DTW-based Adaptive K-means Algorithm for Electricity Consumption Pattern Recognition

Yimeng Shen, Yiwei Ma, Hao Zhong, Miao Huang, and Fuchun Deng

Abstract—**The research on electricity consumption pattern recognition generally encounters some prominent problems such as poor similarity, poor accuracy, and low efficiency of existing clustering algorithms. Therefore, this paper utilizes elbow judgment (EJ), gap statistic (GS), and DTW (dynamic time warping) to develop a DTW-based adaptive K-means (DAKM) clustering algorithm for electricity consumption pattern recognition. The algorithm includes three main aspects. First, the DTW distance with the Sakoe-Chiba band global constraint is used to find the optimal alignment between the two load curves by matching the shapes with local stretching or compression sequences. Second, gap statistic and elbow are used to obtain the optimal number of clusters for high clustering efficiency automatically. Third, a max-min DTW distance (MMDD) method is presented to optimize the initial cluster centers of the K-means algorithm. The comparative experimental results demonstrate that the proposed DAKM algorithm achieved best evaluation values of 0.7055 for DBI, 0.0237 for SSE, 132.0435 for CHI, 0.6649 for SC, and 1.1670 for DI, respectively, which proves that the proposed DAKM algorithm is far superior to other clustering algorithms.**

Index Terms—**Dynamic time warping, K-means, pattern recognition, gap statistic, elbow judgment**

I. Introduction

WITH the development of advanced information technology, a large amount of data is generated in various fields every day. It is crucial to collect and differentiate these big data to help service providers improve their operational performance and gain a competitive advantage in the fierce market. In the competitive electric power market, it is highly desirable for power utilities to know the electricity consumption patterns for improving the demand management service and achieving reliable and economical operation [1]. The widespread development of smart meters has enabled consumers' electricity data to be collected and recorded at regular intervals (15 min., or 5 min.), which provides convenience to reveal the load patterns on the demand side [2]. These load data help to classify the consumption behavior of electricity users, also known as load curve patterns. As a result, electricity end-users can respond to electricity price signals by understanding their own electricity consumption patterns and reduce electricity bills [3], while electricity utilities can achieve effective peak shaving and flexible pricing through electricity consumption patterns [4].

Given the significant time-varying nature characteristics of electricity consumption, it is crucial to select an appropriate clustering algorithm for daily load curve pattern extraction. At present, various classic clustering algorithms are commonly used [5], including K-medoid [6], K-means [7], hierarchical clustering (HC) [7], and Support Vector Clustering (SVC) [8], etc. These classical clustering algorithms suffer from poor accuracy and low efficiency in the electric load curve clustering. Among these algorithms, the K-means algorithm has attracted increasing attention in load curve clustering [9], [10], [11], as it has outstanding advantages such as simplicity, efficiency, and validity. Figueiredo et al. [12] used the K-means algorithm to identify electricity consumption patterns after reducing the dimensionality of the initial dataset through self-organizing mapping, and concluded that the K-means algorithm performed very well in comparative experiments on datasets with continuous attributes. J. Kwac, et al. [13] presented an improved K-means algorithm with a mean squared error threshold to find representative load shapes, and used a traditional hierarchical clustering algorithm to improve the optimal distance between cluster centers. Moreover, Jiang et al. [14] considered the special characteristics of load curves such as high dimensionality and big volume, and proposed a fused K-means algorithm based on discrete wavelet transform (DWT) to identify load curve patterns. This fused clustering algorithm first used DWT to reduce the dimensionality of daily load curves, and then used the K-means algorithm to achieve load curve clustering, which effectively reduced the computational complexity. The optimal number of clusters is determined by the simplified Silhouette width criterion. However, those K-means algorithms mentioned above did not consider the optimization of some key parameters that affect the clustering performance, such as the optimal number of clusters and initial cluster centers. Furthermore, there is no

Manuscript received June 27, 2024; revised November 18, 2024.

This work was supported in part by the National Natural Science Foundation of China (No.61703068), and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN202304206).

Yimeng Shen is a postgraduate student of Electrical Engineering Department, School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (e-mail: S220303010@stu.cqupt.edu.cn).

Yiwei Ma is an associate professor at the School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (Corresponding author to provide e-mail: mayw@cqupt.edu.cn).

Hao Zhong is an associate professor at the Hubei Provincial Key Laboratory for Operation and Control of Cascaded Hydropower Station, China Three Gorges University, Yichang 443002, China. (e-mail: zhonghao022@163.com).

Miao Huang is a senior engineer at the School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (e-mail: huangmiao@cqupt.edu.cn).

Fuchun Deng is a lecturer at the Chongqing College of Finance and Economics, Chongqing 402160, China. (e-mail: 1225735111@qq.com).

comparative validation of the superiority of the K-means algorithm in the literature.

To improve the clustering performance, various improved K-means algorithms have been proposed in the literature. For example, the effectiveness of the K-means algorithm is improved by finding the initial centroid point [15], and an adaptive K-means clustering under-sampling algorithm is presented by considering the variations in dataset type and sampling features to calculate the optimal *K* value [16]. However, these K-means algorithms cannot achieve high-quality pattern recognition of electric load curves, as they are not suitable for time series classification problems. Therefore, this paper proposes a DTW-based adaptive K-means (DAKM) algorithm for electricity consumption pattern recognition. The main contributions of this research are as follows:

(i) This paper presents dynamic time warping (DTW) distance with the Sakoe-Chiba band global constraint as a similarity distance measure for the K-means algorithm to explore the best match between different electric load curves.

(ii) Gap statistic (GS) and elbow judgment (EJ) are used to calculate the optimal K value for the K-means algorithm automatically.

