Telecom Customer Segmentation with K-means Clustering

Luo Ye, Cai Qiu-ru, Xi Hai-xu, Liu Yi-jun, Yu Zhi-min
School of Computer Engineering
Jiangsu Teachers University of Technology
Changzhou, Jiangsu, China, 213001
Lois_ye@163.com

Abstract—**Development of data mining application is very important for the telecommunication enterprise, which is a typical data-intensive industry. Customer segmentation can help analyze customer composition accurately and promote the quality of service and marketing. Using K-means clustering and the commercial automatic data mining tool KXEN, the paper proposes a resolution of customer segmentation for Changzhou telecom in Jiangsu province. Results show that the resolution is effective and successful.**

Keywords- Telecom; Customer segmentation; K-means clustering; KXEN software

I. INTRODUCTION

In these years, with increasingly intense competition in the telecommunication industry, price adjustment strategies have been unable to meet customers' personal needs. The change of supply and demand and the diversification of customer needs require that service provider's business model should turn from "technology-driven" to "market-driven" and "customer-driven", and their competitive field should transfer to the innovation of "business model, customer service, customer management and brand positioning"[1].

Data mining can extract valuable knowledge from large amounts of data set in a human-understandable way, so the development of data mining applications is important for the development of telecommunication corporations. With the development of data mining technology, data mining has a wider application space in the field of telecommunications, such as customer segmentation, analysis of owing fee, pattern analysis of customer calls, dynamic fraud preventing, business forecasting, identification of major customers, and prediction and control of customer loss[2-3].

Customer segmentation is to classify customers into different groups according to one or more attributes. The customers within the same group have greatest similarity, and the ones not in the same group have greatest difference. Through classifying customers into right categories, making a sectional analysation of current and expected customers, and judging the salient features of different sections, we will have an accurate understanding of customer composition, and make service and marketing more targeted. Customer segmentation achieves the following objectives[4]: (1) to understand the customer's overall composition; (2) to understand group characteristics of various valuable customers; (3) to understand group characteristics of loss customers; (4) to understand consumption characteristics of customers; (5) to understand group characteristics of customers with different credit rating.

Clustering technology of data mining can be used to find different groups of customers from the basic customer library, describe the characteristics of different customer groups, and so achieve the purpose of customer segmentation.

II. K-MEANS CLUSTERING METHOD

Commonly used data mining techniques includes association analysis, classification and prediction, cluster analysis, outlier analysis and evolution analysis. Among them, the cluster analysis can be used to solve the problem of customer grouping.

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Clustering techniques are organized into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods[5-7]. Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k<=n$. As a partitioning method, K-means algorithm is one of popular heuristic methods adopted by most applications[7]. This paper adopts K-means clustering algorithm to group customers, due to its following advantages:

a) The algorithm gives a good solution for the clustering problem of data objects with numeric attributes, and often terminates at a local optimum.

b) The algorithm is relatively scalable and efficient in processing large data sets.

c) The algorithm is not sensitive to the input order of data.

978-1-4673-0242-5/12/$31.00 ©2012 IEEE

d) Although the algorithm lacks the ability to process noisy data, the telecom data are relatively complete and the data pre-processing can make up for it.

e) The algorithm is fast in modeling, and its results are easier to understand.

The K-means algorithm is as follows[7]:

```
Algorithm: K-means
Input: The number of cluster k and a database containing n objects.
Output: A set of k clusters that minimizes the squared-error criterion.
1: arbitrarily choose objects as the initial cluster centers;
2: repeat
3: (re)assign each object to the cluster to which the object is the most similar;
4: based on the mean value of the objects in the cluster;
5: update the cluster means, i.e., calculate the mean value of the objects for each cluster;
6: until no change;
```

Let n be the number of all objects, k be the number of clusters, and t the number of iterations. The computational complexity of the algorithm is $O(nkt)$. Usually, $k<<n$ and $t<<n$.

III. KXEN SOFTWARE

As one of three most famous data mining software (SAS/EM, KXEN, SPSS/Clementine), KXEN is different from the other two. Focusing on high-end technologies of data mining, KXEN is results-oriented, rather than for the process. Users do not need a professional background of statistics and machine learning, but only need to know the data and the problem to be analyzed. KXEN provides a simple solution for each problem. As commercial automatic data mining software, KXEN has the following characteristics [8-9]:

a) In the data preparation phase, KXEN can automatically process missing values and outlier values, and code the attribute values. Due to KXEN's unique pre-coding techniques and feature selection methods, modeling time is reduced significantly.

b) When modeling, KXEN needs no additional disk space to store the data, which is processed directly in the data warehouse. KXEN makes good use of the performance of data warehouse and saves hardware cost, coinciding with the current concept of Knowledge Discovery in Database.

c) KXEN has totally four modules, which are robust regression, smart clustering, association rules and time series, to solve all the commercial data mining problems. Users have no need to select the algorithm, because there is only one algorithm for each business problem. All algorithms are based on Vapnik's structural risk minimization theory.

d) In KXEN engine, structural risk minimization theory is used to find the best model automatically, and parameter setting is not required.

e) All components of KXEN are designed to show meaningful and explanatory results to users.

