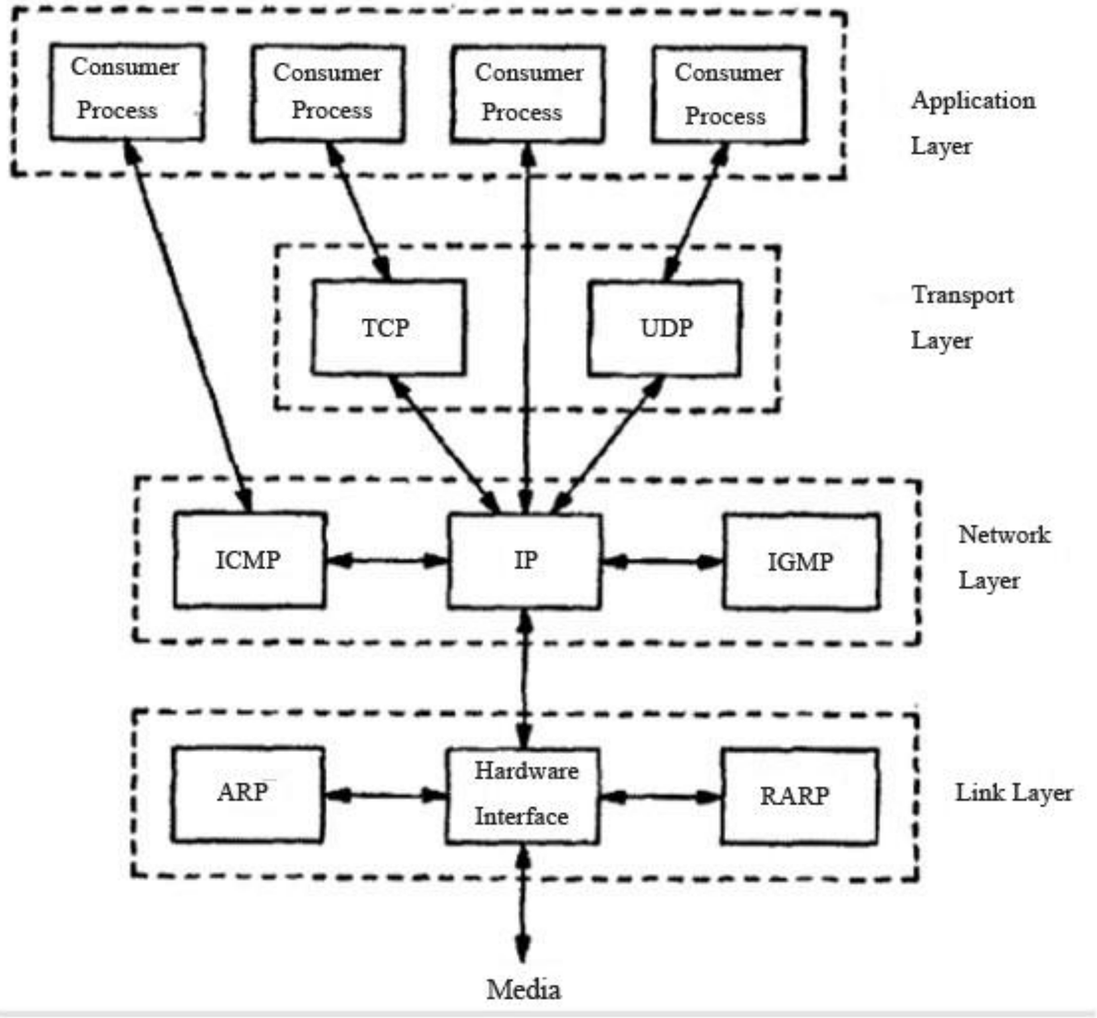


Figure 4.1 Different level protocols in the TCP/IP protocol

For TCP/IP protocols at different levels, the research purpose of traffic classification is different.

(1) Link layer traffic deals analysis mainly with the changes of transfer rate and throughput rate on network cable line. The aim is to reduce the error on the transmission line and improve the transmission speed on the network cable.

(2) Network layer traffic analysis is concerned with the IP packet routing strategy, delay and loss. The aim is to have a certain filtering rules to store and forwarding packets as soon as possible so that it can reduce packets loss.

(3) Since the transport layer and application layer are closely linked, we can put these two traffics together for analysis study. The flow at this level can be defined as: flow is an object and it describes a packet train with the same IP address, port number and protocol (TCP, UDP). It is a five-tuple composed of source address, source port, and destination port and transport layer protocol. This series of IP packet trains can accord this definition to compose a two-way TCP/IP or UDP stream flow. The research purpose of this layer is to identify the application layer protocol.

The project and research aim of the third layer of flow data is to identify application layer protocol. The third layer's five-tuple group coupled with the application layer protocol constitutes the flow in this article.

To sum up, Network Traffic Flow can be defined as follows: Network traffic flow is the amount of data transmitted over the network, and can be seen as the sum of the information that through a network link or device in a specific period, specific, can be seen as an IP packets that go through an observation point on the network at a certain time interval.

