A New Self-Organizing Model and Its Application in Online Trading Customer Classification

Hankun Ye

School of International Trade and Economics, Jiangxi University of Finance and Economics, Nanchang 330013, Jiangxi, China.

Abstract

Correctly and effectively customer classification according to their characteristics and behaviors will be the most important resource for electronic marketing and online trading of network enterprises. Aiming at the shortages of the existing K-means algorithm of data-mining for customer classification, a new online trading customer classification algorithm is advanced based on combination of the K-means Self-Organizing Feature Map (SOM) algorithms. Firstly, based on consumer characteristics and behaviors analysis, the paper designs 21 customer classification indicators including customer characteristics type variables and customer behaviors type variables. Secondly, the limitation of K-means algorithm is analyzed ; Third, SOM & K-means Combination-Based customer Classification algorithms is advanced to overcome the shortage of the K-means classification algorithm and takes advantage of powerful classification ability of the algorithm to classify online trading customer. Finally the experimental results verify that the new algorithm can improve effectiveness and validity of customer classification when used for classifying network trading customers practically.

Keywords

Electronic marketing, Customer classification, K-means algorithm, Consumer characteristics analysis, Customer behaviors analysis.

1. Introduction

Customer relations management is one of the core problems of modern enterprises, whose customer oriented thought requires CRM system to be able to effectively obtain various kinds of information of customers, identify all the relations between the customers and enterprises and understand the transaction relation between customers and enterprises; meanwhile, deeply analyze customers' consuming behavior, find customers' consumption characteristics, providing personalized service for customers, supporting the decisions of enterprises. The three basic problems CRM needs to solve are how to get customers, how to keep customers and how to maximize customer value, among which maximizing customer value is the ultimate purpose, getting customers and keeping customers are both the means for realizing the purpose. The core of analyzing the three problems CRM needs to solve is to classify customers. "Getting Customers" and "Retaining Customers" need to ascertain which customers are attainable, which customers need to be kept, which customers are kept for a long term and which customers are kept for a short term, therefore, customer classification is needed. It is the same case with "Maximizing Customer Value". Due to different values of different customers, "Maximum Customer Value" of different customers should be distinguished. Thus, the core problem of enterprises to correctly implement CRM is to adopt effective method to reasonably classify customers, find customer value, focus on high-value customers with enterprises' limited resources, provide better service for them, keep "High-value" customers for loss prevention; also, establish corresponding customer service system through classification, carry out differential customer service management. Hence, customer classification is becoming a more and more popular research hotspot, also a research difficulty, becoming one of the urgent problems of CRM[1].

2. Summarization of Customer Classification Methods

The widely-used methods of enterprises for customer classification at present are mainly qualitative method and quantitative method. As the qualitative method for customer classification is just to classify all the target customers of enterprises in the macroscopic level, customer classification is carried out according to different value emphasis of different customers. The formation of customer value is simply expressed as: Value = Benefit — Cost. Qualitative classification method classifies customers in a simple way, only offering guidance for customer classification of enterprise in the macroscopic level, unable to provide specific and credible basis for enterprise decisions; furthermore, as there is no strict process of argumentation, the method depends on decider’s subjective inference, there may be certain deviations in the analysis process, easily resulting in faulty decisions. For this reason, to truly provide customer classification information beneficial to enterprises should depend on quantitative technology for customer classification.

Quantitative classification method is to apply quantitative analysis technology to conduct customer classification on the basis of some specific customer variables (credit level of customers, purchasing power of customers, characteristics of demand of customers, etc.). Currently, there are mainly two categories of data mining for quantitative customer classification research, which are traditional statistical method and non-statistical method. The former mainly includes cluster analysis, Bayesian Classification, factor analysis method, etc.; this statistics-based method is unable to process a great deal of sophisticated customer data, and there are some problems on the accuracy of customer classification results, so to fundamentally solve the problem of customer classification needs to rely on non-statistical customer classification method, which mainly includes neural network, fuzzy set method, association rules, genetic algorithm, etc. The classification technology based on neural network is combined with certain information technology, which is a kind of mathematical method applicable to complex variables and multi influencing factors calculation, so it is more effective in solving complex customer classification problems with better classification accuracy, however, the convergence problem of the function itself greatly limits its application value in specific project practice. Secondly, classification is mainly based on such mathematical methods as fuzzy clustering, rough set, association rules, etc., although these methods offer classification reason explanation in a relatively clear way with better classification results under the circumstances of satisfactory data conditions, the modeling process needs to provide specific mathematical equations. As a result, these methods are limited by data conditions in specific application, always having problems like insufficient classification accuracy or poor "robustness", limiting the application in customer classification. Due to lots of influencing factors related to customer classification, more often than not, the complicated relations are difficult to be expressed in mathematical equations[1-6].