Because of above characteristics, KXEN has changed traditional data mining methods. Data preparation costs little time now, in comparison with 70% of the total modeling time previously. However, the KXEN models are as robust and accurate as the models created by traditional tools.

IV. CASE STUDY

A. Definition of the business problem

The paper takes the customers of Changzhou telecom in Jiangsu province as case to atudy. The commercial object defined by telecom service providers is as follows: Group hundreds of thousands of public customers in city from the dimensions of values and behaviors, to understand consumption characteristics of different customer groups, provide an analytical basis of marketing strategies for developing new business, stopping customer losing and competing for users of other networks, and achieve the strategic object of profit improvement. Using the data mining technique of clustering analysis, we can group the target customers, characterize each group and analyze their properties. Furthermore, the target customers will be determined for targeted marketing and appropriate marketing programs will be developed according to customer properties and marketing objectives.

B. Customer clustering with K-means algorithm

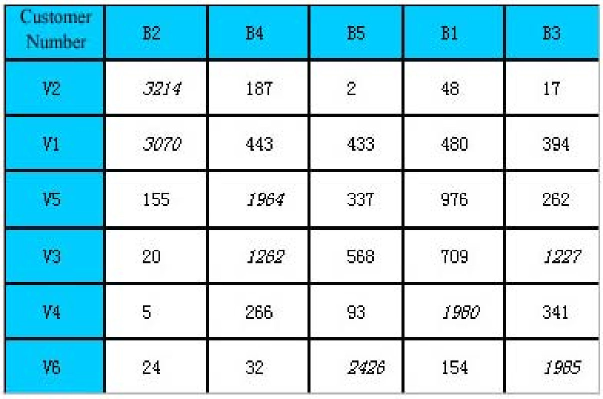
This paper selects small business customers, for whom marketing services are relatively weak, as target customers, and obtains their relevant data set of target customers for nearly a year. There are 23074 small business customers, who have 1-2 telephones or personal handy-phone, but haven’t install BAN (Broadband Access Network). Basic tables cover the following data:

a) Basic customer information: including customer identification, contact methods, product ownership, network access time, the opening of services, benefits package information, customer service information (such as complaints, consultation, fee notice and so on), etc.

b) Value information: including business monthly fees, user fees, concession fees and value-added services, new business, information costs and cards, settlement costs, arrears information.

c) Behavioral information: including call duration, times of calls, hop time, the number of different telephone numbers called by the speaker, concentration of call duration, concentration of times of calls and so on.

KXEN software adopts K-means clustering algorithm based on structural risk minimization to realize the segmentation. Using the KXEN software, the paper divides customers into six value groups from the value latitude (V), and five behavior groups from the behavior latitude (B). There are 21 V variables and 15 B variables participating in segmentation. Then we sort the groups for V variables according to sum of costs, and so VB-crossing matrix are formed, as shown in Figure 1 and Figure 2. From the matrix, eight strategic customer groups SS_1-SS_8 with more than 1,000 people are selected, and the total number of the customers is 17,128, which is 74.23% of all customers.



Customer Number	B2	B4	B5	B1	B3
V2	3214	187	2	48	17
V1	3070	443	433	480	394
V5	155	1964	337	976	262
V3	20	1262	568	709	1227
V4	5	266	93	1980	341
V6	24	32	2426	154	1985

Figure 1. VB-crossing matrix map of customer groups

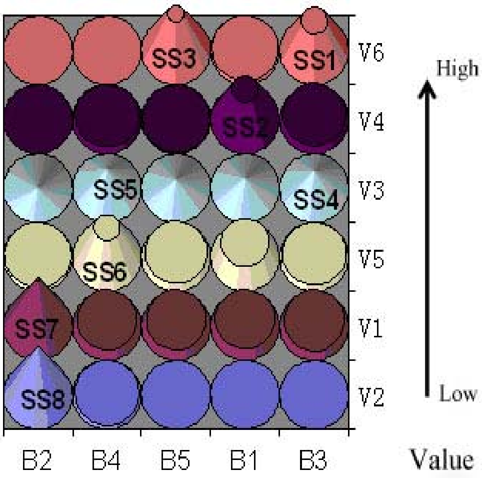


Figure 2. Matrix map of small business customer strategy groups (SS_1-SS_8)

