Application of Teaching Example Based on Clustering Algorithm

Xiang Li[1, a], Xin Su [2], Miao Di [1], Muzi Zhuge [1]

[1]School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China;

[2]Chengxi Electric Power Supply Branch, Tianjin Electric Power Co., Ltd, Tianjin 300192, China.

[a]tutezdh1203@163.com

Abstract. In this paper, K-means and Fuzzy C-means clustering algorithms are used as typical algorithms for data mining courses. In addition, it uses electricity monthly electricity load data for a certain region in domestic as the testing data. With the MATLAB environment, the clustering clusters and different clustering centers for different power consumption behaviors of power consumes are obtained. By analyzing the clustering clusters and clustering centers, the clusters obtained by different algorithms can be roughly divided into three categories, and the power consumers' electricity habits in this area can also be divided into three categories. This paper uses the approach to improve students' interest in learning, expand students' practical applications in clustering algorithms, and cultivate students' creative creativity.

Keywords: Data mining, K- means algorithm, Fuzzy C- means algorithm, MATLAB, power consumer's behavior.

1. Introduction

Data mining is the study of how to mine the knowledge or information hidden in a large amount of data [1-3]. Data mining usually uses computer science, statistics, online analysis, machine learning and many other technologies to achieve the above goals [4]. Faced with the explosion of information and data, it requires a large number of person who are engaged in data mining and analysis to get the society better in social services. As a result, a large number of colleges and universities opens the data mining course. In order to obtain a good classroom teaching effect, it uses case teaching strategies to guide students to learn in the data mining curriculum.

After studying the different consumers' electricity consumption behavior, the clustering algorithm to analyze the power consumption data of consumes is made. Through run with the K-means algorithm and the Fuzzy C-means algorithm, it can be seen that the clustering result is approximately the same: there are three different kinds of electricity usage habits for power consumers.

2. Data Preprocessing

In the paper, the electricity consumption data of the consumers is extracted from the Supervisory Control and Data Acquisition System of the smart grid in a certain area of China in October. Accidents such as missed sampling and wrong sampling in data collection are inevitable which result in the necessity to preprocess the data before analysis. Thus, the data preprocessing is needed. Due to the difference of consumption behavior among customers, the data exist with the characteristics of multiplicity and diversification. In order to reduce the error and enhance the degree of aggregation between the customers of the same consumption behavior, the data of each customer is normalized, that is, the data value is limited to a value varying from 0 to 1.

3. Clustering Model

3.1 K-Means Algorithm

The K-means algorithm is the most popular algorithm to solve the clustering problem conveniently and efficiently based on partition [5]. Based on the calculation of relative distance, the data is divided

into specific number of clusters resulted from K value. The cluster centers are optimized continuously by adjusting the centers by the principle of relative distance. The best clustering effect can be obtained after enough iterations.

The models that can be applied to the calculation of relative distance include similarity rules of Euclidean distance, Jacquard coefficient, cosine, Pearson coefficient, relative entropy and Hollinger distance. Due to the high dimensional nature of the data, similarity model of cosine is selected to calculate the distance between each data and the cluster center.

The steps of the algorithm are as follows:

(1) Determine the value of K, that is, there are K different clustering results;

(2) Initialize K cluster center randomly from the n pieces of data.

(3) According to the similarity model, K pieces of distance between each data and each cluster center are calculated, and the cluster center with the minimum distance replaces the old one to be the new center of the data.

(4) The cluster center is renewed by averaging every piece of data clustered to the same cluster center after all the data is clustered;

(5) Repeat step from 3) to 4) until the preset maxim number of iterations or the error between two adjacent clusters is less than preset value.

3.2 Fuzzy C-Means Algorithm

Fuzzy C-means algorithm is based on the common C-means algorithm and generated by concept of fuzzy theory [6]. It makes the boundaries between different clusters uncertainly, which produces a possibility that one data may have different clusters. It greatly reduces the effect of human factors, so that the clusters become more convinced. The basic idea of Fuzzy C-means algorithm is as follows: C value is taken as the clusters, and the data is partitioned by fuzzy affiliation probability generated by membership function; Through making the optimization and adjustment for the clustering centers, the data can be classified many times and the good clusters can be achieved finally.