K_means algorithm is one of the best information clustering methods in data mining which can extract and find new knowledge and. But it is found that using K-means algorithm to process the data of isolated points has great limitations[7-9]. The paper tries to combines the SOM & K-means algorithms to overcome the limitations of the algorithm and takes advantage of powerful classification ability of the algorithm to classify online trading customer.

3. Selection of Customer Classification Indicators

The selection of reasonable classification variables is the basis of correct and effective customer classification, namely establishing scientific and reasonable classification indicators system. In view of the nature of trading and own characteristics of online trading, this Paper adopts customer characteristics type variable and customer behaviors type variable in the specific selection of customer classification variables[2].

(1) Selection of Customer Characteristics Type Variable

Customer characteristics type variable is mainly used for getting the information of customers’ basic attributes. Such variable indicators as geographical position, age, sex, income of individual customer play a key role in determining the members of some market segment. This kind of variables mainly comes from customers' registration information and customers' basic information collected from the management system of banks, the contents of which mostly indicate the static data of customers' basic attributes, the advantage of which is that most of the contents of variables are easy to collect. But some of the basic customer-described contents of variables are lack of differences at times.

Based on analyzing and summarizing existing literatures, the customer characteristics type variables designed in this Paper include: Customer No., Post Code, Date of Birth, Sex, Educational Background, Occupation, Monthly Income, Time of First Website Browsing, and Marital Status.

(2) Selection of Customer Behaviors Type Variables

Customer behaviors type variables mainly indicate a series of variable indicators related to customer transacting behavior and relation with banks, which are used to define the orientation which enterprises should strive for in some market segment, and are the key factors for ascertaining target market. Customer behaviors type variables include the records of customers buying services or products, records of customer service or production consumption, contact records between customers and enterprises, as well as customers' consuming behaviors, preferences, life style, and other relevant information.

Based on analyzing and summarizing existing literatures, the customer behaviors type variables designed in this paper include Monthly Frequency of Website Login, Monthly Website Staying Time, Monthly Times of Purchasing, Monthly Amount of Purchasing, Type of Consumer Products Purchased, Times of Service Feedback, Service Satisfaction, Customer Profitability, Customer Profit, Repeat Purchases, Recommended Number of Customers, Purchasing Growth Rate.

4. Customer Classification Algorithm Based on K-means

4.1 K-means algorithm principle

Steps for K-means clustering algorithm are[7-9] (see Fig.1):

(1) Select objects as the initial cluster seeds on principle;

(2) Repeat (3) and (4) until no change in each cluster;

(3) Reassign each object to the most similar cluster in terms of the value of the cluster seeds;

(4) Update the cluster seeds, i.e., recompute the mean value of the object in each cluster, and take the mean value points of the objects as new cluster seeds.

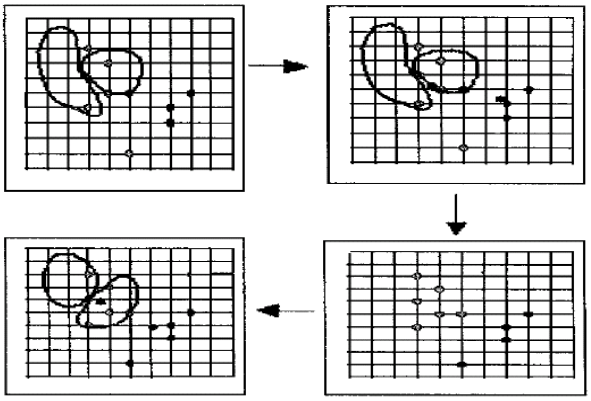

Fig 1. K-means Algorithm Procedures

4.2 Limitation of K-means algorithm

When K-means algorithm is used to cluster data, the stability of the clustering results is still not good enough, sometimes, the clustering effect is very good (when the data distribution is convex-shaped or spherical), while sometimes, the clustering results have obvious deviation and errors, which lies in the data analysis. It is unavoidable for the clustered data to have isolated points, referring to the situation that a few data deviate from the high-dense data intensive zone. The clustering mean point (geometrical central point of all data in the category) is used as a new clustering seed for the K-means clustering calculation to carry out the next turn of clustering calculation, while under such a situation, the new clustering seed might deviate from the true data intensive zone and further cause the deviation of the clustering results. Therefore, it is found that using K-means algorithm to process the data of isolated points has a great limitation.

