QUEBEC UNIVERSITY

THESIS

SUBMITTED TO

QUEBEC UNIVERSITY, CHICOUTIMI
IN FULFILLMENT OF

THE REQUIREMENTS FOR THE

MASTER'S DEGREEE IN COMPUER SCIENCE

BY

Ying Zheng

TITLE OF THESIS
ANALYSIS OF CREDIT CARD DATA
BASED ON DATA MINING TECHNIQUE

APRIL 2009

Bibliothèque
Paul-Émile-Boulet

UQAC
Université du Québec
à Chicoutimi

Mise en garde/Advice

Afin de rendre accessible au plus grand nombre le résultat des travaux de recherche menés par ses étudiants gradués et dans l'esprit des règles qui régissent le dépôt et la diffusion des mémoires et thèses produits dans cette Institution, **l'Université du Québec à Chicoutimi (UQAC)** est fière de rendre accessible une version complète et gratuite de cette œuvre.

Motivated by a desire to make the results of its graduate students' research accessible to all, and in accordance with the rules governing the acceptation and diffusion of dissertations and theses in this Institution, the **Université du Québec à Chicoutimi (UQAC)** is proud to make a complete version of this work available at no cost to the reader.

L'auteur conserve néanmoins la propriété du droit d'auteur qui protège ce mémoire ou cette thèse. Ni le mémoire ou la thèse ni des extraits substantiels de ceux-ci ne peuvent être imprimés ou autrement reproduits sans son autorisation.

The author retains ownership of the copyright of this dissertation or thesis. Neither the dissertation or thesis, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

PERFACE

The main point of the subject is to analyze the potential information in credit card data set. Data mining is a kind of new technique in intelligent information processing. It uses some algorithms to analyze the potential information in database. In the analysis, a serials procedures are to be done in the thesis, it include, data preparation, data mining analysis, and knowledge presentation.

The data set analyzed in the thesis is provided by ICBC Tianjin branch. It contains 2002 credit card transaction data and basic information of a card holder.

In clustering analysis, one of cluster algorithm, *k*-means is used to analyze the credit card consumption data. Category are assign as 2, 4,8,12. In classification analysis, BP network is used to classify the customers into VIP, normal, and stop payment. In association rule analysis, the associations among basic information of cardholder are found.

It is glad to say the after applying the data mining algorithm on the data, some interesting and useful results are gotten. The results can be seen in Chapter 5.

ACHKNOWLEGEMENT

I would like to thank Prof. Cheng Ming. His instructions on the project give me lots chances to practice my computer knowledge.

Thanks to Mr. Zheng Gang for his assistance and help in realizing the implementation.

I would like to thank all my friends and fellow graduate students here at the University of Technology and Science of Tianjin. The assistance and encouragement offered by them has been a great and unexpected aid in my Master's work.

ABSTRACT

In recent years, large amounts of data have accumulated with the application of database systems. Meanwhile, the requirements of applications have not been confined in the simple operations, such as search and retrieval, because these operations were not helpful in finding the valuable information from the databases. The hidden knowledge is hard to be handled by the present database techniques, so a great wealth of knowledge concealed in the databases is not developed and utilized mostly.

Data mining aimed at finding the essential significant knowledge by automatic process of database. DM technique was one of the most challenging studies in database and decision-making fields. The data range processed was considerably vast from natural science, social science, business information to the data produced from scientific process and satellite observation. The present focuses of DM were changed from theories to practical application. Where the database existed, there were many projects about DM to be studied on.

The paper concentrated on the research about data information in credit card by DM theories, techniques and methods to mine the valuable knowledge from the card. Firstly, the basic theories, key algorithms of DM techniques were introduced. The emphases were focused on the decision tree algorithms, neural networks, *K*-means algorithm in cluster and Apriori algorithm in association rule by understanding the background of bank and analyzing the knowledge available in the credit card. A preliminary analysis of credit card information, Industry and Business Bank at Tianjin Department, was performed based on the conversion and integration of data warehouse. The combined databases including information of customers and consumptive properties were established in accordance with the idea of data-warehouse. The data were clustered by *K*-means algorithm to find valuable knowledge and frequent intervals of transaction in credit card. Back propagation neural networks were designed to classify the information of credit card, which played an important role in evaluation and prediction of customers. In addition, the Apriori algorithm was achieved to process the abovementioned data, which could establish the relations between credit information of customers and consumption properties, and to find the association rule among credit items themselves, providing a solid foundation for further revision of information evaluation.

Our work showed that DM technique made great significance in analyzing the information of credit card, and laid down a firm foundation for further research in the

retrieval information from the credit card.

Keywords: Database; Data; Credit card

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANN	Artificial Neural Networks
ART	Adaptive Resonance Theory
BP	Back Propagation
CLA	Clustering Large Applications
CLIQUE	Clustering In QUEst
CURE	Clustering Using Representatives
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENCLUE	Density-based Clustering
DM	Data Ming
ICBC	Industrial & Commercial Bank of China
OPTICS	Ordering Points to Identify the Clustering Structure
KDD	Knowledge Discovery in Database
RDB	Relational DataBase
PAM	Partitioning Around Method
SKICAT	Sky Image Cataloging and Analysis Tool
STING	STatistical INformation Grid

CHAPTER 1

INTRODUCTION

MOTIVATIONS AND OBJECTIVES

Credit card has been widely applied in the social life as a kind of financial service, while its use in China have lasted for about 28 years, beginning to act as an agent of foreign products since 1978. The boom in credit card was prevailing in many provinces in China. For example, "the card was applied in 35 cities among Guangdong province with the amount of 360,000 and its balance above 800 million YUANS, and cumulative expenditure of 350 million YUANS, and the transaction amount was 18 billion YUANS,"(Translation) [1] so it played an important role in the development of bank business. The analysis and study on the credit card had theoretical and practical significances.[2,3]

The daily consumption of credit card could produce a great quantity of data, including many unknown characteristics of customers. Because the card was used by individual, information, such as daily consumptive habits, could be obtained by mining the data stream of credit card, which promoted consumption by assigning credit card to right consumptive mass. The data mining of credit card could be mainly focused on the following aspects:

- The establishment of model for forecasting individual consumption: a large part

of bank profits from credit card originated from expenses of the card. About 2-3 percent of consumptive charge was withdrawn from every customer by using credit card for payment, which caused banks to assign credit card to the active customers when they expanded their business. The amount of credit card was

- Assigned about a few hundred thousands and every customer would receive several consumptive bills, and several advertisements at same time in a year. [2]The same advertisements would be sent to all customers if we did not analyze their consumptive habits in advance, but the customers could be classified in to different kinds of groups based on their consumptive places after we analyzed their habits, which could assign the classified advertisements more exactly to individual, so the earning of banks could increase consequently. This method of analysis and classification were very useful in bank business. Customer classification: the consumptive capacities of customers could be forecasted by analyzing a large amount of business data in credit card. The customers could be classified into three grades, namely VIP, normal and bad, to advise assigning credit card on the basis of different credit grades and information.

- Consumptive analysis of customers: their consumptive behaviors were analyzed on the basis of different credit grades and information to obtain the relationship between the customers and their consumptive properties. The places, date and expenses associated with the customer information were classified into different groups to find the suitable consumptive community.

Our objectives were to analyze data of credit card by mining data information. The project was about the data of credit card provided by Industry and Business Bank at Tianjin Department.[3] The data include 8,000,000 records, namely one year's transaction data. The data mining techniques involved the methods of clustering, classification and association rule. We analyzed and mined a great amount of data to obtain some research results, and guided the practical business of credit card. This technique showed an important practical value in bank business of credit card. The mining technique, belonging to pioneering research in the computer science, had a bright future. Its application in studying on the credit card was just an attempt in financial field, and had many problems to be solved. For instance, the prediction of load venture was one of the important projects, and became the important foundation of bank business, because it involved in the whole management of bank. Our exploratory objectives were expected to provide instructions for establishing the entire analysis system.

1.2 THESIS CONTRIBUTION

The objective of this thesis was to develop flexible products from available data of credit card through mining algorithm, and to set up a solid foundation for the future research. The concrete studies were as follows:

- The mining data system was established by integrating data of credit card and pre-treating essential information of credit card.
- Knowledge types available were analyzed on the basis of associated data of credit card.

- We concentrated on mining cluster, classification and associated rule in order to guide customer classification, characteristics and credit evaluation, and so on.

The research was exploratory in two following parts. One was to study on the mining data of credit card, while the other was to prepare for establishing entire data system of credit card.

1.3 THESIS ORGANIZATIONS

The after mentioned contents belonged to the following chapters.

Chapter two was mainly about the research background. First, concept of mining data, research background and application were presented. Second, process of mining data and main functional module were introduced. Third, main mining techniques involved were discussed. Finally, practical applications of the research were investigated.

Chapter three was mainly about organization format of data in bank business of credit card, and pretreatment of mining and analyzing data.

Chapter four was mainly about principles of cluster, classification and associated rule, methods of applying these techniques into practical data of credit card, and possible results.

Chapter five was mainly about the experimental platform, results and analyses when we put the methods of cluster, classification and association rule into practical data of credit card in bank business

Chapter six was mainly about conclusions of the whole paper and research interests in the future.

CHAPTER 2

LITERATURE REVIEW

Chapter two was mainly about the research background, concept, application and technique process of data mining, principles and application methods of data mining.

2.1 BACKGROUND OF DATA MINING

In recent years, a great amount of information has been produced because of the advance in database techniques and progress in the methods of collecting information, so the total quantity of data in database is most tremendous. "Hundreds of tables, millions of records cause the database capacity to reach GB bytes, even TB bytes."(Translation).[4] If we want to make the data become real resources, we must make the best use of them for our decision-making service, or else the data may become a huge burden, and even rubbish. To face the situation of abundant data and scarce knowledge, it is urgent to develop the theories, methods and techniques about distilling and seeking information.

Traditional database functions included basic addition and deletion, search and statistic and so on, so it was very difficult to process data of database in a deeper level. If we wan to obtain the general characteristics and trend forecast of database, we must have a higher degree of automation and more efficient data processing method to help us to analyze the tremendous data. Data mining was just a kind of needed techniques.

