

Customer Segmentation of Bank Based on Data Warehouse and Data Mining

Shuxia Ren¹, Qiming Sun¹, Yuguang Shi²

¹School of Economics and Management
1 Beijing University of Posts and Telecommunications
Beijing, P.R.China

²Agricultural Bank of China Jilin Branch
Jilin, P.R.China
Email: renshuxia2006@sina.com

Shuxia Ren

School of Economics and Management
Beihua University
Jilin, P.R.China
Email: renshuxia2006@sina.com

Abstract—The problems can not be solved by the traditional method of customer segmentation when banks face with the massive information. Data mining techniques can extract useful information and knowledge that are implicit and unknown in advance but is potential in the practical application. Data mining techniques are broadly used in customer relationship management but there is no unified framework model for customer segmentation by now. Customer segmentation model of bank is built based on data mining which is to define the corresponding mapping relationships between customer attribute and concept attribute in this paper. We apply self-organizing mapping neural network and K-means algorithm to bank customer segmentation and analyze the sample data which we selected from some bank. The outcomes show that we can use a dynamic model of the data mining to describe customer behavior and provide useful information for the managers of banks to decision making.

Keywords—customer segmentation of bank; data warehouse; data mining; self-organizing mapping neural network and K-means

I. INTRODUCTION

In the process of economic development, business management concept evolved from product-oriented idea to market-oriented idea, and then to customer-oriented idea. Whether enterprises can obtain, maintain and develop their own clients or not has become the most critical factor because customers are important strategic resources. Customer Relationship Management (CRM) is based on the respect to customers and it can help enterprises understand the entire life cycle of the customers and enable enterprises to provide the customers with more personalized and more efficient services, and then it can enhance customer satisfaction and loyalty, and improve the competitiveness of enterprises ultimately. Customer segmentation, as the core of CRM, is to classify the customers according to the customer's attributes, behavior, needs, preferences, value and other factors in a clear business strategy and specific market, and it can provide appropriate products, services and marketing models to the customers.

Generally the traditional customer segmentation is using the experience or a simple statistical methods to classify and it is subjective. Therefore it is impossible to understand which customers have higher value customers and which customers have higher credit. In the banking management the traditional method of customer segmentation can not solve the problems when banks face massive amounts of

customer data. So it is essential to establish Bank Data Warehouse. Bank Customer Relationship Management System depends on the Bank Data Warehouse. In order to integrate and analyze information from customers it needs to establish the data mining system to realize automatic collection, automatic transmission, integrated query and analysis. Data mining is the process to extract useful information and knowledge that are implicit and unknown in advance but is potential from a lot of, uncomplete, noisy, vague, and random data of the practical application. It is also called as Knowledge Discovery in Database (KDD) [1]. Data mining of commercial banks can offer many commercial value for banks.

Existing data resources can be used completely in bank customer segmentation based on data warehouse and data mining. Effective data mining and analysis in the bank customer segmentation can help banks find the present and potential customers who can create benefits for the banks accurately, and help banks develop new products to meet the consumer demand and find an effective marketing tool to retain customers. Reference [2] proposed the combined response model based on RFM analysis which is based on BP neural network to solve customer segmentation under the massive data problems. Natter proposed the clustering of the customer segmentation based on the K-mean. Reference [3] used self-organizing and mapping neural network methods for the clustering analysis in the study of on-line business customer segmentation. Reference [4] proposed the method of customer segmentation based on Hopfield-Kagmar neural network. Many scholars also have made use of a variety of data mining technology to research customer segmentation. But studies on customer segmentation are more dispersed and lack of a unified research framework.

II. DATA WAREHOUSE AND BANK CUSTOMER SEGMENTATION

Data warehouse collected data from a unit of historical data to a central warehouse in order to facilitate data processing, and it is subject-oriented, integrated, stable, and lasting data collection over time. Data Warehouse is based on a multi-dimensional model structure. The data warehouse is more like a process of data integration, processing and analysis to the business data within the enterprises.

