The research of improved Apriori mining algorithm in bank customer segmentation

Yang GongXin
Henan Vocational and Technical College
Zhengzhou, Henan, China

***Abstract*—The This paper studies bank customers' segmentation problem. Improved Apriori mining algorithm is a kind of data mining technology which is an important method in bank customers segmentation. In practical application, the traditional algorithm has shortcomings of the initial value's sensitive and easy to fall into local optimal value, which will lead to low accuracy rate of silver class customer classification. According to the shortcomings of traditional algorithm, this paper puts forward a bank customer segmentation method based on improved Apriori mining algorithm in order to improve the bank customer segmentation accuracy. Experimental results show that the algorithm can effectively overcome the traditional algorithm's shortcomings of easy to fall into local optimal value, improve the customer classification accuracy, make mining results more reasonable, lay down different customer service strategies for different client base, improve effective reference opinions of bank decision makers, and bring more benefits for the bank.**

Keywords- Improved Apriori mining algorithm; Customer segmentation; Clustering analysis; Bank;

I. INTRODUCTION

With China's accession to the WTO, financial competition is increasingly fierce in the face of foreign bank entry and the deepening of financial reform. High quality customers gradually become the focus of bank competition. Different types of customers bring obvious value differences to the bank. The bank can guide it to have more reasonable configuration market sales, service and management resources by identifying and distinguishing the difference. With a relatively small input, gain more income. But to solve this problem, customer segmentation is needed [1]. Bank customer segmentation is that in the clear strategy, business model and specific market, the bank classifies clients according to the customers' attributes, behavior, needs, preferences and value, and provides the products, services and marketing mode process [2].

At present, the traditional bank customer segmentation has experienced classification method and analysis method based on statistics. Empirical method's bank customer segmentation is the most primitive classification method, generally the decision makers classify the customers according to their own experience with a very strong sense of subjectivity. The segmentation results are not objective and lack of persuasiveness [3]. Customer segmentation based on statistical method is a kind of quantitative research. Clients are classified according to the customer attribute statistical results. Segmentation results often have strong relevance with classification standard. If classification standard is not reasonable, the results of classification are not reasonable [4]. Along with constantly deepen of our country bank information construction, the bank has accumulated a lot of personal history transaction data and customer information, and at the same time, along with the development of the network, customer data will be accumulated more and more. In the face of mass client data, traditional customer segmentation method will be more ragged [5]. In recent years. Data mining technology has rapidly developed. The fusion of many fields' technologies such as the database, artificial intelligence, and statistics can dig out the useful, reliable, new information and knowledge of the process from a large, incomplete, noisy, and fuzzy original data, in which Apriori mining algorithm is one of the most important data mining method and has been widely used in bank customer segmentation [6]. But in the bank customer segmentation application process of the algorithm, it is easy to fall into local optimal solution and not the global optimal solution, so the algorithm is limited in the application of bank customer segmentation (7).

To solve the traditional algorithm's problems in bank customer segmentation process, this paper proposes an improved customer segmentation algorithm of Apriori mining bank. And experimental simulation results show that the bank customer segmentation clustering results of the improved algorithm are more accurate, and more fit the actual situation of bank customer segmentation.

II. BANK CUSTOMER SEGMENTATION BASED ON IMPROVED APRIORI MINING ALGORITHM

A. Bank customer segmentation process

Along with the continuously improvement of bank information, massive data are accumulated in bank's daily business. A lot of valuable information is hidden behind the data. The traditional database technology is used to input, query and statistic analyze data, but can't find the internal relations and rules in data, cannot effectively appy the data to bank marketing development, and produced " rich data, little knowledge" dilemma. Data mining technology can find useful information and knowledge from mass data, perform customer segmentation, get customer category according to the segmentation, launch different financial products and services to provide personalized service, make its marketing policy more targeted, attract important customers, and improve bank's profit and competitiveness. Apriori mining algorithm method is a strong algorithm in data mining technology. It is very suitable for the customer segmentation. By the use of this analysis technology, bank customers can

Published by Atlantis Press, Paris, France.

be effectively classified. Bank customer segmentation process is shown in Figure 1.