4.2 The properties of network traffic

4.2.1 The definition of the properties

A data set can be seen as a set of data objects. A data object is described by the basic features of the property which is characterized by a set of objects. Property is also called features, variables, fields or dimensions [7]. Property refers to the qualities or characteristics of the objects; it changes with the objects or with the time. For example, skin color is one of the properties for human beings, it is a symbol property and its varies according to individuals, possible values are yellow, white, black.

But property is not number or symbol. However, in order to discuss and analyze more precisely the characteristics of the object, we will give those numbers or symbols. To use a well-defined way to do this we need to measure the scale.

Measurement scale is an associate rule between a numerical or symbolic value with object properties. In many cases in daily life, the various situations of an object's property will be mapped to a numeric or symbolic value.

4.2.2 Property Type

Generally property is classified in to four types: nominal, ordinal, interval and ratio [8].

Nominal and ordinal can be defined as categorical or qualitative properties. Interval and ratio can be defined as quantitative or numeric properties.

Figure 3-2 lists these four properties, and describes each property type, finally there is an example to explain each property type and differentiate each property type.

4.2.3 Network flow property

The research on the network classification usually use the same 5 tuples (source IP, destination IP, source port, destination port, transport protocol) network packets, it is the network flow as the basic processing unit [12]. The statistical abstraction on the network flow, transform them into feature vectors in the feature space. This feature vector contains the basic information of network flow, such as the packets number in the network flow and so on. Maybe it also contains the information after a certain transformation, such as the result of packet arrival sequence after Fourier Transform.

Property type		**Description**	**Example**
Classification	**Nominal**	Nominal value of the property is just a different name, the nominal value only provide enough information to distinguish objects.	Skin color, eye color, identification number
	Ordinal	Ordinal value of the property provide enough information to determine object sequence	Grade , solubility
Value	**Interval**	For the range of attributes, difference value is meaningful , and in the attributes that exists measurement unit	Calendar data , temperature
	Ratio	For a ratio variable, differences and ratio are meaningful	Quality, age

Figure 4-2 Property Type

Structuring a feature set is one of the core missions in data mining. The quality of the feature set will directly influence the data mining result. The structure process on the feature network flow can be divided into one way flow and bidirectional flow. In the one way flow, the packets sequence is strictly valued by the rules according to the 5-tuples. The statistical feature of the one way network flow usually contains: the average number of the packets size, number of the packets, number of the packets with SYN or

FIN label, etc.. The bidirectional flow is the bidirectional packets sequence in the connection, not only contains the independent network flow feature on two direction of communication to the ports, but also the related features between two one way network flow, such as connection duration , connection idle time , etc., so it has stronger describing ability.

4.3 Network traffic classification

4.3.1 The definition of traffic classification

Network Traffic classification means classification according to the internet application type, the two-way TCP flow or UDP flow which is traffic generated in the internet based on TCP/IP protocol, such as FTP, DNS, WWW, P2P, etc ..

The key point of classification is to select the classification method of the TCP flow or UDP flow.

4.3.2 Traffic classification methods and comparison

Today’s traffic classification methods include: port-based identification, signature-based identification, based on BLINC recognition, identification based on statistics features machine learning recognition and so on.

The advantage of the port-based identification classification method is a principle. The implementation is simple. It can meet the real-time requirement of high-speed networks and it does not involve user privacy and can be implemented by hardware without complicated calculations. This approach had a very good identification effects in the early internet development. However, because more and more applications use the non-standard port now, the traditional traffic identification classification method has become more and more difficult. With the rapid development of the Internet, there are more and more application protocols of the application layer, in particular the emergence of P2P network application protocol which uses dynamic ports and imitates a specific port method to camouflage themselves. Then a lot of network bandwidth resources are occupied, and such flows are increasingly accounting for a large proportion of total traffic, even more than a half network. Therefore port identification cannot meet the needs of traffic classification already, it only can be used as a supplement for other traffic identification methods. There is a need for a more effective method of traffic classification.

Compared with the port-based identification classification method, the application layer’s signature identification classification method has a higher accuracy, demonstrate a good ability to identify traffic types and can be used for real-time traffic classification system. Most traffic monitoring system select this method now, but it is questioned because of personal privacy issues. In addition, this method can only identify known P2P applications, but cannot identify new protocols with unknown signature. In fact the update cycle of P2P is very short, new versions are constantly emerging and the point is that the cost for breaking a private protocol signature is expensive, so this technology has no advantage for some encrypted IP packets.

BLINC identification and statistical features identification methods overcome the difficulties which the first two methods cannot solve. The common advantages of them are high accuracy, good completeness, and ability to identify new applications, and remind users to check those suspected virus attack flows. But the disadvantage of the BLINC method designed by Thomas et al (the method designer) is that its accuracy will be interfered by the IP address translation technology or the equipment position testing. In addition, as this method also heuristic proposed, is based on experience, it leaves loopholes and allows attackers to design a new protocol easily to escape this classification. In short, as transport layer behavior is often closely related with the network environment, the transport layer behavior is likely to be quite different if there is the same application in different network environments. This association limits the application scope of this method.