In the process the long-term operations of commercial banks, a lot of data are accumulated in the banks. These data scattering in banking systems is faced for business, not for the

decision-making. Therefore, these disparate data must be extracted, cleansed, transformed and loaded, then the integrated, standardized, so structured banking data warehouse is established. Ultimately a comprehensive, consensus decision-making data formatted. Meanwhile, according to different themes a corresponding data marts can be designed, such as general customer analysis data mart and other large customer analysis data mart, etc. The construction of multi-data mart is benefit to analyze different characteristics of customer behavior.

In recent years, banks and financial institutions to adapt the gradual diversification of customer demand have built the customer relationship management system of bank. CRM system is an integrated multi-technology and generally it is composed of the source of data, data preparation ,metadata, manage and control components and data warehouse tools[5]. It has shown in Fig. 1. Fig. 1 is from [6] and is modified according to this paper.

1) *Operating level(data source)*: Automatic integration of banking business processes.

2) *Analysis level(data storage and management)*: analyzing the data generated by the operational CRM (including data warehouse and data mining techniques).

3) *Customer service level*: for co-operation service system including telephone calling system and e-commerce sites.

When the banks implement customer relationship management they must carry out bank customer segmentation. All this must start with the design of the data warehouse. In the design of logical model of the data warehouse star pattern and snowflake pattern are often used. Star pattern has only one dimension table for each dimension, and in the dimension table there is a primary key, and in the corresponding fact table there is the corresponding foreign key[7]. In this paper we adopt the specimen of loans to corporate clients and use the index, i.e. the degree of contribution and the credit amount to analyze how to build the data warehouse of bank customer segmentation. The data warehouse model uses a common data warehouse-star pattern.

Contribution index include customer contribution analysis, the branch contribution analysis, and product contribution analysis, as well as any combination of the three contribution analysis. The loan amount index for decision makers should also include time, customer, loan types and branch institutions.

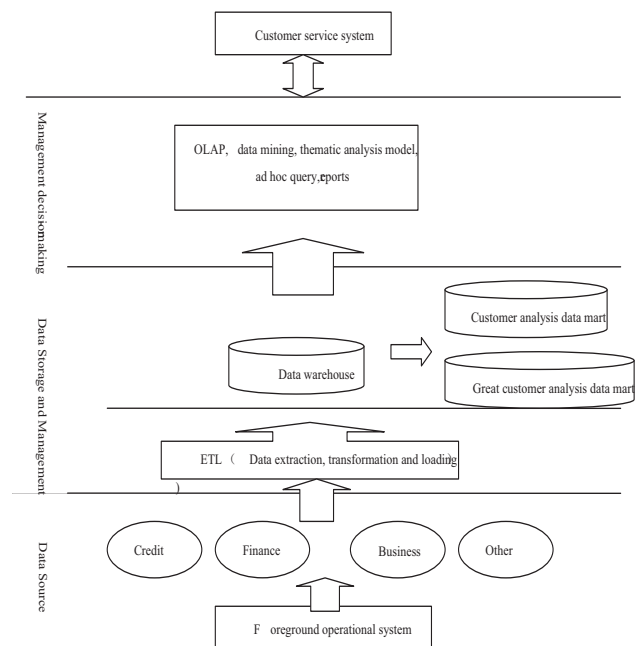


Figure 1. .Banking customer relationship management architecture

Now we adopt bank credit operation data and set up the data warehouse of the star data model. Assessment index in the fact table of central dimension is the contribution index and the amount of loan, and then complemented by customers, branches, products and time as the dimension. The model has shown in fig. 2. Fig.2 is from [5] and is changed according to this paper.