Data mining (abbreviated to DM) was a process to extract implicit or unknown but

potentially usefully knowledge and information from tremendous data. There were many terms expressing the similar meaning, such as knowledge discovery in database (abbreviated to KDD), data analysis, data fusion and decision support and so on. The analyzed data could be organized, such as the data in correlative database, or semi-structured, such as text, graphs, and medium data, even the data on the internet. The methods of KDD might either be mathematical or nonmathematical; deductive or inductive. The knowledge discovered could be used in information management, search optimization, decision making, process control or data self-maintenance. DM aimed at practical applications at its very beginning. The data were intended to discover the correlative associations among affairs by the methods of statistics, analysis, combination and deduction, which could solve the practical problems, and even predict the future activities.[5,6]

"The term, KDD, first appeared on the 11th international artificial intelligence conference in August of 1989,"[7] and now has become one of the highlights of the whole computer filed. Journals or monographs about KDD research were issued in the fields of database, artificial intelligence, information process, and knowledge engineering and so on. For example, Journal of Knowledge and Data Engineering first issued technique monograph about KDD in 1993. [8]

DM was widely used in the following fields. For example, the famous application system in astronomy, namely SKICAT(Sky Image Cataloging and Analysis Tool),[9] information analyses about customer behaviors, inclination and interests in marketing strategy,[10] typical financial assessment about investment evaluation and stock

prediction in the field of financial investment, prevention of banks and business from financial legerdemain,[11] communication management[12] and internet[13] and so on.

REVIEW ABOUT DM TECHNIQUE

DM was a process that established the correlative relations between models and data by all means of analyses. These relations included the classification, associations among data. The relations and models could be applied into information prediction. A preliminary forecast model was established on the given data, and then the model was tested and evaluated by other data. Finally, the model was put into practice. For example, the relations between consumptive behaviors and characteristic information of cardholders, such as age, education, income and so on, were examined to establish consumptive model in analysis system of credit card, and then the model was tested by a large quantity of data in credit card for applying into practical business.

General frame of system

Simply speaking, DM was a process that distilled and mined knowledge from a large quantity of data. DM was known as an essential part of KDD, but in many cases, DM was to be seen equivalent to KDD. From the flowchart given by Fayyad[14], we could know the practical process of KDD and DM.

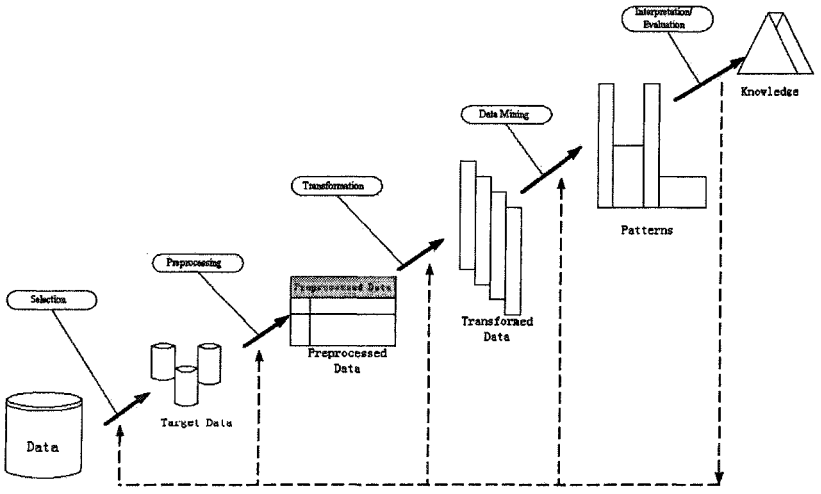


Figure 2-1. General frame of KDD (DM) provided by Fayyad

DM was an entire process executed only after a large amount of data was well prepared. Especially, DM was associated with database, so the work about database must be prepared well in advance. Meanwhile, the mined results must be explained and proved in further researches. The general frame included the following aspects:[14]

- **Determination of objectives:** applications and relative knowledge must be investigated and understood in order to establish the ultimate objective from customer's purposes.
- **Establishment of data set toward objectives:** data set, subset or examples were chosen in the DM process.
- **Data process and pre-treatment** : For example: elimination of data noise, collection of essential model information, determination of strategy for processing data.

- **Data reduction:** useful properties of data were independently chosen on the basis of DM objectives. In some cases, we could reduce the numbers of variables by multi-dimension simplicity, or search irrelative of data to reduce the mining range and improve the mining efficiency.
- **Algorithm matching:** objectives of DM were fitting for the proper algorithm of DM, such as classification, cluster and regression and so on.
- **Selection of DM algorithm** : The specific DM algorithm was selected. The process included the determination of proper models and parameters, and selected proper algorithm for DM.
- **DM** : Namely the practical DM process. The interested modes or similar groups, such as classification rule or conceptive tree, regression and cluster, were selected in a typical table of data.
- **Conversion of DM models:** the mining models were expressed in an explicit and intelligible manner.
- **Evaluation and confirmation:** the mining knowledge was combined with other information, or presented to interested organizations for further confirmation.

2.2.2 Main function modules in DM

2.2.2.1 Pretreatment

"The objectives of pretreatment were to overcome the limitations in present analysis method of DM"(Translation)[15], namely the mining data should be relatively integrated, and had less superfluous information. The interested data in reality might

have some drawbacks, and could not be applied into practice directly. These drawbacks included fragmentary, noise and inconsistent data, therefore, the data should be processed and pretreated in advance, including three following courses: data process, integration and conversion, reduction. [16]

2.2.2.2 Data mining

The DM was the most important part in KDD, so the performance of DM algorithm immediately affected the property of mined knowledge [16]. The present researches were concentrated on the DM algorithm and its applications. People often discriminated loosely DM from KDD, and mixed them up. Generally speaking, DM was a process of deeper analysis. The figure 2-2 [16] was schematic diagram of DM system.

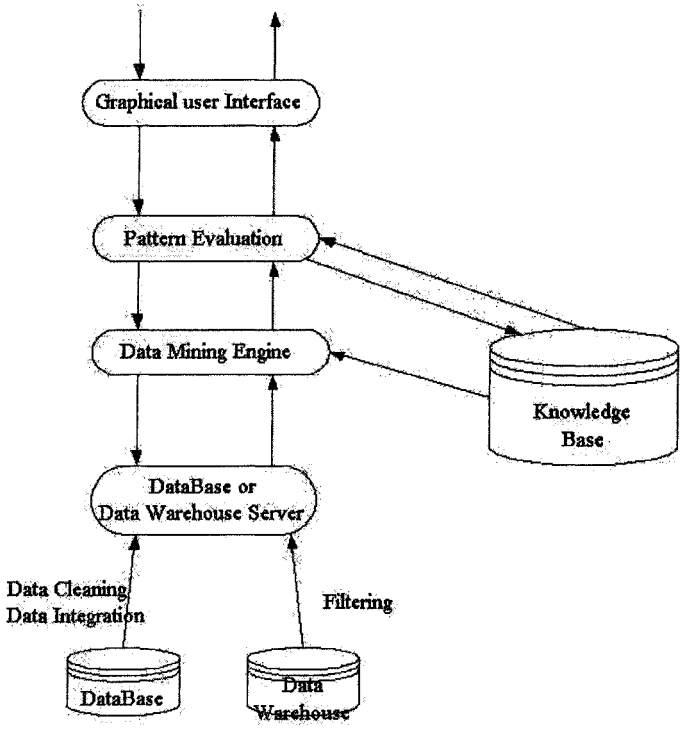


Figure 2-2. Typical schematic diagram of DM system

Some concepts, knowledge base, data mining engine, pattern evaluation module and graphical user interface, must be further explained here. Knowledge base was a degree of interests in searing and evaluating of main parts of knowledge. Data mining engine was a most important part of DM systems, including a series of functional modules, such as summary rule, eigenvector, associated rule, cluster, classification, trend, and evaluation and deviation analysis. Pattern evaluation module aimed at finding the most suitable model mutually associated with data mining engine in order to evaluate and mine the model repeatedly on the basis of a degree of interests. Graphical user interface was a place where users could communicate with DM system through data mining language.

"There were many models of data mining. Based on the different functions, the models were divided into two main categories, predictive and descriptive." (Translation)[17]

Predictive model could determine and predict results on the basis of given values of data, for example, the model was established on the information of animal, if definition of all viviparous animals was mammals, you could judge whether the given animals was viviparous based on this model.

Descriptive model was a description of rule in database, or classifies data in different groups based on their similarities. This model could not be applied into direct prediction, for example, about 70 percent of Earth surface was covered by water, while other 30 percent surrounded by land. This model included statistics and visual

information. For example, we may consider the relation between degree of credit and customer's information, such as age, gender, martial status, or information about the time when you began to work. Therefore, the data in n dimensions of space must be skillfully displayed in two dimensions by the visual instruments. The data were divided into different groups by cluster model with larger differences between groups and smaller differences in groups. Generally speaking, meaning of the term should be understood by knowledgeable people. Association rule were correlative rule among data. The demonstrative example was as follows: those who were unable to repay loan had a month income below 3000 YUANS. Sequence model was similar to associated model, which associated the data with time. In order to establish sequence model, we must know whether specific affairs occurred, more importantly, we must know when the affairs occurred. For example, 60 percent of people who bought color TV might also buy video disc player in three months.

2.2.2.3 Knowledge Presentation

The results of DM should exactly express its requirements, and were easy to understand. Knowledge discovered was examined from different views, and expressed in different ways. The requirements and results of DM were described in high-level language and graphical interface. "Many KDD systems and instruments were short of human-computer interaction at present time, so it was hard to take advantage of knowledge."(Translation) [17]

2.2.3 DM techniques in analysis

Credit card issued by banks has widely used in social life as a kind of financial products. Large amounts of data were produced in daily use of credit card, including many unknown characteristics behind credit card. Because credit card itself was used by individual, people's information, such as their daily consumptive habits, Because the card was used by individual, information, such as daily consumptive habit, could be obtained by mining the data stream of credit card, which could promote consumption by assigning credit card to right consumptive mass. The following techniques were adopted in the data analysis of credit card. [5,12,17]

- Method of Neural network :

  It could establish three kinds of models of neural networks on the basis of MP model and Hebb knowledge rule by simulating nerve cells of human brain. a) front feedback network: it was a typical model that could perceive back propagation and functional networks, which could be applied into prediction and pattern recognition; (b) feedback network: it was a typical model representative of Hopfield discrete model and successive model, which could be applied into associated memory and optimal calculation; (c) self-organized network: it was a typical model representative of ART model and Koholon model, which could be applied into network connection with neural network knowledge. It was a kind of distributed matrix structure.

- Cluster :

It was used in the groups of data set, which made the minimal differences between the groups, while the maximal differences in a group. The cluster discovered could be applied into explaining characteristic of data distribution. In many cases of commercial applications, the method of cluster could obtain the characteristics of different customer groups, which allowed businesses to work out plans in accordance with customers' practical needs, and predicted their consumptive models on the basis of customers' habits. The technique adopted was *K*-means.

- Association rule :

Association rule, one of the main models in DM, aimed at finding the relations among different projects of database. These rule were intended to find the behavior modes of events and people, for example, the correlation between a sold item and another item. These rules could be applied into the design of commodity shelves, storage arrangement, and customer classifications based on purchasing pattern. Data association could play an important role in many cases, such as the analyses of stock, bank deposit. The technique adopted was Apriori algorithm.

