IEEE Transactions on Smart Grid, Volume 11, Issue 2, March 2020, pp. 1809 - 1811
DOI:10.1109/TSG.2020.2965801

Slope-based Shape Cluster Method for Smart Metering Load Profiles

Yue Xiang, *Member, IEEE,* Juhua Hong, Zhiyu Yang, Yang Wang, Yuan Huang, Xin Zhang, *Member, IEEE,* Yanxin Chai, Haotian Yao

***Abstract*—Cluster analysis is used to study the group of load profiles from smart meters to improve the operability in distribution network. The traditional K-means clustering analysis method employs Euclidean distance as similarity measurement, which is insufficient in reflecting the shape similarities of load profiles. In this work, we propose a novel shape cluster method based on the segmented slope of load profiles. Compared with traditional K-means and two improved algorithms, the proposed method can improve the clustering accuracy and efficiency by capturing the shape features of smart metering load profiles.**

Keywords—Cluster analysis, load profile, K-means, similarity.

I. Introduction

With the increasing popularity and penetration levels of smart meters, the load data from individual consumers can be measured to form segmented load profiles. The clustering of these load profiles can transform massive smart metering data into operational information to improve the operability in smart distribution networks, with applications in demand side management, pattern recognition of user behavior, and electricity tariff design [1]. The traditional K-means method is widely used in load profiles clustering. The Euclidean distance is used as the similarity measure between load profiles and their corresponding cluster centers [2]. However, the Euclidean distance is an 'absolute' distance which cannot distinguish the curve slopes and shapes accurately. In the operation of power systems, curve slopes represent behavior of power consumers. They drive load fluctuation which contributes to the deviation in system frequency, and thus deserves more attention. Some improved methods are proposed, such as introducing the cosine similarity as a part of similarity measurement [3], and using the discrete wavelet transform to cluster the load profile features [4]. However, the core of these methods is still the Euclidean distance. Their improved accuracy is achieved at the price of increased computational complexities. Therefore, this letter proposes a simple but efficient cluster method which employs the curve slope as similarity measurement to achieve an alternative way of load profile shape clustering.

II. Methodology

Regarding a segmented load profile $L=\{l_1, l_2,..,l_i,\ldots,l_N\}$, l_i is the load value at the i-th sampling point. The slope of each segment between load value l_i and l_{i+1} is represent as $S=\{s_1, s_2,\ldots s_i,...,s_{(N-1)}\}$. The slope has two properties: direction and steepness. The slope direction t_i for each load segment can be expressed by 1, -1, 0 as follows:

$$t_i=\begin{cases}1, s_i>0\\0, s_i=0\\-1, s_i<0\end{cases} \quad (1)$$

where $T=\{t_1, t_2,..,t_i,\ldots,t_{(N-1)}\}$ summarizes the slope direction for every segment in a load profile L. The number of same slope direction segments e between load profile a and b is calculated as follows:

$$e(T^a,T^b)=\sum_{i=1}^{N-1}\delta(t_i^a,t_i^b) \quad (2)$$

$$\delta(t_i^a,t_i^b)=\begin{cases}0, t_i^a\neq t_i^b\\1, t_i^a=t_i^b\end{cases} \quad (3)$$

It can be observed that the larger e, the higher similarity in the two load profiles. Based on the slope direction and steepness, a novel method is proposed in two main steps: cluster center search and similarity measurement.

A. Cluster center search

The traditional K-means method determines the optimal number of cluster centers by setting the total number k manually, or by the variety of indicators such as error threshold. The former k determination method is empirical thus unreliable, while the latter method relies on different external indicators thus the inconsistent results may present. In comparison, the proposed slope-based cluster method determines the optimal k solely based on load profile data itself, which improves the clustering stability and consistency.

To find the cluster centers among M load profiles, a matrix E that describes the number of same-slope-direction segments between any of two load profiles is established as Fig 1, based on the calculation of e_{pj} by (2) and (3), where p, j are the two load profile indicators, respectively.