Although the classification method based on BLINC identification and on statistical features identification both belong to the probability of classification methods, they are mainly based on the transport layer classification. But the advantage of latter is that it does not rely on IP address or port flow numbers, therefore, it does not interfere with by the NAT technologies. But the disadvantage is that some features are extremely sensitive to the dynamic changes of network, such as packet arrival interval, flow duration. In addition, these methods have a common drawback; the calculation is very large and is not available for the high-speed network in real-time classification yet.

From the implementation process, all the methods above belong to the passive measurement method in network measurement and will not have any impact during the classification process. The common drawback is that the above mentioned methods cannot understand some application’s network behavior, such as the most popular P2P file sharing system now. In addition, because the passive measurement requires the interception and detection on packet, with the rapid development of network speed, the time overhead and space overhead to achieve these methods will be increasingly very high.

Nowadays, for the current network traffic classification method, the statistical features network traffic classification method can effectively overcome the problems in the first three classification methods. So it becomes the main research direction in traffic classification field.

This thesis's research direction is based on flow statistical features, using machine learning algorithms, and application layer protocol identification.

The next section introduces some well-known classification methods based on statistical features.

4.3.3 The classification method based on the statistical introduction

For the data mining method, from the machine learning point view, traffic classification can abstractly use mathematical logic as follows: suppose there is a known type set of network flow $C = \{C_1, C_2, ..., C_m\}$ and a known network flow set of type $X = \{X_1, X_2, ..., X_n\}$, through using the machine learning method to "learn" this network flow set, to structure flow classification model $f : X \to C$, This model can be used to classify and predict the unknown type network flow.

Network traffic classification is a typical multiple classification. Generally, network traffic classification is, through the observation points measuring all the TCP or UDP flow' information or property (such as ports, packets contents, connection information, traffic statistic, etc.) which pass the network link or device. Based on this information, we can speculate the upper network application or layer protocol (such as WWW, FTP, P2P, etc.)

The core work of handling the traffic classification problem by the data mining method mainly contains two aspects:

(1) Selecting the appropriate network flow properties, abstract it to the characteristic vector.

(2) Selecting the appropriate machine learning algorithms to build classification model.

In network traffic classification now, the more widely used data mining method is the

Decision Tree classification method, the Naïve Bayes classification method [4, 13, and 14] and the support network machine classification method [15, 17].

4.3.4 C4.5 Decision Tree traffic classification

Data mining is described as a process with two steps. The first step is to structure a model to describe a known data set. Each item in the data set has a category label to identify the category of the tuples. Because every sample already has the category label this is supervised learning. The second step is to use the model structured before the classification. In this step we need to evaluate the accuracy of the classification method. If the accuracy is acceptable then we can use it to classify the data tuples with unknown category label next. In the classification process, we might need to note some problem. Firstly we pre-process the data according to the characteristics of data, such as data cleaning or feature selection. Secondly, we evaluate on the classification method, we need to select the appropriate method to evaluate the method and the evaluation criteria have strong influence on the final result.

Decision Tree [4, 10, and 11] is a common method to structure the data model. The basic thinking is to select a property that is the most able to distinguish the different type samples, and make properties such as the tree root, and divide the training sample into corresponding pieces, then select the property that has the greatest discrimination in the samples as the second layer node, and so on. The process is terminated, when all the leaf nodes only include one category sample, this tree is called decision tree.

The Decision Tree is similar to the flow chart of the tree structure, and each internal node represents a test on a property, each branch represents the test result, each leaf node represents a given category and the root node is the beginning point of the decision tree.

Handling classification problems using decision tree has two steps generally: the first step is the learning on the training data set to form the decision tree classification model. The second step is to use that decision tree classification model to classify the sample as unknown category.

The key of using a decision tree for classification is to structure an effective decision tree model; the structure process usually has two stages: Tree Building and Tree Pruning. After tree building, the decision tree is not the simplest and most compact, because many branches may reflect the noise or some isolated points of the training data; the tree pruning process tries to detect and remove such branches, to improve the classification accuracy on unknown data sets.

Currently, the most influential decision tree algorithm is ID3 proposed by Quinlan in 1986 and C4.5 proposed in 1993. C4.5 is an improved algorithm compared with ID3, according to the information gain ratio to select the test property, not only can it handle the discrete values property, but also can deal with continuous values property.

For non-discrete network flow attributes, the C4.5 decision tree algorithm uses the strategy to discrete its value space and change it to the discrete form to calculate. The C4.5 decision tree algorithm completes the process top to down, selects the property with the maximum information gain ratio as a test property. In order to remove the abnormal branch caused by the noise point or outliers, the C4.5 decision tree method uses the remaining sample obtained from the training data to prune the initial decision tree and then obtain the final C4.5 decision tree.

In the model construction and sample forecasting process, the C4.5 decision tree method does not rely on the distribution of network flow samples; therefore, this method can effectively avoid the possible impact made by the changes of network flow sample change and has good classification stability. When we use the C4.5 decision tree to treat classified sample to predict classification, we only need to compare top-down according to the property value of the network flow sample, then we can find the appropriate leaf node. This treatment is relatively simple and highly efficient.

4.4 Chapter Summary

This chapter mainly introduces the application in network. It first gives some basic concepts and definition on network and network traffic, describes the properties of data in the traffic, gives the definition of property, analyze the type of property, introduces the process of feature abstraction on the network traffic, derived a concept from the perspective of data mining to do the network traffic classification. At the same time, it discusses several common network traffic classification methods based on data mining, and the application in the network traffic classification.