(iii) A max-min DTW distance (MMDD) method is presented to find the optimal initial cluster centers for the K-means algorithm.

(iv) The DAKM algorithm is presented to extract electric load curve patterns.

The rest of the paper is summarized as follows: Section II introduces the traditional K-means algorithm, and Section III proposes the DTW-based adaptive K-means algorithm. In Section IV, the effectiveness of the proposed DAKM clustering algorithm is demonstrated through comparative experiments. Finally, Section V provides the conclusion.

II. K-means Algorithm

A. Basic Principle

K-means algorithm is a simple and popular unsupervised clustering algorithm for big data analysis. It can divide the dataset by making the data samples within the same class have higher similarity and the data samples between different classes have dissimilarity [9]. Assuming that a numerical dataset $G=\{g_1,g_2,\cdots,g_i,\cdots,g_n\}$ represents the set of clustered objects in a *d*-dimensional Euclidean space R^d , and $C=\{c_1,c_2,\cdots,c_k,\cdots,c_K\}$ is *K* cluster centers. d_{ik}^{ED} denotes the Euclidean distance between g_i and c_k . $U=[u_{ik}]_{n\times K}$ is the affiliation matrix, where u_{ik} is a binary variable (i.e. $u_{ik}\in\{0,1\}$) that determines whether g_i belongs to the *k*-th cluster. Hence, the K-means algorithm can achieve iterative learning by minimizing its objective function $J(U,C)$ to update the equations of cluster centers and membership degrees.

$$J(U,C)=\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}\cdot d_{ik}^{ED} \tag{1}$$

$$d_{ik}^{ED}=\|g_i-c_k\| \tag{2}$$

$$c_k=\sum_{i=1}^{n}u_{ik}\cdot\frac{g_{ik}}{\sum_{i=1}^{n}u_{ik}} \tag{3}$$

$$u_k=\begin{cases}1\,, & if\ (d_{ik}^{ED})^2=\min\limits_{1\le k\le c}(d_{ik}^{ED})^2\\ 0\,, & otherwise\end{cases} \tag{4}$$

The calculation flowchart of the traditional K-means algorithm is shown in Fig. 1, and its detailed clustering process is described as follows.

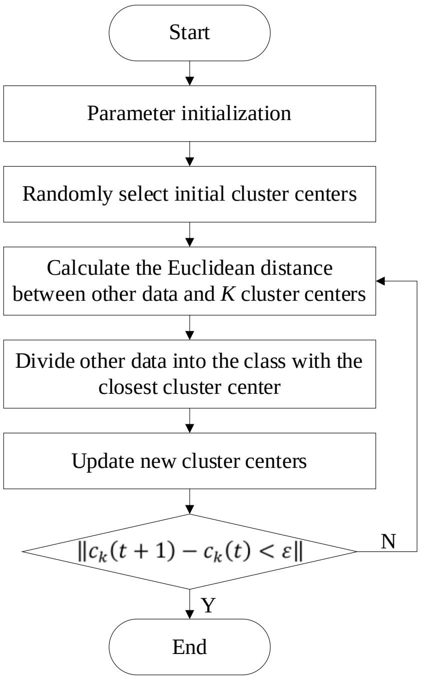
Start
Parameter initialization
Randomly select initial cluster centers
Calculate the Euclidean distance between other data and *K* cluster centers
Divide other data into the class with the closest cluster center
Update new cluster centers
$\|c_k(t+1)-c_k(t)<\varepsilon\|$
N
Y
End

Fig. 1. Calculation flowchart of the traditional K-means algorithm.

1) Firstly, randomly select *K* samples from the dataset as initial cluster centers.

2) Secondly, calculate the Euclidean distance between other samples and initial cluster centers separately, and use these samples as the category of their closest cluster center.

3) Thirdly, calculate the average value of each category for the classified samples mentioned above, and determine their new cluster centroids.

4) Then, compared with the *K* cluster centroids obtained from the previous calculation. If the centroid of the cluster is not changed, proceed to the next Step 5; otherwise, return to Step 2.

5) Finally, end and output the final clustering result when each newly generated cluster is consistent and all sample points will not transfer from one cluster to another.

B. Disadvantages Analysis

In the K-means algorithm, some prominent challenges cannot achieve satisfactory clustering results. It mainly includes three aspects of issues:

(i) Since the K-means algorithm requires a pre-set *K* value for the number of clusters, but it is generally not clear in advance how many clusters the dataset should be divided into. Therefore, choosing the optimal *K* value is very difficult and important.

(ii) The initial cluster centers of the K-means algorithm are randomly selected, which may lead to errors in clustering results or slow convergence. This poses some uncertain risks for good clustering performance.

(iii) The Euclidean distance is commonly regarded as the default distance metric in the K-means algorithm, which makes it difficult to implement time series pattern recognition problems such as the electric load curves.

As shown in Fig. 2, the defect dataset of a certain product based on the K-means algorithm has four distinct clusters with relatively concentrated data distributions. However, due to the influence of cluster centers and the number of clusters, the defect database of this product is divided into three different clusters by the traditional K-means algorithm. It is obvious that the two different clusters at the bottom are mistakenly divided into the same elliptical cluster.

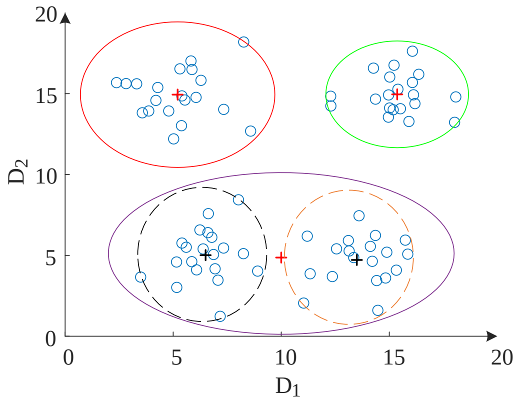

Fig. 2. Clustering of product defect dataset based on K-means.