2.2.4 Present problems

The relative system was not available in analyzing the customer consumption of credit card in our country, especially dynamic system, adjustment with the practical

cases, was much less. The present problems in analysis of credit card were as follows: [1,2,3]

- A large amount of data. for example, "the abovementioned transaction data stored in Industry and Business Bank at Tianjin Department reached 8 million pieces of records just in a year," [2] (Translation) and the document occupied the space about 1 GB. The enormous data made a demanding requirement towards the algorithm complexity of DM and computer performance.
- Fragmentary data. Due to the rapid development in bank business and continual changes in bank operation and rule, the cumulative data were inconsistent and fragmentary, which might cause some difficulties in accurate analyses and calculation. [2]
- Selection and conversion of data. "Properties of different databases should be combined and converted to form new data collection by the methods of recursion and addition."(Translation) [3]
- Standard verification and evaluation of issuing credit card because credit evaluation was a very complicated thing involving in many factors. We must establish the proper rule from a large amount of data to instruct re-establishment of the standard. [3]

CHAPTER 3

BANK REQUIREMENTS AND DATA PRETREATMENT OF CREDIT CARD

Nowadays, there are many uninterested, fragmentary or noise data in the credit card database, so the present data can not be minded directly, we must change the properties, format of data into requisite format stipulated in DM algorithm for further efficient mining operation, therefore, the data must be pretreated in advance. The pretreated process of data includes filter, selection and conversion.

3.1 BANK'S REQUIREMENT ON CREDIT CARD INFORMATION ANALYSIS

Since credit card is widely used in human society, there exists lots of information in it. For example, information of cardholder, credit card transaction records, etc... Banks wanted to know, the real activities of every cardholder when they use their credit card. Since these kinds of information are different among individuals, and it did not represent the group activities, therefore, bank had the requirements on collecting the individual information, summing up them together, use some analysis methods on them. The results were used to predict the future tendency. The requirements given by bank (Industrial & Commercial Bank of China) are list below briefly.

- Consuming model of credit card
- Study of model establishing on individual credit architecture
- Study of credit evaluation standard, classification methods will be used in credit

model

- Classification study on the level standard of customer
- Tendency analysis of customer consuming
- Risk study of credit card information and card holder

3.2 CONSTITUENT ANALYSIS OF INFORMATION

3.2.1 Constitution of bank database

The organization format of credit card information in bank database should be examined. These formats included flowing transaction database of credit card, evaluation database of customers (customer questionnaire of credit evaluation), database of customer information, database of payment lists, database of VIP customers. These databases had reference to each other. The concrete relations were described in figure 3-1. [18, 19]

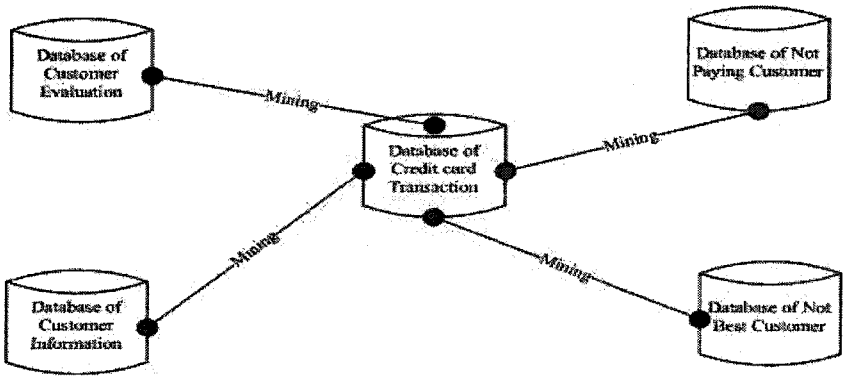


Figure 3-1 Organization structure of credit card database

3.2.2 Field Meaning

- Field meaning of flowing transaction database of credit card. The transaction data of credit card and our DM research were mainly concentrated on the database, which included 20 groups of data. Every transaction data of credit card was recorded in the database[18,19] (seen in table 3-1).

Table 3-1 Description database field of credit card transaction

OPER1	Char(4)	AMTSIGHN	Char(1)
CARD No.	Char(16)	AMT	Decimal(9,2)
ACCNO	Char(11)	BALSIGHN	Char(1)
NAME	Char(8)	BAL	Decimal(9,2)
TXCODE	Char(4)	TXSEQ	Char(4)
TXTYPE	Char(1)	DRITEM	Char(5)
ECFLAG	Char(1)	CRITEM	Char(5)
TXDATE1	Char(8)	TXDATE2	Char(8)
PLACE	Char(8)	TERM	Char(3)
DRCR	Char(1)	ZZACCT	Char(12)
MEMO	Char(1)		

- Field meaning of VIP database. This database was derived from the statistics of transaction data supplied by credit card department of bank (seen in table 3-2).[18] Because of the preliminary analysis and discussion of credit card information, we were only interested in the concrete number of credit card. If we obtained the number of a specific credit card, we could associate VIP database with questionnaire of credit evaluation and transaction database to obtain a preliminary result. The payment database could be processed by the same method. [18]

Table 3-2 Description database field of best customer credit card transaction

CARD No.	Char(16)
TXDATE1	Char(8)
AMTSIGHN	Char(1)
AMT	Decimal(9,2)

- Field meaning of payment database (seen in table 3-3[19]). The payment lists were key information of credit card. If customers owed a debt or postpone the payment after a long period of time for some unknown reasons, the bank would treat them as dangerous customers, and immediately took measures to suspend the customers from credit card. [19]

Table 3-3 Description database field of Stop Payment customer credit card transaction

CARD No.	Char(16)
IDNo.	Char(15)

- Field meaning of customer database. After evaluation of customer credit and information, if banks were willing to send credit cards to customers, they must establish a database document recording the detailed information of the customers for future use. There were 26 types of data in the database. The database of customer information should be associated with the database of customer consumption to discover some laws, to test the existing laws and predict future laws. Especially, some data in database of customer information were very important parts of questionnaire of credit evaluation, so reasonableness of credit evaluation was improved by the test and modification of credit evaluation. [18,19]

- Questionnaire of credit evaluation. It was an important foundation that whether bank issues credit card to a specific customer. The customer credit was determined to evaluate on the basis of experience available and customer information. Our work was to integrate the transaction database, VIP database and payment database, customer database to find the tree relation in questionnaire of credit card. (see Table 3-4) [18,19]

Table 3-4 Questionnaire of credit evaluation

Content		Score
Education 10%	Bachelor	10
	Junior college	8
	High school	6
	Profession school	4
	Junior middle school	1
Profession 20%	Energy Organization	20
	Government	18
	Hospital, University	16
	Foreign company	15
	Native Big company	12
	Service	10
	Small company (Capital under 500000 YUANS)	8
	Non	3
Title 15%	Senior Executive member, or Professor	15
	Middle Executive member	13
	Junior Executive member	11
	Normal Worker	8
	Other	5
Salary 20%	Over 2000 YUANS a month	20
	1500 to 2000 YUANS a month	18
	1000 to 1500YUANS a month	16
	600 to 1000YUANS a month	13
Age 15%	45 and over	9
	36 to 45	13
	26to35	15
	18 to25	12
Warrantor	Organization	15

15%	Income over 1200YUANS	13
	Income over 1000YUANS	12
	Income over 800YUANS	10
	Income over 600YUANS	8
	Property guaranty	6
Working Period 5%	Over 5-15 years	5
	Under 5 years	3

3.3 INTEGRATION OF DATA INFORMATION

The bank database, a kind of nonstandard relational database (RDB), was the source of flowing transaction database of credit card. Only the conversion of database could the data be read and existed in both databases. One was SYBASE for storing customer information; the other was DB2 for storing payment lists. The amount of data was enormous, for example, Industry and Business Bank at Tianjin Department had issued over 200,000 credit cards, and the number was increasing everyday. We should integrate the data by distilling the data in RDB, SYBASE and DB2, and then combining them into a comprehensive database, where we could mine data. In addition, the data was updated everyday, so we must first take into account of given data, namely mining data on the basis of the given data, and then modify the data set for further mining till the needed information was obtained. The obtained information was called transcendental knowledge, and then the intraday data were analyzed. In fact, this case was involved in the concept of data warehouse which could be seen as storage places for converting the data of information system. But we could not further introduce the concept and methods here. Our work was to establish a data environment, warehouse, by

citing these concepts and methods. The concrete way was described as follows:

- Association of VIP database with transaction data to form new data set for analyzing VIP information and properties.
- Association of customer database with transaction data to form new data set by distilling some data from the above two databases for analyzing some concerned problems.

3.4 SELECTION OF DATA

Selection of data was performed on the two dimensions. First, the line or dimension of parameters was selected. The process was an essential part of DM. Secondly, row or dimension of record was selected on the basis of value of fields. In RDB, whether line or row was selected, the general language selected was SQL, or you could use front instruments of data to process. In order to select suitable data, you must understand the problems and basic data in detail. After the selection of data, data should be pretreated before mining process.

- Flowing transaction database of credit card. First, data item and limitation should be selected. We selected the one-transaction data of credit card as pending data. Secondly, some data in the database might be trivial in DM process, so they should be discarded to prevent calculation from affecting. The data could be saved after discussing with experts in credit card.(seen the data item in Table 3-5 [19]).

Table 3-5 Selected Data fields of Credit card transaction Database

OPER1	Char(4)
CARD	Char(16)
TXDATE1	Char(8)
PLACE	Char(8)
TELLER	Char(4)
DRCR	Char(1)
MEMO	Char(1)
AMTSIGHN	Char(1)
AMT	Decimal(9,2)
BALSIGHN	Char(1)
BAL	Decimal(9,2)
TXSEQ	Char(4)
TXDATE2	Char(8)

- Data of customer information. The customer information of credit card played an important role in analyzing the basic data. Some useful data items should be distilled from customer information after discussing with experts in credit card. Seen in table 3-6. [19]

Table 3-6 Selected Data fields of Customer Information Database

NAME0	Char(8)
BIRTH	Char(8)
SEX	Char(1)
POSI	Char(1)
PROF	Char(1)
TITLE	Char(1)
UNIT	Char(1)

- VIP database and payment database. Because we obtained only a small part of data items, the whole data should be saved.

- Questionnaire of credit information. The data item in questionnaire of credit information should be saved entirely, for they were important in the future data analysis.

3.5 DATA CONVERSION

"Data conversion was a process that continued to dispose of the distilled data."(Translation) [20]This process sometimes contained some new data items produced from one or several fields, which meant to replace some fields with a larger amount of information, including the following aspects of conversion.

The conversion of credit card data included the following steps:

- Flowing transaction database of credit card were segmented into small documents. Because data were stored as text format, we divided the database into 110 pieces. The content of every piece was about 9 MB, containing above 60 thousand records. [19]

- Some typical data of credit card were referred to us by bank for treatment. The data included VIP database, normal customer database, and payment database. Through the number of credit card, we could find the proper data from flowing transaction database, customer information database, VIP database and payment database, questionnaire of customer credit and so on.

