Research on Smart Grid Users' Power Consumption Behavior Classification Based on Improved k-Means Algorithm

Yi Sun, Mengyang Jia, and Jun Lu
School of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China
Email: 18811361366@163.com

Baogang Zhang[1] and Wanqing Yang[2]
[1]State Grid Dalian Electric Power Supply Company, Dalian, China
[2]State Grid Liaoning Electric Power Supply Co., Ltd., Dalian, China
Email: C292519601@163.com

***Abstract*—Now, the number of smart gird users is increasing. The classification of smart grid users has become the basis of user behavior analysis, load forecasting and demand response. This paper improves the traditional k-means algorithm that is the most common users' classification algorithm. This improved k-means algorithm uses a new distance calculation method to replace the Euclidean distance. The paper firstly uses two distance calculation methods to process the same set of users load data. Then two different cluster results of smart grid users are acquired. Finally, we uses the cluster validity index Mean Index Adequacy (MIA) to evaluate two results that respectively get from traditional k-means algorithm and improved k-means algorithm. This simulation verifies that the improved k-means algorithm is better than the traditional k-means algorithm. The improved k-means algorithm not only eliminates the normalization of all samples but also makes the clustering result better. And our improved k-means algorithm can solve the smart grid users' classification problem better.**

***Index Terms*—the improved k-means algorithm, the cluster validity index, smart grid, users' classification**

I. Introduction

As the number of smart grid users increasing, users' classification has become more and more important for the users behavior analysis. This classification aims to distinguish different smart grid users' power consumption behavior. After classification, the analysis for one user's power consumption behavior transforms to the analysis for a class of users' power consumption behavior. This can improve the efficiency of analytic work. So classification algorithms of smart grid users' power consumption behavior have been widely concerned [1]-[3].

Currently, the k-means algorithm, the hierarchical clustering algorithm and the Self Organizing Feature Maps algorithm (SOM algorithm) are commonly used in smart grid users' classification. Compared with other algorithms, the k-means algorithm has some features that let this algorithm more adapt to classify smart grid users [4]-[8]. The first feature is that it has high clustering efficiency. This feature indicates that the k-means algorithm can process the large data set. The other feature is that the k-means algorithm can process different types of data. On the other hand, smart grid data have large data volume, variety data types and low-value density characteristics [9]-[11]. So our paper chooses the k-means algorithm to classify residential smart grid users.

The traditional k-means algorithm judges each user's category based on the Euclidean distance. This traditional method needs to normalize all different dimension data before using k-means algorithm. In this paper, the traditional k-means algorithm based on the Euclidean distance is improved. This paper uses a new distance calculation formula to replace the Euclidean distance in our improved k-means algorithm. This new distance calculation method doesn't need to normalize all different dimension data. And this method can greatly reduce the amount of data procession. In additional, this paper's simulation shows the clustering result based on this new distance calculation method is better than that of the traditional method.

This paper firstly introduces the improved k-means algorithm. Then we design a simulation to achieve the classification of smart grid users in a residential place. After the classification, the paper uses the cluster validity index MIA to verify that the improved k-means algorithm has the better clustering result than the traditional k-means algorithm.

II. Improved Users' Behavior Classification k-Means Algorithm

The smart gird users' behavior classification framework is as Fig. 1. The framework divides into three parts. The first part is the data acquisition system. This part responses the data collection. The second part is data

Manuscript received August 20, 2015; revised January 6, 2016.

doi: 10.18178/ijoee.4.1.6-10

processing. This part mainly responses for feature extraction and users classification. The final part is the smart grid users' behavior analysis. The third part responses to analyze the clustering result. And this part can get each class's power consumption features.

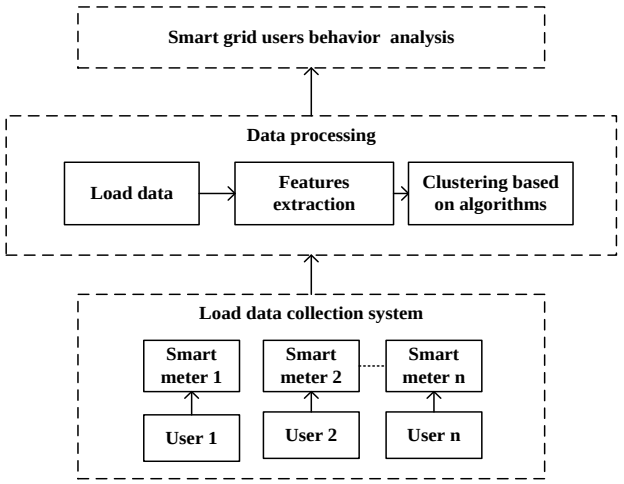

Figure 1. The smart grid users' classification framework.

