Customer Segmentation Analysis Based on SOM Clustering

Ying Li, Feng Lin
Business School of East China University of Science and Technology
Shanghai, China
liying@ecust.edu.cn

***Abstract*—From the angle of customer value and customer behavior, this paper utilizes data mining methods to segment the clients in security industry. Clustering algorithm is a kind of customer segmentation methods commonly used in data mining. In this article, a two-stage integration of K-means clustering algorithm and SOM network is applied to segment customers and finally forms groups of clients with different features. Through analyzing different groups of customers, we try to position the target clients of the company properly.**

Keywords- SOM network; K-means clustering; Customer value analysis; Customer behavior analysis

I. INTRODUCTION

The definition of market segmentation is proposed by Smith in 1956. Market segmentation refers to the process of forming groups of consumers, whereby members in the same group are homogeneous and members in different groups are heterogeneous, and they are accessible by marketing strategies. The value of performing market segmentation analysis includes positioning the product in the marketplace properly, identifying the appropriate segments for target marketing and finding opportunities in existing markets, and gaining competitive advantage through product differentiation.

The method of market segmentation is divided into two branches according to different market opinion, which are the segmentation oriented by product and the segmentation oriented by customer. The second segmentation method mainly includes statistical segmentation, lifestyle segmentation, behavior segmentation and benefit segmentation. The function of behavior segmentation is to classify customers in terms of the behavior pattern of the existed customers in database. Therein, RFM analysis is a method which uses three behavioral variables to describe and distinguish clients, and is presented by Hughes in the year of 1994.

The two-stage method is a popular customer segmentation method, usually being able to outcome relatively better performance than single algorithms. C.-Y. Tsai and C.-C. Chiu proposed a customer behavior segmentation method, which uses genetic algorithm to find the initial cluster centers and the K-means method to realize the final solution. Melody Y. Kiang, Michael Y. Hu and Dorothy M. Fisher segmented the customers in the American Telephone and Telegraph Company (AT&T) by means of a two-stage method which integrates factor analysis and K-means method. The two-stage method produced better performance and provided decision support for the American Telephone and Telegraph Company.

This paper utilizes a two-stage integration of SOM network and K-means algorithm to carry out segmentation analysis for security clients. After demonstration, this kind of two-stage method is able to produce better performance than the integration of factor analysis and K-means method. Availing of above results, the customers are divided into groups according to customer value and behavior, and the detailed features of each customer group are described. Thus, the company can be familiar with their clients and is able to forecast the purchasing behavior of their clients in the near future.

II. BACKGROUND

A. *Self-Organizing Feature Maps*

SOM is an unsupervised algorithm proposed by Kohonen in the year of 1981. SOM maps high dimension data into lower dimension space and maintains original topological structure. Working principle of SOM network is that when the network receives inputs from outside, it will divide the inputs into several regions, shown as Fig.1. Each region responses differently to the input patterns, the inputs owning similar features are closer to each other, otherwise distant. By the way, there is no distribution assumption to the data used to SOM network [2].

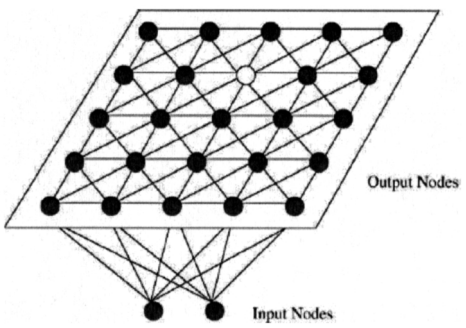

Fig. 1. The Kohonen's feature maps

978-1-4244-2013-1/08/$25.00 ©2008 IEEE

The function of SOM is to lower the dimension of multiple dimension data and keep original relations among data. SOM is commonly used as tool of data preparation. In many two-stage clustering methods, some of them use SOM as the first stage method of data preparation and receive better effectiveness than other two stage methods. Rather, it is unexplored for SOM to be used as clustering tool. This paper utilizes SOM to determine the number of clusters.

