An Improved Spectral Clustering Algorithm Using Fast Dynamic Time Warping for Power Load Curve Analysis

Zhongqin Bi[1], Yabin Leng[1], Zhe Liu[2], Yongbin Li[3(✉)], and Stelios Fuentes[4]

[1] College of Computer Science and Technology,
Shanghai University of Electric Power, Shanghai, China
zqbi@shiep.edu.cn, lyb_0730@126.com
[2] State Grid Shanghai Electric Power Research Institute, Shanghai, China
liuzheacyy@163.com
[3] Office of Academic Affairs, Shanghai University of Electric Power, Shanghai, China
lybin40000@163.com
[4] Leicester University, Leicester, UK
stelios.fuentes@gmx.co.uk

Abstract. Cluster analysis of power loads can not only accurately extract the commonalities and characteristics of the loads, but also help to understand the users' habits and patterns of electricity consumption, so as to optimize the power dispatching and regulate the operation of the entire power grid. Based on the traditional clustering methods, this paper proposes a clustering algorithm that can automatically determine the optimal cluster number. Firstly, Fast-DTW algorithm is used as the similarity measuring function to calculate the similar matrix between two time series, and then Spectral Clustering and Affinity Propagation (AP) algorithm are used for clustering. It is combined with Euclidean distance, DTW and Fast-DTW algorithms to evaluate the algorithm effect. By analyzing the actual power data, our results show that the improved external performance evaluation index ARI, AMI and internal performance evaluation index SSE are significantly improved and have better time series similarity and accuracy. Applying the algorithm to more than six thousands of users, twelve kinds of typical power load patterns can be obtained. For any other load curve, it can be mapped to a standard load by feature extraction. The corresponding prediction model is adopted, which is of great significance to reduce the peak power consumption, adjust the electricity price appropriately and solve the problem of system balance.

Keywords: Cluster analysis · Time series · Fast-DTW · Spectral clustering

1 Introduction

With the continuous and steady development of the social economy, the power load has grown rapidly. In recent years, there has been a phenomenon of power

J. Liu et al. (Eds.): MobiCASE 2020, LNICST 341, pp. 143–159, 2020.
https://doi.org/10.1007/978-3-030-64214-3_10

shortage appears in many areas of China during the peak period of power consumption. At the same time, the traditional method of increasing investment in power generation is not economical enough. Thus, alleviating peak power shortages by exploiting demand-side resources has received more attention. At present, China's electricity market is not yet complete, and the demand-side management (DSM) mainly adopts extensive electricity consumption mode, lacking of serious consideration of load form and low satisfaction with electricity consumption. Therefore, the cluster analysis of power load test data is the cornerstone of DSM and even overall planning of the entire power system.

In order to further investigate the standardized model of power load curve, improve the accuracy of clustering analysis, and provide an effective scheme for the supply-demand side reform of power resource consumption, scholars have applied data mining algorithm to the analysis of power load curve. Familiar algorithms include K-means Algorithm, Fuzzy C-means Algorithm, Hierarchical Clustering Algorithm, and Self-Organizing Feature Map Network Algorithm. Ioannis et al. used two different methods to improve the K-means Algorithm and applied it to the time series analysis of power load curve. The results showed that the clustering accuracy was significantly improved [1]. Gao and Zhao et al. combined Fuzzy C-means Algorithm, Conjugate Gradient Algorithm and Deep Belief Network, and proposed a new combination model for short-term photovoltaic power load forecasting, which achieved ideal results as well [2]. The analysis of the electricity consumption patterns of residential users is helpful to improve the accuracy of load forecasting model and to provide reliable and high-quality electricity supply for electric power enterprises, which is of great practical significance to the reform of electricity price. Load forecasting plays an important role in the planning, dispatching, operation, maintenance and control of modern power system [3]. Moreover, the development of energy industry, the change of load demand and the popularization of smart electricity meters all need a new load model to support the research of power system, so a new method of random load modeling of smart electricity meters is proposed [4]. In addition, two data sets were compared by using clustering algorithm, the commonly used data set reduction techniques and feature extraction methods of load patterns were analyzed, and the existing research on power customer clustering was summarized, with emphasis on the main research results [5]. A new load consumption pattern clustering model is proposed to identify periods with similar load levels, typical load patterns of each customer, and periods at different load pattern levels, so as to provide guidance and suggestions for DSM strategies [6]. In addition, different types of load time series are transformed into mapping models to reduce interference and improve differentiated cluster efficiency of power customers [7].

