Research and Application of Power Consumers Behavior Analysis Based on K-means Algorithm

Genghuang Yang[1, a], Xiayi Hao[1] and Yingmei Zhang[1]

[1]School of Automation and Electrical Engineering, Tianjin University of Technology and Education Tianjin, China

[a]ygenghuang@126.com

Keywords: Cluster Analysis, K-means Algorithm, Power Consumers Behavior Analysis, Monthly Load Curve, Power Service Strategy

Abstract. K-means algorithm is used to analyze pretreated monthly load data of power consumers from a certain domestic area in China. It can get different clusters in three months, and those can be roughly divided power consumers into different types, such as residents, businesses and factories based on power customers' behavior analysis. A measure is introduced to evaluate the reliability of the algorithm. According to the above clustering results, it can be found out that the power consumption characteristics of all kinds of consumers in the corresponding months. In view of these characteristics, this paper can play a guiding role in formulating the corresponding power service strategy for different types of power consumers in power supply companies.

Introduction

With the rapid development of smart grid, the automation in power distribution networks are gradually improved [1,2]. The data collection in distribution network comes down to the common customers. The data sampled from power customers in distribution networks, which is characterized by large data volume, numerous data types and fast growth, becomes the focus of research all over the world [3,4]. Electric power companies can further regulate the operation of the power system by analyzing power customers' consumption based on big data, and then improve the efficiency and service quality of power supply [5].

The premise of realizing the above purpose is to make a reasonable analysis of the recorded power customers' data. By studying the consumption behavior based on big data in a specified area, several types of rules in consumption behavior in the area are summarized to significant guide for electric power companies in power dispatching and supply [6,7].

Data Preprocessing

The data used for analysis in this paper comes from the data collection devices in a common domestic area. Accidents such as missed sampling and wrong sampling in data collection are inevitable which result in the necessity to preprocess the data before analysis.

For the missed sampling data, the average of the nearly last collected data and the nearly next data is calculated to be the substitution. When the number of missed sampling data is large, the specific data will be regarded as a special one and is removed.

For the wrong sampling data, when there is a little negative sampling data, it is feasible to correct the data to zero whereas there is so much negative sampling data, it is excluded as the thoroughly damaged data so as to avoid affecting the stability of the algorithm. Also, it can be checked whether the equipment has been damaged by human/natural way or not.

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.

This paper aims to study the customers' power consumption behavior in a specific area. Changing the value of data or removing some abnormal data does not affect the overall implicit information of the original data.

Clustering Model

The K-means algorithm is the most popular algorithm to solve the clustering problem conveniently and efficiently based on partition. 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, Jaccard coefficient, cosine, Pearson coefficient, relative entropy and Hellinger 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 cluster is less than preset value.

Analysis of power consumption behavior

In order to obtain the characteristics of consumption behavior in monthly load, the data is divided into three parts according to the month of August, September and October. The data of the above months is respectively applied in the clustering algorithm for analysis.

After verifying the clustering result of K-means algorithm, it is found that the customers in the specific region can be divided into 2 clusters in August and September whereas 3 clusters in October.

The preprocessed data is handled by K-means algorithm running on the platform of MATLAB 2015b. The clustering results and cluster centers of the load curve in August are shown in Fig. 1. The results in September are shown in Fig. 2. The results in October are shown in Fig. 3.