B. K-means Clustering

Clustering is a procedure that divides objects into groups, which renders the objects in same cluster be similar and the objects in different clusters be unlike. K-means clustering is a widely used clustering algorithm and finally divides the initial N objects into K clusters. The determination of cluster number is a critical step to carry out clustering and has great influence on clustering results. The standard to weigh a proper K is to minimize the inner-cluster dissimilarity and maximize the among-cluster dissimilarity.

Currently, a widely used rule is that maximized cluster number must not be bigger than square of sample number in data set [7]. The disadvantage of K-means clustering is that it is much sensitive to outliers. Therefore, outliers must be deleted before carrying out clustering. This paper takes K-means as the second stage of clustering which aims to form the objects into required number of clusters. And it utilizes SPSS 12.0 as tool to carry out K-means clustering.

C. Customer Value Analysis and Customer Behavior Analysis

Customer segmentation is to divide current customers in company into different groups in terms of certain standard. This concept is proposed by Wendellr. Smith in 1956. The difference between customer value segmentation and customer behavior segmentation consists in the selection of segmentation dimension.

Customer value segmentation usually takes three dimensions as segmentation dimensions that are customer current value, customer potential value and customer royalty. Primary methods of customer behavior segmentation is RFM analysis, which uses recency, frequency and monetary as dimensions of customer behavior segmentation. Customer value segmentation is a segmentation method commonly accepted by scholars and entrepreneurs. There are a lot of models and algorithms in this field. By contrast, applications of customer behavior segmentation are less, especially in the field of data mining.

III. TWO-STAGE CLUSTERING ANALYSIS

A. The Two-Stage Clustering Algorithm Based on SOM

This paper utilizes a two stage clustering algorithm, which integrates self-organizing feature maps and K-means algorithm. In this paper, we firstly uses SOM network to map the multiple dimension data into a two dimension plot, thus the dimension of data is lowered and the first stage of clustering is performed. Then, merge the clusters in the two dimension space by way of K-means algorithm in order to reach required number of clusters. After that, we suppose a K value. Based on above clustering course, several clusters are obtained finally, which represent different types of customers, respectively. Through analyzing the indexes of customers in each cluster, the customers are all-around understood and it is helpful to find a final target group of customers correctly. The detailed steps are presented in the following.

The first step divides the 1000 customer records into several groups using Matlab 7.1, performing the first stage of SOM clustering. The second step is to calculate the average value of data in each column. And these columns of data are obtained from previous step. In the end of second step, a new set of average values becomes formed. Based on the theory of K-means clustering, we use these average values to represent the original 1000 data and takes them as the starting points of next-stage. In the third step, we use SPSS 12.0 as tool to cluster the points formed in the previous stage by way of K-means algorithm and obtains five clusters in the end. One cluster represents a type of clients. After analysis of these five customer feature indexes, the features of different groups of customers are described.

B. Data Preparation and Selection of Clustering Index

Data were obtained from one security company in Shanghai for the purpose of customer segmentation. There are 1000 records filtered from original database. For market segmentation, we need to select the primary variables to identify customers' value and behavior. Firstly, we choose six indexes as segmentation variables which are contribution, activity, reck, risk, invest return and invest cycle. Secondly, the six indexes are classified into three types, representing customer value, customer risk and customer behavior respectively, as shown in Table I. Therein, the value of frequency is null in database, so we delete this index from the table when doing computational analysis. For value analysis, we choose contribution and activity as segmentation variables. Contribution is commission plus differ. Activity is calculated according to the division of daily match amount and daily capital amount. Customer risk analysis responses to two variables that are customer reck-level and customer risk-level. Here, risk-level is position risk, and reck-level is denoted by the formula, $reck = (1 - marketvalue / \cos t) \times 100\%$. Therein, position risk is a dynamic value. Meanwhile, the implications of behavioral variables are explained as in Table II.