Cluster analysis is an unsupervised learning method in data mining technology. This paper not only proposes a spectral clustering algorithm based on particle swarm, which improves text clustering, but also provides an effective solution for information retrieval, information extraction and document organization [8]. Furthermore, a method of spectral clustering based on iterative optimization is

studied, which solves the problem of spectral decomposition of large-scale high-dimensional data sets and provides an effective solution for spectral clustering [9]. Fully preserving the information integrity of time series is a key link of power load curve clustering analysis. On this basis, an adaptive dynamic time structuring algorithm (ACDTW) is proposed to reasonably arrange the mapping points between two time series. On the one hand, it avoids excessive stretching and compression of time series; on the other hand, it solves the problem that the loss of key feature information will affect the classification accuracy [10]. Time series clustering is a key link in the process of power load curve analysis. It is difficult to fit the similarity of shape and contour of time series adequately. Although many literatures have provided an effective method to extract the standardized model of power load curve and obtained satisfactory results, there is still much room for improvement in the clustering analysis of time series.

Spectral clustering is an algorithm developed from graph theory. Compared with traditional clustering algorithm, it has the advantages of simple implementation and perfect clustering effect. However, the disadvantages are also obvious. Firstly, the clustering center cannot be automatically determined; secondly, the algorithm is prone to local optimization; thirdly, the similarity of time series shape and contour cannot be guaranteed. On this basis, a spectral clustering method without considering the internal characteristics of time series is proposed. Firstly, by comparing Euclidean distance, DTW and Fast-DTW, the influence of similarity measure on time series clustering is studied [11]. Secondly, the clustering effect of K-means Algorithm and AP Neighbor Propagation Algorithm on eigenvectors is compared. Finally, we propose a new Fast-DTW-AP Spectral Clustering Algorithm, which can not only automatically select the optimal number of clusters in arbitrary sample space, but also effectively avoid the algorithm falling into the phenomenon of local optimization. In addition, it has a better affinity for high-dimensional time series data or sparse data.

After the performance test of the algorithm on the standard data set and the real data set, we applied the proposed method to thousands of home power users, and cluster them into 12 clusters. According to the clustering results obtained, the characteristics of each type of load are analyzed, and the standardized model of load is established, which can be applied to different types of load. This article addresses three issues in the definition of a standardized model. The first is to propose a time series load clustering algorithm which is helpful to improve the prediction accuracy in high dimensional space. The second is to use the improved algorithm to cluster the load and then design a load prediction model suitable for this feature. The third is to adjust the electricity price according to the load prediction model, which is of great significance to the safety, economy and stable operation of the power system.

In this paper, a clustering algorithm combining the selection of internal similarity matrix in spectral clustering with AP neighbor propagation is proposed, which can automatically determine the clustering center and effectively avoid falling into local optimization. The Fast-DTW-AP improved spectral clustering algorithm is applied to thousands of households, and 12 standard power models

are obtained. Through the extraction of users' consumption habits and patterns, accurately grasp the law of electricity consumption. On this basis, a more effective electricity price is designed to adjust the residents' demand.

The paper is organized as follows: The second sector is the model building. Sector 2.1 introduces three similarity measurement methods: Euclidean distance, DTW and fast-DTW. Sector 2.2 improves the K-means algorithm of spectral clustering, making the clustering effect of the algorithm more excellent. In the Sect. 3, internal and external indexes are used to evaluate the experiment, and the experimental results of the Fast-DTW-AP improved spectral clustering and other clustering algorithms in time series data sets and standard data sets are analyzed and compared.

2 Model Building

Time series of power demand is the key information source of consumer behavior. Although some scholars have studied the load pattern of extracting a large number of power users, there are few researches about calculation of time series. Therefore, improving traditional clustering techniques, optimizing the number of clustering, and improving the quality of clustering and the similarity of time series have become an important topic.

2.1 Similarity Measure

The core of improved power time series clustering analysis is the similarity measure that constructs the similarity matrix between two power time series curves. In order to study the role of similarity measure in power time series clustering analysis, Euclidean Distance, DTW and Fast DTW are used as similarity measures in the application of spectral clustering, and the final clustering results are analyzed and compared reasonably.

Euclidean Distance. For two power time series curves U and V with length $|U|$ and $|V|$ respectively.

$$U = \{U_1, U_2, \ldots, U_{|U|}\} \tag{1}$$

$$V = \{V_1, V_2, \ldots, V_{|V|}\} \tag{2}$$

The Euclidean distance requires that the sample power time series curve must be equal in length, that is, $|U| = |V|$. The formula for defining the distance between U and V in n-dimensional space is:

$$ED(U, V) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} \tag{3}$$

Euclidean distance is the most commonly used distance measurement method. It measures the absolute distance between two power series curves,

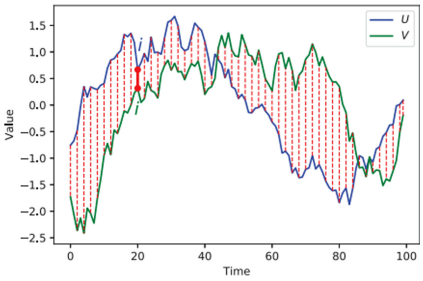

Fig. 1. Euclidean distance between two power time series curves

but it can only measure the time series of the same length. However, the power time series curve generated by power load in the actual power generation process is disordered, so Euclidean distance is difficult to predict whether there is a similar trend between the two power time series. As shown in Fig. 1, the local peaks of the curves and of the power time series do not match, which is caused by the fact that the Euclidean distance can only match the two series point-to-point.