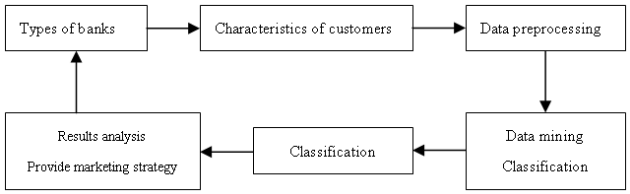

Figure 1. Bank customer segmentation process based on data mining

B. Improved Apriori mining algorithm

Through Apriori algorithm analysis, we can know that when database transaction numbers are little, Apriori algorithm has better performance. When there are large amount of data, a large number of candidate sets would be produced in the connection process, and judge if the candidate set is frequently set needs to repeatedly scanning affairs database, so Apriori algorithm efficiency will be greatly reduced.

Comprehensive consideration all aspects of advantages and disadvantages of Apriori algorithm as well as all kinds of improved method based on the research, this paper puts forward an improved Apriori algorithm from the following aspects.

(1) Reduce the number of candidate sets. Apriori algorithm firstly scans affairs database to get the candidate set C_1, calculates the candidate set's support degree, creates first order frequent itemsets L_1, and then repeated iteration produce C_k by on self-connection L_{k-1}. Improved thought is firstly to count the occurrence number of all projects in L_{k-1}. Using Apriori property reasoning, delete item-sets with project counting less than $k-1$ in L_{k-1}, then generate C_k by connecting the rest sets of the project since connection . Generated C_k contains less candidate item sets than the original quantity, so as to achieve the purpose of optimization.

(2) Becuase project sets in the algorithm are ranked in alphabetical order, in the produce process of C_k by L_{k-1}, first compare, then connect. Making use of the ordering of project sets to reduce the comparison number of judgment.

(3) After pruning C_k, calculate k item set in L_k. If project sets number is less than $k+1$, there is no need to calculate $k+1$ project set, jump out of iteration, and end the algorithm.

(1) The theoretical foundation of the algorithm

Corollary 4.1: To frequent project set L_k, if $|L_k| < k+1$, stop the calculation of itemset $(k+1)-$, in it, $|L_k|$ is the number of frequent $k-$ itemset.

Because there must be a subset in $(k+1)-$ frequent set, if the number of elements in L_k is less than $k+1$, there is no $(k+1)-$ item set in the transaction database.

Corollary 4.2: In ordered arrangement in Apriori algorithm itemsets, if present $(k-1)-$ frequent itemset $L_a = \{L_x[1], L_x[2] \ldots, L_x[k-1]\}$ 和 $L_b = \{L_y[1], L_y[2] \ldots, L_y[k-1]\}$, L_x and L_y satisfy: $L_a[1] = L_b[1], L_a[2] = L_b[2], L_a[k-2] = L_b[k-2], L_a[k-1] = L_b[k-1]$

If a $(k-1)-$ frequent itemset can't connect with L_a, then it also can't connect with L_b and all $(k-1)-$ itemsets after it.

(2) Algorithm description

Improved Apriori algorithm described as follows:

(1) First, scanning affairs database D generates candidate itemsets C_1, and calculates each candidate itemset support degree in C_1.

(2) Delete itemsets less than the minimum support degree and get first order frequent itemsets L_1.

(3) Before generating candidate set C_k by the self-connection of L_{k-1}, first calculate the frequency of all projects in L_{k-1}.

(4) Delete itemsets less than $k-1$ in L_{k-1}, mark L'_{k-1} after pruning of L_{k-1}.

(5) Generate C_k by self-connection of L'_{k-1}, improve the connection judge way of the algorithm by making use of deduciton 4.2.

(6) Cut C_k, get L_k, calculate $k-$ project sets number in L_k. If the number is less than $k+1$, it is no need to operate $k+1$ item set, jump out of the cycle, or turn to step (3) to continue.

(3) Algorithm application analysis

Take sample affairs database in table 4.1 as an example to describe the improved Apriori algorithm.

First of all, scanning affairs database calculates each item set's support number, C_1 = {(A, 3), (B, 5), (C, 4), (D, 3), (E, 3)}; Choose project sets whose project support number is not less than 2 and compose 1 - frequent itemsets L_1 = {A, B, C, D, E}. Then, connect and calculate itemsets support number to get 2 - frequent itemsets L_2 = {AB, AC, AD, BC, BD, BE, CD, CE}.