III. DTW-Based Adaptive K-Means Algorithm

To obtain high-quality clustering results of electricity consumption pattern recognition, DTW distance with global constraint of S-C band, optimal number of clusters based on GS-EJ, and MMDD-based initial cluster centers are used together to create a DTW-based adaptive K-means (DAKM) algorithm.

A. DTW Distance with Global Constraints

DTW is a classical elastic measure of time series similarity measures, which can reflect the overall similarity between different electric load curves. For any given two load curve time series $X=[x_1,x_2,\cdots,x_n]$ and $Y=[y_1,y_2,\cdots,y_m]$, we can define a warping path to reflect their overall similarity between time series. A warping path $W=[w_1,w_2,\cdots,$ $w_l,\cdots,w_r]$ is composed of adjacent elements in the distance matrix $D\in R^{n\times m}$, where r is the total number of elements in the warping path. $w_l=(i,\ j)$ is the coordinate of the l-th element on the path *W*. The following three constraints should be satisfied in the warping path *W* [17].

(1) Boundary conditions: The warping path *W* starts from $(x_1,\ y_1)$ and ends with $(x_n,\ y_m)$. $w_1=(1,1)$ and $w_r=(n,m)$.

(2) Continuity constraint: There must be an upward, downward or diagonal adjacency between the two elements. Denoted by $w_{l+1}-w_l\in\{(0,1),(1,0),(1,1)\}$.

(3) Monotonicity constraint: The elements of the warping path W increase monotonously in the time dimension, which means $i_l\le i_{l+1}$ and $j_l\le j_{l+1},\forall l\in(1,2,\cdots,r)$.

To improve the DTW calculation speed and accuracy, the S-C band is usually introduced to help DTW find an optimal warping path W [18]. The principle is shown in Fig.3, L_{ul} and L_{ll} are denoted as the upper and lower boundary lines in the the S-C band, where λ_u and λ_l are the coefficients of allowed warping.

$$L_{ul}:y=\frac{m}{n}x+\lambda_u m \tag{5}$$

$$L_{ll}:y=\frac{m}{n}x-\lambda_l m \tag{6}$$

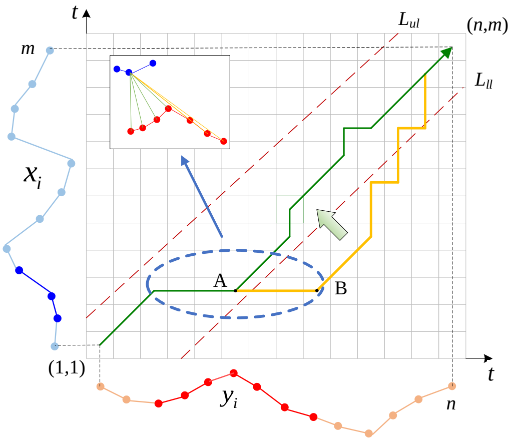

Fig. 3. DTW warping path under the global constraint of the S-C band.

The DTW distance $D_{DTW}^{gc}(X,Y)$ with global constraints of the S-C band can be calculated by constructing a cumulative cost matrix *L* based on the dynamic programming method.

$$D_{dtw}^{gc}(X,Y)=d_{ed}(x_n,y_m)+\min(L_{n-1,m-1},L_{n,m-1},L_{n-1,m}) \tag{7}$$

$$d_{ed}(x_i,y_j)=\begin{cases} d(x_i,y_j) & \lambda_l m\le x_i-\frac{m}{n}y_j\le\lambda_u m \\ +\infty & otherwise \end{cases} \tag{8}$$

$$L_{i,j}=d_{ed}(x_i,y_j)+\min(L_{i-1,j-1},L_{i,j-1},L_{i-1,j}) \tag{9}$$

Where, $D_{DTW}^{gc}(X,Y)$ is the DTW distance with the global constraint of S-C band between load curves *X* and *Y*; $d_{ed}(x_i,y_j)$ is the Euclidean distance between load curves elements x_i and y_j; $L_{i,j}$ represents the element in the *i*-th row and *j*-th column of the cumulative cost matrix L, where $L_{i,0}=L_{0,j}=+\infty$ and $L_{0,0}=0$.

B. Optimized Number of Clusters

It is well known that the optimal number of clusters is very important in clustering algorithms. but artificially preset value may lead to poor performance in load curve patterns. Therefore, an optimized method combining gap statistic (GS) method and elbow judgment (EJ) method is proposed to find the optimal number of clusters automatically. According to Fig. 4, the basic idea of this method is to comprehensively utilize the unique advantages of GS and EJ in determining different numbers of clusters $K=[1,2,\cdots,z]$, where $z\in[1,n]$. Specifically, GS is used to calculate whether the value of *K* is 1 [19], otherwise, EJ is adopted to calculate the optimal number of clusters greater than 1.

Step 1: GS-based optimal number of clusters $K=1$.