Dynamic Time Warping Algorithm (DTW). Dynamic time warping algorithm is a non-linear measure of the minimum distance between two power time series curves [12]. Its purpose is to find the sum of the minimum cumulative distance of all corresponding points of two power time series curves, namely, to find the shortest integration path. It represents the optimal matching between two power time series curves, fully guarantees the shape and contour similarity of the two power time series curves and breaks through the limitation of Euclidean distance for the calculation of unequal length power time series curves.

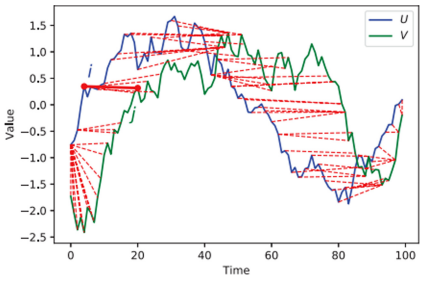

Fig. 2. DTW nonlinear measurement between two power time series curves

As can be seen from Fig. 2, point i in the power time series curve U can be well aligned with point $j(i = j)$ in the series V. The curve integration path start from the start point (1, 1) corresponding to the two sequences and ends at

the end point ($|U|$, $|V|$) corresponding to the two sequences. For the cumulative distance $Dist(i, j)$ is defined as:

$$Dist(i,j) = min\{Dist(i-1,j-1), Dist(i-1,j), Dist(i,j-1)\} + d(u_i, v_j) \quad (4)$$

$Dist(1,1) = d(1,1)$, and $d(u_i, v_j)$ are usually calculated using Euclidean distance. The optimal regular path is determined by $Dist$ ($|U|$, $|V|$). The DTW algorithm can automatically match the peaks and is not limited by the length of the template. It is suitable for clustering analysis of power curve.

However, the DTW algorithm also has obvious defects in the actual power curve analysis, that is, the algorithm complexity is too high. When two power time series curves are relatively long, the efficiency will be slow and the regularity will be too large, which may easily lead to the wrong matching of the power curve.

Fast Dynamic Time Warping Algorithm (Fast-DTW). The Fast Dynamic Time Warping Algorithm is an acceleration algorithm of the classical dynamic time structuring algorithm. The algorithm combines limited search space and data abstraction. On the basis of fully ensuring the accuracy of the algorithm, a reasonable solution is provided for the clustering analysis of power curves with large amount of data in practical application.

The three steps of the Fast-DTW algorithm are as follows: First, the original power time series curve is extracted with coarse-grained data, and repeated iterative optimization is performed. Where, the coarse-grained data points are the average values of the corresponding fine-grained data points. Second, the DTW algorithm is run granularity on coarse-grained power time series curves. Finally, the regular path obtained on the coarser granularity is further fine-grained into a finer-grained power time series curve through a grid [13].

Table 1. The comparison of similarity measures

Algorithm	Complexity	Alignment
ED	O(N)	One-to-one
DTW	O(N^2)	One-to-many
Fast-DTW	O(N)	One-to-many

The algorithm complexity of the three similarity measures is shown in Table 1. The algorithm complexity of the Euclidean distance is N, but it can only meet the requirements of point-to-point, that is, it can only measure power time series curve of equal length. DTW can match different power time series curve, but the algorithm complexity is high. The Fast-DTW algorithm not only meets the needs of unequal length power time series curve, but also reduces the complexity of the algorithm.

2.2 Clustering Algorithms

The key of power load curve clustering analysis is the choice of clustering algorithm. Because of its excellent clustering effect, spectral clustering algorithm is more and more widely used. The Affinity Propagation Algorithm can automatically determine the clustering center, which is not sensitive to the initial power data, and can complete the clustering of large-scale and multi category data sets in a short time, which is suitable for the clustering analysis of power time series curves with large amount of data. In order to obtain more ideal clustering results, this paper makes appropriate improvements based on the two clustering algorithms.

Spectral Clustering. The main idea of spectral clustering is to transform all the power data into points in space. These points can be connected by edges. The edge weight value between the two points with a long distance is lower, while the edge weight value between the two points with a short distance is higher. By cutting the graph composed of all data points, the edge weight between different subgraphs after cutting is as low as possible, and the edge weight sum within the subgraph is as high as possible, so as to achieve the purpose of clustering [14].

There are obvious disadvantages when using original spectral clustering for cluster analysis: (1) The cluster center cannot be determined automatically; (2) The final clustering effect is largely affected by the similarity matrix and feature vector clustering algorithm; (3) It is very sensitive to the choice of clustering parameters.

