Residential power user segmentation based on k-means clustering method in the context of big data

Yongxiu He[1], *Zhe* Jiao[1,*], *Fenkai* Chen[1], *Fengtao* Guang[1], *Peipei* You[2] and *Qing* He[1]

[1] School of Economics and Management, North China Electric Power University, Beijing, 102206, China
[2] State Grid Energy Research Institute CO., LTD, Beijing 100052, China

Abstract. With the deepening of the reform on the selling side of electricity, the selling company must strengthen the analysis on the electricity consumption of users and arrange the purchase and sale of electricity scheme scientifically so as to occupy the target selling electricity market and obtain the profit of selling electricity. Based on the big data background, the paper uses the k-means clustering method to divide the load curve of 2,498 residential power users into 5 categories. On the premise of considering the system load, the above five categories are classified into three types: peak load, partial peak load, and stable power users.

1 Introduction

With the continuous improvement of people's living standards, the proportion of household electricity consumption in the total social electricity consumption is gradually increasing. It has been the research direction of electric power enterprises to better understand residents' consumption habits to develop differentiated services and improve the quality of electric power services [1]. At present, the power household users in China have reached a considerable scale, and it is not feasible to conduct one-to-one load characteristic analysis for these users. Therefore, power user classification is the basis and key link of power user load characteristic analysis.

At present, power load pattern clustering methods mainly include fuzzy clustering method [2,3], Kohonen neural network (KNN) clustering method [4,5] and load clustering method based on Data Mining (DM) [6,7]. Fuzzy clustering method is easy to be interfered by subjective factors, which makes the clustering results have local differences and the algorithm is relatively complex. The practical application of KNN clustering method shows that it can't meet the requirements of load forecasting clustering for curve shape recognition. The k-means clustering algorithm selected in this paper classifies the power load of residents. When processing large data sets, this algorithm maintains scalability and efficiency, and has high accuracy. Sun Yuan (2018) selected daily load data of 96 points per day for major electric power customers in four key industries of ferrous metal smelting and rolling process, ferrous metal mining, non-metallic minerals and metal products in the five cities of northern Fujian. The clustering analysis was carried out by using the electrical load characteristics, and the power consumption behavior was deeply understood through the load characteristics[8]. Jiang Ying (2015) combined with big data platform, based on linear regression, grey prediction and other algorithms to build a data mining model, to achieve the application of residential electricity consumption analysis and electricity load forecast analysis [9].

In this paper, k-means clustering algorithm is used to cluster the daily load curve of multiple residents. The typical load curve of users is extracted, and then the load characteristics of different users are analyzed. The typical user load curve compared with system load curve, the result of the above user classification in the paper, which guides sell electricity company according to different category users targeted put forward the corresponding pricing packages.

2 K - means clustering model

The k-means clustering model under the background of big data mainly includes two parts: data preprocessing and k-means clustering.

2.1 Data preprocessing

If the sample data is missing part of the load point data or if it is wrong, it will affect the correctness of the clustering result. Therefore, the sample data should be processed with missing values before clustering.

(1) Delete invalid records in the table

Since the selected load value is a positive active load, the data in the table should be greater than or equal to zero. Firstly, delete the records with negative values in the table; secondly, in order to make the clustering results representative, delete all records with 0 in the table.

(2) Delete too many data missing in succession

* Corresponding author: 2234923819@qq.com

There are 24 load points per day, and thus the missing samples with more than 12 load values are deleted consecutively.

(3) Individual missing value processing

Since the processing of missing values is to fill in the historical data, and the power usage behavior of each user is a certain regularity, it can be compensated by the mean method according to the same load point data before and after the data of a certain load point.

Due to the different user load size is far, in order to more accurately cluster users with similar characteristics, the original data needs to be dimensionless and then analyzed. In order to ensure the effectiveness of sample data training, power load data should be normalized. In this section, the extremum method is adopted, and the specific formula is as follows.

$$X^* = \frac{X}{X_{\max}} \tag{1}$$

This method can achieve equal scale scaling of the original data, where X^* is the normalized value, X is the original data, and $X_{\max}$ is the maximum value of the original data set. After normalization, the data is limited to [0,1] interval.

2.2 K-means clustering model

K-means clustering is the most classical and simplest clustering algorithm among all clustering algorithms. For a given sample set $D=\{x_1, x_2, \ldots, x_m\}$, we first determine the clustering number (k), First of all, k points were randomly selected as the center of the cluster, and then according to the distance metric find center nearest sample points and classify it into the nearest cluster, and the cluster center will be updated at the same time. Iterate through the above operations until the data is no longer changed or the center of the cluster is no longer changed. At this point, the output divided cluster $C=\{C_1, C_2, \ldots, C_k\}$, the loss function of clustering is as follows.

$$E = \sum_{i=1} \sum_{x \in C_i} \left\| x - \mu_i \right\|^2 \tag{2}$$

Where μ_i is the mean vector of cluster C_i, which is expressed as follows:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{3}$$

The K-means algorithm is described as follows:

Input: iteration termination condition ε, maximum number of iterations maxStep, total number of clusters k and sample set containing N records.