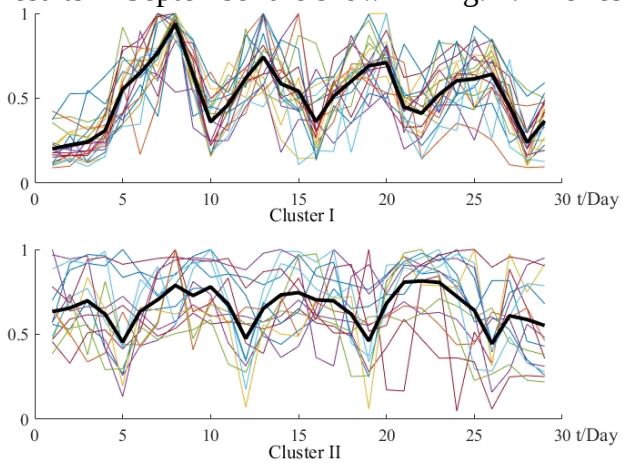


Fig.1 Clusters of behavior in August

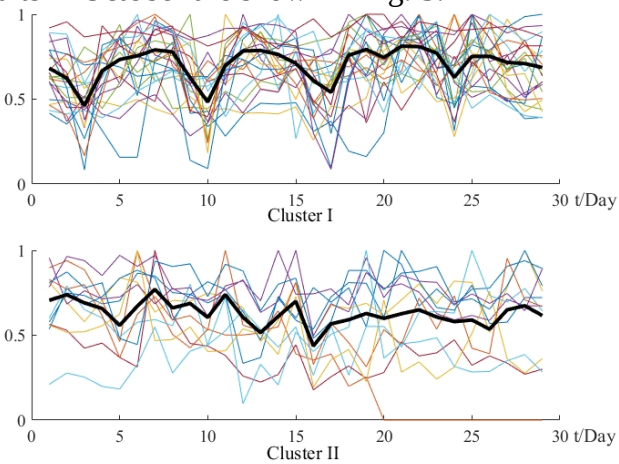


Fig.2 Clusters of behavior in September

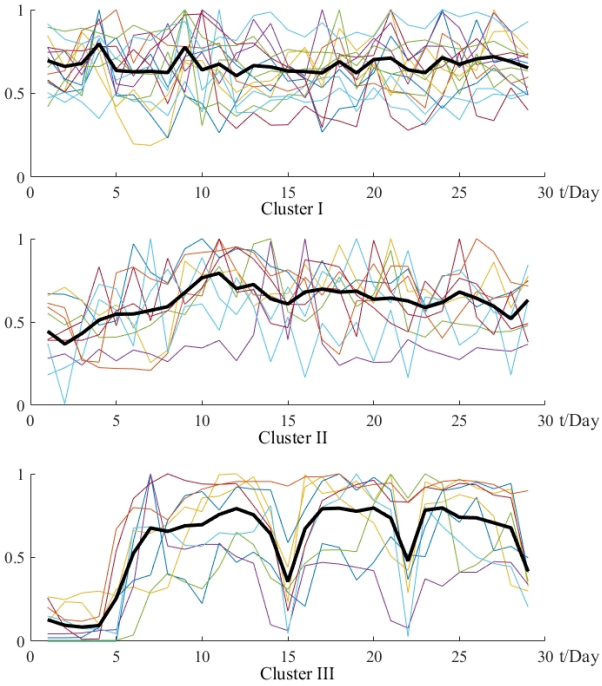


Fig.3 Clusters of behavior in October

Comparing the clustering results in Fig. 1, Fig. 2 and Fig. 3, it is found that the customers' data of different months have the same clustering results and different clustering centers in clustering.

Analysis of clustering results

Table 1 shows the power customers in clusters obtained by K-means clustering algorithm.

Table 1 Customers after clustering monthly

Cluster	Customers(in August)	Customers(in September)	Customers(in October)
I	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22	4,9,10,11,12,13,15,17,19,20,21,22,23,24,25,26,27,28,30,31,32,33,34,35,36,37,38,39,40,42,43,44,45,46,47,48,49,50,51,53,54,55,56,57	5,6,7,10,11,12,13,14,15,17,18,19,20,21,23,24,36,37,47,48
II	23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57	1,2,3,5,6,7,8,14,16,18,29,41,52	1,3,4,8,9,16,22,29,30,33,41,44,52,55
III	\	\	2,25,26,27,28,31,32,34,35,38,39,40,42,43,45,46,49,50,51,53,54,56,57

It can be seen that from the above table, some power customers changed their consumption behavior in September comparing to that in August. One-third of power customers acclimatize oneself to new habit in October. The above transformation may be caused by variation in power supply policy and weather such as air temperature.

Fig. 1 shows that the time when the power customers of cluster I stands at the peak of electricity consumption, the power customers of cluster II stands at the trough. As the load curve is monthly load curve, it can be considered that the power customers of cluster II are restricted use of electricity to ensure the consumption of the power customers of cluster I. Moreover, the cluster center of cluster I varies by certain periodic changes in August shows the difference between week dates and weekends.

Fig. 2 shows that the power consumption behavior of power customers in cluster II are basically smooth and steady in September. Taking into account the seasonal factors, it shows that the power customers of cluster II in September and power customers of cluster I in August should be included in the same consumption behavior of the resident power customers.

Fig. 3 shows that the power consumption behavior of the power customers in cluster III is the same as that of the power customers in cluster I in September and that of the power customers in cluster II in August. Moreover, the power consumption behavior of electricity is characterized by periodic changes. This result shows that the power customers of the above clusters are not affected by factors such as the weather. The power customers of the above clusters stand for the power customers such as enterprises and the factories. Because of the differences between the cluster I and cluster II in October and factor of National Day holiday, the power customers of cluster I stands for business and that of cluster II stands for resident.