The j-th column of matrix E reflects the the number of same-slope-direction segments between the load profile j and others, i.e., $E_j=[e_{1j},...,e_{(j-1)j},e_{(j+1)j},...,e_{Mj}]$. As illustrated in Fig. 1, the elements e_{pj} (p=1...M, $p\neq j$) are sorted from the maximum value (max(E_j)) to the minimum value (min(E_j)) among M-1 elements in E_j. Equation (4) is used to quantify the deviation f_j of data set E_j to its max value. And, the element in E_j would be classified into one of the four blocks: 1) highly similar block A when e_{pj}=max(E_j), 2)similar block B when e_{pj} belongs to interval (max(E_j) – f_j , max(E_j)), 3) dissimilar block C when e_{pj} belongs to interval (min(E_j), max(E_j) – f_j], and 4) highly dissimilar block D when e_{pj}=min(E_j). To unify the influence of other matrix columns on similarity blocks classification, the global average deviation f_{ta} is defined in (5).

This work was supported by the Sichuan Science and Technology Program (2019YFH0171), the Fundamental Research Funds for the Central Universities of China (YJ201654), and the International Visiting Program for Excellent Young Scholars of Sichuan University. Juhua Hong is the corresponding author.

Yue Xiang, Zhiyu Yang, Yang Wang, Yuan Huang, Yanxin Chai, Haotian Yao are with the College of Electrical Engineering, Sichuan University, Chengdu 610065, China. (e-mail: xiang@scu.edu.cn; scu_yzy@sina.com; yangwang@stu.scu.edu.cn; yuanhuang@scu.edu.cn; ziqi_chai@163.com; yaoht97@163.com).

Juhua Hong is with Fujian Power Economic Research Institute, Fuzhou 350012, China. (e-mail:juhuahong@foxmail.com).

X. Zhang is with Energy and Power Theme, School of Water, Energy, Environment, Cranfield University, Cranfield MK43 0AL, U.K. (e-mail: Xin.Zhang@cranfield.ac.uk)

$$f_j = \sqrt{\frac{1}{M-1}\sum_{p=1}^{M-1}(e_{pj} - \max(E_j))^2} \quad (4)$$

$$f_{\text{ta}} = \frac{1}{M}\sum_{j=1}^{M} f_j \quad (5)$$

Let the column j (j=1,…, M) which contains the maximum max(E_j) in matrix E be as the 1st cluster center. The most dissimilar one from the 1st cluster center can be found with the value corresponding to the min (E_j) in column j as the 2nd cluster center. Then, remove the similar load profiles belonging to Block A and B of the 1st cluster center and the 2nd cluster center from the matrix E and then determine the 3rd cluster center from Block D of the 2nd cluster center. Finally, repeat the same cluster process on the remaining load profiles, until there is no load profile in Block C and D of the last center found. The process is shown in Fig.1.

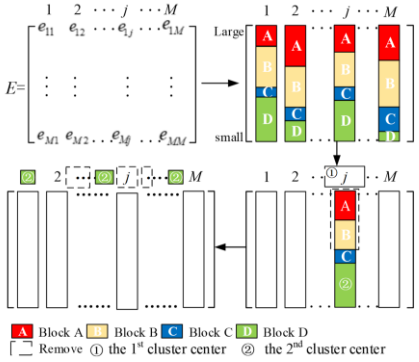

Fig.1 Process of determining initial conditions

B. Similarity Measurement

For a load profile L^a, an initial cluster analysis will compare L^a with all initial k cluster centers in terms of segmented slope direction similarity e_{ak} as illustrated in (2), and produce the similarity blocks from [e_{a1}, e_{a2}, ..., e_{ak}]. It is possible that more than one cluster centers fall into A or B block, therefore, it is necessary to make an additional similarity measurement based on slope steepness of segmented load profile.

Take a cluster center L^q from the similar block A or B and compare with load profile L^a. $|s_i^a|$ shows the slope steepness of segment i in L^a. Based on (3), $|s_i^a - s_i^q| \times \delta(t_i^a, t_i^q)$ compares the slope steepness of the same-slope-direction segments between L^a and L^q. Equation (6) compares the average slope steepness S_{same}^{aq} in same-slope-direction segments, and in different-slope-direction segments S_{diff}^{aq} between load profile L^a and cluster center L^q.

$$\begin{cases} S_{same}^{aq} = \frac{1}{e_{aq}}\sum_{i=1}^{N}|s_i^a - s_i^q| \times \delta(t_i^a, t_i^q) \\ S_{diff}^{aq} = \frac{1}{N-1-e_{aq}}\sum_{i=1}^{N}|s_i^a - s_i^q| \times \left(1 - \delta(t_i^a, t_i^q)\right) \end{cases} \quad (6)$$