- Calculation property: because integrated database still represented the transaction of each record, we must sum up the transaction volume of each

business to comprehensively evaluate the cardholder. In addition, when we analyzed the transaction time, we must convert the format of time, 'year-month-day', into the numerical format.

- Normalization treatment: the sum and balance in flowing transaction of credit card were very enormous, if the data were disposed, other data functions in database might be neglected, so normalization was put into use. The method adopted was that search the maximal sum and balance in the database, and then the selected number was divided by all other data to perform normalization.
- Averaging method: the money in flowing transaction of credit card was average, namely the money in the same card number was added and then divided by the whole record number.

The integrated data in the credit card database formed several data sources of DM, including typical customer information, reduced data by normalization and averaging treatment. By this time, data warehouse was established by DM, including information and consumptive records of cardholders. Based on the treated data, we expected to mine useful knowledge.

CHAPTER 4

MINING INFORMATION IN CREDIT CARD

MINING METHODS SELECTION

There are a lot of data mining techniques can be used in credit card information analysis. Since the information that bank needs were concentrated on tendency, relationships, classification, therefore, we had paid more attention and energy on the analysis of this information. Data mining contains some good algorithms, such as classification, cluster and association rule. The classification of the information was more important than the other methods. Banks wanted to know how many classes in their customers in buying certain merchandises, or the age difference among customers who are intended to buy the certain goods. And cluster is another important method, it does not give the number of how many classes in data, this feature can tell the distribution of every class, the information about consuming behaviors can be presented by this method. The other classical method is associate rule, it can point out the relationship between two or among two more attributes of credit card information, from that, the efficiency distribution of advertisement letter can be improved, in other words, the hit rate of advertisement can be improved. From those reasons, we bring these methods in our short list in credit card information analysis.

4.1.1 Cluster

The information of credit card could be classified into many kinds of data through the method of cluster, from which the model was further extracted.[2,3] There were many types of cluster in the flowing transaction database of credit card. For example:

- Cluster of transaction time: the consumptive time of a year could be divided into several stages by analyzing the database, and then the data were associated with questionnaire of customer credit and information to obtain the consumptive type of different kinds of people in a specific stage.
- Cluster of consumptive property: The flowing transaction property of credit card was clustered to analyze the consumptive characteristics of customers.

4.1.2 Classification

A further study on the data was performed on the basis of given knowledge. The study was composed of two aspects:

- Cluster of credit and information. Bank wanted to know which type specific customers belonged to, such as VIP, normal or overdraft, to mine deeply the real information about consumptive behaviors.
- Cluster of consumptive types. The proper consumptive advertisements were sent to specific customers on the cluster analysis.

4.1.3 Association rule

We could obtain some data that were suitable to association rule in the information of credit card, namely a large amount of data in transaction database of credit card applied to association rule, so we could establish the data relation between the flowing transaction database and questionnaire of credit information.[1,2,3] For example:

- Consumptive abilities:
  - Relation between education background and consumptive abilities
  - Relation between profession and consumptive abilities
  - Relation between age and consumptive abilities
  - Relation between income and consumptive abilities
- Consumptive places
  - Relation between education background and consumptive places
  - Relation between profession and consumptive places
  - Relation between age and consumptive places
  - Relation between income and consumptive places
- Overdraft
  - Relation between education background and overdraft
  - Relation between profession and overdraft
  - Relation between age and overdraft
  - Relation between income and overdraft

Some unknown knowledge might exist in the database, so we must process the given data to extract the knowledge.

4.2 CLUSTER ANALYSIS

4.2.1 Introduction of clustering algorithm

The term of cluster meant to divide the data into groups, which made the minimal differences between the groups, while the maximal differences in a group. Study focused on finding the proper method of efficient cluster for large databases in the field of DM. The active research was concentrated on the flexibility of cluster method. Cluster study was a demanding challenge. The requirements of cluster for DM were described as follows: flexibility, abilities to deal with different types of information and noise data, cluster for random analyses, insensitive to input record, explicable and usability. Following is its description in mathematic:

"For Data set $V\{v_i|i=1,2,...,n\}$, v_i is data object, and separate he. data set into k groups according to the similarity among the data objects, and satisfied:"(Translation)[21]

$$\{C_j|j=1,2,...,k\}$$
$$C_i \subseteq V$$
$$C_i \cap C_j = \Phi$$
$$\bigcup_{i=1}^{k} C_i = V$$

The procedure is called clustering, and $C_i(i=1,2,...,n)$ is called cluster.

The main methods of cluster were as follows:

- Partitioning method: you could set a database containing n objects, and construct k partition of data, so one partition stood for a cluster postulated k was less than or equal to n. In other words, the data were partitioned into k groups meeting the following requirements: (i) each group contained at least

one object; (ii) each object only belonged to one group. The main methods were: *k*-means, *k*-center algorithm, PAM (Partitioning Around Method), CLARA algorithm (Clustering Large Applications) and so on. [21]

- Hierarchical method: the given set of objects was hierarchically partitioned, including two methods of integration and division. The algorithms included CURE (Clustering Using Representatives) . [16,21]
- Density - based method: the partition methods were based on the distance of objects, including DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify the Clustering Structure) and DENCLUE (Density-based Clustering).[17]
- Grid-based method: the space of objects was quantified into finite units to form a grid structure. All operations of cluster, such as STING, CLIQUE and Wave Cluster, were performed in the grids, namely quantified space. [17]
- Model-based method: each cluster was given a hypothetical model to find data for the optimal simulation of model. [17]

4.2.2 Clustering of credit card information

4.2.2.1 Objectives

The data of credit card information were analyzed to cluster the groups, which the administrative department could master the overall situation by analyzing and managing results. The concrete researches were as follows: :

- A large amount of field information about transaction records was clustered to find the similar properties and models by this method.
- The one-year transaction time of credit card was clustered to obtain specific

time range of low and high season, and to determine the time assigning the advertisements of credit card, modifying the policies of credit card and updating new service and so on.

4.2.2.2 Adopted algorithm

In the thesis, *k*-means clustering was adopted. It uses *k* as parameters dividing *n* objects into *k* clusters, which made the minimal similarity in the cluster, while the maximal similarity in the cluster. The degree of similarity was calculated on the basis of mean value of objects (the median point of cluster) in the cluster.

The process of *k*-means algorithm is as follows. First, *k* objects were selected randomly, and every object represented the average value or median point of the cluster. The left objects were assigned to proximate cluster on the basis of their distance to the center of cluster. Then the average value of each cluster was recalculated. This recalculation process was repeatedly performed till the functions converged. The square error generally was adopted for the calculation, and its definition was as follows: [17]

$$E=\sum_{i=1}^{k}\sum_{p\in c_i}\left|p-m_i\right|^2 \qquad (4\text{-}1)$$

E referred to the summation of mean square error from the whole objects in the database; p was the point in the space, standing for the objects of the given data; m_i was the average value of C_i cluster (both P and m_i were multidimensional) . The law attempted to make the resultant cluster close and independent as possible. The algorithm attempted to find the suitable division of groups with the minimal square error. If the

resultant cluster was close, and the difference between clusters was great, the process was rather efficient. As for the database set containing tremendous data, the algorithm was relatively flexible and efficient for its formula of complex degree was O (*nkt*) , where n was the number of objects, k the number of clusters and t the number of iterative. Generally, k was great less than n and t great less than n. seen in figure 4-1.[17]

Algorithm : *k*-means Input : Number of cluster *k* and database of *n* objects Output : *k* clusters , minimize square error
Method : (1) Random select k objects as cluster center (2) Repeat (3),(4) (3) According to the average value of objects in cluster, re-assign the objects to the most likely clusters. (4) Update the average value of cluster , that is , compute the average value of objects in every cluster. (5) Until no change happen

Figure 4-1 Flowchart of *K*-means Algorithm

4.2.2.3 A preliminary result

The data of credit card were processed by classical algorithm of *K*-Means. The cluster analysis of transaction data including field data was introduced, while other analysis methods would be introduced in chapter five. The records here just a part of credit card database amounted to 392. The result was given below by setting categories were 4 and 8.

- The results were shown in table 4-1 when the category was 4.

Table 4-1 Clustering result on 4 category

Category No.	Quality of Records
1	4
2	342
3	34
4	12

- The results were shown in table 4-2 when the category was 8.

Table 4-2 Clustering result on 8 category

Category No.	Quality of Records	Distance to Center
1	5	126.5235
2	39	16.53861
3	29	43.89486
4	8	56.39035
5	4	349.1071
6	12	101.2053
7	79	12.28776
8	216	7.65386

The experiments always showed that there was a kind of category in the cluster with the maximal numbers of species and the proximate average distance to the center of cluster. After associating with the questionnaire of credit and information of credit card, the kind of cardholders might have the following properties: education background, moderate; profession, normal enterprises; office position, below moderate; month income, about 1000 YUANS; age, 35; warrantor, moderate; working period (5-15).

Among the cardholders of credit card, this kind of people accounted for a considerable 55-70 percent of the whole cardholders. The results were in accordance with the cluster analysis. The cluster was analyzed on the basis of transaction date consequently with the cluster number 2, 4, 12.

4.3 CLASSIFICATION

Classification could be applied into distilling the data for the descriptive categories. These basic methods adopted included induction of decision tree, Bayesian classification, neural networks and integration of data warehouse and system classification, and classification based on correlation. In addition, other methods were also adopted for classification, such as *k*-nearest classification, rough set and fuzzy logic techniques. [21]

4.3.1 The basic methods of data classification

Data classification included two steps. First, a model was established to describe scheduled data or concept set. Suppose each element belonged to a scheduled class determined by class label attribute) . If it was accepted to be accurate, the model could be used into classify the unknown data or elements.

4.3.2 BP (Back Propagation) neural networks

Artificial Neural Networks (ANN) were a kind of information system, associated a large amount of processed data set with each other on a specific manner. ANN had the abilities of parallel calculation, nonlinear treatment, self-organization and self-adaptability, and association. Though it had the drawbacks of complex structures, long training time, ANN possessed a unique advantage of low error rate. There were dozens of different ANN models, among which Hopfield network, BP network and ART (Adaptive Resonance Theory) were widely used. [17]

Following is the theory of BP network:

"Cells of feed-forward network can be in several levels. Suppose the network has

M levels. Input cells belong to level 1, output cells belong to level M. Here the average neuron of cell i is σ_i .

$$\sigma_{i\in 1} = S\left(\beta \sum_{j\in\{l-1\}} J_{ij}\sigma_j\right), \tag{4-2}$$

Here $\{l\}$ is the set of level m, S is the response function of cell

$$S = \frac{1}{1+\exp(-x)} \tag{4-3}$$

Let

$$I^{\mu,in} = \{\xi^{\mu}_{m\in l=1}\} \tag{4-4}$$

$$I^{\mu,out} = \{\xi^{\mu}_{i\in l=L}\} \tag{4-5}$$

here $\mu = 1,2,...,p$, it represent the input and output model of system, ξ^{μ} is the input and output, $\xi_{\mu}\square[-1,+1]$.