GS has excellent performance when the number of clusters is 1, so it is only used to determine whether the value of K is 1 by the following equations. Otherwise, it will automatically proceed to the next step.

$$Gap(K) \geq Gap(K+1) - s(K+1) \tag{10}$$

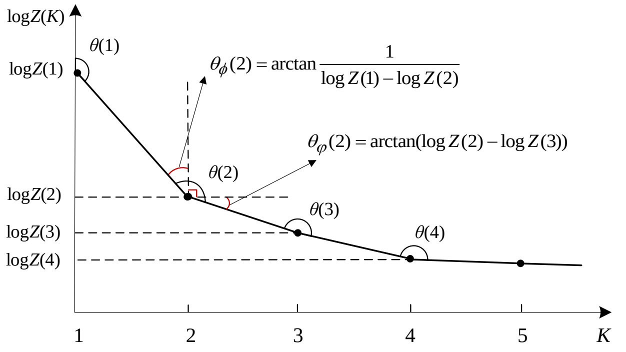

Fig. 4. The optimal number of clusters based on GS and EJ.

$$Gap(K) = \frac{1}{H}\sum_{h=1}^{H} \log Z_h^*(K) - \log Z(K) \tag{11}$$

$$Z(K) = \sum_{K=1}^{z} \frac{1}{2n_r} \sum_{i,j \in c_r} \left\| x_i, y_j \right\| \tag{12}$$

$$s(K) = \sqrt{1 + \frac{1}{H}} \cdot \sqrt{\frac{1}{H}\sum_{h=1}^{H} (\log Z_h^*(K) - \frac{1}{H}\sum_{h=1}^{H} \log Z_h^*(K))^2} \tag{13}$$

Where, $Gap(K)$ is the estimated gap in the cluster K ; $Z_h^*(K)$ and $Z(K)$ are the within-dispersion measurements in the reference and real datasets, respectively; $Z(K)$ is the within-dispersion measurement; h is the number of Monte Carlo sampling, where $h = [1, 2, \cdots, H]$; n_r is the number of samples in the r-th cluster; c_r is the r-th cluster center in the cluster C ; x_i and y_j are two curves in the r-th cluster; $s(K)$ is the simulation error.

Step 2: EJ-based optimal number of clusters $K > 1$.

EJ is used to calculate the optimal number of cluster K when it is larger than 1. Fig. 4 shows that when $K > 1$, the elbow angle $\theta(K)$ at each inflection point of the curve is composed of three parts: $\theta_\phi(K)$, $\theta_\varphi(K)$, and $\pi/2$. The value of K is considered as the optimal number of clusters if it satisfies Eq. (14) [20].

$$\theta(K) < \theta(K+1) \tag{14}$$

$$\theta(K) = \pi/2 + \theta_\phi(K) + \theta_\varphi(K) \tag{15}$$

$$\theta_\phi(K) = \arctan \frac{1}{\log Z(K-1) - \log Z(K)} \tag{16}$$

$$\theta_\varphi(K) = \arctan(\log Z(K) - \log Z(K+1)) \tag{17}$$

C. Optimized Initial Clustering Center

In order to reduce the undesirable effects of randomly selecting initial clustering centers., a max-min DTW distance (MMDD) method is presented to find the optimal initial cluster centers for higher clustering quality. Its basic strategy is to select two load curve samples with the minimum DTW distance and use any one of them as the first initial cluster center c_1^*, and then select the one with the maximum DTW distance from c_1^* as the second cluster center c_2^*. When the value of K is greater than 2, other new initial cluster centers can be obtained by the following formula [21].

$$\max[\min[d_{i,1}^{DTW}, \cdots, d_{i,k}^{DTW}, \cdots, d_{i,K-1}^{DTW}]] > \theta \cdot \overline{D_C^{DTW}} \tag{18}$$

$$\overline{D_C^{DTW}} = \frac{1}{\varphi}\sum_{k=1}^{K} \left\| c_{k+1}^* - c_k^* \right\| \tag{19}$$

Where, $d_{i,k}^{DTW}$ is the DTW distance with the global constraint between the centroid k and sample i; θ is the constant coefficient; $\overline{D_C^{DTW}}$ is the average DTW distance between all different centroids; φ is the total number of combinations of any two cluster centers; c_k^* is the k-th initial cluster center.

D. The Proposed DAKM Algorithm

Based on the three key parameters established above, the DAKM algorithm is developed for electric load curve pattern extraction. The detailed program steps of DAKM algorithm are described in Algorithm 1. In DAKM algorithm, the basic parameters are first initialized, such as the sample sets G, the coefficients of allowed warping λ_u and λ_l, the threshold for initial cluster centers θ, and the number of iterations T. The DTW distance $D_{DTW}^{gc}(X,Y)$ is calculated in advance by Eq. (7)-(9), and then the value of the number of clusters K is increased from 1 and the optimal value is determined by Eq. (10)-(17). Subsequently, the optimal initial cluster center c_k is determined by Eq. (18) and (19). Finally, if the algorithm calculates that the center of clusters is not changing, and then outputs the adaptive cluster number K and load curve pattern assignment matrix S. Otherwise, it continues with the iterative calculation.

```
Algorithm 1   DAKM clustering algorithm
Input:    $G, \lambda_u, \lambda_l, \theta, T$
Output:   $C, S$
Step 1:   Calculate the DTW distance:
          First, perform global constraints $L_{ul}$ and $L_{ll}$ by
          Eqs. (5)-(6), and then Calculate DTW distance
          $D_{DTW}^{gc}(X,Y)$ by Eqs. (7)-(9).
Step 2:   Calculate the optimal number of clusters $K$:
          First, judge if the cluster number $K$ is 1 by Eqs.
          (10)-(13), otherwise calculate the cluster number
          $K$ ($K$>1) by Eqs. (14)-(17).
Step 3:   Calculate the initial cluster center $c_k$ :
          Calculate initial cluster centers $C$ by Eqs. (18)
          -(19).
Step 4:   Run DAKM algorithm
          for $t = 1 \rightarrow T$ do
              Calculate $D_{DTW}^{gc}(x_i, a_k)$ between $x_i$ and $a_k$.
              Divide $x_i$ into the nearest cluster class.
              Update cluster center matrix $C$ and matrix $U$
          by Eqs. (3)-(4).
              if ($t \geq T \,||\, \|a_k(t+1) - a_k(t)\| < \varepsilon$ ) then
                  break
              else
                  Continue to iterate in the loop.
              end if
          end for
Step 5:   Return Adaptive cluster number $K$ and load
          curve pattern assignment matrix $S$.
```