Affinity Propagation (AP) Algorithm. In spectrum clustering, K-means algorithm is used to cluster the eigenvector space. However, K-means algorithm is very sensitive to the selection of the initial clustering center, and its hill-climbing optimization algorithm often fails to obtain the global optimal solution, so the Affinity Propagation (AP) algorithm is introduced. The AP algorithm is a clustering algorithm based on "information transfer" between data points [15]. The algorithm does not need to determine the number of clusters before running it. In addition, because the actual points in the data set are selected, the clustering effect is better with the cluster center as the representative of each class. The basic steps are as follows:

(1) Euclidean distance is used to calculate the similarity between two data points, and a similarity matrix S is constructed. S is a $n \times n$ matrix.
(2) Calculate the attraction matrix:

$$R_{t+1}(i,k) = (1-\lambda) \bullet R_{t+1}(i,k) + \lambda \bullet R_t(i,k) \tag{5}$$

$$R_{t+1}(i,k) = S(i,j) - max\{A_t(i,j) + R_t(i,j)\} \tag{6}$$

Among them, $S(i,j)$ is an element in the similarity matrix S, which indicates the ability of point j to be the clustering center of point i. Generally, a negative

Euclidean distance is used. The larger $S(i, j)$, The closer the distance between the points, the higher the similarity. The degree of attraction $R(i, k)$ indicates the degree that point K is suitable to be the clustering center of data point I, which is a process of selecting point K from point I. The damping coefficient λ is used for the convergence of the algorithm, and the value range is [0.5, 1].

(3) Calculate the membership matrix:

$$A_{t+1}(i,k) = (1-\lambda) \bullet A_{t+1}(i,k) + \lambda \bullet A_t(i,k) \tag{7}$$

$$A_{t+1}(i,k) = min\{0, R_{t+1}(k,k) + \sum_{j \in i,k} max\{0, R_{t+1}(j,k)\}\} \tag{8}$$

The degree of belonging $A(i, k)$ indicates the suitability of point i to select point k as its clustering center. This is a process in which point k selects point i. $A_{t+1}(i, k)$ represents the new $A(i, k)$, and $A_t(i, k)$ represents the old $A(i, k)$.

(4) Iteratively update R and A values.

The selection of the appropriate clustering center is crucial to the quality of the final clustering effect. When the value of $A(i, k) + R(i, k)$ is larger, it means that the probability of K points as the cluster center is greater. In order to find the maximum value, it is necessary to iteratively update the R and A values to obtain the most suitable cluster center. The termination condition of iteration is that the cluster center is not updated to a certain extent or reaches the maximum number of iterations (generally 15 times). After getting the most suitable cluster center, the data set can be classified directly.

Improved Spectral Clustering Algorithm. In order to extract the standardized model of power time series accurately and effectively, more reliable clustering analysis results are obtained. In this paper, a Fast-DTW-AP spectral clustering algorithm (F-A-S) is proposed based on the original spectral clustering. As long as the data points are entered, the cluster center and the number of clusters can be automatically determined. By analyzing the number of clusters, the inherent shortcomings of spectral clustering algorithm are solved, which makes the algorithm not only suitable for arbitrary shape sample space clustering, but also can effectively prevent the algorithm from falling into the local optimal phenomenon. In addition, the improved algorithm has better performance in processing unequal length power time series curves, which fully guarantees the similarity of shape and contour of power time series curves, and its performance is far better than that in processing high-dimensional data and sparse data clustering.

Through comparative experiments, the traditional clustering algorithm finally found that the improved algorithm significantly improved the clustering effect of time series, and the external evaluation indexes ARI, AMI and internal evaluation indexes SSE of the clustering were significantly improved. The specific algorithm steps are as follows:

1. Enter a power time series data set and use the Fast-DTW algorithm to calculate the similarity between each data point.

$$S(i,j) = FastDTW(i,j) \tag{9}$$

The similarity matrix S is generated after calculating the similarity between the data points by using the Fast-DTW distance. S is a $n \times n$ matrix composed of $S(i,j)$. The value of element p in the matrix is 1. The purpose is to preserve the integrity of its own information to the greatest extent.

$$S = \begin{bmatrix} p & S(1,2) & \cdots & S(1,n) \\ S(2,1) & p & \cdots & S(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ eS(n,1) & S(n,2) & \cdots & p \end{bmatrix} \tag{10}$$

2. Apply the following formula to calculate the degree matrix.

$$d_i = \sum_{j=1}^{n} S_{ij} \tag{11}$$

A degree matrix D can be constructed from the similarity matrix. The degree matrix D is a $n \times n$ diagonal matrix composed of d_i. Each element on the diagonal is the sum of the elements of each row of the corresponding similarity matrix, and all other elements are 0.