A. Improved k-Means Algorithm

The thought of k-means algorithm is that firstly given initial cluster centers. Then the algorithm calculates distance between samples and clustering centers. And the algorithm judges each sample's subjection based on these distance. After that, the algorithm calculates each class's sample mean value. And let this mean value be the new clustering center. After repeated iterations, the algorithm gets final clustering centers and the clustering result. In the final clustering result, inter-class distance demands large and the inner-class distance demands small.

The main idea of the k-means algorithm is that divides all data into different classes based on distance between clustering centers and samples. So the method of distance calculation is important. The traditional k-means algorithm uses the Euclidean distance to measure each sample's subjection. In this paper, this distance calculation method been improved. The new distance calculation method considers the size of overall sample values. Detail algorithm steps are as follows [12]-[18].

Step 1: From the input data set $\{x_i\}_{i=1}^{N}$, the algorithm chooses p number of data to constitute the clustering center set $P0=\{\mu_1,\mu_2,...,\mu_p\}$. The number of samples is N. The x_i represents the k-dimensions feature vector of the sample i. This means $x_i=\{x_{ij}\}_{j=1}^{k}$.

Step 2: Calculate distance between each sample and every clustering center μ_j. The new distance calculation method is shown as the following formula (1).

$$d_m(x_i,\mu_j)=\sqrt{(x_i-\mu_j)\sum{}^{-1}(x_i-\mu_j)^T} \quad (1)$$

In the formula (1), $\sum^{-1}$ represents all samples' covariance matrix.

$$\sum{}^{-1}=\begin{pmatrix}\mathrm{cov}(x_1,x_1),\mathrm{cov}(x_1,x_2),...,\mathrm{cov}(x_1,x_N)\\ \mathrm{cov}(x_2,x_1),\mathrm{cov}(x_2,x_2),...,\mathrm{cov}(x_2,x_N)\\ \cdot\quad\cdot\quad\cdot\\ \cdot\quad\cdot\quad\cdot\\ \mathrm{cov}(x_N,x_1),\mathrm{cov}(x_N,x_2),...,\mathrm{cov}(x_N,x_N)\end{pmatrix} \quad (2)$$

The $\mathrm{cov}(x_i,x_j)$ represents the covariance value between two samples.

Step 3: Acquire each sample's category label by comparing distance.

$$\mu_j(i)\leftarrow\arg\min_i d_m(x_i,\mu_j)\ i=1,2,...,N;\ j=1,2,...,p \quad (3)$$

Step 4: Re-calculate clustering centers as following formula (4).

$$\mu_j=\frac{1}{N_j}\sum_{x_i\in\mu_j}x_i,\ \ j=1,2,...,p \quad (4)$$

In this formula, $P=\{\mu_j\}_{j=1}^{p}$ represents new clustering center set. On the other hand, N_j represents the number of users in the j class.

Step 5: Repeat carry out Step 2, Step 3 and Step 4 until the formula (5) is met.

$$\|P-P0\|<r \quad (5)$$

In the formula (5), r is the small threshold.

From above steps, we can see that the improved k-means algorithm is only different with the traditional k-means algorithm in Step 2. The traditional k-means algorithm uses the Euclidean distance [19]-[22]. The Euclidean distance calculation method is as formula (6) shown.

$$d=\sqrt{(x_i-x_j)^T(x_i-x_j)} \quad (6)$$

If we remove the covariance matrix in formula (2), we can find that formula (2) and formula (6) are same. By accessing material, we find that the Euclidean distance is a reference value [23]-[26]. It represents two samples distance when all samples probability appear. But the new distance calculation formula considers every time's input sample population. This reflects at the covariance matrix's adding. The covariance matrix can better solve the different size problem for a set of data. This can let the pattern recognition result becomes more reasonable.

B. Feature Selection

The data acquisition system collects many types of data. The maximum amount of data is users' load data. These load data have low-density value. So we need to extract some features from these load data. These extracted features can more clearly reflect users' power consumption characteristics. This paper chooses four features as following shown.