Output: k clusters satisfying the iteration termination condition, the number of iterations s.

The k-means algorithm is often used in power data analysis. This paper uses this method to cluster the load curves.

3 Case study

Due to the randomness of power users' electricity consumption, the load curve of power users is diverse. Even for a user, the load curve of power users is not similar every day. Therefore, k-means clustering algorithm is adopted in this section to analyze the daily load curve of multiple residents, extract the typical load curve of users, and analyze its load characteristics.

3.1 Cluster analysis

In this paper, 2498 daily load curves of residents in a certain area are selected. The data acquisition time interval is 1 hour, and each load curve contains 24 active power points. First, data is filled in the data samples, and the sample data that is still invalid after the filling is deleted. In this section, 2468 samples after normalization processing are classified based on the k-mean clustering model. When k-means method is used for cluster analysis, the number of categories should be specified in advance. In order to make the classified categories more representative, we respectively specify the number of categories as 3-10, etc. It is found that when the number of classifications exceeds 5, the load curve is similar in shape and the classification is not representative.

Table 1. The number of cases in each cluster.

Item		Number
Clustering categories	1	810
	2	20
	3	955
	4	85
	5	598
Effective		2468
Missing		30

The following section shows the five clustering center curves and their corresponding user differentiated electricity utilization characteristics.

(1) The load curve of the first type user is relatively stable. The load is basically stable with little fluctuation range throughout the day. This type of user has a high load rate of around 0.8. This type of user is suitable for implementing a load rate electricity price package.

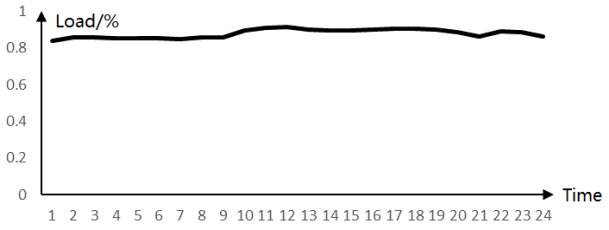

Fig. 1. Center curve of load clustering for the first class.

(2) The daily load curve of the second type of users is one peak and one valley. The electricity load is lower in the daytime and higher in the evening. This type of user is suitable for using the time-of-use tariff package.

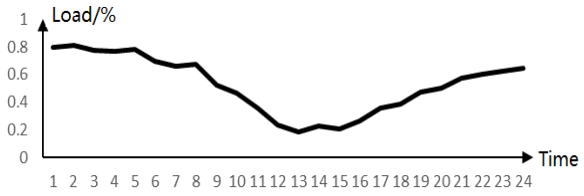

Fig. 2. Center curve of load clustering for the second class.

(3) The third kind of load daily basic load curve is represented by two peaks and one valley, which is a typical peak load curve. and the peak of electricity consumption appears around 11 and 20 points respectively. The peak of electricity consumption is around 11 and 20 points respectively, and the proportion of electricity consumption in the low period is small. Therefore, the load of this kind of users has a great capacity of load control. Education should be promoted, time-sharing tariff package should be implemented to increase the willingness of residents to participate in power demand response management.

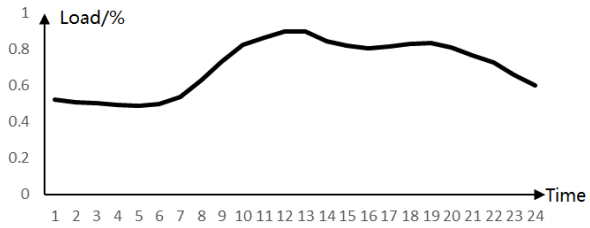

Fig. 3. Center curve of load clustering for the third class.

(4) The fourth type of load is similar to the first type of load and is relatively stable. But unlike the first, the load tends to fall after 8 o 'clock.

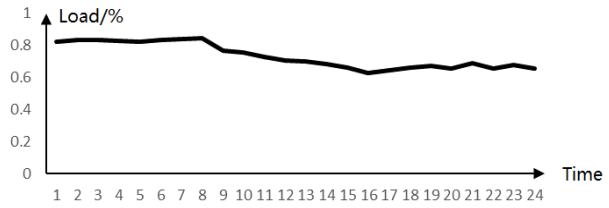

Fig. 4. Center curve of load clustering for the fourth class.

(5) The daily load curve of the fifth type of residents is similar to that of the third type of users. It has great potential of load control. It is characterized by one peak, one flat and one valley.

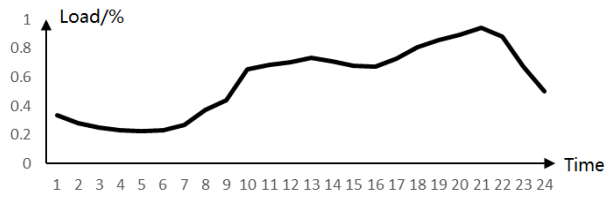

Fig. 5. Center curve of load clustering for the fifth class.

