International Journal of Hybrid Information Technology
Vol.8, No.10 (2015), pp.311-318
http://dx.doi.org/10.14257/ijhit.2015.8.10.28

Research of Customer Loyalty Based on the Improved K-means Algorithm

Li Min[1], Liu Wei[1] and Chen Ming[1]

[1]*School of Computer and Information Engineering, Harbin University of Commerce*
Harbin, 150028, China
E-mail:lm81612@163.com

Abstract

During the iteration process, the traditional K-means algorithm is easily fall into local optimal solution. In order to solve this problem, this paper proposed an improved K-means algorithm, and used the method of maximum distance equal division to select the initial cluster centers. We preset k cluster centers, and avoid it falling into local optimal solution. Apply this improved algorithm into e-commerce customer loyalty analysis, this paper put forward a customer loyalty analysis model using the parameters of shopping recency, shopping frequency, shopping monetary, customer satisfaction and customer attention, and used the improved K-means algorithm to analyze the RFMSA customer loyalty model. The studies show that the improved K-means algorithm and RFMSA model can effectively divide the loyalty of the e-commerce customer, it also can fully reflect the customer's current value and potential value-added ability, and provide the basis that the e-commercial enterprises can adopt different marketing strategies for different target customers.

Keywords: *customer loyalty, customer satisfaction; customer attention, K-means algorithm, initial cluster centers*

1. Introduction

With the vigorous development of the Internet and e-commerce, the e-commercial enterprises face more fierce market competition, and the competition for customers becomes the decisive factors of the survival and development of the enterprises. How to strengthen the loyalty of the old customer and potential customer satisfaction is the new problems which are need to be solved by e-commerce. At the same time, divide the customer loyalty using the data mining technology is also a hot research topic in data mining application field.

In order to ensure the division accuracy of customer loyalty, we need to select the appropriate classification method. In the traditional customer classification method, RFM model was widely applied. The three important behavior index parameters of the customer of RFM model are shopping recency (R), shopping frequency (F) and shopping monetary (M). In recent years, some experts and scholars at home and abroad do continuous studies on customer classification method. Paper [1] introduced total profit property on the basis of RFM model, and proposed RFP model. Paper [2] increased the analysis on value matrix of the customer, and created AFH model. However, both RFP model and AFH model didn't consider the Customer Satisfaction and Customer Attention. In e-commerce, Customer Satisfaction and Customer Attention are the significant elements that can affect the customer loyalty. In terms of classification method, we usually use the simple and effective K-means algorithm. K-means algorithm needs to iterate data repeatedly in the process of processing data, it is easy to fall into

ISSN: 1738-9968 IJHIT

local optimal solution when we select the initial cluster centers, and the clustering results easily have large fluctuation. In actual application, the clustering effect is not ideal. Therefore, many scholars improved its random selection method of initial clustering centers from different perspective. Paper [7] proposed data segmentation method to determine the initial cluster centers. Paper [8] proposed distance estimation method to determine the initial cluster centers. Paper [9] put forward a method based on minimum spanning tree to divide the data points into k initial data set, and calculate the initial cluster centers. Because the traditional K-means algorithm is easily fall into local optimal solution when we select the initial cluster centers. In order to solve this problem, this paper proposed an optimized K-means algorithm using the segmentation method. We use the Euclidean distance to find out the farthest two points in the data set, and calculate their distance in each dimension. According to the k value, we divide it into k segments, and use the boundary points of each dimension as the initial cluster centers. Finally, we use the RFMSA model and improved K-means algorithm to analyze the current customer's loyalty of e-commerce website, and give the corresponding marketing strategies for different customer loyalty.

2. Build RFMSA Customer Loyalty Model

Customer loyalty refers to that the customer has a good feeling for a particular product or service, forms an "independent" preference, and a tendency to repeatedly purchase. According to the analysis of research status of customer loyalty at home and abroad, this paper will do the research based on the most representative RFM model.