1) Peak hours power consumption rate

This feature represents the ratio of peak hour total power consumption and one day's total power consumption.

2) Flat hours power consumption rate

This feature represents the ratio of flat hour total power consumption and one day's total power consumption.

3) Valley hours power consumption rate

This feature represents the ratio of valley hour total power consumption and one day's total power consumption.

4) Daily load quantity

This feature represents one day's total power consumption.

From above four features' definitions, this paper gets the feature vector $x_i = \{x_{i1}, x_{i2}, x_{i3}, x_{i4}\}$ of the user i. This feature vector consists of four feature values. So this paper's input data set for the improved k-means algorithm is $\{x_i\}_{i=1}^{30}$.

III. Simulation and Analysis

This paper designs a simulation to verify that the improved k-means algorithm can get better clustering result than the traditional k-means algorithm in smart grid users' classification problem.

A. Data Source

This simulation's original data come from a residential place. These original data are 30 residential users' daily load data. The sampling time is March 3, 2015. And the sampling interval is one hour. Finally, each user gets 24 points load data. But this simulation's input data set doesn't directly consist of these load data. As the second part introduction, the input data set for the improved k-means algorithm are each user's feature vectors. This simulation classifies 30 users into five classes. So $p = 5$, $N = 30$ and $k = 4$.

B. Clustering Result

Using the improved k-means algorithm described in part two achieves the classification of 30 users. The clustering result is shown in the Fig. 2. In this figure, the X-axis represents sampling time in one day. The Y-axis represents Power Consumption (PC) in each hour.

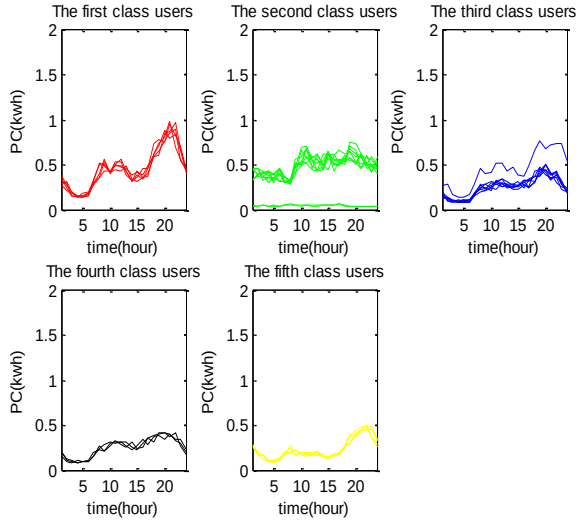

Figure 2. The clustering result based on improved k-means algorithm.

Fig. 2 shows each class users' daily load curves. And Table I shows 30 users categories.

TABLE I. Each User's Category Label

Category Label	Users ID
The first class	7,8,9,10,12
The second class	1,2,13,14,15,16,17,18,19,20
The third class	11,21,22,23,24,25
The fourth class	26,27,29
The fifth class	3,4,5,6

For analyzing each class users' power consumption behavior characteristics, this paper uses mean method to acquire each class typical daily load curve. Five classes' daily load curves are shown as following Fig. 3.

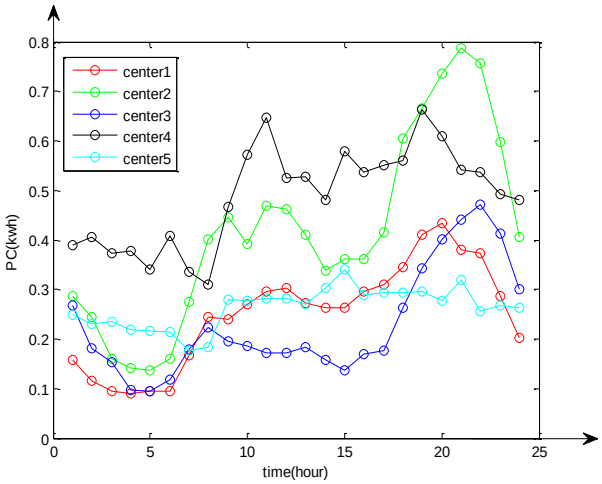

Figure 3. Five classes' daily load curves.

From Fig. 3, we can see that five classes users' power consumption wave between 0.1kwh-0.8kwh. These users are all residential customers. This simulation classifies 30 users into five classes.