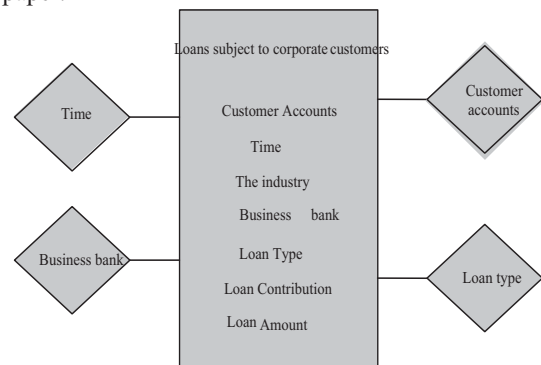


Figure 2. Model of fact table

III. BANK CUSTOMER SEGMENTATION BASED ON DATA MINING

A. Data Mining Techniques

Data Mining(DM) is a kind of new technology to discover and extract useful information hidden in large databases or data warehouse. The valuable and potential knowledge, models or rules can be extracted from the large amounts of data to help companies find business trends, reveal the facts known and predict the results of the unknown [8]. The methods of data mining include: concept description, association analysis, classification and prediction analysis, clustering analysis, outlier analysis, evolution analysis, complex types of data mining, etc.

B. Customer Segmentation Model of Bank

Although customer relationship management system has been widely used, the studies on the customer segmentation in current, still lack a unified model. Managers often use some abstract concept to describe customers, such as "high-income customer", "low-income customer", "high-risk customers", "low-risk customers" and so on when they formulate marketing strategies or carry out the customer relationship management decisions. Connotation and extension of some concepts are more clear and are more easy to define while connotation of more concepts often define a structured description hardly and their extension has also been affected by a number of subjective and objective factors. The main task of customer segmentation of bank is to define the corresponding relationships between these concepts and customers. In customer data we can look on attribute of each customer as a space and each customer is a dot in the space, then all customers in the customer database of the company can constitute a multi-dimensional space. We define the space as the customer attribute space. The mapping relationships between the customer attributes and the concepts are mined through customer data of the known concept type in the database. The mapping relations between the customer attribute space and concept space induced automatically.

We set $A = \{A_1, A_2, \dots, A_n\}$ as a set of attribute describing the customer characteristics and behaviors, which constitute n dimensional attribute space A^n of the customers. Each dimension is an attribute dimension. The value of this n attribute dimensions of each customer determines the location of customer in the attribute space.

We set $G = \{G_1, G_2, \dots, G_p\}$ as P different concept dimensions, the P concept dimensions constitutes P dimension concept space G^p of the customer. A can determine a set of concepts G , $C' \subseteq C$, C' is the known customer collection of concept type. We determine a mapping relationship P , $C \rightarrow G$. Make $\forall c \in C'$, if $c \in G_i$, then $p(c) = G_i$. $\forall c \in C$, we can determine the

concept type what it is through $p(c) = G_i$. So customer segmentation refers to establish mapping relationships between the customer attribute space and concept attribute space $A^m \rightarrow G^p$.

Customer segmentation process model based on data mining has shown in Fig.3. Fig.3 is from [9] and is modified according to this paper. We use the corresponding algorithm to identify classification model or classification functions describing and distinguishing data classification or concepts classification. The classification model obtained can be expressed in various forms, such as association rules, neural networks, decision trees and so on. In the building process of the classification model the classification is needed to label the known customer data. This data source is data warehouse of customer relationship management system. Data Source is mainly consist of by the customer data storage and concept dimensional data. Among them, customer data storage includes all intrinsic attributes, extrinsic attributes and behavior attribute data of the corporate customers. The concept dimensional data involves various concepts related to customer analysis in the concept space. The methods section is the core of customer segmentation model which will include a variety of data mining methods and the integration of these methods. Data mining is to build mapping relations between the customer data and the concept dimension through the algorithm of data in the customer data storage. Functional analysis includes customer value analysis, customer loyalty analysis, credit risk analysis, marketing mix selection whose features are based on the establishment of mapping relations between the client and concepts. [9] [10]