5 Data mining application in business project

5.1 The practical application of data mining

What data mining addresses is finding valuable hidden events in a huge database, and then analyze them to obtain meaningful information, summarize them into a useful structure, as the basis for decision-making in enterprise. Thus data mining application is very extensive. As long as the enterprise has the analysis value and demand, it can use tools for a purpose excavation analysis. Common application cases occur in the retail, manufacturing, finance and insurance, communications and medical services.

5.1.1 Typical business problems solved by data mining

It should be stressed, from the beginning that data mining technology is application-oriented. At present, in many areas, data mining is a very fashionable word, especially in banking, telecommunications, insurance, transportation, retail (e.g. supermarkets) and other commercial areas. The typical business problems that can be solved by data mining include: Data Marketing, Customer Segmentation & Classification, Profile Analysis, Cross-selling as well as market analysis Churn Analysis, Credit Scoring, Fraud Detection etc..

5.2 Data mining applications in marketing

Data mining technology has some more common applications in the enterprise marketing, it is based on the principles of market segmentation in marketing. The basic assumption is that “a consumer’s past behaviors is the best explanation for his propensity to consume in the future”.

Through the collection, processing and handling a lot of consumer behaviors information, determine the interest, consumption habits, consumer trends and consumer demand of some specific consumer groups or individuals, then infers the corresponding consumer groups or individuals for the next consumer behavior. On this basis, we can take target marketing with specific content to the identified consumer groups, if we compare this with the traditional mass marketing which does not distinguish the consumers characteristics, this method offers significant saving marketing costs, improves marketing effectiveness, thus bring more profits for the enterprise.

5.3 Success cases

(1) Telephone charges and management approach

A telephone company in the BC province in Canada, asked the KDD research group in Simon Fraser University to look at their ten years old customer data and summarize, analyze and propose new phone charging and management practices. The research group developed policies conducive to the company and also helped the customers.

(2) American Auto Trader.com is the largest car sales site in the world; there are large numbers of users visiting their site to seek information. They use SAS software to do data mining, analyze the data everyday to find the user’s access patterns, evaluate the degree of liking of the products, and set a specific server then get the success.

(3) Bass Export is one of the largest beer importers and exporters in the world, engaged in transaction overseas in more than 80 markets, sending 23000 orders each week, Bass Export needs to understand each customer’s habits, such as brand preferences, etc., Bass Export uses Intelligent Miner developed by IBM to solve these problems very well.

6 Data mining test on WEKA

6.1 Introduction of WEKA system

WEKA's full name is Waikato Environment for Knowledge Analysis, the abbreviation of this software also is a unique bird in New Zealand. Interestingly, the main developer of WEKA just comes from New Zealand's University of Waikato. WEKA is completely open software for data mining work provides a unified interface, collects the most classic machine learning algorithm and data preprocessing tools. As a complete knowledge acquisition system, it includes the data preprocessing, classification, clustering, and association rules, attribute selection, and achieves visualization in a new interactive interface. We can compare the result obtained from the different methods, to find the best algorithm for solving the problem.

The implementation of WEKA from the accumulation research in machine learning field was carried out by Eibe Frank et al (the developers); the WEKA version before 1998 was implemented by using C++. After 1998, Eibe Frank started a program using JAVA. For this move, he was assisted by the other members in the project team and some free software developers at that time. In August 2005, in the 11th ACM SIGKDD International Conference, the WEKA group of the University of Waikato won the supreme service award in data mining and knowledge discovery field. The WEKA system has been widely recognized by society, and is known as a milestone in the history of data mining and machine learning filed. It is now the most comprehensive data mining tools in the world, and so far it has 11 years of development history.

6.2 The characteristics of the WEKA system

WEKA is a free for academic license, not integer with other systems. As a typical representative of the academic data mining, it has the following characteristics:

(1) Cross-platform, it supports Windows and Unix, and many other operating systems;
(2) It supports the structures text file, the data mining format (C4.5), and provides database interface (JDBC);
(3) It can handle the data types of continuous, discrete, characteristic, date types.
(4) It provides the missing value treatment, elimination noise, standardization, data discretization, attribute structure, transform variable, split data, data balance, sample sorting, sample shuffle, data clustering, dimensional reduction, value

reduction and sampling operation;

(5) It can complete preprocessing, classification, clustering, association, visualization and other tasks;

(6) It supports machine learning and neural networks;

(7) It provides algorithm combinations, users embedded algorithm, algorithm parameter settings (basic, advanced);

(8) It can generate basic reports, test reports, output format, implementation model explained, model comparison, data score function;