When input is $\xi^{\mu,in}$, the neuron of cell in level M-1 is

$$\sigma_{j\in L-1}(I^{\mu}) = S\left[\beta \sum_{k\in L-2} J_{jk} S\left[\beta \sum_{h\in L-3} J_{kh} S\left[\beta \cdots \left[\sum_{m\in L-1} J_{hm}\xi^{\mu}_{m}\right]\right]\right]\right] \tag{4-6}$$

Judgment function is:

$$H = \sum_{i\in L}\sum_{\mu}^{p} g(x_i^{\mu}) \tag{4-7}$$

$g(x^{\mu}_i)$ is defined as:"(Translation) [22]

$$g(x_i^{\mu}) = \left\{\xi^{\mu}_{i\in L} - \sigma_{i\in L}\left(I^{\mu}\right)\right\}^2 \tag{4-8}$$

In the analysis, level of network is M=3, $\beta = 0.5$ and $\varepsilon = 0.5$.

Among the learning strategies and algorithms, BP (Back Propagation) algorithm, used in multilayer front feedback network, was mostly widespread, seen in figure 4-2. [21]

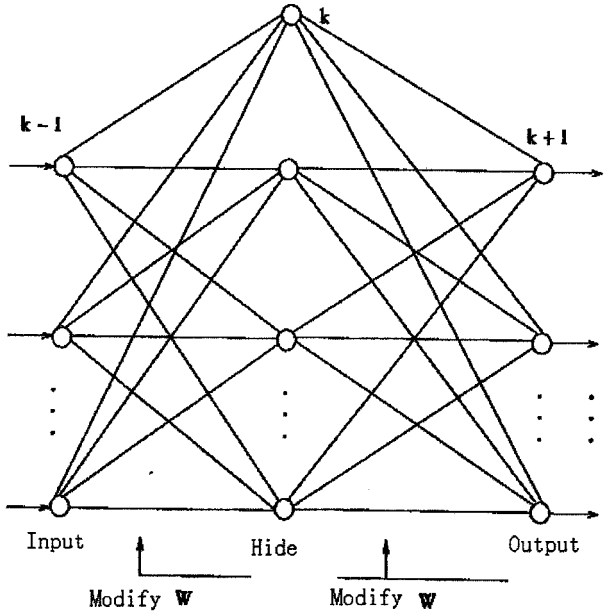


Figure 4-2. Multilayer front feedback network

The main ideas of BP algorithm were divided the study into two courses: forward and backward propagation processes. The process was described briefly as follows:[23,24,25,26]

- Forward propagation: the input samples passed through the hidden layers from the beginning of input layer till the end of output layer gradually. In the process of gradually propelling layer, every layer of nerve cells only had influence on the next nerve cells. The output layer compared the real value with expected value, if the output layer did not obtain the expected value, the backward propagation was adopted.

- Backward propagation: error signals were transferred on the contrary process of forward propagation. The error was obtained on the basis of real

value and expected value by gradually recursive calculation. Parameters of every hidden nerve cell were regulated on the calculation that made the error reached the minimum.

BP algorithm was actually a question calculating the minimal error of the functions. The algorithm was adopted the quickest descending method belonging to nonlinear programming, and modified the weighted coefficient by negative grads of error function. The process was given in figure 4-3. [21]

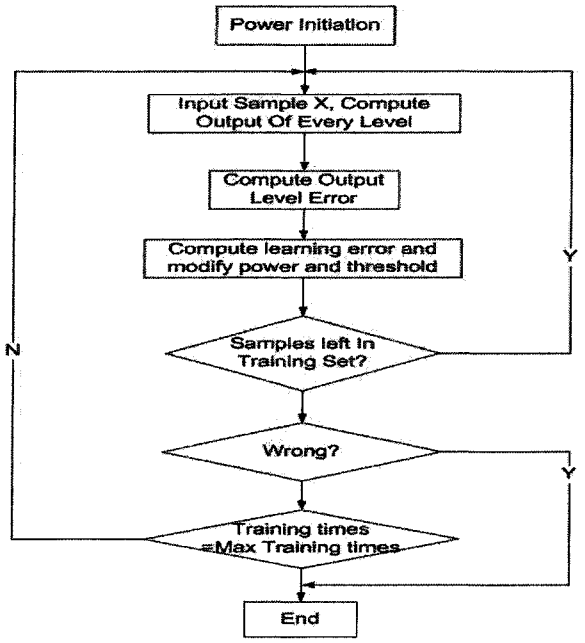


Figure 4-3. Schematic diagram of BP algorithm

4.3.3 Classification of credit card data

- Objectives

Through the classification of flowing transaction database and questionnaire of credit and information, we expected to obtain the classification of credit grades, namely which kind of people had good credit. This work was done by experiential or simple statistical methods before. We took advantage of DM method and the given data to classify the customers into three types, VIP, normal and dangerous (stop payment).

- A preliminary result of neural network classification:

We classified the credit card into two groups by making use of neural network method. First, transaction data in the credit card should be treated to sum up the transaction money in a specific number of the credit card, and then to be divided by the transaction number, so the average consumptive money of customers was acquired. The 240 pieces of data records were selected randomly, given the tested samples were 190, 140 and 90 respectively. The accurate rates were shown in figure 4-4. In the following calculation, the parameters about month income and the numbers of transaction were introduced to improve the accurate rate of calculation.

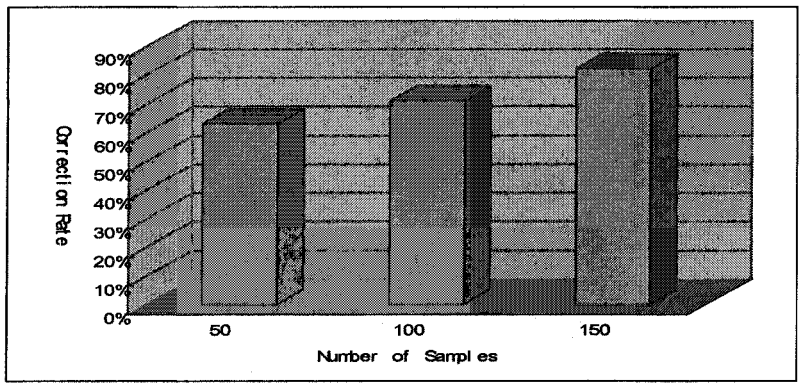


Figure 4-4. A preliminary result of accurate rate in classification calculation

4.4 MINING KNOWLEDGE BY ASSOCIATION RULE

"Association rule, one of the most important models in DM",[26] aimed at finding the relations among the different projects in database. These rules were intended to find the behavior modes of events and people, for example, the correlation between a sold item and another sold item. This rule could be applied into the design of commodity shelves, storage arrangement, and customer classifications based on purchasing pattern. "Data association also played an important role in many cases, such as the analyses of stock, bank deposit."(Translation)[27] The technique adopted was Apriori algorithm.

The association rule of mining the transaction database of customers was firstly presented by Agrawal in 1993[28]. Subsequently, a good deal of study was performed on the problems by many scientists. Their researches included the optimization of present algorithm to improve the efficiency of DM algorithm, for instance, introduction of the ideas about random samples and parallel calculation and so on; applications of the rule

were extended. In addition, hundreds of, or even more association rule were obtained easily in the database. DM brought so many rules that an idea of knowledge management was subsequently proposed. [29,30,31,32]

4.4.1 Basic concepts of association rule

Given $I=\{i_1, i_2,\ldots, i_m\}$ was a binary system set of characters, the elements in the set were called items. Given D was a set of transaction, where I was the set of items, and $T\subseteq I$. Each correspondence T had sole label, such as the number of transaction, marked as TID. Given X was a set of items in I, if $X\subseteq T$, we could say that T contained X.[17]

An association rule was a function as $X\Rightarrow Y$, where $X\subset I$, $Y\subset I$, and $X\cap Y=\Phi$. The support of association rule ($X\Rightarrow Y$) in T database was the ratio of the transaction number in X and Y sets to total transaction number, marked as support($X\Rightarrow Y$), namely

$$\text{support}(X\Rightarrow Y)=|\{T: X\cup Y\subseteq T\ ,\ T\in D\}|/|D|$$

The confidence of association rule ($X\Rightarrow Y$) in T database was the ratio of the transaction number in X and Y sets to the transaction number of X, marked as confidence($X\Rightarrow Y$), namely

$$\text{confidence}(X\Rightarrow Y)=|\{T: X\cup Y\subseteq T\ ,\ T\in D\}|/|\{T: X\subseteq T\ ,\ T\in D\}|$$

Given a transaction set as D, the purport of DM association rule was to bring forward an association rule with support and confidence greater than minimal support and confidence of the rule stipulated by customers.[17]

From the association rule, we could draw the conclusion that gender = 'female' =>profession = 'secretary', gender = 'female'=>average (income) =2300.

4.4.2 Typical Apriori algorithm of association rule

A basic algorithm, based on the idea of two stages of frequent item set, was first designed by Agrawal in 1993, so the association rule could be divided into two affiliated problems.[28]

- All the sets with the support greater than minimal support of item set, called frequent item set, were found.
- The expected association rule were produced by using the item sets found in the first stage.

The second stage was relatively simple. Given a frequent item set, $Y=I_1I_2...I_k$, $k\geq 2$, $I_j \square I$, the association rule of all the set in the item set, $\{I_1, I_2, \ldots, I_k\}$, were produced with the largest number of k. Each rule had only one item set in the right part, for instance, $[Y-I_i] \Rightarrow I_i$, $\forall 1 \leq i \leq k$), once the association rule were formed, only the rule with the confidence larger than the given minimal confidence could be saved. [17]

In order to form all the frequent item sets, the recursive algorithm was adopted. Its key ideas were shown in figure 4-5. [17]

```
(1)    L1 = {large 1-itemsets};
(2)    for (k=2; Lk-1≠Φ; k++) do begin
(3)      Ck=apriori-gen(Lk-1);
(4)      for all transactions t∈D do begin
(5)              Ct=subset(Ck,t);
(6)          for all candidates c∈ Ct  do
(7)          c.count++;
(8)      end
(9)       Lk={c∈ Ck |c.count≥minsup}
(10)   end
(11)   Answer=∪kLk;
```

Figure 4-5. Key ideas of Apriori algorithm

First, one frequent item set containing one item, labeled L_1, was produced, then frequent item set containing two items, labeled L_2, till the an empty set, L_r, appeared, the algorithm stopped. In the k circulation, the candidate item set, C_k, was first produced, in which two L_{k-1} item sets with only one item difference associated with each other through the connection step of k-2. The item sets in C_k were seen as candidate item sets for establishing frequent item sets. The last frequent item set, L_k, must be the subset of C_k. Every item in the C_k was confirmed whether it joined the L_k item set in the transaction. The confirmation process was a bottleneck of the algorithm performance. The algorithm might require several calculations of transaction database with a huge amount of data, for example, if the frequent item et included 10 items, 10 scanning processes were needed, which demanded a great load of I/O. [17]

4.4.3 DM in association rule of credit card information

- Objectives

The knowledge in credit card information needed further mining. Our mining work

focused on the association dataset between flowing transaction data of credit card and customer credit and information to find the relation between customer credit and consumption through association rule, which was the foundation to regulate the grade of customer credit.[2,3]

- Method

Typical Apriori algorithm was adopted to find the frequent item sets of consumption properties and establish association rule from customer credit and information and flowing transaction data of credit card. Our affair set included the following items: (card number, transaction date, transaction place, expense, education background, profession, position and title, month income, age, working year). The frequent item set was found by Apriori algorithm. Because our work was the foundation for the further research, all possible information should be collected by all means. Therefore, we expected to find the strong association rule, meanwhile we should pay much attention to the weak rule, which might also contain important knowledge.

