Improved K-Means Cluster Algorithm in Telecommunications Enterprises Customer Segmentation

Jinghua Zhao, Wenbo Zhang
Computer College
Jilin Normal University
Siping, China
E-mail: zwb@jlnu.edu.cn

Yanwei Liu
College of Computer Science and Technology
Jilin University,
Changchun, China

***Abstract*—According to the actual requirements of telecommunications enterprises customer segmentation, in this paper an improved K-Means algorithm was introduced. The algorithm was used to initialize cluster center in the cluster analysis phase of Data Mining technology. The experimental results proved that the novel algorithm has greater improvement than the original one in efficiency and accuracy. The results applying the novel method to telecommunications enterprises customer segmentation showed that the segmentation results obtained can be used as the data basis in differentiated services for customers and have positive significance for product design and phone packages recommendation.**

Keywords—Customer Segmentation; Data Mining; K-Means Cluster Algorithm; Telecommunications Enterprise

I. INTRODUCTION

In face of increasingly competitive market economy, customer has become the most important resources associated with the enterprises' development. Understanding the preferences of different consumer groups, their shopping attitudes and the concept of price are the keys to marketing success. Based on this condition, customer segmentation[1,2] can help enterprises to find the proper market positioning for their products. It can also be used to select appropriate customer groups in target market to find the marketing opportunities in current customer groups and to gain competitive advantages through product diversification.

Traditionally, marketers often differentiate their customers by one-dimension attribute. For example, telecom business always divide their customers as great customers，business customers and public customers according to revenue contribution. This method has its merits in simplicity, convenience and approachability, however, it can't keep pace with customers' needs diversification and technical progress. In face of this question, it is necessary to found a new method, such as customer behavioral segmentation. This new segmentation methodology can handle hundreds of variables, and its results can make marketers understand their customers better. This new method is also called "segmentation based on Data Mining"[3].

Data Mining is a process of exploring some unknown and potential useful information and knowledge from a large amount, incomplete, noisy, indistinct and random data. In this paper an improved K-Means algorithm was used as cluster method in the process of applying Data Mining technology to customer segmentation. A model of telecommunications enterprises customer segmentation was designed. Based the model the segmentation process was executed and a features analysis was processed on 14 segmentation customer groups obtained and some marketing proposal were put forward.

II. IMPROVEMENT AND VALIDATION OF K-MEANS ALGORITHM

K-Means algorithm[4] is a classical algorithm to solve the clustering problem. The idea of specific algorithm is that the k sample points selected randomly are taken as the center of initialized cluster and then performing iterative operations. Clustering results are affected by the choice of initial point, and therefore the solutions obtained are always local optimum, not global optimum.

Therefore, the main a breakthrough of this algorithm improvement is that finding a set of data objects reflecting the data distribution and taking it as the cluster center, namely, the first step of improved K-Means algorithm. During the choice of the initial cluster center, selecting k samples in which their distances are relatively far away more representative than the random sample of k. With introducing improved method distances of k cluster centers were made farther as far as possible in order to assign them to different classes. The process of initialized cluster center as follows:

- Setting m=1, a sample is taken randomly as the first cluster center in the data set;
- Setting m=2, taking a farthest sample as the second cluster center by calculating;
- Judging whether the m is greater than or equal to the number of clusters k-1,if it is true take the arithmetic to the fourth step, otherwise, setting m=m+1 and making
  $\max(d(x_i, z_1) + d(x_x, z_2) + \wedge + d(x_i, z_{m-1}))(m \le i \le n)$
  satisfy the sample x_i , and setting x_i as the mth center z_m ;
- If k=2 take the arithmetic to the fifth step, otherwise, calculating $z_{centroid}$ of top k-1 cluster centers and calculating each sample $x_i(i = 1, 2, K, n)$ in data set in order to make $d(x_i, z_{centroid})$ minimum. The x_i obtained is kth center;
- Determining k initial cluster centers of the algorithm.

978-1-4244-6943-7/10/$26.00 ©2010 IEEE

By above processes k cluster centers were initialized and after applying the 2nd-4th algorithm steps the cluster results were obtained.

To verify the effectiveness of the improved algorithm, in this experiment a set of test data Iris[5] was selected, which come from UCI database. The following is a comparison of the two algorithms results, which were shown in table 1.