The process that L_{k-1} connects and produces C_k of improved Apriori algorithm is divided into two steps. The first step, deduction 4.1 is made use of to trim and get $L_k^{'}$'s support number through calculating all items of frequent item L_k in $k-$. The second step, deduction 4.2 is made use of to change the judgment connection way of frequent itemsets in $L_k^{'}$ in the process of $L_k^{'}$'s connection produce process of C_{k+1}.

When $L_2^{'}$ connects and generates C_3, take frequent itemsets AB as an example. Because AB cannot connect with BC, here stop judgement of whether AB can connect with BD and BE after BC. Thus reduce 2 steps judgment. In Apriori algorithm, when generating C_3, need $7*(7+1)/2=28$ times comparison, and the improved algorithm only need $(4+3+2+3+2+1+1)=16$ times comparison.

Through the connection of $L_2^{'}$ to form C_3, calculate all project sets' support number, get frequent 3-itemsets L_3, as is shown in Table 4.3.

Calculate each project's support number in L_3. As is shown in Table 4.4, use deduction 4.1, delete frequent itemsets with occurrence number less than 2, get $L_3^{'}$, as is shown in Table 4.5, similarly, $L_3^{'}$ connection and get C_4 and finally get 4 - frequent itemsets L_4, as is shown in Table 4.6 and Table 4.7. Because there is only one 4- frequent itemsets element less than $k+1$ (i.e., 5), so skip the next step connection, directly judge $L_5=\phi$, and the algorithm end.

Through algorithm analysis, we can see that improved Apriori algorithm has same idea with Apriori algorithm, both proceed according to Apriori steps. First is to scan transaction database, calculate each item set's support number, and compare with the minimum support number and get frequent 1 - itemsets L_1, successively connect and get candidate set C_2, also, scan again transaction database and get frequent 2 - itemsets L_2, so repeated, get the candidate set C_k and frequent k - itemsets L_k.

Improved Apriori algorithm has the advantage of clip before the L_{k-1} connection and generation of C_k, delete project set of occurrence number less than $k-1$, get $L_{k-1}^{'}$, reduce the number of candidate project sets in the next step C_k; Again in the process of $L_{k-1}^{'}$ connects and forms C_k, change connection judgment way of the original algorithm, greatly reduce the unnecessary judgment, and finally calculate the frequent itemsets number, judge whether the iteration would stop, and reduce the iterative times. In practical applications, the efficiency of data mining can be obviously improved.

III. SIMULATION EXPERIMENT

A. Experimental data

Simulation experiment adopts data from customer classification data of the personal financial management business system of domestic city bank. After the pre-process of original data, generate 1000 customer records. Each record includes field: the customer numbers, age, working years, customer monthly salary, bank account number, bank use frequency, lending conditions and housing situation. The bank customers are divided into five categories, specifically see Table 1.

B. Data predict treatment

In the process of original data collection, database contains the incomplete, containing noise data because of human deviation, and at the same time, each field recorded in the database represents different characteristics, often use different measure units, which value differences are very wide. Therefore, it is necessary to preprocess the original data in order to improve the data quality, so that the data mining process is more effective and the classification is more accurate. This paper's prediction process adopts transformation method of centralization and standardization. Centralization's aim is to let each field value have the same base point, specifically as in Formula (4) :

$$x_{ij}^{'} = x_{ij} - \frac{\sum x_{ij}}{n} \qquad (4)$$

In it, x_{ij} is number i article recorded number j field.

On the basis of centralization, transformate it through standardization, make each field's transformation range unified, adopt zero – mean value standardization. Its field mean value and standard deviation are used for standardization, its specific expressed is as Formula (5) :

$$(x_{ij}^{'})' = \frac{(n-1)x_{ij}^{'}}{\sqrt{\sum (x_{ij} - x_j)}} \qquad (5)$$

After the processing of predicting data, each field has same basepoint and range. The standard deviation is 0, and mean value is 1.