- Basic results

The association rule were mined through Apriori algorithm, aiming to assorting and setting the 70000 records in questionnaire of credit information, given the minimal support of 21.1% and minimal confidence of 50%. The preliminary results were shown in table 4-3

Table 4-3 The preliminary results of Credit cart based on association rule

No.	Support	Confidence	Association Rule
1	.257	.853	Age (25-35) →Title (Junior)
2	.925	.734	Title (Junior) →Profession(Foreign Company)
3	.332	.866	Education (Junior college) →Profession (Service)
4	.425	.964	Salary (600-1000) →Profession (Service)
5	.425	.875	Salary (600-1000) →Warrantor (Income over 800YUANS)
6	.701	.922	Profession(Foreign Company) →Warrantor (Income over 800YUANS)

The first line represented the mined association rule, the second the confidence, the third line association rule, for example, the rule in the first row was shown in the table 4-4.

Table 4-4 A association Rule

1	.257	.853	age (25-35) →Title (normal worker)

The explanation was as follows: the title of cardholders with age from 25 to 35 was normal worker. The rule had the support of 25.7%, and confidence of 85.3%. More mined association rule would be further analyzed in the chapter five.

CHAPTER 5

RESULTS OF CREDIT CARD INFORMATION

5.1 EXPERIMENT DESIGN OF INFORMATION ANALYSIS IN CREDIT CARD

5.1.1 Objectives

- To verify the efficiency of DM in analyzing credit card data[1,2]

  The efficiency of DM in analyzing credit card information was verified through mining the data in it. We tested the following algorithms, *k*-means in cluster, front feedback neural network in classification, and Apriori algorithm in association rule.

- The foundation was established for further study. [3,19]

  At the beginning of mining data in credit card information, some contents could be certain, while others uncertain. But the uncertain factors might contain a large amount of our interested contents for determining the further mining research and establishing the entire system. Our work was experimental and exploratory, so we could not expect to obtain complete and accurate results, but just a preliminary research.

5.1.2 Experiment design

- Database of customer categories: the number, information and consumptive properties of sample customers, such as VIP, normal and suspending customers,

were contained in the database, which could be used in the verification of algorithm. The database objects mainly associated with VIP and normal customers discriminated by BP networks.

- Flowing transaction data of credit card: the analyses of consumptive places, data and money were applied to verify the cluster algorithm by the method of *K*-means in deal with consumptive information of cardholders.
- Questionnaire database of credit information was applied to find the association rule between properties. The mined rule could be applied to verify the Apriori algorithm and to analyze the inner relations of customer properties by typical algorithm.

5.2 RESULTS

5.2.1 Cluster

The flowing transaction data in only one transaction date were clustered with the number of 2, 4, 12 respectively, and 3184 pieces of records. The results were given below:

- **Cluster into two categories**

Table 5-1. The results at the cluster number of 2

Category	Period	Days	Transaction number	Transaction Number per day	Center point	Distance to center
1	Jan.1—July 1	181	1465	8.094	May 8	4.638626E-2
2	July 2—Dec.27	180	1718	9.544	Oct. 2	4.679509E-2
Sum	361		3183	8.817		

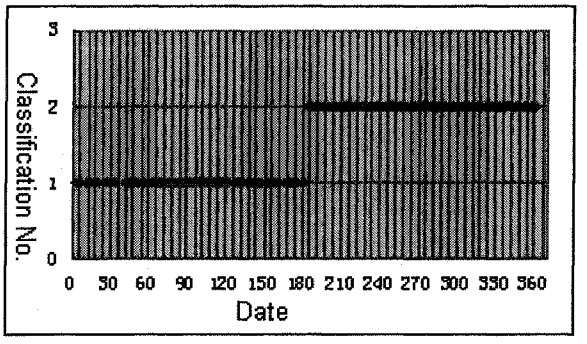


Figure 5-1. The results at the cluster number of 2

From the figure 5-1, we could explicitly find the two parts. In the figure 5-2, we also gave the further explanation. The abscissa only had the symbols of classification besides date. The date of first transaction center was August 2, and the second was October 2. The consumptive centers, concentrating on the shopping climax day about May Day and National Day, were just in accordance with real situation, which proved the cluster to be true.

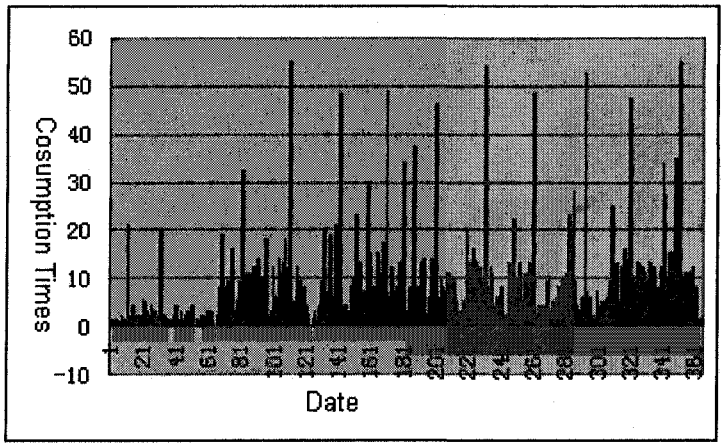


Figure 5-2 Data classification graph at the cluster number at 2

➢ **Cluster number at 4**

The results were shown in the figure 5-3, 5-4 and table 5-2.

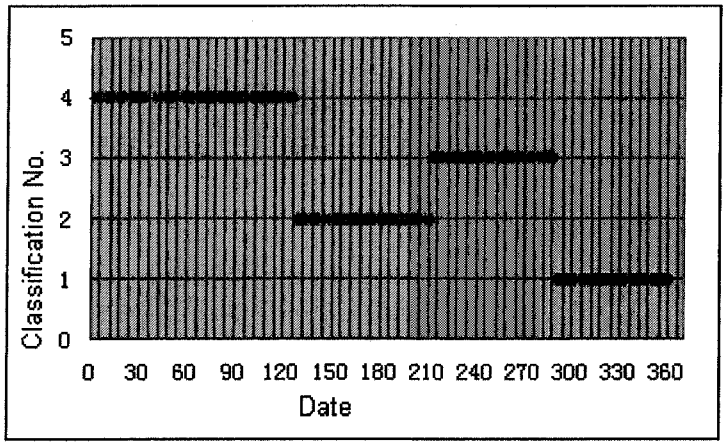


Figure 5-3. The results of cluster number at 4

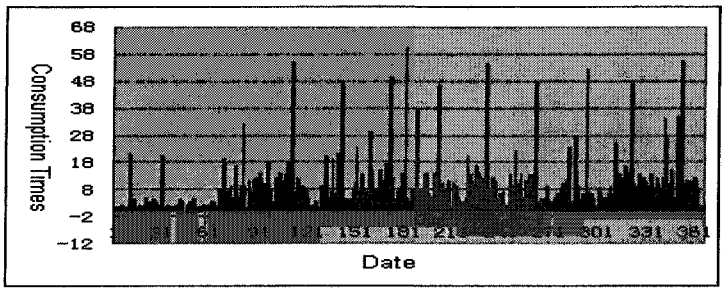


Figure 5-4. Data classification graph at the cluster number at 4

Table 5-2. The results of cluster number at 4

Category	Period	Days	Transaction number	Transaction Number per day	Center point	Distance to center
4	Jan.1—May 6	126	632	5.016	Mar. 23	6.687437E-2
2	May 8—July.30	84	1118	13.309	Jun. 22	4.679509E-2
3	July 31—Oct.16	78	647	8.295	Sep. 7	5.469961E-2
1	Oct.17—Dec.27	72	787	10.931	Nov. 30	4.638626E-2
Sum			3183	8.817		

From two figures and one table, we could see that winter, spring, the beginning of summer was the lowest shopping periods, when the cluster number was at 4. The frequent intervals were late in the March with the center time of March 23, just the rhythm seasons from cold winter to warm spring. But just after the Spring Festival, purchasing power of people was still in low level. The frequent intervals of summer were concentrated on the June with the center time of June 22, also the rhythm season. Because the clothes in winter and spring could not be dressed, they must be changed. The most of commodity could not be serviced, so the impending summer made a prosperous shopping season. The periods from end of August to beginning of September were a normal transaction interval with the center time of September 7. A shopping climax came into being at the end of November, just for the rhythm season and impending festival, such as Christmas Day and so on. The analyses of four-season shopping data, we could find the correspondence situation in the reality, which showed that the cluster method was right.

➢ **Cluster number at 12**

The results were shown in the figure 5-5, 5-6 and table 5-3.