3. Laplace matrix L is calculated according to similarity matrix S and degree matrix D.

$$L = D - S \tag{12}$$

The degree matrix D and the similarity matrix S are different to generate a Laplacian matrix L. In order to obtain better clustering results, after obtaining the Laplacian matrix, this paper uses a symmetric normalization method to normalize the Laplacian matrix.

$$L = D^{-1/2}(D - S)D^{-1/2} \tag{13}$$

4. The eigenvalues of Laplace matrix L are calculated, the eigenvalues are sorted from small to large, and the eigenvector $u_1, u_2, \ldots, u_n$ corresponding to each eigenvalue is calculated. The matrix $U = \{u_1, u_2, \ldots, u_n\}$ (n rows and n columns) is composed of n column vectors.
5. Let Y_i be the i-th row vector of $U (i = 1, 2, \ldots, n)$, and then form a new matrix $Y = \{y_1, y_2, \ldots, y_n\}$.
6. Use the Affinity Propagation (AP) algorithm to cluster the new sample point $Y = \{y_1, y_2, \ldots, y_n\}$ and divide it into k clusters.

Algorithm 1. The Fast-DTW-AP Spectral Clustering Algorithm

Input:

Sample dataset$X = \{x_1, x_2, \ldots, x_n\}$;

Output:

Cluster set $C = \{C_1, C_2, \ldots, C_k\}$;

1: Calculate the Fast-DTW distance of X, and obtain the similarity matrix S;
2: Compute the standardized laplacian using Eq. (13);
3: Compute the eigen vectors $U_1, U_2, \ldots, U_n$ of Laplacian;
4: Let $X \in R^{n \times n}$ which contains the vectors $U_1, U_2, \ldots, U_n$ as a column;
5: Let $Y \in R^{n \times n}$ be the vector corresponding to the ith row of U;
6: Group the points $Y_i \in R^{n \times n}$ with the AP clustering algorithm into$\{C_1, C_2, \ldots, C_k\}$

3 Experiment and Analysis

3.1 Basic Description of the Experimental Environment

All the experiments in this paper are carried out on a computer equipped with windows 10 operating system, Intel (R) core (TM) i5-3230m 2.60 GHz CPU and 8 GB RAM, and the algorithm is implemented with Python 3.7 software.

The Comprehensive Control Chart Time Series (SCCTS) data set used in this paper is a standard time series test data set in the UCI database. As shown in Table 2, the data set contains 6 different classes with a total of 600 rows and 60 columns [16].

Table 2. Classification and abbreviations of SCCTS

Number	Abbreviation	Class
0–99	N	Normal
100–199	C	Cyclic
200–299	I	Increasing trend
300–399	D	Decreasing trend
400–499	US	Upward shift
500–599	DS	Downward shift

The real power consumption data comes from the smart electricity customer behavior test conducted by Energy Regulatory Commission (CER) during 2009–2010. Smart meters measure power consumption in KW per half hour. The data set contains more than 6000 customer records, of which 66% are residents, 7% are small and medium-sized enterprises, and 27% are other customers [17].

3.2 Clustering Quality Evaluation Index

The clustering quality evaluation index consists external evaluation index and internal evaluation index. Among them, the external evaluation index is to apply the clustering algorithm to the standard test data set with clear classification, and then calculate the clustering accuracy of the algorithm to the data set with relevant indexes. While the internal indicators refer to pre-defined evaluation criteria, which are usually used to describe some intrinsic characteristics and quantitative values of clustering after clustering in order to evaluate the quality of clustering results.

External Evaluation Indicators. The adjusted clustering method was evaluated by using the Adjusted Mutual Information (AMI) [18] and Adjusted Rand index (ARI) [19] as external standards for evaluating the quality of clustering. Given two sets $U = \{U_1, U_2, \ldots, U_i\}$ and $V = \{V_1, V_2, \ldots, V_j\}$ i is the number of clusters in U, and j is the number of clusters in V. ARI represents the number of paired samples belonging to the same classification or different classifications in two sets, and the expected value between U and V is defined as:

$$E(U,V) = \left| \sum_{u_i} \binom{n_i}{2} \sum_{v_j} \binom{n_j}{2} \right| / \binom{n}{2} \tag{14}$$

ARI is defined as:

$$ARI(U,V) = \frac{\sum u_i \sum v_j \binom{n_{ij}}{2} - E(U,V)}{\frac{1}{2}[\sum_{u_i} \binom{n_i}{2} + \sum_{v_j} \binom{n_j}{2}] - E(U,V)} \tag{15}$$

AMI is a clustering evaluation index based on the degree of correlation between two random variables, called mutual information MI, that is, the amount of information about one random variable contained in one random variable. AMI is defined as:

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{max\{H(U), H(V)\} - E\{MI(U,V)\}} \tag{16}$$

Both the AMI and ARI indexes have a value range of [0, 1], and the larger the value, the more consistent the clusters divided by the standard clusters.