We consider that the more similarity between load profile L^a and cluster center L^q, the smaller values should be observed in both S_{same}^{aq} and S_{diff}^{aq}. The criteria have been set to find the most suitable cluster center for load profile L^a in (7). It is assumed that there are t cluster centers fall into the A and B blocks of L^a based on the similarity of segmented slope direction. The S_{same}^{at} and S_{diff}^{at} are calculated between L^a and all the t cluster centers, from which their minimum values are denoted as min (S_{same}^{at}) and min (S_{diff}^{at}). Based on (4) and (5), the average deviations against minimum values are also calculated as $u_{s,ta}$ and $u_{d,ta}$ respectively.

$$\begin{cases} L^q \in X, S_{\text{same}}^{aq} < \min(S_{same}^{at}) + u_{s,ta} \\ L^q \in Y, S_{diff}^{aq} < \min(S_{diff}^{at}) + u_{d,ta} \end{cases} \quad (7)$$

Set X contains the cluster center(s) with the most same-slope-direction segments, whose average slope steepness is close to load profile L^a, while Y corresponds to the different-slope-direction. As a matter of fact, the difference of their different segments reflects different electricity consumption behavior. Thus, when it is difficult to be classified, the smallest difference of different segments should be taken as a measure index, so that the similar electricity consumption behavior could be classified as well. Let $Z=X\cap Y$, if $|Z|$=1, the intersection is the most suitable center; if $|Z|$ =0, the center is indicated by min (S_{diff}^{at}) among t cluster centers; if $|Z|$>1, the center is denoted as min (S_{diff}^{at}) among the intersections.

III. Case Study

The proposed cluster method is tested with the load profile data from [5], covering 2000 customers at half hour resolution in 2010. The number of the load profiles is about 730000. Then 6 initial cluster centers could be obtained base on the method proposed in II.A, as shown in Fig. 2.

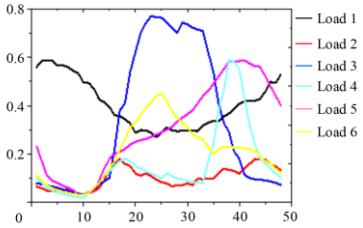

Fig.2 Initial clustering center

The load profiles are clustered into six groups. The load profiles of Load 1 are diverse, and there is no fixed power consumption mode, but the average power consumption is high. Load 2 has two closer load peaks in the morning and evening, which is consistent with the workers' load pattern. Load 3 has one load peak during the day time and last for a long time; Load 4 also has two peaks with different load levels, and the higher appears at night; Load 5 has one load peak in the evening, and one peak appears in the morning for Load 6.

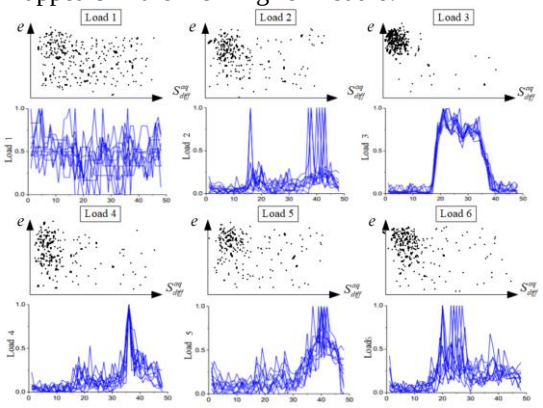

Fig.3 Clustering results

Take the average slope steepness in different-slope-direction segments as the abscissa and the number of same slope direction segments as the ordinate, each cluster center as the origin. The load profiles are expressed by (e , S_{diff}^{aq}). As shown in Fig. 3, the load profiles are mainly concentrated on the upper left part of the coordinate diagram and Load 1 are more dispersed. The larger e, the smaller S_{diff}^{aq}, the more similar the load profiles are. In addition, the higher the intra class similarity, the higher the density of points. Therefore, the validity of load profiles clustering by these two indexes is verified.

In this case, we compare the accuracy of the proposed method with the traditional K-means and improved K-means [4]. The cluster results based on three different cluster methods for the same load profile are shown in Fig.4. The black curve represents an original load profile L^a taken from a smart meter. Blue curve is the center of a cluster where the L^a is assigned based on the slope-based cluster method, While the red and green curves are cluster centers based on K-means and advanced K-means methods. In order to compare the clustering accuracy, the indices [6]: Peak Magnitude Error Index (PMEI), Maximum Magnitude Error (MME), Mean Absolute Percentage Error (MAPE), Root Mean Square Error(RMSE), Peak Time Error (PTE) are calculated and compared in Table I.