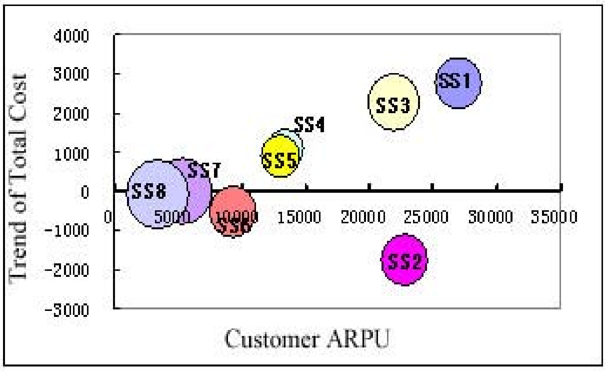


Figure 3. Bubble distribution diagram of small business customer strategy groups (SS_1-SS_8)

According to segmentation results, the ARPU of each group is calculated for customer value analysis:

$$ARPU = \frac{\text{Total revenue of a customer group}}{m}$$

Where m is the group size, i.e. the number of customers in a group, and ARPU (Average Revenue Per User) is average revenue of every customer. Figure 3 is the bubble graph of group distribution. Each bubble stands for one group, and the bubble size presents the customer number of the corresponding group. SS_7 and SS_8 are the two largest customer groups. The groups on the right-hand of horizontal direction have higher value than the ones on the left-hand. The groups of SS_1, SS_2 and SS_3 are highly valuable small business customers, and the groups of SS_1, SS_2 and SS_3 are low valuable ones. The vertical axis shows the customers' consumption trends. The horizontal axis below shows a downward trend, and the more deviation, the greater decline of value.

C. *Decision Analysis*

Here an analysis of the middle valuable declining group SS_6 is given. Due to space limitations, only the competitive characteristics are shown in figure.

1) *Characteristics of the total cost*

As the middle valuable group, SS_6 has 1,964 customers, the 8.5% of all small business customers, and its customer ARPU is 93.35 Yuan. The declining trend of total costs is -5.28 Yuan, which is the second lowest among groups. There are only a few customers with rising trend of total costs. In the first half of the year, the mean total cost is 106 Yuan, and the trend is steady.

2) *Characteristics of long-distance call*

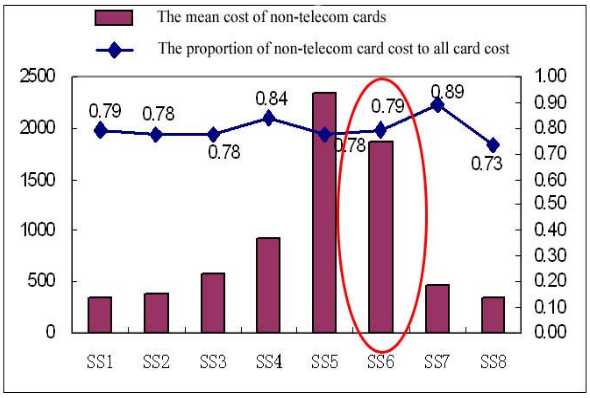
In this group, the mean cost of the customers' long-distance call is 9.97 Yuan, which is a relatively low percentage of 11% of total cost. The long-distance charges drop 1.96 Yuan monthly on average, which is the second highest among groups. There are a low percentage of customers whose long-distance charges are more than 50 Yuan or on the increase, and the customers using the traditional call are the largest proportion of all groups. In addition, the customers of this group rarely use long-distance calls, among which they mainly use the traditional call. It is the largest proportion of all long-distance calls.

3) *Characteristics of Local calls*

The mean local cost of this group is a middle value, 41.67 Yuan. On average, it drops 0.37 Yuan each month. The mean interval cost is 2.92 Yuan, and the interval cost drops 0.37 Yuan monthly on average.

4) *Competitive characteristics*

According to Figure 4, the customers of this group have a habit of using card business, the main of which is other networks. But the decline is most evident.