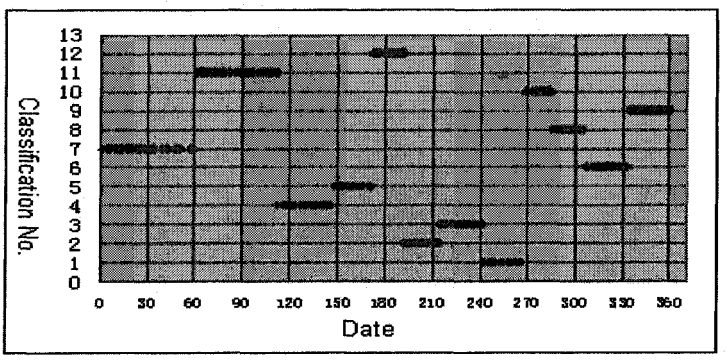


Figure 5-5. The results of cluster number at 12

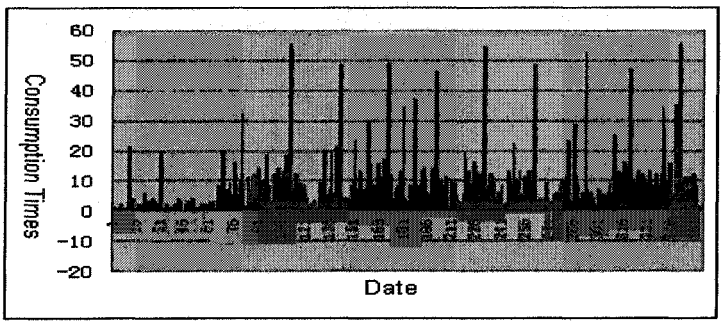


Figure 5-6. Data classification graph at the cluster number at 12

Table 5-3. The results of cluster number at 12

Category	Period	Days	Transaction number	Transaction Number per day	Center point	Distance to center
7	Jan.1—Feb.28	59	115	1.949	Jan.27	3.2616 E-2
11	Mar.2—Apr.23	53	452	8.528	Apr.2	3.6301 E-2
4	Apr.24—May 27	34	246	7.235	May.12	2.2023 E-2
5	May 28—Jun.20	24	256	10.666	Jun.10	1.7945 E-2
12	Jun.21—Jul.10	20	478	23.9	Jul.1	5.3370 E-3
2	Jul.11—Aug.2	22	218	9.909	Jul.21	1.2717E-2
3	Aug.3—Aug.29	27	259	9.592	Aug.17	1.4635 E-2
1	Aug.30—Sep.23	25	201	8.04	Sep.13	1.490512E-2
10	Sep.24—Oct.10	17	108	6.353	Oct.6	1.1137 E-2
8	Oct.11—Oct.31	21	166	7.905	Oct.20	1.2279 E-2
6	Nov.1—Nov.28	28	313	11.178	Nov.17	1.6592 E-2
9	Nov.29—Dec.27	28	371	13.25	Dec.14	1.6488 E-2
Sum	361		3183	9.875		

Through the abovementioned cluster, we could see that the shopping in January and February were the lowest months in a year, when the cluster number was at 12. The main reasons might be due to the decreasing consumption of credit card for many companies distributed festival goods or consumptive cards to their employees, and for many customers bought goods most by cashes without using the credit card. The highest shopping appeared in June and July. The main reasons might be due to bank settling interests on July 1. Our cluster results also showed that July 1 was just the period of shopping center. The lowest shopping appeared in September and October for the long intervals of National Day. The center period was October 6, demonstrating that most of the people went traveling during golden week, so the consumption of credit card was much less during this period. In addition, the consumption with cashes in other cities grew hugely, when the customers return home, the purchasing power dropped. The higher shopping period might be mainly due to the festival consumption, such as Christmas Day and New Year' Day. The data through cluster analysis were in right accordance with real situation.

5.2.2 Classification

We selected data of consumptive money, balance, consumptive number respectively from flowing transaction database, and selected data about customer month income from questionnaire of credit information, to establish a new data set, where we combined the data of money and balance, and summed up, and then divided by consumptive number to obtain average consumptive money and balance and additional consumptive number, so three items in a dataset were established. Then, the month income of customers was

found by the credit card number combined with the dataset to form new dataset for classification.

The classification was achieved by the method of BP neural networks, which were divided into three layers. The four properties were classified and recorded from the input layer, namely the first layer, and then calculated in the middle layer, the second layer, and obtained the results at output layer. The results were shown in figure 5-7.

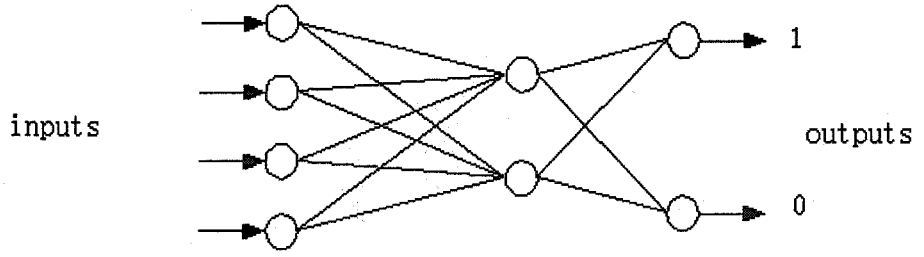
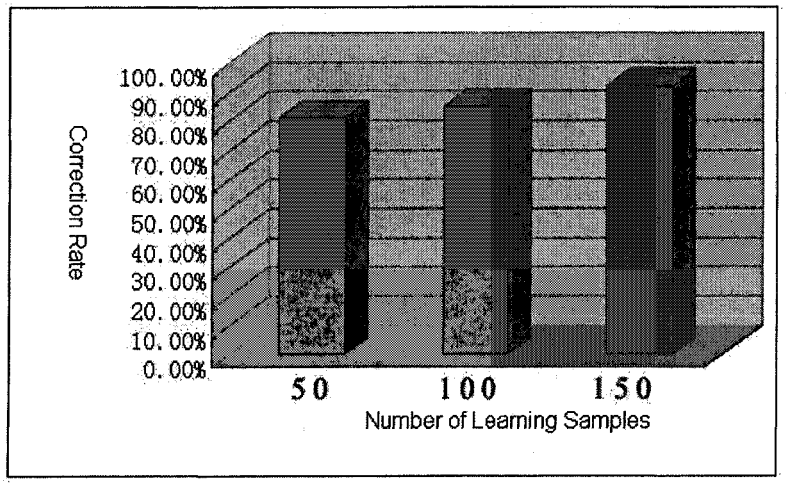


Figure 5-7 Classification process of neural networks

The record number of classification was 238, sample numbers were 188, 138, and 88 respectively. We introduced the two items of month income and transaction number on the basis of abovementioned classification. The transaction number introduced was helpful in discriminating the frequent consumptive customers from infrequent consumptive customers, which might improve the discriminative abilities from average transaction money. This classification method improved the accuracy of results, reaching 92.04%. The results in figure 5-8 showed that the accuracy in this method improved greatly compared with former classification.



.Figure 5-8. The accuracy after introduction of month income item

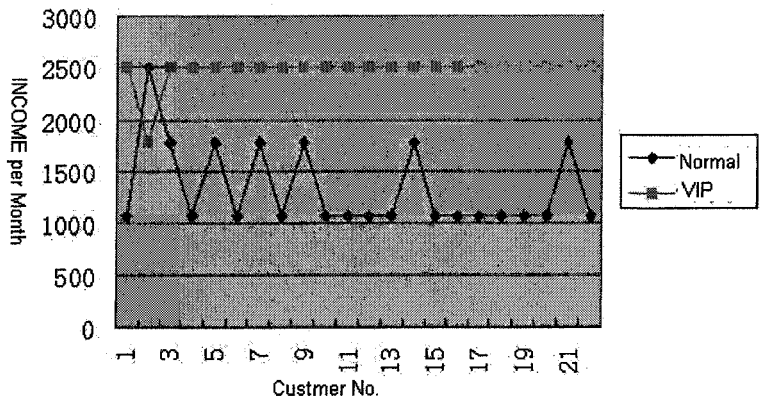


Figure 5-9. The comparison between VIP and normal customers

We obtained the relative scientific measure to discriminate VIP from customers through classification. The former measure just took into account the whole consumptive money and statistical ranking of a specific customer, which made the method one-sided, and could not represent the whole situation. All properties of a few databases were combined together to analyze the overall results, so we could look at all sides of the general situation and winds of change of VIP customers. Meanwhile, the sales strategies for credit card were changed accordingly to improve the efficiency of credit card, which made more normal customers to convert to VIP. For example, the month income of VIP

was much more than that of normal customers. The visual results were shown in figure 5-9.

5.2.3 Association rule

We selected randomly 800 customers to form a property database for establishing the association rule by DM technique. The DM algorithm directed mainly at the resultant table set composed of properties of 800 customer credit information. These properties included age, profession, surety, position, and month income. The records in the table set were about 7000. The results were shown in table 5-4 after the further mining of Apriori algorithm.

Table 5-4. Analysis results of credit card information based on association rule

No.	Support	Confidence	Association Rule
1	.251	.842	Age (25-35) → Warrantor (Income over 800YUANS)
2	.282	.748	Age (36-45) → Profession (Foreign Company)
3	.925	.914	Profession (Service) → Warrantor (Income over 800YUANS)
4	.332	.803	Education (Junior college) → Profession (Foreign Company)
5	.332	.945	Education (Junior college) → Warrantor (Income over 800YUANS)
6	.425	.648	Salary (600-1000) → Profession (Foreign Company)
7	.701	.968	Profession (Foreign Company) → Profession (Service)
8	.908	.931	Warrantor (Income over 800YUANS) → Profession (Service)
9	.908	.712	Warrantor (Income over 800YUANS) → Profession (Foreign Company)
10	.282	.748	Age (36-45) → Profession (Service), Profession (Foreign Company), Warrantor (Income over 800YUANS)
11	.332	.735	Education (Junior college) → Profession (Service), Profession (Foreign Company), Warrantor (Income over

			800YUANS)
12	.425	.605	Salary (600-1000) → Profession (Service), Profession (Foreign Company), Warrantor (Income over 800YUANS)
13	.257	.821	Profession (Service), Salary (600-1000),Profession (Foreign Company) , Warrantor (Income over 800YUANS) → Age (36-45)
14	.282	.748	Age (36-45) , Profession (Service), Warrantor (Income over 800YUANS) → Profession (Foreign Company)
15	.284	.744	Profession (Service), Salary (600-1000), Warrantor (Income over 800YUANS) → Age (36-45), Profession (Foreign Company)

These typical rule selected were divided into three types, which could express the main ideas.

(1) Association rule as verifying function

Some association rule in the abovementioned table set played an verifying function in analyzing the credit card information, such as the rule of 1, 3, 5, 8.

Table 5-5. Association rule for verifying function

No.	Support	Confidence	Association Rule
1	.251	.842	Age (25-35) → Warrantor (Income over 800YUANS)
3	.925	.914	Profession (Service) → Warrantor (Income over 800YUANS)
5	.332	.945	Education (Junior college) → Warrantor (Income over 800YUANS)
8	.908	.931	Warrantor (Income over 800YUANS) → Profession (Service)

The abovementioned rule involved the four properties of age, education, position, surety, which represented the information of cardholders to demonstrate the mutual relations among the properties. Especially, the rule 3 and 8 could support each other.

(2) Association rule as illuminating function

One type of the mined rule was illuminating. The contents of rule were not obtained accurately from our experience, such as the rule of 10, 11, 12 and 13, seen in table 5-6. The rule represented novel or unimaginable information of cardholders. For example, the cardholder information of age, education and month income could deduce the profession, position, surety of cardholders. Meanwhile, the combination of abovementioned information could deduce the age of cardholders, which could be verified each other. After consulting with experts in credit card information, these rules were proven to be true.