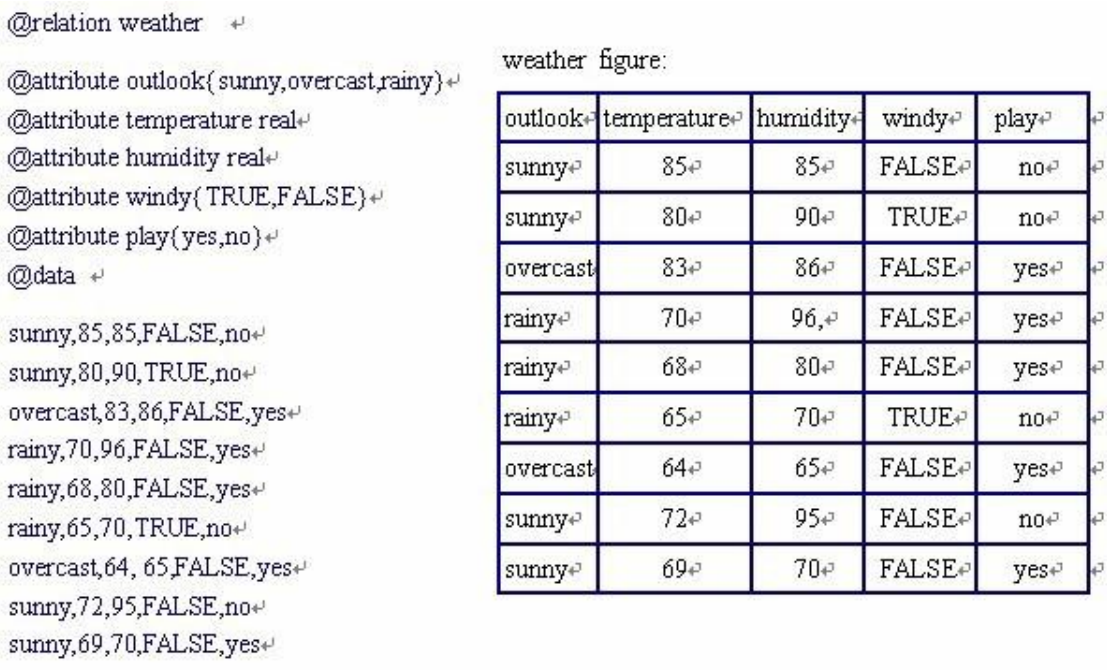
(9) It achieves data visualization, mining process visualization, and the mining result visualization (comprehension, evaluation).

Many characteristics of WEKA can also reflect the function of WEKA. The WEKA data mining platform completely, practically and at a high level achieves a number of popular learning programs; these programs can be directly applied to practical data mining or research. In addition, it also provides a framework for the form of JAVA class libraries; this framework supports the embedded machine learning applications, and even the implementation of new learning programs.

6.3 The file format of the WEKA system

The WEKA system supports three types of data file to open, respective that imports from the local data file, data site or database to be tested. However, whichever way the slide to open, WEKA always has a certain limit on the format of the imported data.

WEKA uses a data format called ARFF (Attribute-Relation File Format), this is an ASCII text. The ARFF file is composed by a set of examples; the weather data in Figure 6.1 corresponds to the ARFF file is shown below: in the form, a transverse called an instance is equivalent to a sample in statistics, or a record in a database. The vertical line is called an attribute, is equivalent to a variable in statistics, or a file in the database.



```
@relation weather

@attribute outlook{sunny,overcast,rainy}
@attribute temperature real
@attribute humidity real
@attribute windy{TRUE,FALSE}
@attribute play{yes,no}
@data

sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64, 65,FALSE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
```

weather figure:

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96,	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	FALSE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes

Figure 6.1 A data file sample for WEKA

It can be seen from Figure 6.1 that the ARFF data format is relatively simple. Specific instructions are as follows:

The ARFF file can be divided into two parts. The first part gives the Head information, including a statement of relations and attribute declarations. The second part shows the Data information, the given data in the data set.

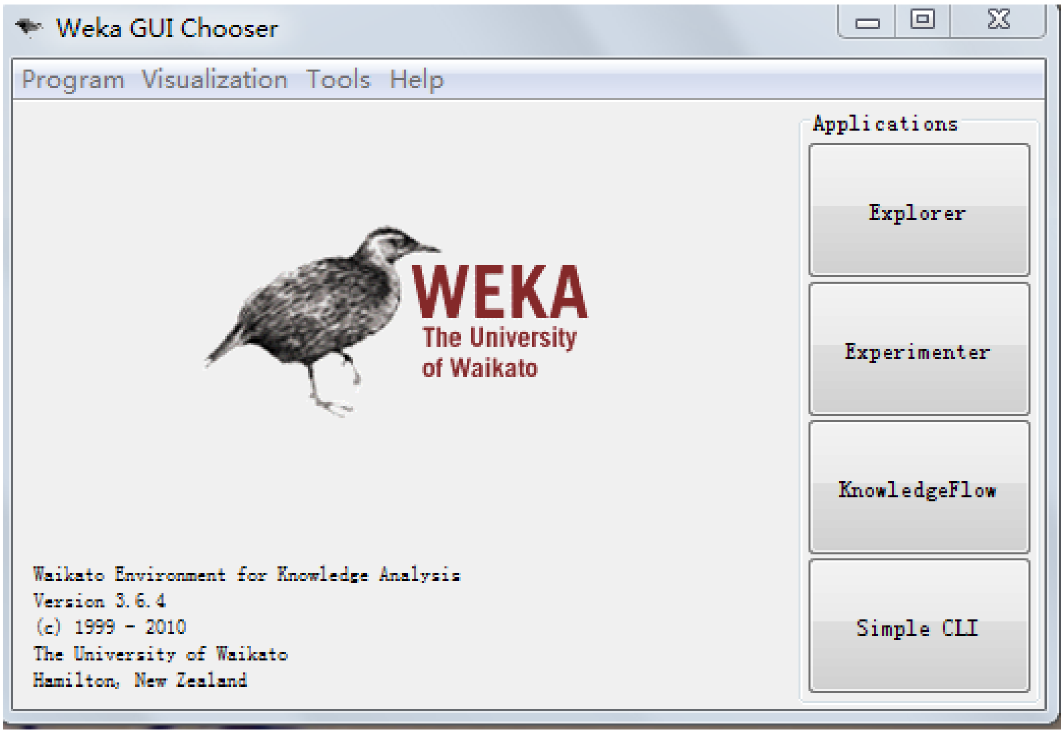
(1) Head information: @ relation defines the data set name, equivalent to the data table name. @ Attribute defines the data set attribute; it contains the attribute name and possible values of attribute or the attribute type.