IV. Experimental Analysis

A. Data Description

The electric load data of a certain zone in Chongqing city from April 23, 2023 to July 31, 2023 were selected as experimental data. In this zone, there are two small kitchenware factories, 25 electric vehicle (EV) charging points, and 2000 residential users. A sample of 82 daily electric load curve profiles is shown in Fig. 5. Data preprocessing is required before data clustering, which involves removing erroneous data and filling in missing data.

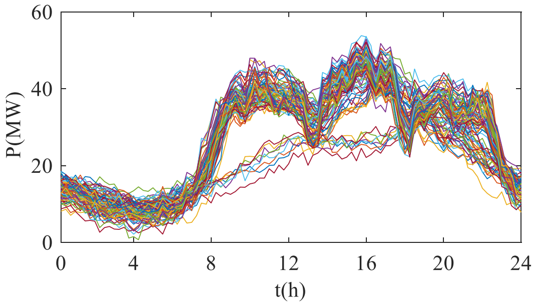

Fig. 5. 82 sampled daily electric load curve profiles.

To verify the effectiveness of the proposed DAKM algorithm more comprehensively, three comparative experiments were conducted: (i) Clustering results and comparative analysis between DAKM and traditional DTW-based adaptive K-means (TDAKM), (ii) Comparison of clustering validity, and (iii) Comparison of clustering efficiency. In the experiments, we initialized the relevant parameters based on experiential knowledge: $\lambda_u, \lambda_l = [0.1, 0.2, \cdots, 1]$, T=100, and $\theta \in [0.5, 1)$. The experimental environment was a computer equipped with an Intel® Core (TM) i5-8300H CPU @ 2.30GHz and 16.0 GB memory, running with MATLAB R2020a.

B. Results and Analysis

(1) Clustering Results and Comparative Analysis

In the load clustering experiment, GS-EJ was used to calculate the optimal number of clusters K. Firstly, GS method was used to determine whether K was 1 or not. When K = 1, the value of $Gap(1)$ was smaller than the difference between $Gap(2)$ and $s(2)$, so the optimal number of clusters was not 1. Then, EJ method was used to calculate K between 2 and 8. From the elbow curve in Fig. 6, when K = 3, the decline of the elbow curve became significantly gentle. And the elbow angle satisfies Eq. (14). In summary, the optimal number of clusters was chosen as 3.

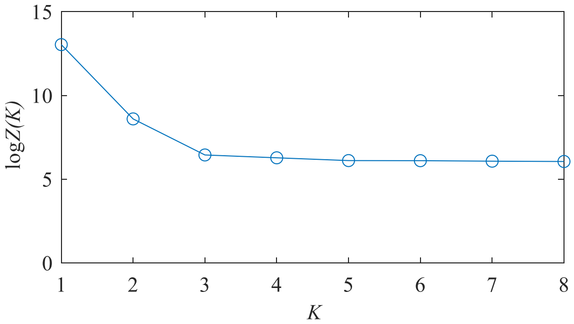

Fig. 6. Optimal number of clusters based on EJ method.

The corresponding clustering results of the daily electric load curve were shown in Fig. 7, and three types of load curve patterns were workday pattern (cluster “a”), weekend pattern (cluster “b”), and holiday pattern (cluster “c”), respectively. In Fig. 7, the cluster “a” was workday electricity consumption pattern with 54 daily load curves, and the main characteristics of the load curves included obvious three different peaks. The three peaks appeared at 10:00 am, 3:50 pm, and 8:00 pm, respectively. In this electricity consumption pattern, the electric power obviously fluctuated, which clearly reflected that the three time periods electricity consumption characteristics were mainly caused by EV charging and factory operation. The cluster “b” was weekend electricity consumption pattern with 20 daily load curves, and this load pattern characteristic had two peaks occurred at 11:00 am and 5:00 pm. This load curve pattern clearly reflected two time periods electricity consumption characteristics of the entire regional electric users during the weekend period. The cluster “c” was holiday electricity consumption pattern with 8 daily load curves, which had obvious low-stability peak power consumption characteristics. This pattern showed that the factories complied with national holiday requirements and stopped production, which greatly reduced electricity consumption. In summary, the proposed DAKM algorithm achieved accurate clustering results of daily electric load curves with high-dimensional and complex time series characteristics.

To verify the accuracy of the proposed DAKM algorithm in electricity consumption pattern recognition, Fig. 8 illustrates the clustering results of three load curve patterns based on TDAKM algorithm, as well as the differences compared with the clustering results using the proposed DAMK algorithm in Fig. 7. To clearly explore the specific differences in clustering results between DAMK algorithm and TDAKM algorithm, each cluster in Fig. 7 was set as a benchmark, and two different colors (blue and red) were used to indicate the differences between the clusters (“a” and “b”) in Fig. 7 and Fig. 8. Four load curves highlighted in blue within cluster “a” in Fig. 7 were surprisingly recognized as clusters “b” in Fig. 8. Then, two load curves highlighted in red within cluster “b” in Fig. 7 were surprisingly recognized as clusters “a” in Fig. 8. The comparison results fully demonstrated that the DAMK algorithm was far superior to the TDAKM algorithm. In addition, it also verified the proposed DTW distance with global constraints had the greatest impact on improving clustering performance.

