An Approach of Electrical Load Profile Analysis Based on Time Series Data Mining

YING SHI†1, TAO YU†1, (Member, IEEE), QIANJIN LIU†1, (Member, IEEE), HANXIN ZHU†1, FUSHENG LI†1, AND YAXIONG WU‡2

†College of Electric Power, South China University of Technology, Guangzhou 510640, China
‡Grid Planning and Research Center, Guangdong Power Grid Company Ltd., CSG, Guangzhou 510030, China

Corresponding author: Tao Yu (taoyu1@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51777078, and in part by the Technical Projects of China Southern Power Grid under Grant GDKJXM20172939.

ABSTRACT In the current electrical load profile analysis, considering the shortage of traditional methods on the typical load profile extraction of single consumers and the load profile feature extraction, this paper proposes an approach based on time series data mining. Firstly, this method reduces the dimension of the load profile of a single consumer based on the Piecewise Aggregate Approximation (PAA), and re-expresses the load profile of the consumer over a period based on the Symbolic Aggregate Approximation (SAX), representing the consumer’s load profile with a symbolic string to extract the typical load profile. Then, combined with the load characteristic indices and time series-based features, the typical load profiles of different consumers are clustered based on the K-means algorithm to analyze the power consumption behaviors. Finally, this paper performs a case analysis with a UCI test data set, and the results show that the proposed approach can excavate typical power consumption behaviors of consumers and improve the electrical load profile analysis efficiency and the clustering quality.

INDEX TERMS Cluster analysis, electrical load profile analysis, symbolic aggregate approximation, time series data mining.

I. INTRODUCTION

With the development of smart grids and the construction of advanced metering infrastructure (AMI), a massive amount of fine-grained electricity consumption data has been collected, which contains a wealth of consumers’ information, such as the load profiles [1]–[4]. In recent years, power big data has become a research hotspot [5], attracting an increasingly number of scholars. Analyzing the correlation between electricity consumption data in the big data context, exploring the consumer’s electricity consumption behavior habits hidden in the electricity consumption data, and studying effective data mining algorithms to classify different consumers accurately, can help electric utilities to carry out consumers’ energy-saving work and provide consumers with differentiated service [6]–[9].

An electrical load profile is a graph of the variation in the electrical load versus time. Typical load profile (TLP) of an individual consumer can reflect his typical power consumption behavior [10], [11]. To extract the TLP of a single consumer over a period of time, the typical conventional method manually selects a load profile that is close to the annual average load rate with no distortion. However, this method ignores the difference between load characteristics of different consumers, making it not versatile and accurate. Reference [12] proposes a method based on Fuzzy C-Means to determinate consumer’s TLP, which can efficiently assign TLP to the consumer. Reference [13] discusses several approaches and proposes a general framework for the extraction of TLP. However, the consumer load data is easily affected by power system faults or devices’ metering errors, producing abnormal load profile. Besides, as the load profile of different consumers are quite different, their abnormal load profiles are also different. The above research has certain defects, which cannot reduce the impacts of abnormal load profile effectively.

In addition, data mining techniques have been widely applied to the electrical power industry. The load profile classification aims to separate enormous load profile into several typical clusters. With the development and widespread popularity of AMI, most meters will be capable of generating data with high temporal resolution [14], [15], which
means we can grasp the fluctuation of the load profile more precisely. Performing cluster analysis on these high resolution load data is extremely complex and time-consuming. Therefore, it is necessary to reduce the dimension of the load data and extract the key feature of the consumer load profile to enhance the effectiveness and computing performance of clustering. Reference [16] analyzes the impacts of different feature selections for clustering load profiles, and proposes a novel feature construction method to generate processed data as inputs for clustering algorithm. Reference [17] has proposed several indices to capture relevant information about consumer load profile, so as to achieve the purpose of dimension reduction. Meanwhile, a variety of clustering methods have been proposed to analyze consumer load profile, such as K-means, hierarchical clustering, network-based clustering algorithm of self-organizing map [18], [19] and so on. However, the above studies are mostly based on raw load data or load indices, which lack consumer load profile extraction based on time series data mining. Most studies ignore the time series feature of load profile, failing to capture the information of load profile accurately.

In summary, the current research on electrical load profile analysis method mainly has the following two shortcomings: 1) There is no effective method to extract the TLP of consumers, and the effect of eliminating abnormal load profile is not good, which will have a certain impact on the analysis of typical power consumption behaviors of consumers; 2) Conventional load feature indices cannot reflect the inherent characteristics of the load profile, and the feature extraction of the load profile is insufficient, leading to poor clustering results and huge errors.

