An Improved K-means Algorithm for Online Trading Customer Classification

Hankun Ye

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

407706483@qq.com

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, this paper advances a new customer classification algorithm through improving the existing K-means algorithm. First the paper designs 21 customer classification indicators based on consumer characteristics and behaviors analysis, including customer characteristics type variables and customer behaviors type variables; Second, limitation of K-means algorithm is analyzed; Then corresponding improvements for K-means algorithm are advanced including improvement of K-means algorithm principle, improvement of initial classification centers selection and improvement of the flow of K-means Algorithm. 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].

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. 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 present some improvements to overcome the limitations of the algorithm and takes advantage of powerful classification ability of the algorithm to classify online trading customer.

2. Customer Classification Algorithm Based on K-means

2.1 K-means algorithm principle

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

(1) Select n 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.

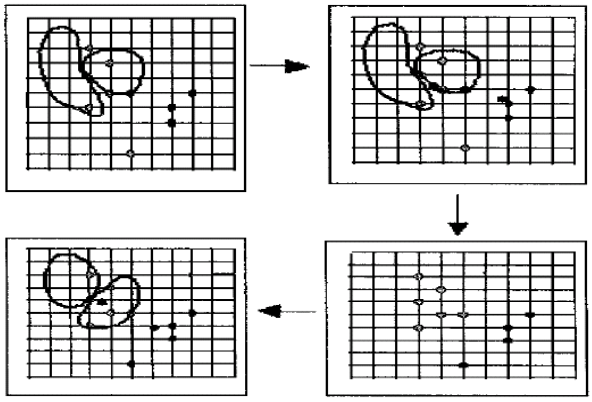


Fig. 1 K-means Algorithm Procedures

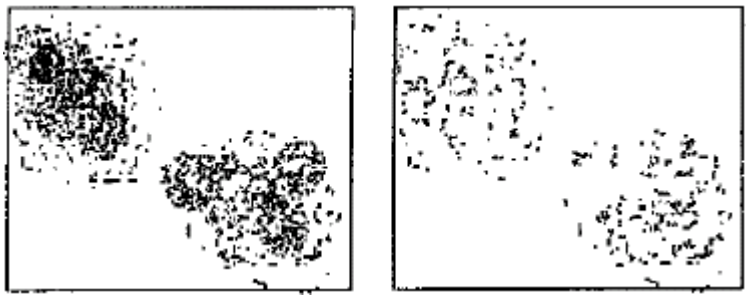
2.2 Limitation of K-means algorithm

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.

2.3 Improvement of K-means Algorithm Principle

The original K-means algorithm selects k points as initial cluster centers, and then the iterative operation begins. Different selection of initial point can achieve different clustering result. For the reduction of the clustering result's dependence on the initial value and the improvement of the clustering stability, better initial cluster centers can be achieved by the search algorithm of the cluster center[7-9].

In the search process, the sampled data tries to be undistorted and is able to reflect the original data distribution through the random data sampling, as shown inFig. 2. , among which, (a) original data distribution, (b) sampled data distribution.



（a）Original data distribution （b）Sampled data distribution

Fig. 2 Comparison of data distribution before and after sampling

The sampled data and the original data are clustered by K-means algorithm respectively, and little change of final cluster centers is found. Therefore, the sampling method is suitable for the selection of the initial cluster centers. In order to minimize the sampling effects on the selection of the initial cluster centers, the sample set extracted each time should be able to be loaded into the memory, and do best to make the sum of the sample sets extracted J times equivalent to the original data set. Each extracted sample data is clustered by K-means algorithm and one group of cluster center is produced respectively; the samplings J times produce J groups of the cluster centers in all, and then the

comparison of clustering criterion function values is conducted for J groups of cluster centers, and one group of minimum cluster center in J_c value is given as the optimal initial cluster center.

For the protection against segmenting large clusters into small clusters by the criterion function, the algorithm takes the initial cluster as K' and $K' \succ K$. According to the quality requirements and the time, K' value does the compromise selection. Larger K' value is able to expand the solution search scope, and the phenomenon of no initial value near certain extremal vertexes is diminished. The utilization of the searched initial cluster center clusters the original data by another K-means algorithm and outputs K' cluster centers, and then the reduction of each cluster quantity to the specified K value is studied.

2.4 Improvement of Selection of Initial Classification Centers

Euclidean Distance has a very intuitive significance for clustering, thus, it is used in the paper to express the distance between the sample points, and the distance between sample $X=(x_1,x_2,...,x_n)$ and $Y=(y_1,y_2,...,y_n)$ is calculated according to Formula 1.

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_n-y_n)^2} \quad (1)$$

The distance between a sample point and a sample set is defined as the nearest distance between the sample point and all sample points in the sample set. The distance of a sample point X and a sample set V is defied as Formula 2[7].

$$d(X,V)=\min(d(X,Y),Y\in V) \quad (2)$$

Proposed that the sample set U has n samples, clustered as category k, and the initial value of m is 1, and the improved algorithm is described as follows.

(1) The distance $d(X,Y)$ between any two samples is calculated, the nearest two points in the set U are found and compose the set $Am(1\le m\le k)$, and then these two points are deleted from the set U.

(2) The point is found that is nearest to the set Am, added into the set Am and deleted from the set U.