C. Algorithm convergence speed comparison

In the hardware environment: CPU P4 3.0 G, 4G memory, 200 G hard drive, the operating system is Windows XP, programming language is c + + environment, and realize the simulation contrast experiment of traditional algorithm and improved algorithm in this paper. The results is shown in Figure 3.

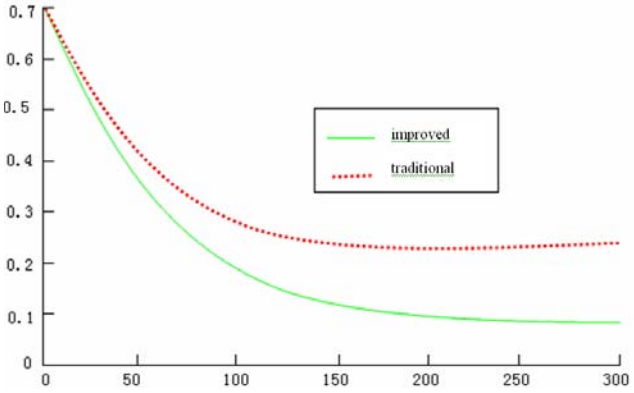

Figure 2. The convergence speed comparison of improved algorithm and traditional algorithm

In Figure 3, it is known that the proposed improved algorithm has faster convergence speed, and get global optimal value. Under the same conditions, traditional algorithm's convergence speed is relatively fast, easy to fall into local optimal value, and the result is very unstable.

D. Customer segmentation accuracy comparison

Divide customer data in customer classification data is divided into training set and testing set two parts, the customer categories is shown in Table 1.

TABLE I. CUSTOMER CATEGORY POINT SCALE

Customer type	**percentage**	**clustering accuracy (%)**	
		traditional algorithm	***this paper algorithm***
Senior customer	5.96	82.15	85.75
Big customer	14.56	85.52	88.15
General customer	60.25	80.98	95.15
Small customer	11.46	85.87	90.07
Potential customers	7.77	90.14	95.14

From table 1's classification results, it can be seen that the improved algorithm this paper proposed overcomes the local optimum resulting from random selection of initial point of the traditional algorithm. The user must give the faults of cluster number formed before hand. And this paper's improved algorithm reduces the clustering dependence of algorithm. the mining accuracy is 80.98% after algorithm improvement, the accuracy of mining is 95.15% after algorithm improvement, and misses almost 20% key customers before improvement. It can be seen that the mining effect of improved algorithm is practical for the actual demand of bank customer segmentation. It provides a new way for customer segmentation by making use of the existing mass data.

IV. CONCLUSION

It is urgent to effectively make use of modern information technology to establish customer management system and enhance the competitiveness of bank. Through data mining technology, find out useful information from huge customer data, thus bank customer segmentation is the key way to solve the problem. This paper analyzes classical mining algorithm in detail in data mining, points out the shortcomings of this algorithm, and puts forward the corresponding improvement methods. And contrast test of the algorithms before and after improvement. The results show that the proposed algorithm optimizes the parameters of the algorithm, speeds up silver class customer segmentation's speed, improves customer segmentation's accuracy, provides decision basis for bank account management and service, thus improves our country's bank " rich data, little knowledge" situation.

REFERENCES

[1] Shao Fengjing, Yu Zhongqing. Data mining theory and algorithm [M]. Beijing: China Water Conservancy and Hydropower Press, 2003.

[2] Zhang JianPing, Liu Xiya. k-means algorithm research and application based on clustering analysis [J]. Computer Application Research, 2007, (5) : 166-168.

[3] Liu YingZi, Wu Hao. Customer segmentation method study review [J]]. Journal of Management Engineering, 2006, 16 (l) : 53 -.

[4] Ma HuiMin, Lu YiQing, Yin HanBin. Customer segmentation method based on the customer share [J]. Journal of Wuhan University of Technology, 2003, 25 (3) : 184-187.

[5] Zhao Faxin, Wang GuoYe. Clustering analysis research in data mining algorithm of [J]. Journal of Tonghua Normal College, 2005 (3) : 11-13

[6] Kuo R J, Ho L M, Hu C M. Cluster Analysis in Industrial Market Segmentation Through Artificial Neural Network[J].Computers and Industrial Engineering, 2002, 42(2): 391-399.