In view of the shortcomings of the existing research, this paper proposes an approach of electrical load profile analysis based on time series data mining. The method first reduces the dimension of the load profile of a single consumer based on the Piecewise Aggregate Approximation (PAA), and re-expresses the load profile of the consumer over a period of time based on the Symbolic Aggregate approximation (SAX), and represents the consumer’s load profile with a symbolic string to extract the TLP of the consumer. Then, combined with the load shape indices and time series features of the TLPs of different consumers, based on the K-means algorithm, the TLPs of different consumers are clustered to analyze the power consumption behaviors of different types of consumers. Finally, an experiment shows that the proposed method can mine the typical power consumption behaviors of different consumers and improve the electrical load profile analysis efficiency and clustering quality. The main contributions of this paper are summarized as follows:

- Use the PAA and SAX method in time series data mining to analyze and filter the abnormal load profiles of the consumer, and then extract the typical load profile of the consumer.
- A time series analysis method is proposed to extract the load profile features of different consumers, which combines with the load characteristic indices and time series-based features. The extracted features can effectively improve the clustering quality.

The remaining of this paper is organized as follows. Section II proposes a TLP extraction method for a single consumer. Then, section III introduces cluster analysis of TLPs of different consumers based on the proposed method. Next, an experiment is carried out in Section IV. Finally, Section V briefly summarizes the paper.

II. TYPICAL LOAD PROFILE EXTRACTION FOR A SINGLE CONSUMER

The process of extracting the TLP of a single consumer mainly includes five steps: input raw load data, raw load data preprocessing, load data dimension reduction, load data re-expression, consumer typical load profile extraction. The diagram of these steps is shown in Figure 1.

A. RAW LOAD DATA PREPROCESSING

Data preprocessing is an essential step for any data mining method. In this paper, we consider a point as an outlier and remove it, which falls outside of 3 standard deviations, $3\sigma$, of the mean, $\mu$, of the daily load data $x(t)$. And then, we use the Z-score to normalize the daily load data to have
an approximate 0 mean and a standard deviation of close to 1 before converting it to the PAA representation. The main purpose of this step is to eliminate the effects of certain gross influences [20], so as to have a better comparison with load profiles of different consumers.

Assuming that the raw daily load data is \( X = \{x_1, \ldots, x_n\} \), outliers in the raw daily load data are removed firstly, and then the processed data is transformed into \( X' = \{x'_1, \ldots, x'_n\} \) with a mean of 0 and a standard deviation of 1. The \( i \)-th element of \( X' \) is calculated by the following equation:

\[
x'_i = \frac{x_i - \mu}{\sigma}, \quad i = 1, 2, \ldots, n
\]

where \( x_i \) and \( x'_i \) are the actual load data and Z-normalized load data at the \( i \)-th moment; \( n \) is the number of load sampling points in a day; \( \mu \) and \( \sigma \) represent the mean and standard deviation of daily load data, respectively.

### B. LOAD DATA DIMENSION REDUCTION

The Piecewise Aggregate Approximation (PAA) method is used to reduce the dimension of the preprocessed load data. This method is intuitive and highly efficient, and it can effectively reflect the overall trend of the time series. The PAA method is a feature representation method of time series.

First, a Z-normalized load data \( X' \) is divided into \( w \) equal-length sub-sequences. Then the mean value of each sub-sequence becomes the representation of data-reduced value. Performing PAA dimension reduction on the Z-normalized load data \( X' \) can obtain the PAA representation of daily load data, \( \tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_w\} \). The \( i \)-th element of \( \tilde{X} \) is calculated by the following equation:

\[
\tilde{x}_i = \frac{w}{n} \sum_{j=\frac{i}{w}(i+1)+1}^{\frac{i}{w}} x'_j
\]

where, \( w \) is the dimensions of the PAA representation. Typically, the number \( w \) is much smaller than the number \( n \) of the raw load sampling points. \( \tilde{x}_i \) is the value of the PAA representation of daily load data.

### C. LOAD PROFILE RE-EXPRESSION

The re-expression of the load profile is to use the symbolic aggregation approximation (SAX) method [21] to assign symbolic strings to the PAA dimension-reduced data, that is, using a discrete string to represent the consumer load profile. By means of re-expressing the load profile based on SAX, we can find Discords (in frequent load profiles) and Motifs (the most frequent load profiles) of the consumer. So we can get the TLP of consumer without the impacts of Discords, which means the TLP is much more accurate to reflect the typical power consumption behaviors of consumers.