In RFM model, R (Recency) refers to the customer purchasing time interval from last time to now. The smaller the time interval is, the greater the likelihood to buy again. F (Frequency) refers to the number of purchasing within a specific period of time. The more the number of purchasing is, the higher the customer loyalty will be. M (Monetary) refers to that the customer expends the total money within a certain period of time. The greater the M is, the higher the customer value will be. The drawback of the RFM model is that after classification, there are too many groups of customer, and there are linear defects between F and M. Due to the randomness of customer shopping in e-commerce industry, the customer can easily read other customer's satisfaction evaluation of some commodities. Therefore, the method of using RFM model to analyze the customer loyalty has some limitations. On the other hand, although some customers didn't buy commodities in this e-commerce enterprise, they may collect the commodities and add them to the shopping cart to focus these commodities, so they have the potential to buy the commodities. Therefore, this paper added the S (Satisfaction) and A (Attention) as the standard of measuring customer loyalty on the basis of RFM model, and put forward a customer loyalty analysis model based on RFMSA for e-commerce enterprises.

3. Improved K-means Algorithm

3.1 K-means Algorithm

K-means algorithm is a kind of indirect cluster method based on similarity measurement among samples. According to the principle of similar inside the class and dissimilar outside the class, we can put the data set Y into k classes (j=1,2,…,k). It belongs to an unsupervised learning method. Select the Euclidean distance as the judgment of similarity and distance, the closer the distance is, the higher the similarity will be. For a given sample set containing nth d dimensional data points , we have $Y=\{y_i|i=1,\ldots,n\}$, where $y_i \in R^d$, and clustering number is k. K-means algorithm can find the map $\Phi : Y \rightarrow \{1,\ldots,k\}$, guarantee each sample y_i can map on class $A_j(1 \leq j \leq k)$, and make sure that the criterion function J is minimum:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{M_j} \|y_i - z_j\|^2$$

Where M_j represents the sampling number of class j, y_i represents the sample in class j, z_j is the cluster center of class j. The basic procedure of K-means algorithm is shown in Figure 1.

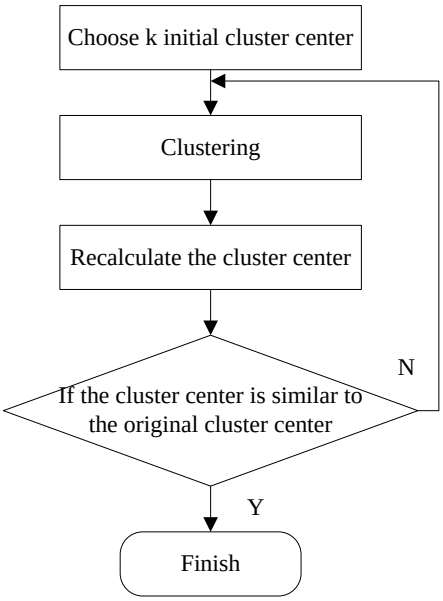

Figure 1. The Flow Chart of K-means Algorithm

(1) Select k initial cluster centers randomly from n data objects.

(2) Calculate the minimum average distance of each object and its clustering object, and distribute them to k cluster centers.

(3) Recalculate the cluster centers of each class.

(4) If the new cluster center is equal to or less than the original preset threshold, then the calculation is over. Otherwise, repeat step (2).

3.2 The Improved Selection Method of Initial Cluster Center

This paper proposed a segmentation select method of initial cluster center based on the analysis of current K-means cluster algorithm. According to the farthest two points and clustering number of sampling data set, we can calculate the cutting distance, and find the initial cluster center of most data sets, the steps are as follows:

First, calculate the Euclidean distance of different data points in data set Y using formula (1), and find the two points with farthest distance. In formula (1), v and w represent the two points with farthest Euclidean distance, irepresents the dimension of data points, and d represents the Euclidean distance of two data points.

$$d(v-w)_{=}\sqrt{\sum_{i=1}^{n}(v_i - w_i)^2}\ \mathrm{i} = 1, \cdots, \mathrm{n} \quad (1)$$

The reason we want to find the two points with farthest distance is that the difference of the two points is the biggest. That is to say, the data of the two points can cover the most of the data points in data space. According to cluster number k and formula (2), we can calculate the cutting distance, the formula is as follows:

$$d_{cut}^{i} = \frac{v_i - w_i}{k-1}\ \mathrm{i} = 1, \cdots, \mathrm{n} \quad (2)$$