Internal Evaluation Indicators. In order to better evaluate the clustering effect of the algorithm, The sum of squares due to error (SSE) is used as the internal standard to evaluate the clustering quality.

$$I_{SSE} = \sum_{i=1}^{k} \sum_{x \in G_i} \|x - o_i\|^2 \tag{17}$$

SSE is the sum of the squares of the distances from the data points in all subclasses to the corresponding cluster centers after clustering.

3.3 Experiment Analysis

By comparing the performance of standard test data and actual power consumption, the validity and effectiveness of the proposed algorithm are proved. First, in order to evaluate the effectiveness of the proposed algorithm, the Fast-DTW-AP spectral clustering algorithm (F-A-S) is used for clustering on the standard SCCTS data set, and the external evaluation standards ARI, AMI and internal evaluation standard SSE are used to measure the final result. And the clustering effect is compared with other improved spectral clustering algorithms based on DTW and AP, such as the original spectral clustering algorithm (S), Fast-DTW spectral clustering algorithm (F-K-S), DTW-AP spectral clustering algorithm (D-A-S), etc.

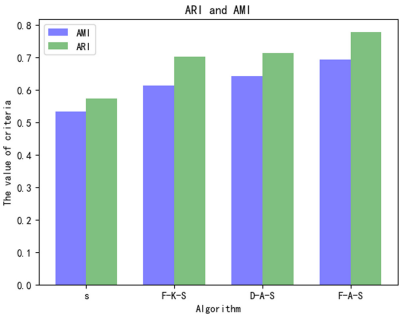

Fig. 3. External evaluation indicators of the Fast-DTW-AP spectral clustering algorithm and other improved spectral clustering algorithms

As can be seen from the comparison results in Fig. 3, the improved Fast-DTW-AP spectral clustering algorithm (F-A-S) has significantly improved the performance of AMI and ARI indexes, indicating that the algorithm proposed in this paper has higher fitting accuracy for time series and can better realize the classification of standard data. As Euclidean distance was used as the similarity measure in the original spectral clustering, it could not match the unequal time series. Therefore, compared with the traditional spectral clustering algorithm, the AMI and ARI evaluation indexes of the Fast DTW-AP spectral clustering algorithm were improved by 16.2% and 18.4%, respectively. Since AP algorithm has better processing effect than K-means algorithm in dealing with complex time series, compared with Fast DTW and K-means (F-K-S) combined algorithm, AMI and ARI of Fast-DTW-AP spectral clustering algorithm are improved by 7.1% and 8.6% respectively. Compared with combination algorithm of DTW algorithm and AP algorithm (D-A-S), AMI and ARI of Fast-DTW-AP spectral clustering algorithm increased respectively 6.4% and 7.3%, which is because fast DTW solves the problem of excessive regularity of DTW. The above three comparative experiments fully show that our proposed algorithm has better clustering effect.

Table 3. Internal evaluation index of Fast-DTW-AP

Algorithm	SSE
S	1.0833×10^5
F-K-S	1.0422×10^5
D-A-S	1.0359×10^5
F-A-S	1.0297×10^5
Standard time series set	1.0195×10^5

As can be seen from Table 3, the SSE index value of the improved Fast-DTW-AP spectral clustering algorithm (F-A-S) is lower than the other three comparison algorithms, and the intra cluster variance is closer to 1, which indicates that the improved Fast-DTW-AP spectral clustering algorithm has achieved excellent clustering results. At the same time, it is noted that the intra class variance of this algorithm is closer to the result of standard data. Experimental results show that the proposed algorithm has better performance in processing time series.

In order to further verify the feasibility of the algorithm, K-Medoids algorithm and AP algorithm were used to cluster the same data set, and AMI and ARI evaluation indexes were used to evaluate the final clustering effect.

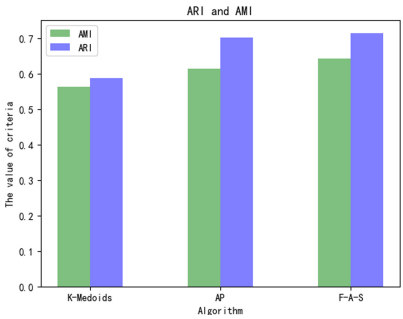

Fig. 4. Evaluation indicators of K-Medoids, AP algorithm, and improved spectral clustering algorithm

The experimental results are shown in Fig. 4. Compared with the improved k-Medoids algorithm, AMI and ARI are improved by 8.2% and 13.8% respectively. Compared with the AP algorithm, the improved spectral clustering The algorithm increased by 6.8% and 3.3% respectively, which shows that the proposed algorithm also performs well when compared with some other data mining algorithms.