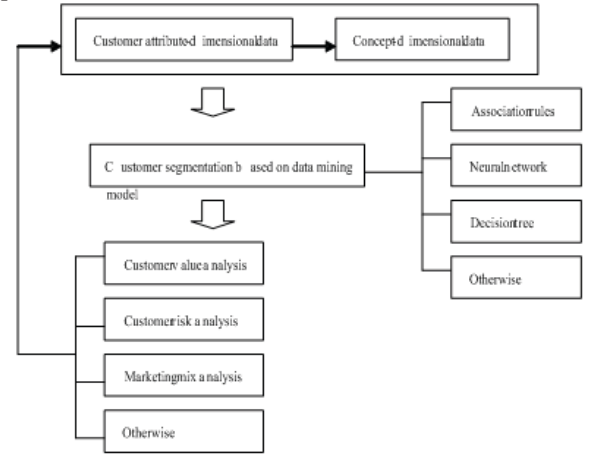


Figure 3. Customer segmentation model of bank based on data mining

IV. THE APPLICATION OF BANK CUSTOMER SEGMENTATION BASED ON SOM AND K-MEANS ALGORITHM

It is necessary to collect user information comprehensively, to store user information, and to use data mining techniques to classify the users by the inherent characteristics of the user before customer segmentation of the user groups. In the customer segmentation the behavior

characteristics is the main and the natural attributes are secondary features.

A common clustering algorithm in data mining i.e.K-means was proposed by MacQueen that based on the error squares in 1967 [11]. The algorithm is convenient, fast and able to deal effectively with large databases. But there is inherent weakness in K-means algorithm[12]--e.g. the initial value impacts on the clustering results greatly,it can easily be trapped into local optimum, and it depends on experience to determine the number of optimal classification. These shortcomings limit its effectiveness and scope of application significantly. Self-organizing and mapping neural network is an unsupervised learning mode.It can move data to map from high-dimensional space to low-dimensional space which it is through the dimension reduction to find the major statistical characteristics of multi-dimensional data.It can classify the data automatically into different categories according to the similarity among data.So it can achieve the results that enhance useful customer information and reduce the noise.But precise clustering information after classification can not be provided by the network while known K-means has a very high accuracy in the known number of clustering and the center[13]. We combine advantages and disadvantages of the two algorithms [14] to carry out application in bank customer segmentation.

The algorithm belongs to a two-stage method: in the first phase of clustering, the clustering of massive data samples begins by the neural network, and the vectors with similar features belongs to the same type ,then the sample data gather into different categories and obtained the number of type and focal point for the various classes; in the second stage, the results in the first phase is used as the initial value to be input in K-means algorithm through further clustering to form the final clustering results, the algorithm is described below [15]:

The first phase is to obtain the number of clustering and types of center by SOM network and to output the number of clusters K and the clustering center Z.

The second stage is to select threshold and the output results in the first phase (the number of clusters K and the clustering center Z) is as the initial input in the K-means algorithm, then iteration, until convergence. Step 1: Initialized. Set the number of categories K, assign initial value to clustering center of each category. Step 2: Sample divided. Divide vectors of all the samples. By taking this step each sample vector can associate with one of the categories. Step 3: Calculate the new clustering center. Use the new-typed collection of all members established in step 2 to recalculate the center value of each type so that the distance Z from each vector in the category to a new clustering center minimizes. Step 4: Check convergence. Step 5: Output clustering information.

SOM-K-means algorithm should comply with data mining process in the application of customer segmentation. It mainly includes five steps such as selection of segmentation variable, data preparation, segmentation modeling, customer profiling and marketing mix. In this paper we take loans to corporate clients from banks as example.

1) *variable selection*:Multi-dimensional segmentation variables is selected in this paper, including customer value, behavior characteristics, and demographic characteristics. In this case the variables like the amount of loans and the degree of the contribution are selected to describe the behavior and value of segmentation customers, and these variable data is as objects of clustering to collect the data.

2) *data preparation*:Data preparation includes data cleansing, data integration, data filtering and data conversion.

3) *segmentation modeling*:In this paper we select loan to corporate customers of some bank as a sample by the mentioned SOM-K-means algorithm to clustering analysis, and data subjects are divided into three types.The clustering results has shown table 1.

customer profiling and marketing mix:In three types of customers from table 1 loan tends to focus on large customers obviously. 61% of the total number of customers has less loans and contribute little to the bank.For this part the managers can not pay attention to them.26% of the total number of customers is in the medium in the proportion of the total loan and they belong to the basic customer on contribution to the bank.For this part the managers should attach importance to strengthening the management and specific guidance. First, they must strengthen supervision and management and window guidance to prevent the credit deterioration and result in new non-performing loans.Second is to do its best to improve service level for the customer,while it is required to concern about the risk of changes in some of these customers. Only 13% customers account for larger loan. Although nowadays it is a larger contribution to the bank, it may cause the bank to concentration and systemic risk.The managers must attach great importance to the operational risks of big clients under the premise of prevention and strengthen the maintenance of large customer relationships in order to achieve maximum profits from the major clients effectively.