The purpose of data preparation is to integrate, select and transform the data from one or more databases into the data required for the proposed methodology. In this paper, we firstly delete the null and fill the remaining blank with average value. Due to part of the reck value is negative, so plus an absolute minimum to each value, then the data are all transformed to be positive. After that, put all the data into an interval between 0 and 1, thus it can satisfy the requirement of the adopted algorithm. The formula (1) used to calculate the index P_k is as following. Where, k=1, 2, …, q.

$$P_k = \frac{p_k}{\|p_k\|} = \frac{\left(p_1^k, p_2^k, \ldots, p_3^k\right)}{\left[\left(p_1^k\right)^2 + \left(p_2^k\right)^2 + \ldots + \left(p_3^k\right)^2\right]^{\frac{1}{2}}} \quad (1)$$

TABLE I DATA SELECTION

Index classification	Indexes selection
Customer value analysis	Contribute level
	Activity level
Customer risk analysis	Reck level
	Risk level
Customer behavior analysis	Recency: null
	Monetary: Invest return
	Frequency: Invest cycle

TABLE II IMPLICATIONS OF RFM IN APPLICATION

	Monetary	Frequency	Recency
RFM analysis	The aggregated expenses of one consumer over a certain period	The times a customer purchases over a certain period	The time interval from last purchasing to now
RFM applied in security industry	Capital: the quantity a client invests in one year	Invest cycle: the days a client invests in one year	The time interval from open date to close date for one client.(invalid)

C. *Computer Implementation of Clustering Based on SOM*

The first step is to put 1000 data into an interval between 0 and 1. Using Matlab 7.1, the 1000 customer records are derived from 49 neurons. The neurons locate in a two dimension plot, and each neuron contains several dozens of customer records. Then, we calculate the averages of the records in each neuron respectively, and gain 49 average values for each index. Finally, the 49 average values are taken as the initial points for the next stage of clustering. In the course of data preprocessing for next stage clustering, there is 5‰ error due to the fact of manual calculation.

Taking 49 average values calculated in previous stage as initial clustering objects, the 49 objects are clustered once again by means of K-means clustering using tool of SPSS 12.0. The standard to weigh whether a clustering result is correct include cases number in each cluster, distance between two cluster centers and cluster number. When carrying out clustering, the number of cases in each cluster should not be too much and it is much better if the distance between one cluster center and another is larger. Meanwhile, the number of clusters needs to be suitable to usage requirement. Considering the problem in reality, it is anticipated that the objects will spread evenly in each cluster and there is no existence of the fact of little object in one cluster. By the way, the cluster number should not be bigger than seven according to the maximized rule of cluster number in part II B.

TABLE III DISTANCES BETWEEN FINAL CLUSTER CENTERS

Cluster	1	2	3	4	5	6	7
1	0						
2	1.77	0					
3	8.02	7.41	0				
4	1.20	1.53	8.08	0			
5	3.41	3.49	8.21	3.37	0		
6	22.16	22.41	22.84	22.47	22.36	0	
7	7.41	7.93	10.17	8.32	6.55	23.27	0
Cases number	11	16	1	15	4	1	1

TABLE IV FINAL CLUSTERING RESULT

Cluster	Contribution	Activity	Monetary	Frequency	Reck	Risk	Cases
1	-0.22	-0.22	-0.063	1.09	-0.045	0.268	195
2	-0.23	-0.15	-0.067	-0.01	-0.013	0.244	123
3	0.06	-0.01	-0.052	-0.91	-0.111	0.275	183
4	-0.26	0.99	-0.070	0.96	-0.023	0.279	201
5	-0.05	1.01	-0.058	-0.63	-0.030	0.260	164

Firstly, we need to preprocess the 49 sets of data once again. Delete the outliers in data set because K-means algorithm is sensitive to outlier. Suppose K=7, the clustering result is shown in table III.