(2) Comparison of Clustering Validity

To better illustrate the advantages of the proposed DAKM algorithm, five evaluation indicators, such as Davies-Bouldin index (DBI), Sum of Squared Error (SSE), Calinski-Harabasz index (CHI), Silhouette coefficient (SC), and Dunn Index (DI) [22], [23], [24], were used to compare it with five other clustering algorithms, such as TDAKM, adaptive K-means (AKM), DTW-based K-means (DKM), traditional DTW-based K-means (TDKM), and traditional K-means (TKM). The above five indicators comprehensively compared six methods from their respective perspectives, thus verifying the effectiveness of DTW distance with global constraints and adaptive algorithm in load curve classification.

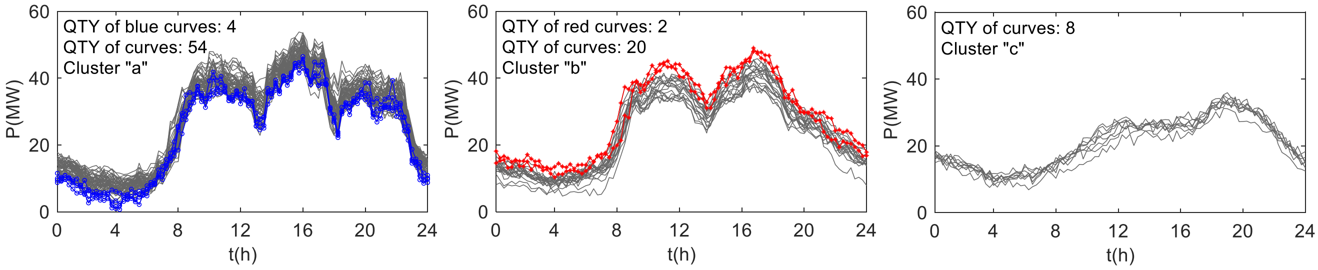

Fig. 7. The optimal three load curve clustering results based on DAKM.

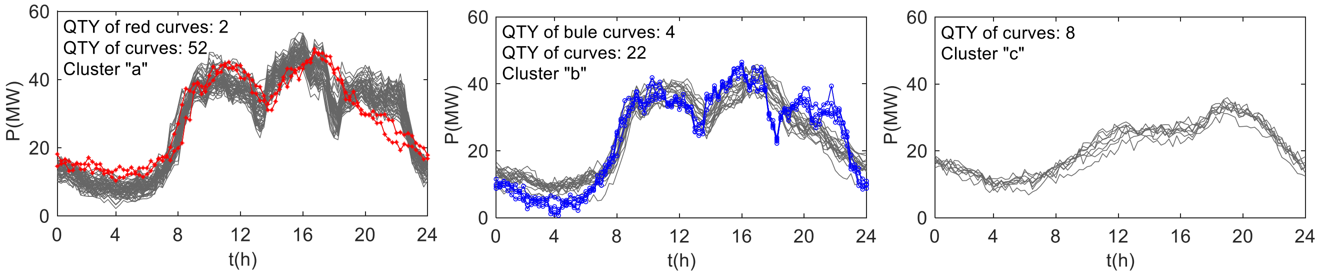

Fig. 8. The optimal three load curve clustering results based on TDAKM

Table I shows a detailed clustering performance comparison of six different K-means algorithms by using five evaluation indicators. For evaluation indicators, the symbol ↑ represents that the higher the evaluation value, the better the clustering performance, while the symbol ↓ indicates the opposite. The value highlighted in bold represents the best data obtained for the optimal algorithm.

TABLE I
PERFORMANCE EVALUATION USING DIFFERENT VALIDATION INDICATORS

Algorithm	DBI ↓	SSE ↓	CHI ↑	SC ↑	DI ↑
DAKM	**0.7055**	**0.0237**	**132.0435**	**0.6649**	**1.1670**
TDAKM	0.7241	0.0942	115.2167	0.6630	1.1104
AKM	0.7983	0.6482	81.3077	0.6581	0.9293
DKM	0.7834	0.1344	102.3546	0.6612	0.9633
TDKM	0.8042	0.4649	96.2456	0.6588	0.5631
TKM	1.2704	2.7957	41.1549	0.3357	0.4256

Table I clearly indicates that the performance of DAKM algorithm in electricity consumption pattern recognition is superior to the other five algorithms, because it achieved the best values on five clustering evaluation indicators, namely 0.7055 for DBI, 0.0237 for SSE, 132.0435 for CHI, 0.6649 for SC, and 1.1670 for DI. From the comparison of DAKM, TDAKM, and AKM in five clustering indicators, it can be directly obtained that the clustering performance of DTW distance measures with global constraints is better than traditional DTW distance and Euclidean distance. DAKM algorithm with adaptive algorithms performs better than DKM algorithm without adaptive algorithms. In addition, Fig. 9 also shows the intuitive differences between six different K-means clustering algorithms in terms of each evaluation indicator, which also proves that the proposed DAKM algorithm outperforms the other five algorithms. Combining Table I and Fig. 9, it is more clearly demonstrated that the proposed DAKM algorithm is the best, followed by TDAKM algorithm, then DKM algorithm, and the worst is TKM algorithm, while AKM and TDKM have similar clustering validity scores. This fully demonstrates that the adaptive algorithms and DTW distance measures with global constraints effectively improved the clustering performance, while only adaptive algorithms or DTW distance measures with global constraints did not achieve the best performance.