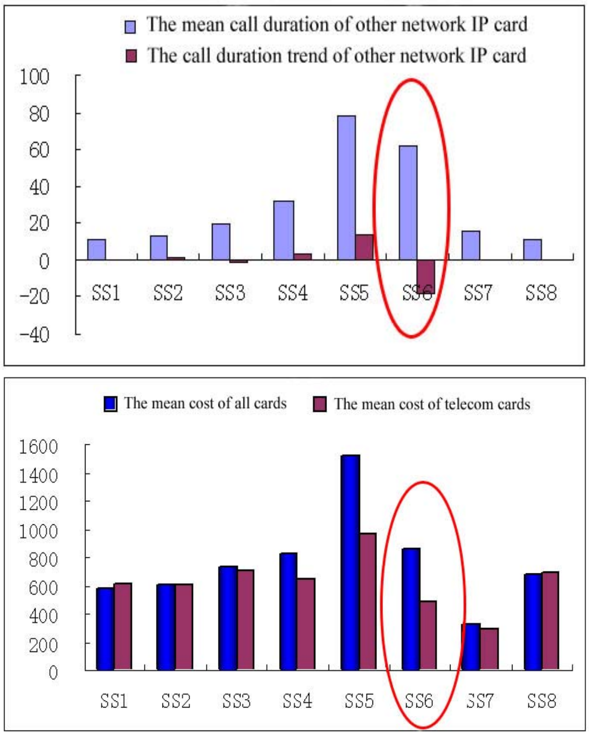


Figure 4. Competitive features of SS_6 group

5) Other characteristics

The cost of narrowband Internet accessing in this group has dropped, and the mean discount cost is 13 Yuan, which is the second lowest among all groups.

In summary, the characteristics of the group SS_6 are as follows. (1) With a downward tendency, the ARPU of this middle valuable group is 93.35 Yuan. The total cost has an obviously descending trend of 5.28 Yuan every month. (2) The group is low valuable in long-distance calls, most of which are traditional calls. (3) Other network cards are used more, but the using of other network IP cards has a downward trend. The mean non-telecom cost of this group is the highest, and there is a serious loss of long-distance calls.

V. CONCLUSIONS

With the telecom reform and further opening up of the telecommunications market, the domestic telecom service providers are facing a new and more intense market competition. While the marketing strategies are transforming product-oriented to customer-oriented to market-oriented profoundly, the corporations are commencing to focusing on customer assets rather than products and business. Data mining can automatically analyze large amounts of data sets and learn new potential patterns. Cluster analysis is applied to target marketing of telecommunications to solve the problems of customer segmentation.

With data mining tools KXEN, this paper uses K-means clustering method to give a solution for telecom customer segmentation. By segmentation, consumption characteristics of each group can be showed in figure, and so we may understand user groups deeply. According to the results of customer segmentation and business characteristics of different groups, market analysts recommend customers their favorite business, which to some degree support targeted marketing strategies carried out by telecom. Practice results show that the target marketing solution of customer segmentation provided by this paper is successful and effective.

ACKNOWLEDGMENT

This work was supported by National education information technology research fund (116230340), and Modern education technology research fund of Jiangsu province (2011-R-18859). Furthermore, we are indebted to the support and encouragements received from the staff and colleagues of the school of computer engineering.

REFERENCES

[1] Tan Jun. The Analysis and Design of the Telecom Customer Segmentation Model Based on Data Mining of the CRM[D]. Chongqing: Chongqing University, 2005.5.

[2] Wang Li. The research and application of customer market segmentation of a certain Netcom corporation. Beijing: University of international business and economics, 2007.4.

[3] Jiang Xin, Li Yi-jie, Liu Ming-yi. Application of Clustering Algorithm in Cross Selling of Telecommunication[J]. Computer Simulation, 2009,26(9):261-263.

[4] Tao Lu-jing. Design and Implementation of Telecom customer segmentation based on data mining[D]. Nanjing: Nanjing University, 2005.6.

[5] Zhang Jian-ping, Liu Xi-yu. Research and application of K-means algorithm based on clustering analysis [J]. Application research of computers, 2007(5):166-168.

[6] WU Dong-yang, YE Ning, SHEN Li-rong. Clustering Method for Automatic Timber Defects Detection Based on the Color Moment[J]. Journal of Jiangnan University (Natural Science Edition),2009,8(5):520-524.

[7] Han Jia-wei, Micheline Kamber. Data mining-concepts and techniques[M]. Higher education press, 2001.

[8] Liu Wen. KXEN commercial data mining[EB/OL].(2008-7-9). www.datom-i.com.

[9] Liu Wen. A tutorial market segmentation with KXEN [EB/OL].(2007-7). http://www.amteam.org/print.aspx?id=606090.