An Exploration of Customer Segmentation Methods Based on Clustering Algorithm in the Context of Big Data

Wenbo Zhao *

College of Computer and Control Engineering, Northeast Forestry University, Haerbin, Heilongjiang,150040,China

zt16792023@163.com

Abstract. Accompanied by the continuous development of big data technology, various industries are well aware of the advantages of big data, which are widely used in customer service work, especially in the support of customer segmentation work, and have achieved good results. In this paper, for the problems of large fluctuation of clustering results and low clustering purity in the traditional data mining process, the big data precision mining technology with improved clustering algorithm is proposed. And it is applied in the field of customer segmentation, and the experimental results show that the improved clustering algorithm is applied in customer segmentation, the result curve fluctuation amplitude is small, and the clustering purity is significantly higher than the traditional algorithm.

Keywords: K-means algorithm; big data; customer segmentation; data mining.

1. Introduction

In the era of big data, the establishment of data mining model is to mine the information among the complicated data, in order to achieve the purpose of more efficient information screening and acquisition[1]. In the data mining model, clustering algorithm, as a commonly used algorithm, mainly divides the data into multiple clusters, and selects the data by comparing the similarity of multiple different clusters. In this paper, we study the fuzzy improved clustering algorithm of data mining model in big data, i.e. incremental fuzzy clustering algorithm, which is mainly based on the minimum weight threshold.

2. Overview of clustering algorithms

2.1 FCM clustering algorithm

The basic principle of FCM (Fuzzy-c-Means algorithm, FCM) clustering algorithm is fuzzy theory, so it is also known as fuzzy C-means algorithm[2].FCM clustering algorithm is to take n user data as n vectors xi, and the value of fuzzy affiliation of FCM clustering algorithm is [0, 1], and categorize them by calculating the fuzzy affiliation of each vector.The essence of FCM clustering algorithm is to construct fuzzy clusters. The essence of the clustering algorithm is to construct a fuzzy matrix U, each element of the matrix is the fuzzy affiliation of each vector, so its value is of the magnitude of [0, 1], and the sum of the fuzzy affiliation of each element of the clustering is 1. The center of the clustering is found by setting the non-similarity function and taking the minimum of the function as the target value.

The expression of nonlinear constraints of FCM clustering algorithm is shown in equation (1).

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \forall j = 1,2,...,n \qquad (1)$$

Where u_{ij} is the affiliation matrix.

The expression of the objective function of FCM clustering algorithm is shown in equation (2).

$$J(U, c_1, c_2, ..., c_c) = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \quad d_{ij} = \left\| c_j - x_j \right\| \qquad (2)$$

Where: u_{ij} takes the value of [0, 1]; c_i is the clustering center of fuzzy class i; d_{ij} is the Euclidean distance between the i th clustering center and the j th vector; and m is the weighting index, which takes the value of [1, ∞][3]. In order to minimize the objective function, the paper makes the following improvement, whose expression is shown in Equation (3).

$$J(U,c_1,c_2,...,c_c,\lambda_1,\lambda_2,...,\lambda_n)=J(U,c_1,c_2,...,c_c)+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right)$$

$$=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}d_{ij}^{2}+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \quad (3)$$

Where: λ_j is the Lagrange factor of n constraint equations.

The necessary condition to make the objective function obtain the minimum value is shown in equation (4).

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \quad u_{ij}=\frac{\left(d_{ij}\right)^{\frac{2}{m-1}}}{\sum_{k=1}^{c}\left(d_{kj}\right)^{\frac{2}{m-1}}} \quad (4)$$

The basic steps of FCM clustering algorithm are as follows.

1) Calculate the fuzzy affiliation degree of each vector and construct the initial fuzzy matrix u such that each element of the matrix takes the value of [0, 1] so that it satisfies the sum of affiliation degrees of the vectors in each class is 1[4].

2) Calculate the clustering center c_i for C fuzzy classes.

3) Calculate the objective function and set the threshold of the objective function.

4) Calculate the new fuzzy matrix U. Then return to step 2 and iterate until the condition is satisfied.

The flowchart of the FCM clustering algorithm is shown in Figure 1.