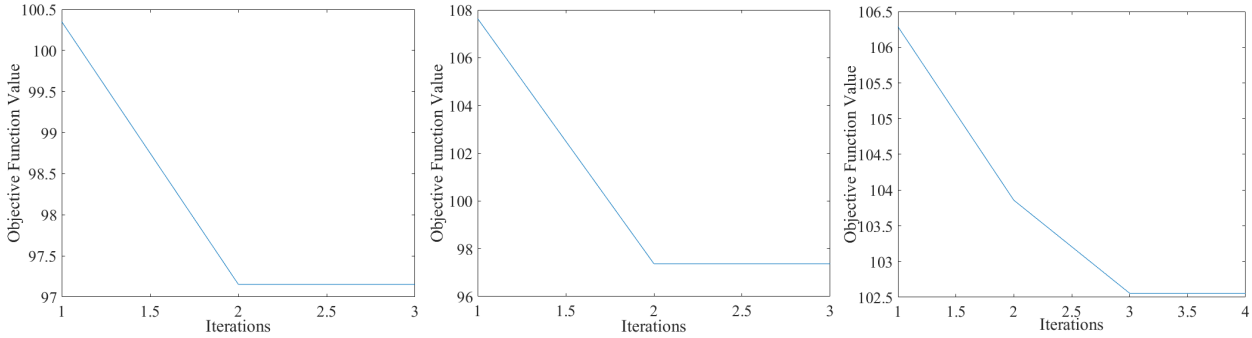
The above clustering result helps the power company build up or revise the corresponding service strategy. For example, restricted use of electricity can be forced upon the customers who consume too much electricity at peak periods to reduce the power consumption. By implementation of the preferential policies for power customers who consume electricity at trough, the efficiency of power supply is enhanced to make the power system runs steadily.

Analysis of Algorithms

The objective function as the Eq. 1 shows is formed by the total error difference after data clustering by K-means algorithm. When the total error between adjacent points of the objective function is set to be less than 1×10^{-5}, the algorithm end and the result is obtained.

$$dist = \sum_{j=1}^{k}\sum_{i=1}^{n} d_{ij}^{2} \tag{1}$$

In the above Eq. 1, k represents the clustering result; n represents the number of data clustered to j-th cluster; d represents the Euclidean distance of i-th data to the cluster center in the j-th clustering result.



(a) Data of August (b) Data of September (c) Data of October

Fig.4 Total error- number iterations monthly

It can be seen from Fig. 4 that K-means algorithm obtains the same clustering results in the 3rd iteration processing with data of August, the 3rd iteration of processing with data of September, and the 4th iteration of processing with data of October. The above curve shows that the K-means algorithm can cluster data in high speed. However, the experimental results show that the algorithm has the disadvantage of instability in runtime, and non-fixed probability in clustering effect. The deficiency result from the selection of the initial clustering center of K-means algorithm.

Guild for electricity service strategy

The conclusions obtained by analyzing the above load curve and data of power customers result in the guild. Firstly, the power company could propose the scheme of price for residents, factories and business in August, so as to reduce power consumption in the period of peak load. Secondly, the

electricity dispatching in this area should be enhanced in the appropriate period so that the customers in the cluster II would be able to use the electricity normally when the power customers in the cluster I consume so much electricity. Thirdly, power customers in different enterprises and factories were encouraged to staggered electricity to ensure the stability of the power grid.

Conclusions

In this paper, K-means algorithm for clustering analysis is used to divide power customers into different types with different characteristics in power consumption, so as to distinguish the power consumption behavior of different power customers.

From the above describes, K-means algorithm makes a credible distinction among different power customers. After distinguishing power customers, the type of power customers based on clustering can be definitely confirmed according to their respective power consumption characteristics and practical significance. The load curve can clearly express the current running status of power transmission in the grid, which has a guiding significance for large-scale regional power scheduling. According to the information contained in the monthly load curve of power customers, the power company can propose different strategies in sales for different type of customers in different months. Through these sales strategies, a large number of fluctuations in the power grid will be reduced, and the win-win benefit of the company and the power customers will eventually be gained.

Acknowledgements

This work was financially supported by the Science and Technology Project Funded by Zhangjiakou Power supply Co. Ltd, State Grid Jibei Electric Power Company (SGTYHT/16-JS-198).

References

[1] J.Y. Hu, E.G. Zhu, X.G. Du, et al. Application Status and Development Trend of Power Consumption Information Collection System [J]. Automation of Electric Power Systems, 2014, 38(2): 131-135.

[2] China Electrical Engineering Society Information Specialized Committee. China Power big data development white paper [R]. 2013.

[3] Y.Q. Song, G.L. Zhou, and Y.L. Zhu. Present Status and Challenges of Big Data Processing in Smart Grid [J]. Power System Technology, 2013, 37(4): 927-935.

[4] Q.D. Feng, Z.Y. He. Analysis and Comparison for the Development of Smart Electricity Consumption in Domestic and Foreign [J]. Electrical Measurement & Instrumentation, 2012, 554(49): 1-6.

[5] J.Y. Lu. Study on Multidimensional Classification on Method of User Behavior in Smart Grid and Its Application [D]. Beijing Jiaotong University, China, 2016

[6] W.L. Zhang, Z.Z. Liu, M.J. Wang, et al. Research Status and Development Trend of Smart Grid [J]. Power System Technology, 2009, (13): 1-11.

[7] L.J. Xu, D.X. Zhang, and J.L. Wang. Analysis of User Classification and Electricity Consumption Behavior Based on Cluster Analysis [J]. Shanxi Electric Power, 2016, (10): 23-27.