In order to make the effectiveness of the algorithm more rigorous and ensure that it not only has a better clustering effect on this data set, the staff absence standard time series data set (AAW) was selected from the UCI database to

conduct clustering effect test. The data set is derived from real data and contains a total of 740 instances and 21 attributes [20].

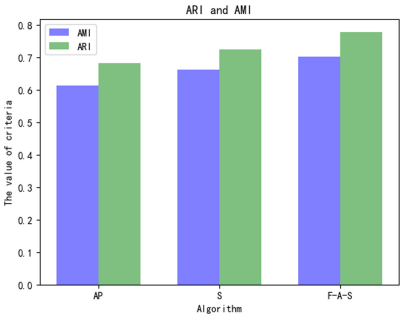

Fig. 5. Evaluation indicators of spectral clustering, AP, and improved spectral clustering algorithm in AAW dataset

As shown in Fig. 5, after replacing the AAW data set, the clustering effect obtained by using the improved spectral clustering algorithm also has obvious advantages. Compared to the AP algorithm, the improved spectral clustering algorithms AMI and ARI The index is increased by 8.8% and 9.7%, respectively. Compared with the traditional spectral clustering algorithm, the AMI and ARI indexes of the improved spectral clustering algorithm are improved by 3.9% and 5.4% respectively, indicating that the improved spectral clustering algorithm has better performance. The same performance is achieved when applied to other data sets.

Finally, the actual efficiency of the algorithm is verified by using actual energy consumption data from THE Energy Council (CER). In this paper, the power data set is preprocessed, including deleting users who lose data, deleting data that is not suitable for analysis near zero, and using minimum-maximum normalization to map the data uniformly to the interval [0, 1].

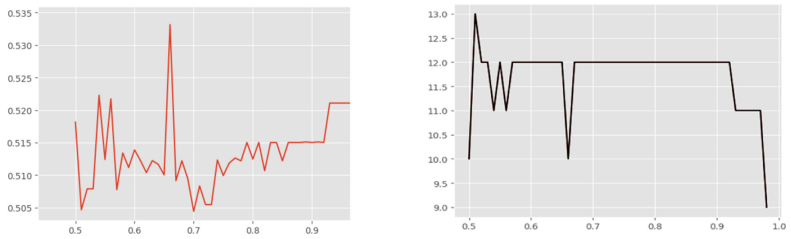

Fig. 6. Relationship between the number of clusters and the contour coefficient

On the basis of data processing, we apply the Fast-DTW-AP spectral clustering algorithm to cluster the power load curves of thousands of users' power

consumption patterns. Figure 6 shows the process of automatically determining the optimal number of clusters in the analysis, and the most accurate cluster number is obtained by calculating the optimal damping coefficient. Therefore, the contour coefficient is used to evaluate the clustering effect, and the damping coefficient is taken as a parameter, whose variation range is between (0.5, 1). The relationship between the damping coefficient and the profile coefficient is obtained by taking the damping coefficient as the horizontal axis and the profile coefficient as the vertical axis. The higher the contour coefficient is, the better the clustering effect of the corresponding damping coefficient is. It is of great significance to find the optimal damping coefficient for the final clustering effect. As shown in Fig. 6, the highest contour coefficient Y = 0.533382 is obtained when X = 0.658032, indicating that the optimal damping coefficient of the data set is 0.658032, and the optimal number of clusters is 12. In order to achieve the best clustering effect, the optimal number of clusters is obtained by calculating the optimal damping coefficient.

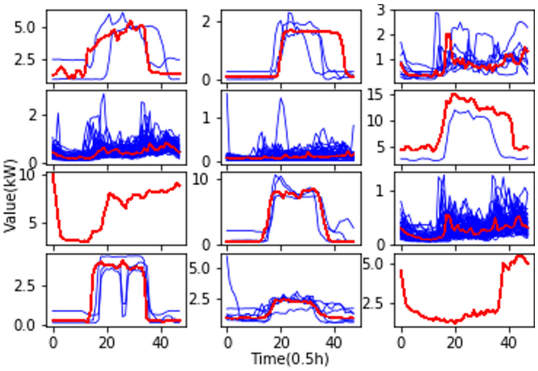

Fig. 7. Clustering effect of CER dataset (Color figure online)

Finally, 12 types of power load curve standardized models were obtained. For each clustering result, a power load curve is drawn as Fig. 7, and a typical load curve is extracted by taking time as the abscissa and electricity load power as the ordinate. In this way, the electricity consumption curve (blue line) and typical load curve (red line) of each household can be obtained, so as to accurately extract the commonness and difference of load.

4 Conclusion

In order to effectively extract valuable information from power data, optimize power dispatch and regulate the operation of the entire power grid, this paper proposes a Fast-DTW-AP improved spectral clustering algorithm based on time series. The main contributions are summarized as follows: First, the external

index of AMI, ARI and internal index of SSE were used to evaluate the clustering results of the Fast-DTW-AP improved spectral clustering algorithm and other three time series clustering methods. The spectral clustering algorithm can effectively retain the morphological and contour similarities between time series.