(3) Comparison of Clustering Efficiency

Regarding the clustering efficiency, Fig. 10 shows a comparison of iterative convergence among six different

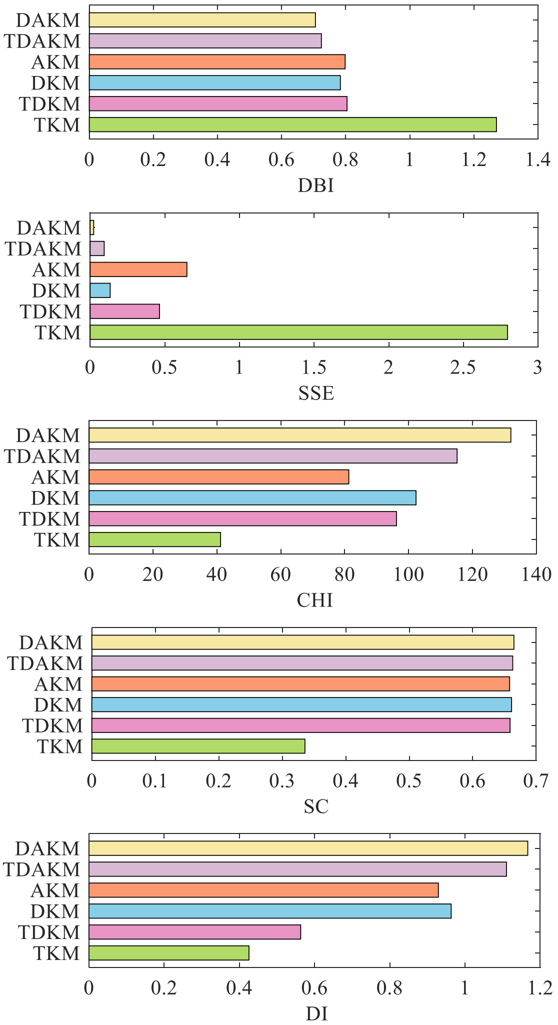

Fig. 9. Comparison of different algorithms using evaluation indicators.

algorithms. The specific data are shown in Table II, the minimum number of convergence iterations (6), while the iteration times of other algorithms were 8 for TDAKM, 12 for AKM and DKM, 15 for TDKM, and 25 for TKM. However, the iteration time required for the proposed DAKM algorithm was 0.2297 seconds, which was slightly slower than 0.1249s of TDAKM, 0.0767s of AKM, and 0.0934s of TKM, but still faster than DKM and TDKM algorithms. Overall, the iteration time of these algorithms was less than 1 second, which had a very fast speed, especially for the DAKM algorithm with an iteration time of less than 0.3 seconds. The root cause for the difference in clustering efficiency between DAKM and other algorithms are that the MMDD method, GS, and EJ are very helpful in improving the iteration times of K-means, but the DTW with global constraints increases the iteration time.

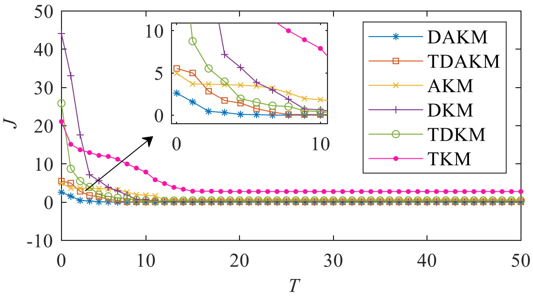

Fig. 10. Comparison of convergence among different algorithms.

TABLE II
COMPARISONS OF CONVERGENCE AMONG DIFFERENT ALGORITHMS

Algorithm	Iteration number	Iteration time (s)
DAKM	**6**	**0.2297**
TDAKM	8	0.1249
AKM	12	0.0767
DKM	12	0.5205
TDKM	15	0.3376
TKM	25	0.0934

From the previous experimental results illustrated in Tables I and II, as well as Fig. 9 and 10, it can be observed that the proposed DAKM algorithm is far superior to the other five algorithms, which also fully demonstrates that the DAKM algorithm is effective and efficient for electricity consumption pattern recognition. Therefore, the DAKM algorithm in this paper is a new choice of clustering algorithm for electric load curve pattern extraction, which can play a significant role in electricity demand response design and others.

V. CONCLUSION

A novel DAKM algorithm is proposed for electricity consumption pattern recognition, which is different from the traditional K-means algorithm. Its main contributions include four aspects:

(i) To improve the similarity measurement of time series, DTW distance with the global constraint of the S-C band is presented to replace the Euclidean distance in traditional K-means algorithm.

(ii) To reduce the adverse effect of random initial cluster centers, the MMDD method is adopted to obtain the optimal initial cluster centers.

(iii) To improve the clustering performance, GS and EJ are used to calculate the optimal number of clusters K automatically.

(iv) the DAKM algorithm is presented for electric load curve pattern extraction.

The experiment results clearly indicated that the proposed DAKM algorithm obtained the best values in five clustering evaluation indicators of DBI, SSE, CHI, SC, and DI, with values of 0.7055, 0.0237, 132.0435, 0.6649, and 1.1670, respectively, which proves that the DAKM algorithm is superior to AKM and TKM algorithms. The DAKM algorithm is also effective and feasible in solving the clustering problem of electric load curves.

REFERENCES

[1] J. Kwac, J. Flora and R. Rajagopal, “Household energy consumption segmentation using hourly data,” *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp420-430, 2014.

[2] F. A. Borges, R. A. Fernandes, I. N. Silva, and C. B. Silva, “Feature extraction and power quality disturbances classification using smart meters signals,” *IEEE Trans. Ind. Informat.*, vol. 12, no. 2, pp824-833, 2016.

[3] T. Teeraratkul, D. O’Neill, and S. Lall, “Shape-based approach to household electric load curve clustering and prediction,” *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp5196-5206, 2017.