4.3 Derivation of Classification Algorithm based on SOM and K-means

4.3.1 Self-Organizing Model (SOM) and Algorithm

SOM is the Self-Organizing feature Map proposed by Kohonen. Kohonen believed that, a nervous network's outer input receiving model was to divide the nervous network into different regions, these regions have different corresponding features to the input model, and such an input process is finished automatically[2]. The connecting weights of various neurons have a certain distribution, the nearest neurons excite each other, while the distant neurons inhibit each other, and the more distant neurons have a relatively weak inter-inhibition effect. In a word, Self-Organizing feature Map method is a teacher-free clustering method, and compared with the traditional model clustering methods, its former cluster centers could be mapped on a contour or plane, with the topological structure maintained original. Competitive Study refers to that various neurons at the same neuron level compete with each other and the winner neurons modify the connecting weights related with them. Competitive Study is a kind of study without supervision, and only some studying samples are required providing for the network during the study process, rather than the ideal output. The network finishes the self-organization according to the input samples and partitions them into the corresponding model categories. Due to no demand for the presentence of the ideal output samples, the supervised model classification method is promoted. Fig. 2 represents the structural model of a competitive network.

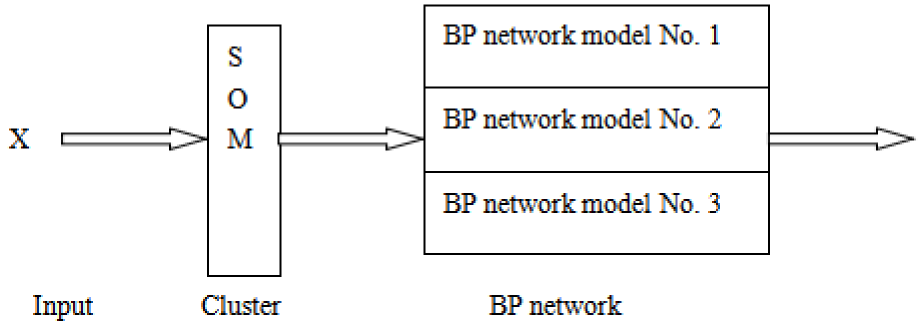

Fig 2 Competitive Network Topology

The competitive network consists of two levels, respectively input level and competition level. The input level of the competitive study network is used to receive the input samples, while the competition level is used to finish the classification of the input samples. The neurons at these two levels are fully interconnected, that is, each neuron at one level is connected separately with all neurons at the other level. At the competition level, the neurons compete with each other and eventually only one or several neuron activities adapt(s) to the current input samples. The neurons winning in the competition represent the classification model of the current input samples, each input node and each output node are connected through the connecting weight w, and the connecting weight $w_{i,j}$ of the node j at the output level to the node $x_i (i = 1,2...N)$ at the input level is the cluster center of the j category. The model studying sample is composed by the actually measured samples of N classification indications. Proposed these study samples are all the points in N-dimensional space, it is obvious that the samples in the same categories or having some similar features are relatively close to each other in N-dimensional space. These relatively close samples compose a category and form a cluster in N-dimensional space. If the input samples belong to various categories, N-dimensional space will have a feature of several-cluster distribution. Each cluster represents one category, and the center of the cluster is the cluster center. The distance between the samples in the same category and the cluster center of this category is smaller than that between these samples and the cluster center of another category. Eyclid distance could be used to represent the distance, see Formula (1). In Formula (1), D_j represents Eyclid distance, x_i represents classification indication, w_{ij} represents the cluster center of the j category, and k represent iteration times.

$$D_{j(k)} = \sum_{i=1}^{N} (x_i - w_{ij(k)})^2 \tag{1}$$

The clustering combination algorithm of SOM and K-means (SOMK) is described as follows: (1) Firstly perform SOM algorithm, input the data objects to be clustered into SOM network, and output a group of weights through network training. The training times in this phase could be reduced, and it is unnecessary to completely converge SOM, for example, to finish 300 cycles of SOM. (2) Secondly initialize K-means algorithm with the weights obtained through SOM's clustering results as the initial cluster center. Such a clustering combination algorithm not only maintains the self-organizing features of SOM network, but also makes up the disadvantages of SOM network's overlong convergence duration and the bad clustering effect caused by the inadequate selection of K-means algorithm's initial cluster center.