Second, comparing Fast-DTW-AP improved spectral clustering algorithm with other two commonly used data mining algorithms, we found that the external evaluation indexes AMI and ARI were significantly improved, which further proved the robustness and practical feasibility of the algorithm.

Third, the Fast-DTW-AP spectral clustering algorithm was tested on SCCTS, AAW and CER of Irish smart meter. Multiple experimental results show that the Fast-DTW-AP improved spectral clustering algorithm has achieved the best performance.

Fast-DTW-AP spectral clustering algorithm has certain advantages compared to other clustering algorithms when processing time series. In general, a reasonable power model is designed to help adjust the appropriate electricity price, minimize peak power consumption, and solve the problem of system balance.

Acknowledgment. This work is supported by the National Nature Science Foundation of China (No. 61972357, No. 61672337).

References

1. Panapakidis, I.P., Christoforidis, G.C.: Implementation of modified versions of the K-means algorithm in power load curves profiling. Sustain. Cities Soc. **35**, 83–93 (2017)
2. Gao, Z., Li, Z., Bao, S.: Short term prediction of photovoltaic power based on FCM and CG DBN combination. J. Electr. Eng. Technol. **15**, 333–341 (2020)
3. Fu, X., Zeng, X.J., Feng, P., Cai, X.: Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. Energy **165**, 76–89 (2018)
4. Khan, Z.A., Jayaweera, D., Alvarez-Alvarado, M.S.: A novel approach for load profiling in smart power grids using smart meter data. Electr. Power Syst. Res. **165**, 191–198 (2018)
5. Rajabi, A., Eskandari, M., Ghadi, M.J., Li, L., Zhang, J., Siano, P.: A comparative study of clustering techniques for electrical load pattern segmentation. Renew. Sustain. Energy Rev. **120**, 109628 (2019)
6. Charwand, M., Gitizadeh, M., Siano, P., Chicco, G., Moshavash, Z.: Clustering of electrical load patterns and time periods using uncertainty-based multi-level amplitude thresholding. Int. J. Electr. Power Energy Syst. **117**, 105624 (2020)
7. Motlagh, O., Berry, A., O'Neil, L.: Clustering of residential electricity customers using load time series. Energy **237**, 11–24 (2019)
8. Janani, R., Vijayarani, S.: Text document clustering using spectral clustering algorithm with particle swarm optimization. Expert Syst. Appl. **134**, 192–200 (2019)
9. Zhao, Y., Yuan, Y., Nie, F., Wang, Q.: Spectral clustering based on iterative optimization for large-scale and high-dimensional data. Neurocomputing **318**, 227–235 (2018)

10. Wan, Y., Chen, X.-L., Shi, Y.: Adaptive cost dynamic time warping distance in time series analysis for classification. J. Comput. Appl. Math. **319**, 514–520 (2017)
11. Han, T., Peng, Q., Zhu, Z., Shen, Y., Huang, H., Abid, N.N.: A pattern representation of stock time series based on DTW. Phys. A Stat. Mech. Appl. **550**, 124161 (2020)
12. Kang, Z., et al.: Multi-graph fusion for multi-view spectral clustering. Knowl.-Based Syst. **189**, 105102 (2019)
13. Salvadora, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. **11**, 561–580 (2007)
14. Cao, Y., Rakhilin, N., Gordon, P.H., Shen, X., Kan, E.C.: A real-time spike classification method based on dynamic time warping for extracellular enteric neural recording with large waveform variability. J. Neurosci. Methods **261**, 97–109 (2016)
15. Han, Y., Wu, H., Jia, M., Geng, Z., Zhong, Y.: Production capacity analysis and energy optimization of complex petrochemical industries using novel extreme learning machine integrating affinity propagation. Energy Convers. Manag. **180**, 240–249 (2019)
16. Alcock, R.: Synthetic control chart time series data set (1999). http://archive.ics.uci.edu/ml/machine-learning-databases/synthetic_control-mld/. Accessed via UCI
17. CER smart metering project-electricity customer behaviour trial (2017). http://www.ucd.ie/issda/data/commissionforenergyregulationcer/. Accessed via the Irish Social Science Data Archive
18. Xie, J., Gao, H., Xie, W.: Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors. Inf. Sci. **354**, 19–40 (2016)
19. Xie, J., Zhou, Y., Ding, L.: Local standard deviation spectral clustering. In: IEEE International Conference on Big Data and Smart Computing, vol. 143, pp. 242–250 (2018)
20. Martiniano, A., Ferreira, R.P., Sassi, R.J.: Absenteeism at work Data Set (2010). http://archive.ics.uci.edu/ml/datasets/Absenteeismatwork/. Accessed via UCI