The first class users and the third class users are office workers. Because these two classes users have lower power consumption. After 17:00, the power consumption increases significantly. This phenomenon coincides the working ending hour. The difference between these two classes' users reflects in the power consumption feature at 12:00. The first class users' power consumption increases at about 12:00. This indicates that the first class users are at home in 12:00. On the other hand, the third class users only have one peak time in the afternoon. This indicates that the third class users come back home at 17:00.

The fifth class users are elderly people living alone. This class's user has stable load curve in one day. And the whole day's power consumption maintain at 0.2kwh-0.3kwh. This situation coincides that the elderly concern to save electricity. So we estimate that the fifth class users are elderly living alone.

The second class users' load curves are similar to the sum of the first and fifth class users load curve. So we estimate that these users are workers and elderly people living together. And the power consumption is increasing after 17:00.

The fourth class users are business users. These users have higher power consumption than that of other four classes' users at night. And the whole day's power consumption is higher than other users too. So we estimate that these users are small shops in the residential place.

C. Evaluation of the Clustering Result

This paper calculates the cluster validity index MIA to verify that the result of improved k-means algorithm is better than that of the traditional k-means algorithm. The calculation formula of MIA is as the following formula (7) shown [27]-[30].

$$MIA = \sqrt{\frac{1}{p}\sum_{j=1}^{p} d^2(x_c, x_j)} \tag{7}$$

In the formula (7), the number of classes is p. And the clustering center set is x_c. The $d(x_c, x_j)$ represents the distance mean value that gets from each class's users load data and each class clustering center.

Final demands of clustering are inner-class distance as small as possible and inter-class distance as large as possible. And the MIA value is smaller, the clustering result is better.

From the Table II, we can see that the improved k-means algorithm has smaller MIA value. So we can think that the improved k-means algorithm this paper proposed is better than the traditional k-means algorithm in solving smart grid users' classification problem.

TABLE II. MIA VALUES BASED ON TWO ALGORITHMS

Methods	MIA
The traditional k-means algorithm	0.1229
The improved k-means algorithm	0.1118

IV. CONCLUSION

This paper uses a new distance calculation method to replace the Euclidean distance in traditional k-means algorithm. Using this improved k-means algorithm achieves 30 smart grid users' classification. Advantages of the improved k-means algorithm are reflected in two aspects. One point is that the improved k-means algorithm eliminates the normalization for different dimension features. The other advantage is that the clustering result of the improved k-means algorithm is better than that of the traditional k-means algorithm. So the improved k-means algorithm is more suited to solve the smart grid user's power consumption behavior classification problem.

ACKNOWLEDGMENT

This paper gets supported by three projects. One is 863 National High Technology Research and Development Program of China (No.2015AA050203). Other two projects are 2015 Science and technology project of State Grid China "research on smart grid users behavior theory and interaction patterns" and Fundamental Research Funds for the Central Universities of China "Key technology research for Power Grid Energy Saving based on Demand Response".