(2) Data information : @ data defines the start of data set record, the following is all the data sets record, the record is unordered, every data item between each row is separated by comma “,” .Also for the missing data items, we use “?” to express the missing value. But there is no missing value in the sample.

Certainly, when we import the data file, we will find that we can also import the file form with the file extension name. csv (which may be exported by Excel or Matlab); the instance of the C4.5 original file with extension file name is .names and .data, and has been serialized the extension file name is .bsi’s . That is because the WEKA system comes with three kinds of file format converters were: CSVLoader, C45Loader and SerializedInstanceLoader so when the WEKA ARFF file could not be loaded, the system will automatically call the file format converter automatically converter to the additional types of files to ARFF format for testing.

6.4 The system interface

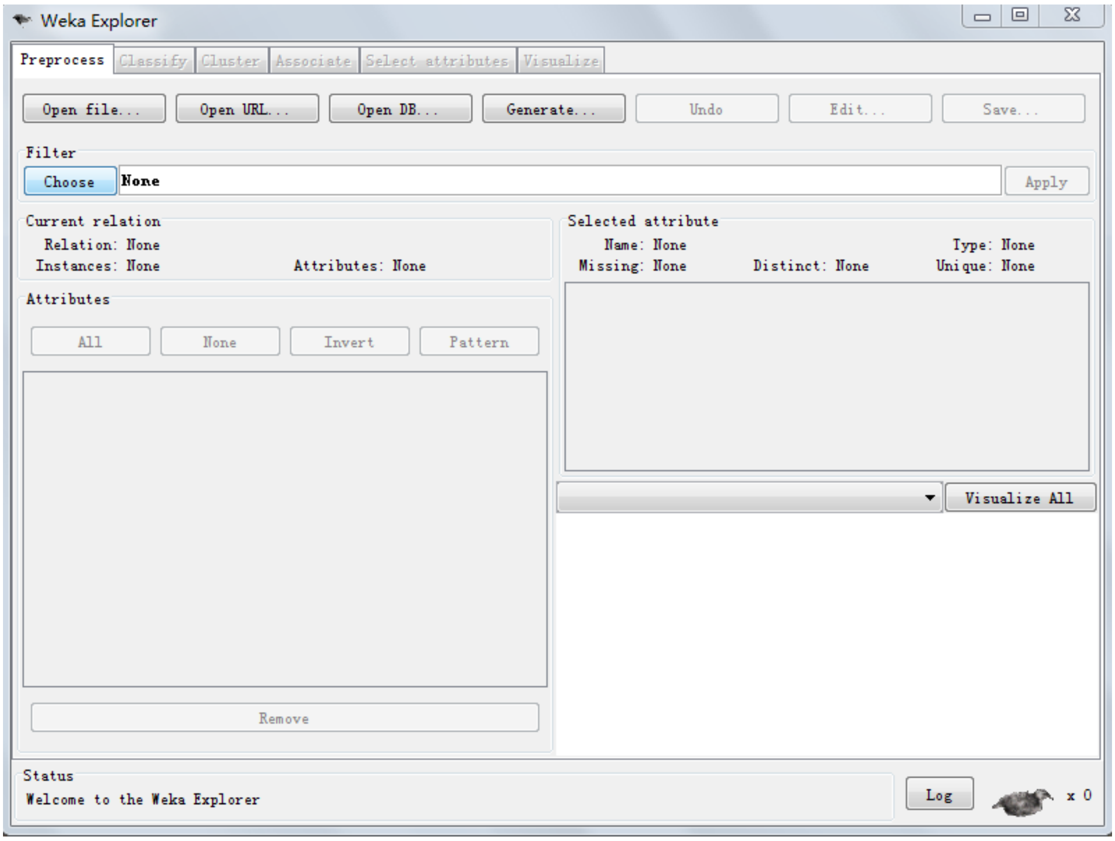
WEKA uses a series of standard machine learning techniques that is unified graphical user interface (GUI), to combine with many pre-processing and post-processing methods, apply many different learning algorithms into data sets, and assess the corresponding results. When the user runs WEKA, the WEKA GUI Chooser interface will appear, as shown in Figure 6.2, including the Simple CLI, Explorer, Experiment, Knowledge Flow.