From table III, we can see the number of cases in each cluster and the distances between one cluster center and another. Majority of the objects occur in cluster 1, 2 and 4. Cluster 3,6 and 7 have only one object. Distribution of the data objects is not even. But the distances between the center of cluster 3, 6 or 7 and other cluster center are obviously larger than that of cluster 1, 2, 4 or 5. Therefore, we delete the objects in cluster3, 6 and 7 as outliers. Then, we carry out clustering once again using the data after deleting outliers. Set K=5, 6 and 7 respectively, clustering results are all similar to that in table III when K is seven. Among three clustering results, the distribution of objects in each cluster remains to be uneven. Whether K is 5, 6 or 7, the phenomenon that one cluster only contains only one object is always existed. Rather, some objects overly converge in one cluster. It means that the existence of outliers renders K-means method ineffective, so we must go on deleting the outliers in data set.

Repeating above steps for four times, we deleted 16 outliers in 49 sets of data including two null outputs, remaining 31 sets of data in the end. There are three kinds of clustering results. The final clustering result we choose is the occasion when k is five. Because there existed the phenomenon of outliers and uneven distribution all the same when cluster number is six or seven. Merging into five clusters is a satisfying result by comparison. According to the result of K-means clustering, we find out the real number of clients existing in each cluster from original data set, listed in table IV.

IV. CUSTOMER SEGMENTATION ANALYSIS

Order the values of all the clusters for every index in table IV, "1"represents a low level, "2" represents a medium level and "3" represents a high level, as shown in table V. Therein, the frequency is divided into four ranks. The "one", "two", "three" and "four" depicts "low", "medium", "relatively high" and "high", respectively. In table V, the values in the right row are calculated through multiplying M and F together. Therein, M and F are positive values turned from monetary and frequency in table IV, the multiplication result depicts clients' behavior.

As shown in table V, contribute-level is commission the clients charged plus differ, representing profit the clients bring to security company. Activity-level is an index that weighs the level one client participates in investment. The higher the activity-level is, the more active the client invests in securities. Here, the index of monetary is denoted by investment-area and it implies the amount of money one client invests. Frequency adopts number of days one client invests during a certain period. The longer the investment period is, the more the client pays attention to securities investment. Two indexes of reck-level and risk-level are to identify the existing risk of one client. Due to the reck values in table IV are all negative and risk values are prone to be closer, so we will ignore the effect of the two risk indexes when doing customer segmentation analysis.

The most obvious feature of clients in cluster one is that frequency of these clients is highest, that is to say, the number of days one client holds stock is very long. This means that this type of clients is willing to spend time and power on securities investment. But the contribution-level and activity-level of this kind of clients are both low, and invest-area is not very large. So, we can say that they will not bring high-profit to enterprise.

The features of clients in cluster one and two are similar. The difference between cluster one and two mainly consists in the value of frequency, the front is higher than the back at a percentage of 1.2.On the contrary, invest-area and investment days are both a little small. Therefore, similar sales strategy can be applied to these two kinds of clients.

The contribute-level of clients in cluster three is very high, who is a type of clients that can bring higher profits to security company. The clients in cluster three belong to profitable type of clients. The activity-level of these clients is mediate and invest-area is relatively larger, but number of days the clients invested in securities is not too long. Therefore, we believe the clients of cluster three are a kind of high profitable clients, and own better capability of returning from investment. So, we should choose this group of clients as company's sales target.

The most outstanding feature of the clients in cluster four is that the frequency and the activity-level of them are both very high, belonging to the type of active engagement. But, such kind of clients has a low contribute-level and invest-area, difficult to bring real income to security company.

The feature of clients in cluster five is that the activity-level is very high and the invest-area is relatively larger. But, the contribute-level and the frequency are both in the middle. This is a kind of clients who own more capital to invest and abundant time to engage in investment. Meanwhile, they are capable of gaining a certain level of profits by way of investing in securities. We can consider turning them into high profitable clients.

RFM analysis is a way to forecast customers' behavior according to the values of recency, monetary and frequency. The larger the R value is, the bigger the possibility to reach new business is. The higher the trading frequency is, the larger the possibility to reach new business is. If the M value is bigger, the customers are more likely to response to company's product and service. Due to the recency value in existing data is ineffective, so we only consider the two indexes of monetary and frequency when analyzing.