REFERENCES

[1] S. X. Zhang, J. M. Liu, and B. Z. Zhao, "Cloud computing-based analysis on residential electricity consumption behavior," *Power System Technology*, vol. 37, no. 6, pp. 1542-1546, 2013.
[2] X. G. Peng, J. W. Lai, and Y. Chen, "Application of clustering analysis in typical power consumption profile analysis," *Power System Protection and Control*, vol. 42, no. 19, pp. 68-73, 2014.
[3] Z. Liu and L. Feng, "Fuzzy-Rule based load pattern classifier for short-term electrical load forecasting," in *Proc. IEEE International Conference on Engineering of Intelligent Systems*, 2006, pp. 1-6.
[4] F. Rossi, *et al.*, "Clustering functional data with the SOM algorithm," in *Proc. ESANN*, 2004, pp. 305-312.
[5] I. Khanchouch, K. Boujenfa, and M. Limam, "An improved MULTI-SOM algorithm," *International Journal of Network Security & Its Applications*, vol. 5, no. 4, 2013.
[6] Y. Takemura, *et al.*, "Development of SOM algorithm for relationship between roles and individual's role adaptation in rugby," in *Proc. IEEE World Automation Congress*, 2014, pp. 1-6.
[7] L. Lin, *et al.*, "Differential game based centralized clustering algorithm for wireless sensor networks," in *Proc. First International Conference on Future Information Networks*, 2009, pp. 1713-1723.
[8] J. Wang, *et al.*, "A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 8, no. 3, pp. 607-620, 2011.
[9] G. Hamerly and C. Elkan, "Alternatives to the K-means algorithm that find better clusterings," in *Proc. Eleventh International Conference on Information and Knowledge Management*, 2002, pp. 600-607.
[10] J. Xu and H. Liu, "Web user clustering analysis based on K-means algorithm," in *Proc. IEEE International Conference on Information Networking and Automation*, 2010, pp. V2-6-V2-9.
[11] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571-588, 2012.
[12] D. S. Modha and W. S. Spangler, "Feature weighting in k-means clustering," *Machine Learning*, vol. 52, no. 3, pp. 217-237, 2003.
[13] A. W. Zhou, B. L. Chen, and Y. Wang, "Research and improvement of K-means algorithm," *Computer Technology & Development*, 2012.
[14] Y. Jian and H. Y. Jia, "K-Means algorithm based on data field," *Application Research of Computers*, vol. 27, no. 12, pp. 4498-4501, 2010.
[15] K. Krishna and M. N. Murty, "Genetic K-means algorithm," *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, vol. 29, no. 3, pp. 433-439, 1999.
[16] S. Saatchi and C. C. Hung. "Hybridization of the ant colony optimization with the k-means algorithm for clustering," *Lecture Notes in Computer Science*, pp. 511-520, 2005.
[17] A. Shafeeq and K. S. Hareesha, "Dynamic clustering of data with modified k-means algorithm," *International Proceedings of Computer Science & Information Tech*, pp. 221-225, 2012.
[18] X. U. Hui and L. I. Shi-Jun, "A clustering algorithm integrating particle swarm optimization and k-means algorithm," *Journal of Shanxi University*, vol. 34, no. 4, pp. 518-523, 2011.
[19] S. R. Chen and X. H. Zhu, "Research of K-means algorithm by fuzzy logic," *Computer Engineering & Science*, vol. 34, no. 12, pp. 155-159, 2012.
[20] A. Zhou, *et al.*, "The application of AGNES in K-means algorithm," *Microcomputer & Its Applications*, 2011.
[21] T. Liu, *et al.*, "An improved genetic k-means algorithm for optimal clustering," *Mathematics in Practice & Theory*, vol. 37, no. 8, pp. 793-797, 2007.

[22] H. Zhang, T. B. Ho, and M. S. Lin, "An evolutionary k-means algorithm for clustering time series data," in *Proc. International Conference on Machine Learning and Cybernetics*, 2004, pp. 1282-1287.
[23] Y. J. Bian, *et al.*, "New clustering algorithm based on improved shuffled frog leaping algorithm and k-means algorithm," *Computer Systems & Applications*, 2014.
[24] F. Y. Li and L. F. Nong, "Improved k-means algorithm based on vector semantic similarity," *Information Science*, vol. 2, p. 6, 2013.
[25] H. Ouyang, *et al.*, "Two times K-means algorithm based on grid," *Journal of Guangxi University of Technology*, vol. 1, p. 5, 2012.
[26] A. K. Jain, "Data clustering: 50 years beyond k-means," in *Machine Learning and Knowledge Discovery in Databases*, Springer Berlin Heidelberg, 2008, pp. 3-4.
[27] G. Chicco and J. S. Akilimali, "Renyi entropy-based classification of daily electrical load patterns," *IET Generation Transmission & Distribution*, vol. 4, no. 6, pp. 736-745, 2010.
[28] K. Y. Huang, "Applications of an enhanced cluster validity index method based on the fuzzy C-means and rough set theories to partition and classification," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8757-8769, 2010.
[29] I. Prahastono, *et al.*, "Electricity load profile classification using fuzzy C-means method," in *Proc. 43rd International Universities Power Engineering Conference*, 2008, pp. 1-5.
[30] L. Sun, *et al.*, "An optimized approach on applying genetic algorithm to adaptive cluster validity index," in *Proc. IEEE Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2007, pp. 582-585.

Yi Sun was born in Liaoning, China, on April 1, 1972. He is now with the Department of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China. He has worked on wireless sensor network, IoT (Internet of things), and communication in power system.

Mengyang Jia was born in May 1992. He is now a postgraduate in North China Electric Power University, research in the field of smart power. He is studying in big data and Clustering Algorithm.

Jun Lu was born in August 1976. Jun Lu is now with the Department of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China and teaches at North China Electric Power University, research in the field of smart power.