Waikato Environment for Knowledge Analysis
Version 3.6.4
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Figure 6.2 The interface of WEKA

We click the Explorer button, go into the Explorer graphical user interface, as shown in Figure 6.2.



Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Status: Welcome to the Weka Explorer

Figure 6.3 The interface of WEKA

In Figure 6.3, there are six labels at the top of the WEKA Explorer interface, separately corresponding to different data mining methods supported by WEKA. These include: Process, Classify, Cluster, Associate, Select attributes, Visualize. Through this user interface, all the WEKA functions can be completed by menu selection and form filling. This is does by changing the option into menu, setting the not applicable option as not available, and designing the user options as the form filling shape, to guide the user step by step to completely explore the algorithm in proper order. At the same time, it also gives the tools usage tips in the pop-up window, which is a great help for the users, and the reasonable default values allows the users to achieve the desired results with minimal effort .

In addition, WEKA also contains three graphical user interfaces, as follows:

(1) Experiment interface: It is designed to help users answer the basic problem encountered in the practical application, that is, what methods and parameters can achieve the best result. Although the explorer can also interactively compare different learning techniques, the Experiment interface can make the process more automate and simple.

(2) Knowledge Flow interface: It enables users to set up how to handle the data flow by themselves. It allows users to drag the box on the screen which is express learning algorithms and data source, and get them together to set. This enables the users to combine all parts which separately present the data sources, processing tools, learning methods, assessing tools and visualizing modules together, form a data stream, then realize the incremental batch read and treatment of large data sets, the Explorer can only handle small and medium-scale datasets problems.

(3) Simple CLI: Through running the Simple CLI interface, users can achieve the basic functions of Explorer. Knowledge Flow and Experimenter by WEKA. When the user types a program without any command-line options in the edit box at the bottom of the interface, the panel above the edit box will show all available options: first, general options, then options associated with the program. Through entering the appropriate operation command, the corresponding function can be achieved.

6.5 Project Test

The data mining process in the WEKA system

Before the experiment of WEKA data mining, we should first take a look as the WEKA data mining system process. Each level's brief description of data mining process is described as follows:

(1) Data input layer: This is the preparation phase of the whole data mining. There are three ways of data input, opening the local files, site download, database import. Open the local files can import ARFF, CSV, C4.5, BSI formats.

(2) Data mining layer: This includes preprocessing, classification, clustering and other functions; the preprocessing is the most important part. In this layer, we take the preprocessing on data firstly, and then place the processed data sets into learning programs to carry out appropriate mining tasks.

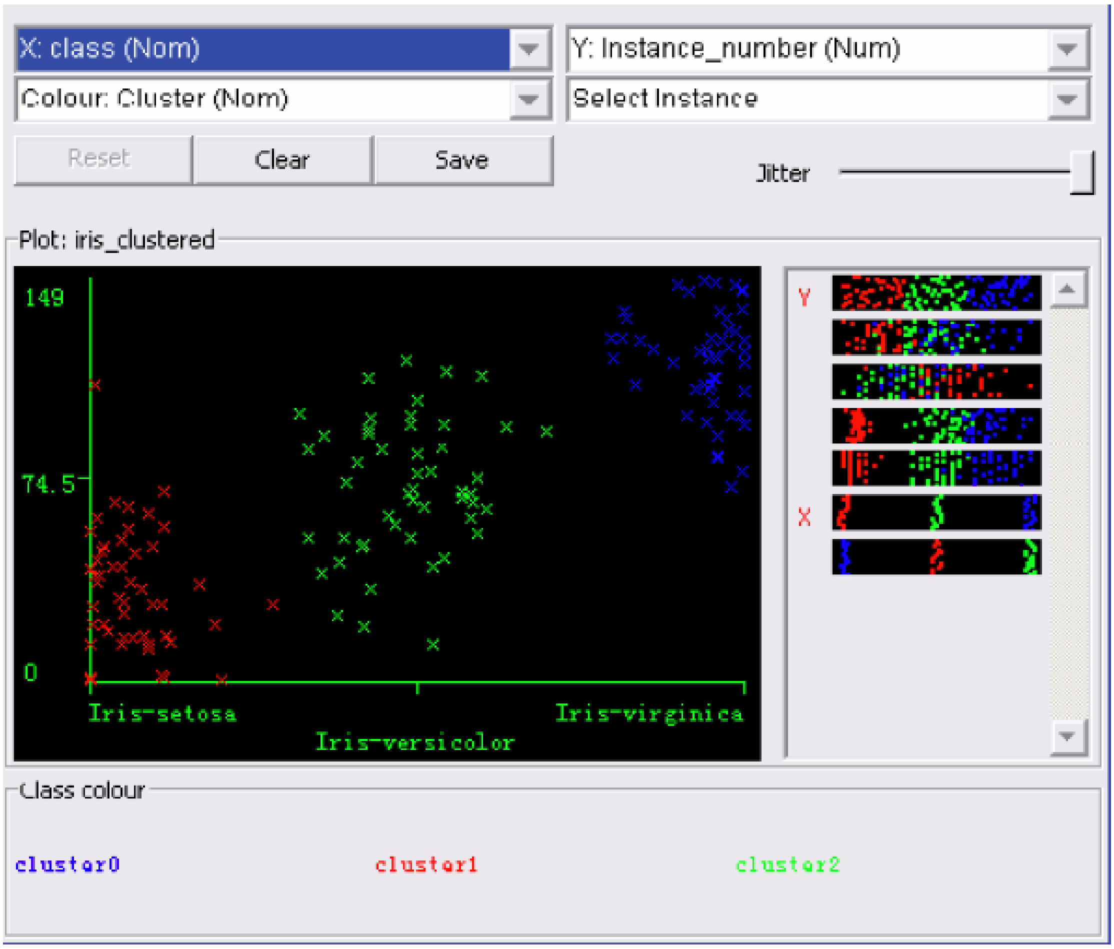
(3) Model evaluation layer: It takes model assessment on the result of data mining, analyzes and studies the results of data mining.