Where irepresents the dimension, d_{cut}^{i} represents the cutting distance of this dimension, k is the cluster number. When the cluster number is 2, the system doesn't need to calculate, the cutting distance of the farthest two points is the initial cluster

center. If k>2, we need to calculate the initial cluster center according to the cutting distance. Because each dimension's calculation is independent, the cutting distance is also independent. Through formula (3) we can calculate the initial cluster center, the formula is as follows:

$$z_j^i = z_{j-1}^i + d_{cut}^i \mathrm{j} = 2, \cdots, \mathrm{k}-1 \quad (3)$$

Where z represents the cluster center. For example, the Euclidean distance with the farthest two points are V and W in data sets, the cluster number is 4. According to the above formula we can calculate the cutting distance and initial cluster center of each dimension, as is shown in Table 1:

Table 1. The Cutting Distance of Each Dimension

k=4	Dimension 1	Dimension 2	Dimension 3	Dimension 4
V	38	20	27	44
W	2	65	48	5
Cutting distance	12	15	7	13

After we calculate the cutting distance, we can calculate the k initial cluster centers using V, W and cutting number k, as is shown in Table 2.

Table 2. Initial Cluster Center

	Dimension 1	Dimension 2	Dimension 3	Dimension 4
z_1	38	20	27	44
z_2	26	35	34	31
z_3	14	50	41	18
z_4	2	65	48	5

4. Algorithm Implementation

4.1 The Customer Loyalty Analysis of Improved K-means Algorithm

Use the improved K-means algorithm, use the R, F, M, S, A as the index, and divide the customer with similar loyalty, the specific procedure is as follows:

(1) Use the analytic hierarchy process to determine the weight of R, F, M, S and A.

(2) Standardize the indexes of R, F, M, S and A, and calculate the weighted integral total integral of each index.

(3) Determine the number of maximum cluster class K_{max}.

(4) Cluster the data using improved K-means algorithm, and get k clustering results.

4.2 Weighted Analysis of RFMSA

Due to the difference of shopping recency, shopping frequency, shopping monetary, customer satisfaction and customer attention, we need to use scientific method to analyze. Therefore, we determine the weighted index based on analytic hierarchy process combining with expert consultation method. The distribution of the weight can directly affects the final classification results, the use of analytic hierarchy process can accurately determine the weight of each parameters.

First, determine the evaluation factor set U according to the five elements in the model.

$$U = \{u_1, u_2, u_3, u_4, u_5\}$$

Where u_1, u_2, u_3, u_4, u_5 represent evaluation factors of Shopping Recency (R), Shopping Frequency (F), Shopping Monetary (M) Customer Satisfaction (S) and Customer Attention (A) respectively. u12 represents the ratio between weight of shopping recency u1 and shopping frequency u2. If u12=1, then it means that u1 and u2

are equally important. If u12=3, then it means that u1 is slightly important than u2. If u12=9, then it means that u1 is more important than u2. If u12=1/3, then it means that u2 is slightly important than u1.

Second, we need to the procedure of expert consultation and evaluation. Through the expert consultation method, we can compare different factors, get the important data, and build the evaluation matrix of the data.

$$P = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} & u_{15} \\ u_{21} & u_{22} & u_{23} & u_{24} & u_{25} \\ u_{31} & u_{32} & u_{33} & u_{34} & u_{35} \\ u_{41} & u_{42} & u_{43} & u_{44} & u_{45} \\ u_{51} & u_{52} & u_{53} & u_{54} & u_{55} \end{bmatrix}$$

After getting the estimation matrix, process each column with normalization.

$$\overline{u}_{ij} = \frac{u_{ij}}{\sum_{n=1}^{5} u_{nj}} \qquad (i, j = 1,2,3,4,5)$$

Sum up the normalized matrix according to the row sequence.

$$\overline{w}_i = \sum_{n=1}^{5} \overline{u}_{ij} \qquad (i, j = 1,2,3,4,5)$$

Finally, normalized the feature vector $\overline{w}_i = (\overline{w}_1, \overline{w}_2, \overline{w}_3, \overline{w}_4, \overline{w}_5)^T$.