3.2 Residential user load characteristics classification considering system load

After extracting the typical load curve of the user, we can compare the power consumption characteristics of the relative individual with that of the group. Furthermore, a classification rule is proposed to facilitate the design of retail tariff package for the future power selling companies.

The typical daily load curve in this region is in the shape of three peaks and one valley. As shown in the figure below, three peaks can be seen in the morning, afternoon and evening. The early peak occurs between 9:00am and 11:00am. The midday peak occurs between 14:00pm and 18:00pm and lasts for a long time. The late peak occurs between 19:00pm and 21:00pm, and the duration is relatively short. The trough occurred between 1:00am and 6:00am and lasted a long time.

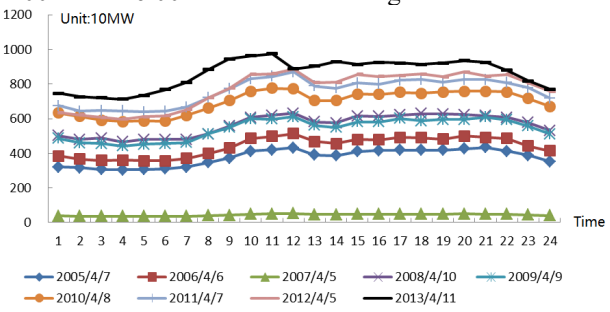

Fig. 6. Typical daily system load characteristic curve of a region.

By analyzing the relationship between the typical load characteristic curve of residents and the load characteristic curve of the system, we can divide residents into three categories: peak load, partial peak load, and stable power users.

(1) The peak load type users mainly correspond to the third and fifth users mentioned in the previous section. The curve peak time interval is highly consistent with the regional total load curve, and he error is within 30 minutes. The maximum of the morning and afternoon peaks is significantly lower than the maximum of the late peaks.

(2) Partial peak load type users mainly correspond to the second users in the previous section. The curve peak time interval is basically consistent with the regional total load curve, and there is a peak-avoidance power behavior relative to the overall load.

(3) The smooth power users correspond to the first class and the fourth class of users in the previous section, and the fluctuation of power consumption during the day is not large.

The above classification is mainly based on the difference between the individual curve shape and the group curve shape. The analysis of the user load characteristics considering the system load characteristics can better meet the needs of the power company, and provide a basis for the formulation of retail electricity price packages and value-added services.

4 Conclusion

The clustering results show that k-means clustering algorithm can well distinguish the characteristics of users' load curve, so as to realize the load characteristic analysis of users' load curve.

Through the cluster analysis of 2,498 power load curves of power households, it is found that there are 5 distinct power consumption characteristics. On the premise of considering the system load, the above five categories can be classified into three types: peak load type, partial peak load type and stable user type. In the open market environment of the selling power side, the selling power company designs differentiated electricity services for different types of users to occupy more electricity selling market.

Acknowledgment

This work is supported by the Science Technology Project Fund of the State Grid Corporation of China "Investigation of Retail Electricity Price Policies and its Application under the Reform of Electric Sale Side" (No. SGNY0000CSJS1800046).

References

1. J. Ying, W. Zhiqiang, D. Bo. Research and Analysis of Residential Electricity Consumption Behavior Based on Big Data [J]. ELECTRIC POWER ICT, 2015, 13(11) :7-11.
2. Y. X. Li, S. L. Fang, L. Q. Yu. et al. Power system short-term load forecasting based on fuzzy clustering analysis and BP neural network[J] Power System Technology, 2005, 29(1) :20-23.
3. Tranchita C., Torres. Soft computing techniques for short term load forecasting[C]. 2004 IEEE PES Power Systems Conference & Exposition, New York, USA, 2004, :1497-502.
4. Osowski S., Siwek. The selforganizing neural network approach to load forecasting in the power system[C]. International Joint Conference on Neural Networks. Washington DC, USA, 1999, 5 :3401-3404.
5. Beccali M., Cellura M., Brano V. L.. Forecasting daily urban electric load profiles using artificial neural networks[J]. Energy Conversion and Management, 2004, 45(18) :2879-2900.
6. Fatima R, Jorge D, Vera F.. A comparative analysis of clustering algorithms applied to load profiling[C]. The Third International Conference on Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany, 2003 :73-85 .
7. Duarte F. J., Rodrigues F., Figueiredo V.. Data mining techniques applied to electric energy consumers characterization[C]. Proceedings of the Seventh International Conference on Artificial Intelligence and Soft Computing, Banff, Canada, 2003 :105-110.
8. Y. Jiang, Z. Q. Wang, B. Dai. Research and Analysis of Household Electricity Consumption Habits Based on Big Data [J]. ELECTRIC POWER ICT, 2015, 13(11): 7-11.
9. Y. Sun, T. T. Zang, F. Jiang. Analysis of Enterprise Power User Load Characteristics under the Background of Big Data [J].Statistics & Decision, 2018, 34(08): 186-188.