Table 5-6. Association rule as illuminating function

No.	Support	Confidence	Association Rule
10	.282	.748	Age (36-45) → Profession (Service), Profession (Foreign Company), Warrantor (Income over 800YUANS)
11	.332	.735	Education (Junior college) → Profession (Service), Profession (Foreign Company), Warrantor (Income over 800YUANS)
12	.425	.605	Salary (600-1000) → Profession (Service), Profession (Foreign Company), Warrantor (Income over 800YUANS)
13	.257	.821	Profession (Service), Salary (600-1000),Profession (Foreign Company) , Warrantor (Income over 800YUANS) → Age (36-45)

(3) Weak association rule

The rule given below had weak association after being tested by concrete cases.

Table 5-7. Weak association rule

No.	Support	Confidence	Association Rule
14	.282	.748	Age (36-45) , Profession (Service), Warrantor (Income over 800YUANS) → Profession (Foreign Company)
15	.284	.744	Profession (Service), Salary (600-1000), Warrantor (Income over 800YUANS) → Age (36-45), Profession (Foreign Company)

The objectives of mining the questionnaire of credit information were aimed at revising the content and property weighing in questionnaire of credit information. We analyzed an easy problem on the basis of mined association rule. The close relation existed among the properties of credit information. One hand, it was easy to make use of the properties, but on the other hand, the close relation must cause the similarity in two properties, so in the investigation of credit information, we should pay more attention to the setting of every property (target item), which would play a vital role in the investigation of credit information. If the properties were so close that they could support each other, just as rule of 1 and 8. The probability of cardholders was 84.2%, with the following properties, aged from 25 to 35, month income about 800 YUANS; while the probability of their ages between 25 and 35 was 93.1%, when the properties of cardholders were as follows: sureties were normal workers and month income below 800 YUANS The properties and ages of sureties had their own weighing, which made the situation similar, so the evaluation results about credit and information were affected. This was the reason why the questionnaire was revised. Our work aimed at establishing a solid foundation for further research in the near future.

CHAPTER 6

CONCLUSIONS

6.1 OUR CONTRIBUTIONS

Data mining, just as a further extension of database, showed flourishing vitality, which could play an important role in dealing with a great amount of data and information. Preliminary and exploratory results were obtained by aiming at analyzing the data in credit card, which might paved the way for fulfilling the information management system in bank business. Our main contributions could be summarized as follows:

1. Pretreatment of data in credit card

The given databases were reorganized to perform a set of database program, which could be especially applied into searching and combining of the data in flowing transaction database of credit card. Such pretreatment technologies as data filter, selection, conversion and reduction were adopted by distilling the five databases to form a new dataset for mining data. The key technique of the data mining belonged to the domain of data warehouse.

2. Analyses of DM in database of credit card

Some potential knowledge and information were analyzed and mined to present correspondence solutions on the basis of DM in database of credit card, including consumptive modes, such as time, places and comprehensive properties, and prediction classification of credit card, and reevaluation of credit information on the basis of consumptive records.

3. Preliminary analyses of credit card data to obtain valuable results

By utilization of the data warehouse concept, many databases involved were reorganized and converted to form a comprehensive database including credit information, consumptive information of customers. Cluster was performed by the method of *K*-means algorithm to find some valuable knowledge. Meanwhile, the transaction time was analyzed by cluster method to obtain the frequent periods. BP neural networks were designed to classify the data of credit card, which played an important role in the analysis of credit evaluation and prediction of cardholders. In addition, the abovementioned combined data were processed by Apriori algorithm to mainly produce the association between credit information and consumption, meanwhile, the association of credit information itself was also achieved, which made great significance in the further work.

6.2 PRESENT PROBLEMS AND FURTHER RESEARCH

6.2.1 Problems

Some preliminary results of information analyses in credit card were obtained on the basis of DM technique, but for a variety of reasons, the entire analysis of card information was not fulfilled. The existent problems were as follows:

1. The efficiency of pretreatment data should be improved

Not all the data over the years could be applied into data mining. Because of storage patterns, some data or some contents were not suitable for data mining. Meanwhile, some particular data were not stored to make great harm to our mining work. The data available must be pretreated in advance, so we divided the data into several groups based on the premise that data were not damaged. The present pretreatment of data depended on the text format, so we had to divide the data, transaction business in one year, into 100 databases or above, which made great difficulties in searching and partitioning, and the process was not automotive.

2. Efficiency of DM should be further improved

Classical algorithms were adopted to analyze the data, but the database in bank business had its own characteristic, so the algorithms should be revised in accordance with the variation of data, which could improve the efficiency and make the practical applications become true.

3. Our exploratory work was just the beginning of the project, credit card information system based on data mining technique. The further research aimed at establishing the DM analysis platform of credit card information, and expecting the combination between the platform and databases of credit card to achieve dynamic analysis, which made the work more important in social life.

6.2.2 Future works

In future study of bank credit card will be focused on the efficiency delivery of credit card accounting letter with advertisements. That will include following contents:

1. Classification by customers consuming behaviors

   In the credit card information, it normally contains consuming time, consuming place, consuming amount, etc... Lots of classification can be done, such as,

   (1) Consuming date: predicate consuming behavior in half year, season, or month.

   (2) Consuming amount: predicate customers' spending potential amout.

   (3) Consuming place: where they usually spend money.

2. Relationship between customer group and advertisement type

   Different customers can get there most interesting advertisement along with the accounting letter.

   (1) Consuming date and merchandise promotion

   (2) Customer group and consuming place

(3) Personal information of customer and their consuming, such as, age, gender, incoming, etc…

REFERENCE

[1] Personal Finance Dept., "Proceedings of Personal Consumption Records and Analysis", ICBC, Beijing, China, 2004.8

[2] Credit Card Dept. "Credit Card Analysis in 2002", ICBC, Beijing, China, 2003.1

[3] Credit Card Dept. Tianjin, "Credit card transaction records—Tianjin in 2002", ICBC Tianjin, Tianjin, China, 2003.4

[4] Su Deng, Hongbin, Huang, *et al.*, "Design And Realization of Knowledge Discovery System Based On Database", Computer Engineering and Applications, Vol.36, No.6, 2000.6, pp119-121

[5] Weimin Ouyang, "Discovery of Association Rules with Temporal Constraint in Databases", Journal of Software , Vol.10, No.5,1999.10, pp 527-532

[6] Yubao liu, Zhiqing Meng, "Ming of Association Rules with Spatial Constraint in Sales Database", Computer Engineering and Applications, Vol.36, No.9, 2000.6, pp110-111

[7] Intl' Conf., "Proceedings of 11th international joint conference on AI", Detroit, MI, Vol.1 and 2, 1989.8

[8] IEEE Computer Society, "Special Issue on Learning and Discovery in Knowledge-Based Database", IEEE Trans. On Knowledge and Data Engineering, Vol.5,No.6, 1993.12

[9] Fayyad U.M, Weir N., Djorgovski S.,"SKICAT: A Machine Learning System for Automated Cataloging of Large Scale Sky Surveys", Proceedings of Tenth International Conference on Machine Learning (ICML), 1993, pp112–119

[10] Michael J.A. Berry, Gordon S. Linoff, "Data Mining Techniques", China Machine Press, Beijing, China, 2006.7

[11] SooCheon Kweon, Sawng, Y., *et al.*, "An Integrated Approach Using Data Mining & Genetic Algorithm in Customer Credit Risk Prediction of Installment Purchase Financing", 2006 International Symposium on Collaborative Technologies and Systems, 2006.6, pp125-131

[12] Yi Du, Mao Tian, Yuhao Wang, "An approach of intelligent optimization system for mobile communication networks", Proc. of 2005 international Conference on Wireless Communications, Networking and Mobile Computing, Vol.2, 2005.9 pp1125-1128

[13] Borzemski L., "Data mining in the analysis of Internet performance as perceived by end-users", Proceedings of 18th International Conference on Systems Engineering, Vol.1, 2005.8, pp34-39

[14] U. Fayyad, G. P. Shapiro, P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining , Menlo park, CA, AAAI Press, 1996.pp82-88

[15] Ke Luo, Jie Wu, "Apply Constraint and Multidimensional Technique to Data Mining", Computer Engineering and Applications, Vol.36, No.4, 2000,pp95-97

[16] Jiawei Han, Micheline Kamber, "Data Mining_Concepts and Techniques", China Machine Press, BeiJing, China, 2001.8

[17] Zhongzhi Shi, "Knowledge discovery", TsingHua University Publishing house, BeiJing, China, 2002.1

[18] Computer Dept. "System design of Banking Software", ICBC Tiangjin, Tianjin ,

China, 2000.1

[19] Computer Dept. "Banking Deposit Software Design", ICBC Tiangjin, Tianjin , China, 1999.2

[20] Qinbao Song, Junyi Shen, "The Research of Data Preparation for Data Mining with Neural Networks", Computer Engineering and Applications,Vol.26, 2000.12, pp102-104

[21] Yuntao Zhang, Ling Gong, "Data mining theory is technique", Publishing house of Electronic Industry, BeiJing, China, 2004.4

[22] Tom M. Mitchell, "Machine Learning", China Machine Press, BeiJing, China, 2004.9

[23] Cherkassky V., Vassilas N., "Performance of back propagation networks for associative database retrieval", Proc. Of International Joint Conference on Neural Networks, Vol.1,1989.6, pp.77-84

[24] Kwo*k*-Wah Hung, Wing-Chung Chan, "Stroke encoded Chinese handwriting input system based on bac*k*-propagation networks", Proc. Of IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering, Vol.2, 1993, pp. 1106 - 1109

[25] Cherkassky V., Shepherd, R. "Regularization effect of weight initialization in back propagation networks", Proceedings of the 1998 IEEE International Joint Conference on Neural Networks & 1998 IEEE World Congress on Computational Intelligence, Vol.3, 1998.5, pp2258 - 2261 vol.3

[26] Zhaohui Zhang, Yuchang Lu, "An Algorithm for Mining Quantitative Association Rules", Journal of Software, Vol.9, No.11, 1998, pp801-805

[27] Senmiao Yuan, Xiaoqing Chen, "Clustering Method for Mining Quantitative Association Rules ", Chinese Journal of Computers, Vol.23, No.8, 2000, pp866-871

[28] R. Agrawal,T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD Conference on Management of data, 1993, pp.207-216,

[29] Yiwen Liang, Xia Cao, "The Heuristic Ways of Finding the Association Rules", Computer Engineering and Applications, Vol.36, No.12, 2000, pp116-117,178

[30] Rastogi, R., Kyuseok Shim, "Mining optimized association rules with categorical and numeric attributes", IEEE Transactions on Knowledge and Data Engineering, Vol., No.1, 2002.1, pp. 29-50

[31] Tung, A.K.H., Hongjun Lu, Jiawei Han, Ling Feng, "Efficient mining of intertransaction association rules", IEEE Transactions on Knowledge and Data Engineering, Vol., No.1, 2003.1, pp. 43-56

[32] Sung, S.Y., Zhao Li, Tan, C.L., Ng, P.A., "Forecasting association rules using existing data sets", IEEE Transactions on Knowledge and Data Engineering,