The basic steps of the algorithm are shown as follow:

(1) Determine the value of C, which is represented as the number of clusters.

(2) Generate the initial fuzzy matrix according to the requirement, as shown in Eq.1.

$$\sum_{j=1}^{c} u_{ij} = 1, \ \forall i = 1,...,n \tag{1}$$

Wherein, u_{ij} represents the membership degree of the i-the data in the j-the cluster center, which is equal to probability.

(3) Calculate the initial cluster center using Eq.2.

$$cluster_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{2}$$

Wherein, $cluster_j$ represents the j-the cluster center; m is a weighted index, generally taking 2; x_i represents the i-the data vector.

(4) The data are classified based on the affiliation within the fuzzy matrix.

(5) After all the data are classified, the fuzzy matrix is optimized by Eq.3, and updated.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}^2}{d_{ik}^2})^{\frac{1}{m-1}}} \tag{3}$$

Wherein, d_{ij}^2 represents the square of "relative distance" between the i-the data and the j-Th cluster center.

(6) Calculate the objective function, as shown in Eq.4;

$$J(U,cluster)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}d_{ij}^{2}=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}\left\|cluster_{j}-x_{i}\right\|^{2} \quad (4)$$

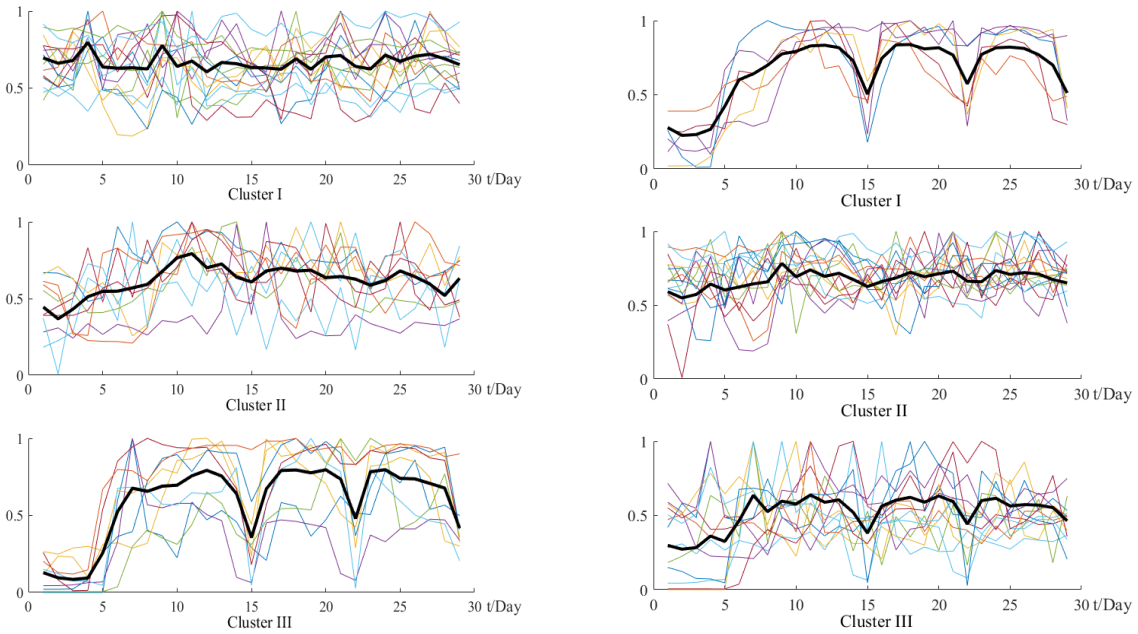
Wherein, U is fuzzy matrix?

(7) Repeat steps 3) to 6) until the number of iterations overflows or the objective function error is less than the preset parameter.

4. Case Demonstration

In order to realize the classification of power usage behavior of power consumers and achieve the optimal teaching purpose of clustering algorithm analysis, the data is brought into the clustering algorithm for analysis. The algorithms are run on the MATLAB 2015b version.

The cluster clusters and cluster centers of the load curve in different algorithms in October are shown in Fig. 1.