[4] K. Mets, F. Depuydt, and C. Develder, “Two-stage load pattern clustering using fast wavelet transformation,” *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp2250-2259, 2016.

[5] Y. F. Lv, “Cloud computation-based clustering method for nonlinear complex attribute big data,” *IAENG International Journal of Computer Science*, vol. 49, no.3, pp736-744, 2022.

[6] F. McLoughlin, A. Duffy, and M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data,” *Appl. Energy*, vol. 141, pp190-199, 2015.

[7] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” *Energy*, vol. 42, no. 1, pp68-80, 2012.

[8] Gianfranco Chicco, and Irinel-Sorin Ilie, “Support vector clustering of electrical load pattern data,” *IEEE Trans. Power Syst.*, vol. 24, No. 3, pp1619-1628, 2009.

[9] P. Wang, Y. W. Ma, Z. Q. Ling, and G. H. Luo, “Peak-valley period partition and abnormal time correction for time-of-use tariffs under daily load curves based on improved fuzzy c-means,” *IET Gener. Transm. Distrib.*, vol. 17, pp5396–5409, 2023.

[10] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, “K-means and alternative clustering methods in modern power systems,” *IEEE Access*, vol. 11, pp119596-119633, 2023.

[11] K. P. Sinaga and M.-S. Yang, “Unsupervised K-means clustering algorithm,’’ *IEEE Access*, vol. 8, pp80716-80727, 2020.

[12] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, “An electric energy consumer characterization framework based on data mining techniques,” *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp596-602, 2005.

[13] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly data,” *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp420-430, 2014.

[14] Z. G. Jiang, R. H. Lin, F. C. Yang, and B. D. Wu, “A fused load curve clustering algorithm based on wavelet transform,” *IEEE Trans Ind. Inform.*, vol. 14, no. 5, pp1856-1865, 2018.

[15] H. D. Chiang, T. S. Xu, X. L. Lv, and N. Dong, “Hierarchical trust-tech-enhanced K-means methods and their applications to power grids,” *IEEE Open Access Journal of Power and Energy*, vol. 9, pp560-572, 2022.

[16] Q. Zhou, and B. Sun, “Adaptive K-means clustering based under-sampling methods to solve the class imbalance problem,” *Data and Information Management*, vol. 8, no. 3, 2024.

[17] T. Teeraratkul, D. O’Neill, and S. Lall, “Shape-based approach to household electric load curve clustering and prediction,” *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp5196-5206, 2018.

[18] H. Li and C. Wang, “Similarity measure based on incremental warping window for time series data mining,” *IEEE Access*, vol. 7, pp3909-3917, 2019.

[19] P. Wang, Y. W. Ma, Z. Q. Ling, and G. H. Luo, “A modified k-means

clustering algorithm based on FMF-GS-DD," *Engineering Letters*, vol. 31, no. 4, pp1518-1525, 2023.

[20] F. You and Y. Wang, "Application of optimized GSA algorithm in bad data detection of power dispatching system", *Techniques of Automation and Applications*, vol. 38, no. 07, pp33-36, 2019.

[21] Yan Liu, "Parallel K-means clustering algorithm based on sampling and maximum ＆minimum distance method," *Intelligent Computer and Applications*, vol. 8, no. 6, pp37-39, 2018.

[22] R. Xu, J. Xu, and D. C. Wunsch, "A comparison study of validity indices on swarm-intelligence-based clustering," *IEEE Trans. Syst. Man. Cy. B.*, vol. 42, no. 4, pp1243-1256, 2012.

[23] Q. T. Bui *et al.*, "SFCM: A fuzzy clustering algorithm of extracting the shape information of data," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 1, pp75-89, 2021.

[24] S. D. Nguyen, V. S. T. Nguyen and N. T. Pham, "Determination of the optimal number of clusters: a fuzzy-set based method," *IEEE Trans. Fuzzy Syst*, vol. 30, no. 9, pp3514-3526, 2022.

Yimeng Shen received the B.Sc. degree in electrical engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2022. She is currently working toward the M.S. degree at School of Automation, Chongqing University of Posts and Telecommunications. Her research interests include demand-side management, and load pattern clustering involved in the fields of vehicle-grid integration, microgrid and new energy power system.

Yiwei Ma received the M.S. degree in control engineering (2007) and the Ph.D. degree in electrical engineering (2015) from South China University of Technology in China. In 2015, She joined Chongqing University of Posts and Telecommunications, where she is currently an Associate Professor with School of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China. Her research interests include optimization design, operation control, and artificial intelligence in the fields of microgrids, smart grids, vehicle-grid integration, and power internet of things.

Hao Zhong received the M.S. and Ph.D. degrees in electrical engineering from Hunan University, Changsha, China, in 2008 and 2011. In 2011, he joined China Three Gorges University, where he is currently an Associate Professor with College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, China. His research interests include power system operation and optimization techniques.

Miao Huang received the M.S. and Ph.D. degrees in electrical engineering from Chongqing University, Chongqing, China, in 2006 and 2011. In 2011, he was employed by State Grid Chongqing Electric Power Co. Electric Power Research Institute in China. In 2015, he joined Chongqing University of Posts and Telecommunications, where he is currently a Senior Engineer with School of Automation and Industrial Internet, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include power system operation and simulation.

Fuchun Deng received the M.S. degree in control engineering (2021) from Chongqing University of Posts and Telecommunications in China. In 2021, He joined Chongqing College of Finance and Economics, where he is currently the professional leader of the Intelligent Product Development at Chongqing College of Finance and Economics, Chongqing, China. His main research fields are smart grid dispatching, intelligent control, and uncertainty research.