4.3.2 SOM & K-means Combination-Based customer Classification

In the algorithm, the commonly used vector spatial model representation method is used to represent the customer information, that is, use the characteristic items and their weights to represent the customer information. The vector $d = (w_1, w_2, ...w_m)$ represents the characteristic items and corresponding weights of the customer d, of which m represents the numbers of all items in the customer set, $w_i (i = 1,2,...m)$ represents the weight of the item t_i in the customer d. To obtain the characteristic items, it is needed to firstly delete the unusable words from the customer's characteristic set and then simplify the characteristic items according to TF-DF rules. In order to avoid the situation that an item obtains a large weight only due to its high appearance frequency (a tf larger value) in one customer, Formula (2) is used to calculate the weights. In Formula (2), w_{ij} represents the weight of the j item in the i customer, and $coef_{ij}$ could be obtained through Formula (3). While in Formula (3), tf_{ij} represents the appearance frequency of the j item in the i customer.

$$w_{ij} = (coef_{ij}) \cdot (\log N - \log df_i) \tag{2}$$

$$coef_{ij} = \left\{ \begin{matrix} 1 & if & tf_{ij} = 1 \\ 1.5 & if & 1 \prec tf_{ij} \le 5 \\ 2 & if & 5 \prec tf_{ij} \le 10 \\ 2.5 & if & tf_{ij} \succ 10 \end{matrix} \right\} \tag{3}$$

Thus, a group of vectors representing the customer set is obtained, that is, the model sets to be classified. The distance between the customer vectors is represented by adopting the cosine distance, defined as Formula (4).

$$d(doc_i, doc_j) = 1 - sim(doc_i, doc_j) \tag{4}$$

In Formula (4), $sim(doc_i, doc_j)$ could be calculated through Formula (5). $sim(doc_i, doc_j)$ is called as cosine similar function, and the bigger its value is, the more similar the customer i and j are, thus, the smaller the cosine distance between these two customer are.

$$sim(doc_i, doc_j) = \frac{\sum_{k=1}^{m}(w_{ik} \cdot w_{jk})}{\sqrt{\sum_{k=1}^{m}(w_{ik})^2 \cdot \sum_{k=1}^{m}(w_{jk})^2}} \tag{5}$$

The main process to cluster the customers by using the clustering combination algorithm of SOM and K-means (SOMK) is below: firstly apply the commonly used vector spatial model to represent the customer information, delete the unusable words with the conventional method, and simplify the characteristic items according to TF-DF rules to obtain the customer's characteristic set, secondly calculate the weights of various characteristic items and express the customer in the form of vectors, thirdly input the vectors of the customer set for SOM algorithm and cluster the customer s through SOM (the number of SOM network's input nodes equals to the dimension of the customer vectors, while the number of SOM network's output nodes equals to the number of the customer s' categories) to obtain a group of output weights, and finally initialize K-means algorithm's cluster centers with this group of weights and implement K-means algorithm to cluster the customer sets.

5. Experimental Verification

5.1 Object of Experimental Verification

The instance data of the experiment conduct empirical research on the customer data of the B2C transaction of certain enterprise website of the recent three years (totaling data of 41351 customers, 21 attributes in the data table are listed in the third part of the paper including customer characteristics type variables and customer behaviors type variables), making statistics on attribute values like annual transaction frequency, total amount, product cost, etc. of certain customer according to customer transaction records in information base, forming an information table (among which the decision attribute set D is null) [4].

5.2 Process of Experimental Verification

The process of the experimental verification can be listed as follows[5].

First, what is to be processed during the classification is the numeric data, so the numeric coding on character data should be conducted first;

Second, if the value number of certain attribute is equal to sample number, it means that it has little effect on classification, hence, remove such attribute first. Three attributes as Customer No., Post Code and Date of Birth are removed in this case.