(4) Visualization layer: It achieves data visualization, mining process visualization, and mining result visualization, provides a good support tool for the mining and improves the mining efficiency.

(5) Storage layer: It uses a specific format to store the mining results.

Because this test requires a lot of real data to test, I chose the experimental data and experimental result from the team project I did before during the work placement. Here follows the clustering function test and analysis.

In the modules of clustering function, we chose iris flowers as examples of the test data set, it contains 150 samples of examples, each sample has four attributes, sepal length, sepal width, petal length, petal width, and they are numeric. As we already know in advance that the iris has three categories, setosa, versi color, virginica, so we use the SimpleKMeans algorithm in this clustering experiment. At the same time, we change the number of cluster (numClusters) to 3 in the cluster object edit box, and then we run and see a visual graph clustering. As shown in Figure 6.4, data sets in this figure are divided into three categories, red represents iris-setosa, green represents iris-versicolor, blue represents iris-virginica. Each category of iris has 50 samples; we can click every point in this two-dimensional graph, to see the specific attribute values and iris category that instance to this point.



X: class (Nom)

Y: Instance_number (Num)

Colour: Cluster (Nom)

Figure 6.4 The experimental result

7 Conclusions

This thesis describes the current network environment we are now living, simply analyzes the development and mature status of network technology, discuss the next hot technology spot that can promote the progress of human society, and obtain the current phenomenon “data explosion but lack of knowledge”. We found that people hope to analyze higher level data to make better use of these data, leading to the data mining and knowledge discovery techniques, and made a detailed elaboration and introduction on the data mining method which was proposed in the 80’s of 20th century. The chapter three and chapter four introduce details of the data mining application in network and business, and introduce several successful cases and data mining methods based on statistical features, and a typical algorithm based on this method which is the decision tree algorithm. Finally, the thesis introduces the WEKA software and some interrelated knowledge, and the test process, and a simple test on data mining based on the WEKA platform.

Certainly, because this topic is a field that I have never touched before, there must be some shortcomings either views from a research point or a practical application point. In the future I would like to study deeper, research data mining applications.

As a mining tool, data mining is convenient for analysis. Not only can be used to mine knowledge, but also it can be used for decision support and prediction analysis, it greatly facilitates database management.

Data mining is the same with other technologies, its applications have advantages and disadvantages. Because data mining can extract a kind of knowledge that is not easy to find, if it is not used correctly, it may pose a threat to privacy and information security. To solve this problem, we need to further develop the methodology, in order to ensure privacy protection and information security during the mining process.

References

[1] Zhangwei, Liao Xiaofeng, Wu Zhongfu. A new clustering method based on generic algorithm[J]. Computer science. 2002, 29(6): 114-116.

[2] Lin Sin, Xu Peng, Liu Qiong. Traffic classification based on support vector machines [J]. Computer research and development 2008, 25(8): 2488-2490.

[3] Kim H, Claffy K, Fomenkov M, Barman D, Faloutsos M, Lee K. Internet traffic classification demystified: myths, caveats, and the best practices[A]. In: ACM CoNEXT Conference[C]. ACM: Madrid, Spain, 2008, 1-12.

[4] Data Mining Applications introduction website:

http://wenku.baidu.com/view/594457cda1c7aa00b52acb38.htmlAccessed:2011-04-02

[5] Network Traffic introduction website:

http://baike.baidu.com/view/411702.htmAccessed:2011-04-02

[6] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers: San Francisco, USA, 2000.

[7] Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining[M]. Addison-Wesley: Michigan State, USA, 2006.

[8] Sun Guijie, Liu Jie, Zhao Lianyu. Clustering algorithm research[J]. Software Journal 2008, 19(1): 48-61

[9] Nguyen TTT, Armitage G. A survey of techniques for Internet traffic classification using machine learning[J]. IEEE Communications Surveys and Tutorials 2008, 10(4): 56-76.

[10] Xu Peng, Lin Sin. Traffic classification based on C4.5 decision tree[J]. Software Journal 2009,20(10): 2691-2074

[11] WEKA official website:

http://www.cs.waikato.ac.nz/ml/weka/Accessed:2011-03-01