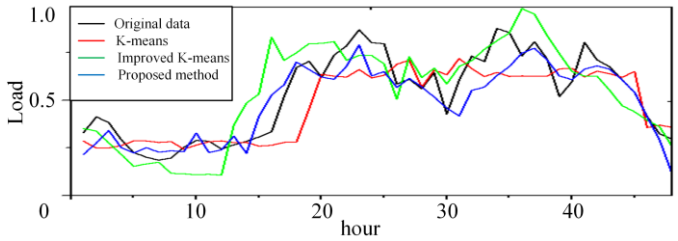

Fig.4 Comparison of cluster centers with a real load profile

TABLE I
COMPARISON OF THREE CLUSTER METHODS BY THE ERROR INDICES

Algorithm	PMEI (%)	MME (%)	MAPE (%)	PTE (hour)	RMSE
Traditional	18	58	0.19	3	0.76
Improved [4]	12	62	0.26	1	0.79
Proposed	18	34	0.15	1	0.57

Fig.4 shows that the cluster center calculated from traditional K-means method (red line) cannot reflect the peaks and truffles of original load profile L^a. The improved K-means method (green line) is able to capture the spikes of original load profile, however the cluster center does not follow the load profile well. The center from the slope-based cluster method (blue line) not only replicates most of spikes in load profile L^a, but also correctly reflects the load profile change (the shape) in both direction and steepness. Such observations are supported by the results from table I.

From Table I, it can be seen that the improved algorithm proposed in [4] only has certain advantages in some aspects. It is because that the algorithm partially improves the traditional K-means and does not involve its distance measurement. Therefore, for the sake of fairness, the proposed algorithm is compared with another algorithm [7], which replaces the Euclidean distance formula with the Pearson correlation coefficient. The specific process is detailed in literature [7]. The comparison results are shown in Table II.

TABLE II COMPARISON OF TWO CLUSTER METHODS BY THE ERROR INDICES

Algorithm	PMEI (%)	MME (%)	MAPE (%)	PTE (hour)	RMSE
Improved [7]	14	60	0.17	1	0.71
Proposed	18	34	0.15	1	0.57

The results show that the performance of the two algorithms is similar, but the proposed algorithm still has certain advantages. This is because the two methods are essentially shape clustering based on vector direction and angle rather than value of Euclidean distance in traditional clustering. The difference is that the object of the proposed algorithm is not the whole curve, but focuses on the similar and different segments, so that the classification of curves can be more precise.

The computation efficiency of the proposed method is compared with the traditional K-means in number of iterations and computation time in table III.

TABLE III COMPARISON OF THE COMPUTATION EFFICIENCY

Algorithm	Number of Iterations	Computation time
Traditional	78	764s
Proposed	23	461s

Table III shows that the proposed method achieves clustering faster than the traditional K-means in this case, which indicates the higher computation efficiency of the proposed method.

IV. CONCLUSION

This letter proposes a novel cluster method based on segmented slope of load profiles. Case study verifies that the proposed method has improved accuracy and computation efficiency in clustering the smart metering load profiles, particularly in shape clustering.

REFERENCES

[1] Wang Y, Chen Q, Hong T, Kang C, "Review of smart meter data analytics: applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125-3148, 2019

[2] Yu Z, "A temperature match based optimization method for daily load prediction considering DLC effect," *IEEE Transactions on Power Systems*, vol. 11, no. 2, pp. 728-733, 1996.

[3] Wang X, Chen Z, Peng X, "A new combinational electrical load analysis method based on bilayer clustering analysis," *Power System Technology*, vol. 40, no. 5, pp. 1495-1501, 2016.

[4] Li R, Li F, Smith ND, "Load characterization and low-order approximation for smart metering data in the spectral domain, " *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 976-984, 2017.

[5] CER Smart Metering Project. [Online][access in 2018]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[6] Zhong S, Tam K. S, "A frequency domain approach to characterize and analyze load profiles," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 857-865, 2012.

[7] Bu F, Chen J, Zhang Q, et al, "A controllable refined recognition method of electrical load pattern based on bilayer iterative clustering analysis," *Power System Technology*, vol. 42, no. 3, pp. 904-910, 2018.