Third, establish training sample set according to domain (prior) knowledge. Times of purchasing and total amount of purchasing of each customer are two major factors of customer classification (this is the prior knowledge of domain), so select 400 pieces of typical data among all the customers to form training sample set. And divide them into five types as Gold Customers, Silver Customers, Copper Customers, General Customers and Negligible Customers according to ABC management theory.

Fourth, use the customer classification algorithm above-mentioned, and the customer classification results can be expressed in Table 1. In the specific algorithm realization, this Paper simultaneously realizes ordinary K_means algorithm and customer classification algorithm based on BP neural network. The performance comparison of these three algorithms can be expressed in Table 2.

Table 1 Customer Classification Result of Some Website

Customer Type	Number of Customers	Percentage%	Profit Contribution Proportion
Gold Customers	2866	6.93	53.1
Silver Customers	5977	14.45	31.1
Copper Customers	10201	24.67	12.3
General Customers	13787	33.34	5.4
Negligible Customers	8520	20.60	-1.9
Total	41351	100.00	100.00

We can see from Table 1 that in the autonomous learning of algorithm of this Paper, such five factors as the educational background, income, occupation, times of purchasing, and total amount of purchasing of customers have a relatively great influence on customer classification. Through the classification result in Table 1, it can be seen that Gold Customers take up 6.89% of the total number of customers, while the profit takes up 52.1% of the total profit. These customers play a significant role in the existence and development of enterprises. However, the negligible customers account for 17.7%, who not only do not bring profit to enterprises, but also make enterprise lose money. These customers should be either further cultivated or eliminated according to the actual situation.

Table 2 Classification Performance Comparison of Each Algorithm

Algorithm	Algorithm in This Paper	Ordinary K-means Algorithm	BP Neural Network Algorithm
Accuracy Rate	99.6%	86.39%	93.12%
E Value	105.66	160.73	117.84

We can see from Table 2 that the cluster accuracy rate of algorithm in this paper is the highest, reaching 99.7 %, obviously higher than ordinary K-means algorithm and BP Neural Network algorithm; the square errors and E values on customer classification of three algorithms are 104.33, 159.81 and 119.96 respectively. The smaller the E value is, the smaller the possibility of wrong classification is. Thus it can be seen that the square error and E value of the algorithm in this paper during the classification are far more less than ordinary K-means algorithm[4] and BP Neural Network algorithm[6]. Therefore, it shows that the improvement on K-means clustering algorithm in this paper turns out to be a success, with reasonable classification results.

6. Conclusion

Customer relations management of online trading is still developing. But to correctly and effectively classify online trading customers is the critical issue for reforming network marketing mode, improving customer management and service level and enhancing competitiveness of network enterprises[5]. On account of the shortcomings of the typical K_means clustering algorithm in data mining, this Paper puts forward several improvement measures, and applies them into the classification of online trading customers. Simulation results indicate that the improved online trading customer classification has higher accuracy rate on customer classification and more reasonable classification results.

References

[1] Liu Zhaohua. Study on Model of Customer Classification Based on the Customer Value, A Dissertation of Huazhong University of Science and Technology,2018.

[2] Zhou Huan, Study of Classifying Customers Method in CRM, Computer Engineering and Design, 2020,Vol 29, No.3, pp.659-661.

[3] Deng Weibing, Wang Yan, B2C Customer Classification Algorithm Based on Based on 3DM, Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019,Vol 21, No.4, pp.568-572.

[4] Guan Yunhong, Application of Improved K-means Algorithm in Telecom Customer Segmentation, Computer Simulation, 2021, Vol 28, No.8, pp.138-140.

[5] Qu Xiaoning, Application of K-means Based on Commercial Bank Customer Subdivision, Computer Simulation, 2018, Vol 28, No.6, pp.357-360.

[6] Yang Benzhao, Tian Gen, Research on Customer Value Classification based on BP Neural Network Algorithm, Science and Technology Management Research, 2019, Vol 23, No.12, pp.168-170.

[7] Bradley P S, Managasarian L. k-plane Clustering.Journal of Global Optimization,2020,16(1)23-32.

[8] Tang Yong, Rong Qiusheng. An Implementation of Clustering Algorithm Based on K-means. Journal of Hubei Institute for Nationalities, 2018,Vol.22 No.1,pp.69-71.

[9] Zhang Y.F., Mao J. L., An improved K-means Algorithm, Computer Application, vol.23. no.8, pp. 31-33,2019.