TABLE I. COMPARISON OF CLUSTER RESULTS BETWEEN TWO THE ALGORITHMS

	K-Means			Improved K-Means		
experiment	initial cluster center	time of iterations	accuracy	initial cluster center	time of iterations	accuracy
1	120, 72, 11	9	85.33%	94, 118, 15	3	85.33%
2	19, 6, 79	11	68.67%	55, 14, 119	7	84.00%
3	48, 63, 94	13	85.33%	90, 118, 23	7	84.00%
4	6, 71, 16	9	68.67%	84, 23, 118	7	84.00%
5	77, 10, 62	6	85.33%	22, 119, 107	5	84.00%
6	36, 33, 136	6	69.33%	110, 14, 61	11	85.33%
7	148.52, 114	6	84.00%	40, 119, 107	5	84.00%
8	77, 33, 1	8	68.67%	76, 14, 119	7	84.00%
9	53, 111, 131	10	84.00%	31, 119, 107	5	84.00%
10	139, 41, 9	5	69.33%	58, 118, 15	4	85.33%
average		8.9	76.87%		6	84.53%

It could be illustrated from comparative results that improved K-Means was advanced, compared with original algorithm in terms of time of iterations and accuracy, and particularly in terms of accuracy, improved K-Means was more stable and difference between the maximum and minimum was only 1.33%, compared with 16.66% of original algorithm. At the same time, sometimes two centers were selected in the same group to makes less effective in the K-Means algorithm, for example, in the 2nd, 4th, 8th and 10th experiment, but in the improved K-Means algorithm, most of the cluster centers were assigned in three different groups, thus better results could be got.

III. IMPLEMENTATION AND RESULTS ANALYSIS OF TELECOM CUSTOMER SEGMENTATION

CRISP-DM divides Data Mining process into six phrases: business understanding, data understanding, data preparation, modeling, evaluation, deployment[6]. This paper references the Data Mining process of CRISP-DM, combining with the characteristics of telecommunications customer data, dividing the customer segmentation process into six phrases: definition of operational issues, data acquisition, data processing, cluster analysis, evaluation of cluster results and deployment.

During the cluster analysis phase consists of the following[7]:

- Confirmation of the number of clusters K

  The choice of k has a great influence on clustering results. By one time it is difficult to determine and need to be adjusted based on the clustering results. According to the characteristics of telecommunications enterprises customers and the past experience of operational staffs, the optimal number of the cluster was determined by repeated experiments, namely, k=14.

- Choice of the initial cluster center

  In this study the improved K-Means algorithm was used.

- The output of clustering results

  A part of variable about the overall feature description of customers was shown on figure 1 after the clustering.

Name of Segmentation-groups	The Main Age	Major Occupation	ARPU	MOU	Average Voice Price
Medium Economical Group	20-40 year old	Individual owners	48.7	386.6	0.17
Medium Long-Call and Short-Message Group	20-25 year old	student	45.1	254.8	0.17
...					
Higher Native Busy Group	30-40 year old	Business people	99.1	1124.4	0.22
Higher Business Group	30-40 year old	Business people	123.4	744.1	0.31
...					
lower Native Group	30-40 year old	Employees	28.9	192.8	0.19
lower Night Chat Group	30-40 year old	Employees	25.6	180.2	0.17

...

Name of Segmentation-groups	The Total Number	Proportion	Total Income	Percentage of Revenue
Medium Economical Group	52514.00	8.89	8303444.00	5.62
Medium Long-Call and Short-Message Group	42981.00	7.27	9912834.00	6.70
...				
Higher Native Busy Group	36467.00	6.17	2.430721E7	16.43
Higher Business Group	49192.00	8.32	4.7686195E7	32.25
...				
lower Native Group	50644.00	8.57	7796839.00	5.27
lower Night Chat Group	36262.00	6.14	4330383.00	2.93

Figure 1. overall feature description of customers

Based above results and the telecommunications business rules it was necessary to make some reasonable explanation in order to find out some undiscovered potential laws to guide practices. If reasonable explanation could not be done, the cluster analysis was unsuccessful and essential to repeat these steps.