$$w_i = \frac{\overline{w}_i}{\sum_{n=1}^{5} \overline{w}_i} \qquad (i, j = 1,2,3,4,5)$$

The feature vector $w_i = (w_1, w_2, w_3, w_4, w_5)$ is the weight value of each element in RFMSA. According to the calculation, the corresponding weights of R, F, M, S and A are $[W_R, W_F, W_M, W_S, W_A] = [02,\ 0.2,\ 0.3,\ 0.2,\ 0.1]$. E-commercial website is mainly considering the profit, so the weight value of M is the biggest. The experts believe that the value of the shopping monetary is a significant element that can affect the customer value.

4.3 The Pre-processing of Experiment Data

The experiment data comes from an actual database of e-commerce website, it includes commodity data, website data, sales data, supplier data, customer data, etc. Make the RFMSA data table from database based on RFMSA model, the table includes customer number, shopping recency, shopping frequency, shopping monetary, customer satisfaction and customer attention. The five indexes have different meaning and data range, they need to be standardized, and convert them into integral. This paper uses the minimum-maximum standardization method [12], and gets the maximum value Vmax and minimum value Vmin, then we use formula (6) to convert the five indexes into standardized integral. After the conversion, if the F, M, S and A become larger, it means that the probability of shopping recency will be higher. If the R is bigger, it means that the customer didn't buy anything for a long time, so the integral of R can be represented as (100-R). Finally we calculate the corresponding weighted integral of the five indexes, sum up and get the total integral that can represent customer loyalty.

$$\mathrm{V} = \frac{(V - V_{\min})}{(V_{max} - V_{min})} \times 100 \qquad (6)$$

4.4 Clustering Result and Strategy Analysis

Use the improved K-means algorithm to clustering the R, F, M, S and A, the result is shown in Table 3. The maximum value of R, F, M, S and A are 20, 20, 30, 20, 10, respectively.

Table 3. Customer Loyalty Classification Situation after using the Improved K-means Algorithm

	Class 1	Class 2	Class 3	Average value
R	6.33	17.44	15.86	13.21
F	0.86	9.76	1.56	4.06
M	5.79	14.56	1.64	7.33
S	0.66	5.89	3.32	3.29
A	0.48	4.36	2.96	2.60
Customer value R+F+M+S+A	14.12	52.01	25.34	30.49

Define the customer current value as the sum of shopping frequency (F) and shopping monetary (M), and define the value-added potential of customer as the sum of customer satisfaction (S) and customer attention (A), the result is shown in Table 4.

Table 4. Customer Value Table of RFMSA Model

	Class 1	Class 2	Class 3	Average value
Current value	6.65	24.32	3.20	11.39
Value-added potential	1.14	10.25	6.28	5.89
Customer value	14.12	52.01	25.34	30.49
Customer number	96	158	47	

After analyzing Table 3 and Table 4, we can get the level of customer loyalty and its proportion, as is shown in Table 5. Class 2 is the customer with high loyalty, the current value and value-added potential are the biggest, it is the most valuable customers of e-commercial website. The website needs to invest major resources to keep good touch with these customers, and keep their satisfaction and confidence to this website. Class 3 is the customer with potential high loyalty, it belongs to the potential customers, the website needs to fully get these customers, protect these customers, regulate the corresponding marketing strategies and make them become a high loyalty customer. Class 1 is the customer with lower loyalty. Although these customers have very low current value, they may be the new customers. We need to keep an attention on these customers, invest more resources, promote the further development, thus improve the customer loyalty.

Table 5. The Proportion of Customer Loyalty

Class	Loyalty level	Number	Ratio%
Class 2	High loyalty	158	52.5
Class 3	Potential high loyalty	47	15.6
Class 1	Low loyalty	96	31.9

This paper proposed the RFMSA customer loyalty model, the e-commercial website can evaluate the customer's loyalty according to their current value, and classify the customer with different loyalty according to the above experiment. Then the enterprises can distribute their resources according the value, and accomplish the long term maximum value of the enterprises. The e-commercial enterprises can use the RFMSA model to accomplish the above strategies, the steps are as follows:

(1) Use the RFMSA model to get the current customer loyalty classification situation. According to the current value and value-added potential of the customer, the customer can be divided into three classes: high loyalty customer, potential high loyalty customer and low loyalty customer.