TABLE I. CLUSTERING RESULTS ATTRIBUTE

attribute Categories		Loan amount	Contribution	Propotion of customer
Cluster1	mean	0.006108	0.000132	0.61
Cluster2	mean	0.052654	0.098536	0.26
Cluster3	mean	0.139856	0.253590	0.13
Total	mean	0.026363	0.028896	1

V. CONCLUSIONS

In this paper a structural model of customer segmentation is established based on data mining techniques. A customer classification prediction model is proposed by analyzing historical data from known and classified information by data mining. Because of non-deterministic and non-uniformity of the customer behavior a dynamic model of customer behavior descriptions should be built in order to objectively reflect the real features of customers. In a customer-oriented management system, the results of

customer segmentation will be related to a variety of decision making of banks. Of course, it is not limited to customer relationship management, customer segmentation, and application of data mining is also not limited to customer relationship management. This article attempts to a only try in data mining applying in the field of customer segmentation of bank.

REFERENCES

- [1] W.H.Lnmon,Data Warehouse,3rd ed., Beijing: Mechanical Industry Press,Aug.2005,pp.68-75.
- [2] Suh E.H, Noh K.C, and Suh C.K.,“Customer list segmentation using the combined response model,” Expert Systems with Applications, vol. 17,Feb.1999,pp.89-97.
- [3] Alfredo V.A.,Methodology for the Characterization of Business-to-Consumer E-commerce,Liverpool, UK: John Moores University, 2000,pp.78-83.
- [4] Boone D.S,and Michelle R.,“Retail segmentation using artificial neural networks,”International Journal of Research in Marketing, vol. 19,Mar.2002,pp.287-301.
- [5] Lu Haiyan, Fu YingLiang,and Xing Cuifang, “Data Warehouse in the banking customer relationship management application,”Journal of Dalian Maritime University vol.33,Jun. SupPI.2007,pp.181-186.
- [6] Zhang Rong,“An application of Data warehouse and data mining in customer relationship management of bank,” Computer and Information Technology,Feb. 2006,pp.79-81.
- [7] Li Yibo, Ji Hongyu,and Bai Xiaoming, “A reasonable choice of data mining tools,” Computer and Information Technology,Jan. 2005,pp.68-70.
- [8] Jiawei Han, and Michelline Kamber,Concept and Technology of Data Mining,Beijing: Mechanical Industry Press, 2002,pp.137-140.
- [9] Zhang Guozheng,“ Customer segmentation in customer relationship management based on data mining,”Commercial Research, vol.345,2006(13),pp153-155.
- [10] Fang Anru, Ye Qiang,Lu Qi,and LI Yijun, “Customer segmentation framework model based on data mining,” Computer Engineering, vol.35,Oct.2009,pp.251-253.
- [11] Tou J.T, and Conzalez R.C., Pattern recognition, principle. AddisonWesley, Reading, 1974,pp.110-112.
- [12] Liu Jingming, Han Lichuan, and Hou Liwen, “Based on Particle Swarm K-means clustering algorithm,”Engineering Theory and Practice,Jun.2005, pp.54-58.
- [13] Balakrishnan P.V. (Sundar), and Cooper, M.C, & Jacob, V.S., “A study of classification capabilities of neural networks using unsupervised learning: A comparison with K-means clustering,” Psychometrika, Vol.59.2004 (4),pp.509-525.
- [14] Zhou Huan, and Huang Liping, “SOM neural network-based K-means algorithm,” Computer Applications, Jun.2007,pp.51-52.
- [15] Zhou Huan, and Liu Yabin “Clustering analysis in the application of customer segmentation,”Jiangsu Business,Aug.2008,pp.77-79.