(a) K-means Algorithm (b) Fuzzy C-means Algorithm

Fig 1. The clusters of behavior of power consumers in October

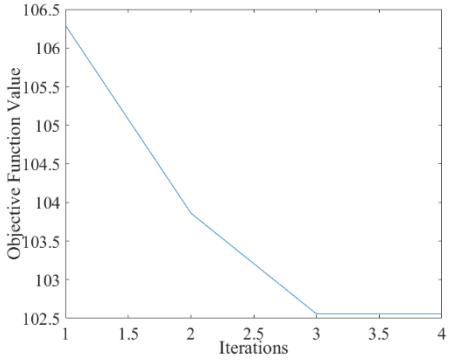
5. Algorithms Analysis

For the K-means algorithm, an objective function needs to be defined first as the total error value of the clustering data in the K-means algorithm as shown in Eq.5. When the total error between the adjacent target function is less than 1×10-5, end the algorithm and output the result.

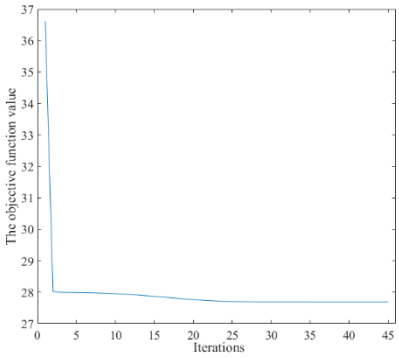
$$dist=\sum_{j=1}^{k}\sum_{i=1}^{n}d_{ij}^{2} \quad (5)$$

Where k represents the clustering result; n represents the number of data clustered to j-the cluster; d represents the Euclidean distance of its data to the cluster center in the j-the clustering result.

Since the Fuzzy C-means algorithm has its own objective function, it can be used as an evaluation method for the algorithm. When the total error between the adjacent target function is less than 1*10-5, end the algorithm and output the result. The results of the two programs running are shown in Fig.2.



(a) K-means Algorithm



(b) Fuzzy C-means Algorithm

Fig 2. The Relationship between the Number of Iterations and Total Error from Data in October

6. Conclusions

By using K-means algorithm and Fuzzy C-means algorithm for cluster analysis of power consumer data, it is found that both algorithms classify the power users of the region in October into three types of electricity usage. The analysis of cluster clusters and cluster centers shown in Fig. 1 reveals that there are some differences between other two cluster clusters and cluster centers, while the left cluster clusters and cluster centers have certain similarities. Through continuous experimental demonstration of the clustering results, it can be found that the K-means algorithm has the disadvantage of instability at runtime and there is a certain probability that the clustering effect is not fixed. This is related to the selection of the initial clustering center of the K-means algorithm. The result of Fuzzy C-means algorithm is relatively stable. Through the above example teaching, it not only helps the students to focus on the clustering algorithm and the clustering of the algorithm itself, but also advantages to cultivate the students' ability of the optimization algorithm and the desire to contact the new clustering algorithm so as to achieve a good teaching effect.

Acknowledgements

Support by Tianjin Natural Science Foundation (17KPHDSF00290, 16KPHDSF00050), Tianjin Science and Technology Project (17JCTPJC48300, 16ZXHLSF00200), Tianjin Education Commission Scientific Research Plan Project (JWK1606), Tianjin University of Technology and Education talent start-up project (KYQ01614), Tianjin Education Science "13th Five-Year Plan" Plan (VEYP5044), and University Student Innovation Training Program (20171066167) University Student Innovation Training Program (201810066056).

References

[1]. Hangdog Li, Tainting Yao. Pattern classification. Beijing: China Machine Press, 2003, p. 415-481.

[2]. Hexing Zhang, Chennai Sun, Guying Shu. Neuron-fuzzy and software calculations. Xi'an: Xi'an Jiao tong University Press, 2000, p.203-304.

[3]. Bedeck J. C. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press, 1981:65-80.

[4]. Lijiang Wang, Zheng Hai, Reich Cain, et al. Case Study of K-means Clustering Algorithm. Computer Education. (2016) No. 8, p. 152-157.

[5]. Lily Liu: Research and Improvement of K-Means Clustering Algorithm (Master Degree, Queue Normal University, China 2015). p.14-16.

[6]. Yun he Zhu: Research of the Related Problems on Fuzzy C-Means (Master Degree, Ocean University of China, China 2011). P.14-17.