(3) Step 2 is repeated until the number of the sample points in the set is not less than $\alpha_{n/k}(0\langle\alpha\le 1)$, and the value of α varies from the experimental data. If the value of α is too small, it might have several initial cluster center points obtained in the same region, while if too big, it might have the initial cluster centers deviating from the intensive zone, thus, based on the experimental situation, it is suitable to value it as 0.75.

(4) If $m\langle k$, $m=m+1$, two points nearest to the set U are found to form a new set $Am(1\le m\le k)$ and are deleted from the set U, then the Step 2 is repeated.

(5) The sample points in k sets formed eventually are calculated for their means, so as to form k initial cluster centers.

(6) Based on k initial cluster centers, K-means clustering algorithm is used to form the final clustering.

2.5 Improvement of the Algorithm Flow

The general K-means algorithm is a gradient ascent iteration algorithm, each time of iteration could cause the corresponding increase of the target function values, and the iteration might be ended in the limited steps. However, such an algorithm also has some disadvantages, for example, the algorithm is easily trapped in the local maximum solution and such a solution depends on the selection of initial partition. Therefore, the means algorithm is used as local searching process to be inlaid in the local search structure of the iteration in order to obtain better text clustering results through the relationship between balancing the reinforcement of the local search and extending the searching range[8].

In the text clustering problems, D neighborhood of a partition refers to the partition obtained through randomly selecting D different texts in a certain partition and redistributing them into other clusters.

In other words, the neighborhood of the current partition means the partition obtained through randomly selecting one text and redistributing it into other cluster. The calculation flow of the K-means-based iteration clustering algorithm of local searching texts is displayed in the following.

Input: the number k of the results' clusters, containing the data set of N texts.

Output: k clusters, ensuring that the texts in all clusters are similar or correlated.

Step 1, Randomly select an initial partition $P_k = \{C_1, C_2, ...C_k\}$ and calculate the corresponding concept vector $c(C_i), i = 1,2,...k$, then initialize the current maximum target function value $fopti$ and determine the ending conditions of the algorithm, the parameter value $\varepsilon(\varepsilon \succ 0)$ receiving the conditions and the maximum iterating times n that the target function value is not improved any more.

Step 2, Repeat.

Step 3, Perform the local search on P_k with the means text clustering algorithm to obtain a local maximum target function value $fopt$ and its corresponding partition P_k^*.

Step 4, If $fopt \succ fopti$, the current best partition is $P_k^{'} = P_k^*$, $fopti = fopt$, the current partition is not improved any more, and the iteration times $t := 0$.

Step 5, Repeat.

Step 6, Randomly generate a text $x_i (i = 1,2,...N)$ and repeat the following processes:

(1) If x_i is beyond the tabu list, it will be redistributed into other cluster to calculate the increase Δf of the target function value and the times of iteration without improvement is $t : t = t + 1$; while if x_i is in the tabu list, Step 6 will be repeated.

(2) If $\Delta f \succ \varepsilon$, P_k is the partition of redistribution, the target function value is $f_{opt} = f_{opt} + \Delta f$, x_i is added into the tabu list, and the tabu length of other tabu objects is deducted 1.

(3) If $fopt \succ fopti$, $fopti = fopt$, $P_k^{'} = P_k$, and the times of iteration is $t := 0$.

(4) If x_i is tested throughout all clusters and the times of iteration without improvement is $t \prec n$, Step 6 will be repeated.

Step 7, "Until $t = n$" means there is no improved partition generated in the successive n times of iteration.

Step 8, Randomly select several texts from P_k and redistribute them into other clusters to obtain the new partition P_k.

Step 9, Until the ending conditions are met.

3. Experimental Verification

3.1 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].

3.2 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	2849	6.89	52.1
Silver Customers	5921	14.32	30.1
Copper Customers	10193	23.65	13.1
General Customers	13751	36.44	6.1
Negligible Customers	7319	17.70	-1.4
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.

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.

Table 2 Classification Performance Comparison of Each Algorithm

Algorithm	Algorithm in This Paper	Ordinary K-means Algorithm	BP Neural Network Algorithm
Accuracy Rate	99.7 %	88.47%	94%
E Value	104.33	159.81	119.96

4. 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,2015).
[2] Zhou Huan:Study of Classifying Customers Method in CRM, Computer Engineering and Design,Vol 29(2013), No.3, p.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, Vol 21(2015) No.4, p.568-572.
[4] Guan Yunhong:Application of Improved K-means Algorithm in Telecom Customer Segmentation, Computer Simulation, Vol 28(2016) No.8, p.138-140.
[5] Qu Xiaoning:Application of K-means Based on Commercial Bank Customer Subdivision, Computer Simulation, Vol 28(2016), No.6, p.357-360.
[6] Yang Benzhao, Tian Gen:Research on Customer Value Classification based on BP Neural Network Algorithm,Science and Technology Management Research, Vol 23(2017), No.12, p.168-170.
[7] Bradley P S:Managasarian L. k-plane Clustering, Journal of Global Optimization(2015), 16 (1)23-32
[8] Tang Yong and Rong Qiusheng:An Implementation of Clustering Algorithm Based on K-means,Journal of Hubei Institute for Nationalities ,Vol.22(2014) No.1,p.69-71
[9] Y.F. Zhang and Mao J. L.:An improved K-means Algorithm, Computer Application, vol.23. (2016) No.8, p. 31-33.