TABLE V SEGMENTATION ANALYSIS

Cluster	Contribute	Activity	Monetary	Frequency	M*F
1	1	1	2	4	1.0767
2	1	1	1	3	0.03267
3	3	2	3	1	0.00432
4	1	3	1	4	0.0588
5	2	3	3	2	0.01554

From table V, we know that the frequency and monetary of cluster one is highest and they owns the biggest possibility to purchase in the near future. On the contrary, the frequency and monetary of the clients in cluster three is lowest, who is a type of clients with lowest possibility to purchase in the near future. Therein, purchasing possibility of the clients in cluster four is lower than that of the clients in cluster one. Also, purchasing possibilities of the clients in cluster two and five locate in the back of cluster four.

To sum up, security company should consider both of their resources and advantages to select a group of target clients suitable to the company. If the company is able to response rapidly to client's behavior, the company may consider taking the first kind of clients as target clients. And try their best to reach new trade in short time. If main purpose of the company is to increase profit, then the company should take the third group of clients as target clients, rendering it be consistent with enterprise's goal.

V. CONCLUSION

This article divides the clients into five number of groups in terms of customer value segmentation and customer behavior segmentation method. In the course of clustering, there are many disadvantages. During the second stage of clustering, we utilized K-means algorithm. But the serious disadvantage of K-means algorithm is that it is much sensitive to outliers, thus it may have great effect on final clustering result. The best way to solve this problem is to use K-center method instead of K-means algorithm. Due to the consideration of time, this paper deletes the outliers in data sets. The amount of data decreases from 1000 to 866.The decrease of data at a large extent may delete some meaningful information in client group, having influence on final segmentation result.

REFERENCES

[1] G XIONG Xiong, ZHANG Wei, "Off-site Commercial Banking Regulation Based Self- organizing Feature Map Neural Networks," Theory and Practice of Systems Engineering. Vol 6, pp26-33, 2002.

[2] LIU Lin, YU GuoPin, "Application of Discovering the Potential Customers Based on Self-Organizing Feature Map Neural Network," NanChang University Periodical, Vol 5,No 30,pp507-510, 2006.

[3] LIN Sheng, XIAO Xu, "A method of telecom consumer market segmentation based on the RFM model," Journal of Harbin Institue of Technology, Vol 38,No 5, pp758-760,2006.

[4] XIA WeiLi, WANG QingSong. "Customer Segmentation and Retention Strategy Based on Customer Value," Management Science In China, Vol 19,No 4 ,pp35-38,2006.

[5] LIU YingZi,WU Hao. "A Summarization of Customer Segmentation Methods," Journal of Industrial Engineering/Engineering Management. Vol 20, No 1,pp53-57,2006

[6] LI HongMei, LI ShiYu, LIN WeiQiang. "Sustainable Development Evaluation Using Self-organizing Feature Map Neural Network," ACTA Scientiarum Natulium Universities Sunyasenti.Vol 43,No 6,pp159-162,2004

[7] ZHANG WeiJiao, LIU ChunHuang, LI FangYu. "Method of Quality Evaluation for Clustering," Computer Engineering.Vol 31, No. 20, pp10-12. 2005

[8] Tsai, Chiu. "A purchase-based market segmentation methodology". Expert Systems with Applications. Vol 27, pp265-276,2004.

[9] R.J. Kuo, Y.L. An, H.S. Wang, W.J. Chung. " Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation," Expert Systems with Applications. Vol 30,pp313-324,2006.

[10] Melody Y. Kiang, Michael Y. Hu & Dorothy M. Fisher. "An extended self-organizing map network for market segmentation-a telecommunication example," Decision Support Systems.Vol 42,pp36-47,2006.

[11] R.J. Kuo, L.M. Ho, C.M. Hu. "Integration of self-organizing feature map and K-means algorithm for market segmentation," Computers & Operations Research, Vol 29, pp1475-1493, 2002.