(2) Regulate the appropriate marketing strategies for different kinds of customer loyalty. Keep a strong relationship with high loyalty customer, build the strategic relationship, and maintain its loyalty. For the potential high loyalty customer, keep the marketing relationship with them, regulate the corresponding marketing strategies to attract their attention, and make them become the high loyalty customer. For the low loyalty customer, analyze the specific reason, improve the relationship with the customers, prevent them to turn into the competitors and become the lost customers.

(3) Distribute the resources of the enterprises reasonably according to the different kinds of loyalty. The e-commercial enterprises need to focus on the high loyalty customers, and guarantee the long term cooperation with the customers. For the potential high loyalty customers, they also need to invest more, promote the bilateral relationships from potential high loyalty to high loyalty customers, and increase their shopping time. For the low loyalty customer, the enterprises need to invest the corresponding resources, and make sure that the customers will not turn into their competitors. For the low loyalty customer, if the customer is valuable, the enterprises need to keep them, and invest the resources appropriately. If the customer has no value, the e-commercial enterprises don't need to invest more resources.

5. Conclusion

This paper proposed a RFMSA customer loyalty analysis model on the basis of deeply analyzing the RFM model, and used the shopping recency, shopping frequency, shopping monetary, customer loyalty and customer attention as the indexes of customer loyalty. Through improving the cluster center selection method to ameliorate the K-means algorithm, this paper proposed a segmentation method based on maximum distance equal division, and avoided the fluctuation caused by the different input of clustering result. This paper used the improved clustering analysis algorithm to analyze the customer loyalty. The experiment results show that the RFMSA model has strong practicability, and can fully reflect the current value and potential value-added ability of the customers. The improved K-means algorithm can effectively analyze the customer loyalty, divide them into different customer classes. The e-commercial enterprises can market according to different customer loyalty, this paper provides a scientific basis for them.

ACKNOWLEDGEMENTS

This research was supported byNature Science Foundation of Heilongjiang Province (F201348).

References

[1] X. Xiangbin, W. Jiaqiang, T. Huan, *et al.*, "Customer classification of E-commerce based on improved RFM model", Journal of Computer Applications, vol. 32, no. 5, **(2012)**, pp. 1439－1442.
[2] W. Kefu, "Applied research on AFH customer classification based on data mining technology", Techno economics & Management Research, no. 11, **(2012)**, pp. 24－28.
[3] D. Tian, X. J Zeng and J. Keane, "Core-generating approximate minimum entropy discretization for rough set feature selection in pattern classification", International Journal of Approximate Reasoning, vol. 52, **(2011)**, pp. 863-880.
[4] G. J. P. van Breukelen and M. J. J. M. Candel, "Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient", Journal of Clinical Epidemiology, vol. 65, **(2012)**, pp. 1212-1218.
[5] A. T. Bida, D. Gil and A. G. Schrum, "Multiplex IP-FCM (immunoprecipitation-flow cytometry): Principles and guidelines for assessing physiologic protein–protein interactions in multiproteincomplexes", Methods, vol. 56, **(2012)**, pp. 154-160.

[6] D. Le, "Research on customer classification of B2C e-commerce enterprises", North China University of Technology, **(2014)**.
[7] X. Lili, "Study on the relationship between customer value and customer loyalty", Lanzhou University of Finance and Economics, **(2014)**.
[8] X. Zhonghai and W. Ling, "The concept of CRM based on customer life cycle. The concept of CRM", scientific research management, vol. 24, no. 6, **(2013)**, pp. 94-102
[9] O. Ibrahim, M. Nilashi and K. Bagherifard, "Application of AHP and K-means Clustering for Ranking and Classifying Customer Trust in M-commerce", Australian Journal of Basic and Applied Sciences, vol. 5, no. 12, **(2011)**, pp. 1441-1457