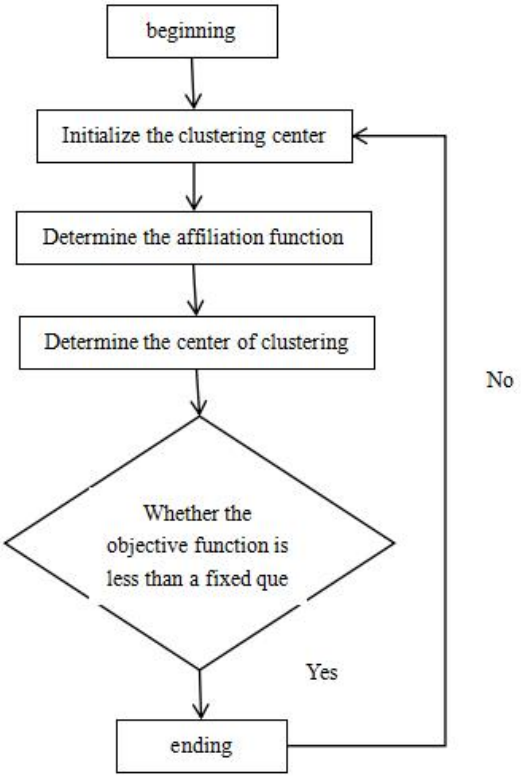

Fig. 1 Flowchart of FCM clustering algorithm

2.2 Description of K-means algorithm

K-means is a classical clustering method based on division, generally using the Euclidean distance as a measure of similarity between two data points, the greater the similarity, the smaller the distance[5]. The core idea of the algorithm is: first determine the number of clusters K and K initial clustering centers. According to the distance between the data points and the clustering center, the location of the clustering center is constantly updated, making the sum of squared error (SSE) of each cluster smaller. When the SSE no longer changes or the objective function converges, the SSE value reaches the minimum, the iteration stops, and the final clustering results are obtained. The algorithm flow is as follows.

1) Initialize the clustering center to determine the number of clusters K, and randomly select K points from the data set as the initial clustering center $C_i(1 \le i \le k)$.

2) Allocate samples. Calculate the Euclidean distance between the remaining data points and the clustering center C_i , find out the shortest distance and assign all the samples to the clusters corresponding to the clustering center C_i . The Euclidean distance formula is

$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - C_{ij})^2} \qquad (5)$$

In equation (1), x is the data object, C_i is the i th clustering center, m is the dimension of the data object, and x_j , C_{ij} are the j th attribute values of x and C_i .

3) Update the clustering center. Calculate the average value and square error of all points in each cluster, take the average value as the new clustering center and repeat step 2). The squared error is calculated as

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |d(x, C_i)|^2 \qquad (6)$$

4) Until the clustering center no longer changes or reaches the maximum number of iterations, the loop ends and the final clustering result is obtained[6].

3. Conclusion

K-Means is one of the most commonly used and widely used clustering algorithms. The algorithm is relatively simple and does not require high computer performance, so it is also suitable for cluster analysis of a large number of data samples. The algorithm promotes the formation of the optimal solution of the distance between different classes through the form of iteration, and finally obtains the clustering center. Through the research in this paper, it can be confirmed that the use of K-Means algorithm to analyze the customers in the enterprise, can be taken for the customers to take targeted services, in order to effectively improve the competitiveness of the product, and promote the customer to have a good product experience.

References

[1] Hua-Ping Wu; Weiwei Feng; Lin Li. RFM-based clustering algorithm in retail market customer segmentation research[J/OL]. Journal of Chongqing University of Technology (Social Science),1-16.

[2] Bai Zhao; Wang Shenwu; Wang Mengyang. Research on civil aviation customer segmentation based on improved RFM model and clustering algorithm[J]. Science and Industry,2023,23(19):200-203.

[3] Zhu Weiguang. Research on smartphone demand preference discrimination and customer segmentation model construction based on online reviews[J]. Computer Age,2023,(09):132-135+141.

[4]Tang Xin. Research on the application of optimized K-means clustering algorithm in customer segmentation[J]. Intelligent Computer and Application,2023, 13(09):194-197.

[5]Wei Jianbing. Research on customer segmentation of RFM model based on K-means algorithm[J]. Computer Knowledge and Technology,2023,19(13):73-75.

[6] Li Xiaoli; Su Qin; Wu Bo; Li Winzhou; Li Qingqian. User value analysis of department store based on K-means clustering algorithm[J]. Journal of Shanxi Normal University (Natural Science Edition),2023,37(01):7-13.