Using higher business group as an example, a part of basic behavior as shown in figure 2:

Basic Behavior of Higher Business Group	
Proportion is 8.32%; Large customers proportion is 15%; Online time is longer; Mobile roaming time is longest; Number of local calls and Talk Time are highest;	
General Characteristics	The Total Number: 49192.00 Proportion: 8.32
Cost characteristics	Average Points: 2700 The average call costs: 65 The average percentage of the month arrears: 0.01
Call characteristics	Average talk time: 2.07 Average call times: 265 Provincial long-distance time: 22 ...
New Business characteristics	Average number of Short-Message: 148 Flow of GPRS: 351 ...

Figure 2. a part of Basic Behavior of Higher Business Group

By the analysis of the characteristics of high-end business groups, according to the customer groups, the marketing strategy can be divided: with regards to products, local calls, telephone roaming, long-distance telephone are

important; with regards to price, local calls consumption flat rate, long-distance discount, telephone roaming discount, business function discount are important. Moreover, in face of the potential of becoming successful business man, it is necessary for companies that not only emphasizing auxiliary function, but also care for these customers, especially complaint and changing of menu, are significant.

According to the results of grouping customer segmentation, the consumer behaviors can also be analyzed, such as the analysis of APRU value/MOU value, shown in figure 3.

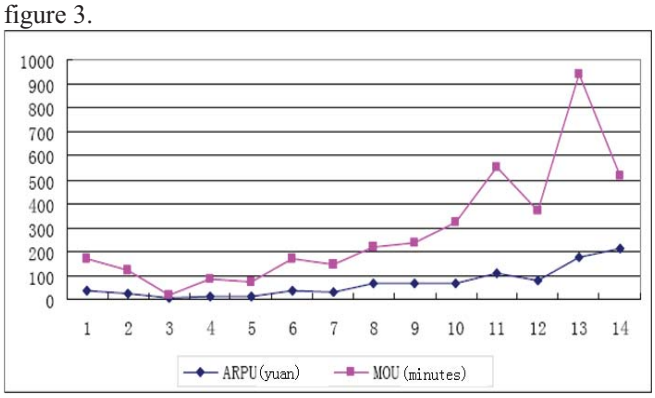

Figure 3. ARPU and MOU value analysis

ARPU, average revenue Per User, is the important indicator for measuring business income of telecom operators. The more high-end users are, the higher the value of ARPU is. MOU means minutes of usage. In figure 3, the amounts of middle-end users and high-end users are much larger than low-end users; especially the value of ARPU in the 14th groups of high-end is the highest. This result also means high-end users are major. Moreover, the value of MOU of 13th and 14th groups is higher than average level. The reason of this result about 13th group is that because of being high-end busy group, its local call timing is the longest. Both the value of ARPU and MOU are the important indicator for measuring business income, therefore, from the figure, high-end and middle-end groups are the main sources of company income.

IV. CONCLUSION

Based on the analysis of disadvantage of the traditional method of telecommunication enterprise customer segmentation, "Data Mining" technology was put forward. The advanced K-Means algorithm was used in cluster analysis of customer segmentation models, and finally the analysis of the results of segmentation was made. The final results showed that the segmentation results obtained can be used as the data basis in differentiated services for customers and have positive significance for product design and phone packages recommendation.

REFERENCES

[1] Suzanne Donner. "What Can Customer Segmentation Accomplish." Bankers Magazine,Mar./Apr.1992,Vo1.175 Issue 2:72-81.

[2] Wang Tao, Wang Yong. "The Design of a Specific Protocol for Embedded Systems' Access into Internet," International Conference for Embeded Systems 2001:288-291.

[3] Zhang Zhiqing. "A Customer Segmentation Research Based on Data Mining," Jinan University,2007.

[4] He Ling,Wu Lingda. " Survey of Clustering Algorithms in Data Mining, " Application Research of Computers, 2007(1)55-57.

[5] Jiawei Han, Micheline Kamber." Data Mining: Concepts and Techniques," Beijing: Machinery Industry Press,2007:251-253.

[6] Sheater C. "The CRISP-DM model: The new blueprint for data mining," Journal of Data Warehousing, 2000,5(4):13-22.

[7] Sambasivam S, Theodosopoulos N. "Advanced data clustering methods of mining Web documents," Issues in Informing Science and Information Technology, 2006(3):563−579.