The SAX representation is a time series symbolic representation method based on PAA dimension reduction. Having transformed a load profile into PAA, we can symbolize \( \tilde{X} \) into a discrete string using SAX. The value of each sub-sequence of PAA has been calculated and alphabetic character is assigned according to where the value lies within a set of vertical breakpoints, \( B = \{\beta_1, \ldots, \beta_{w-1}\} \). These breakpoints are determined by a chosen alphabet size \( \alpha \) that is the number of equiprobable areas under Gaussian distribution curve, so we can look them up in a Gaussian distribution table, as seen in Table 1. Finally, we can obtain a SAX word, \( \hat{X} = \{\hat{x}_1, \ldots, \hat{x}_w\} \), to re-express the load profile of consumer.

### TABLE 1. Gaussian distribution table.

| \( \beta \) | 3   | 4   | 5   | 6   |
|------------|-----|-----|-----|-----|
| \( \beta_1 \) | -0.43 | -0.67 | -0.84 | -0.97 |
| \( \beta_2 \) | 0.43 | 0 | -0.25 | -0.43 |
| \( \beta_3 \) | 0.67 | 0.25 | 0.84 | 0.43 |
| \( \beta_4 \) | - | - | - | 0.97 |

To illustrate the process of re-expression, an example transforming a load profile with 96 points into a SAX word is seen in Figure 2. We set the dimensions of PAA representation \( w = 8 \) and the alphabet size \( \alpha = 4 \) (the alphabet = \{a, b, c, d\}), so the load profile is mapped to the SAX word “caabdcbe”.

### FIGURE 2. Schematic diagram of load profile re-expression.

### D. CONSUMER TYPICAL LOAD PROFILE EXTRACTION

Based on the PAA method, the consumer load profiles within a certain period of time are reduced, and then the SAX method is used to transform the reduced consumer load profiles into the SAX words. Once the SAX words are created, we can observe the discords and motifs intuitively and identify whether a load profile belongs to discords or motifs. The consumer’s discords mean these load profiles are infrequent power consumption behaviors, and they cannot represent the typical power consumption behaviors of the consumer. So, in the paper, we propose a method to extract the consumer TLP from the motifs.

Assuming that after the consumer load profile is re-expressed, the number of SAX words that appear most is \( r \),
that is the number of the motifs is \( r \), and the load profile corresponding to the SAX words is considered to be the motifs of the consumer. Denote that the consumer TLP data is \( \tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_n] \), and the \( i \)-th element of \( \tilde{X} \) can be obtained from equation (3).

\[
\tilde{x}_i = \frac{1}{r} \sum_{k=1}^{r} x_{ki}, \quad i = 1, \ldots, n
\] (3)

Take the load profiles of a consumer for two weeks as an example, extract the TLP of the consumer. Here, we set the dimensions of PAA representation \( w = 8 \) and the alphabet size \( \alpha = 4 \). Figure 3 shows the process of converting the consumer’s load profiles for two weeks into SAX words in the form of a Sankey diagram. As can be seen from the figure, the most frequent SAX word is “bcaa” and the SAX word corresponding to the load profiles is what we call the motifs. Therefore, the TLP of the consumer for the two weeks can be extracted from equation (3).

III. CLUSTER ANALYSIS OF TYPICAL LOAD PROFILES OF DIFFERENT CONSUMERS

Time series clustering can be simply based on the raw data, or based on time series features. The feature selection method can map the time series in the high-dimensional space to the low-dimensional feature space, thus achieving the purpose of data dimensionality reduction, and meanwhile the reduced data can effectively reflect the information of the raw time series [22].

The consumer TLP reflects the consumer electricity consumption behaviors, however the consumer’s behaviors are greatly affected by the weather, type of consumer, electricity price policy and so on, leading to the diversity of the consumer TLP. Selecting an effective TLP feature of a consumer can reflect the inherent characteristics of the consumption’s electricity consumption behavior, and at the same time can reduce the complexity of clustering calculation and improve the clustering analysis performance.

In this paper, we purpose the features of consumer load profile are combined with conventional load shape indices and time series-based features. And then the TLPs of different consumers are clustered based on K-means algorithm as a demonstration. Clustering validity indices are used to evaluate the clustering quality of different feature selection method.

A. LOAD PROFILE FEATURE SELECTION

1) LOAD SHAPE INDICES

The load shape indices refer to the load characteristic indices commonly used in the power system, such as load factor, peak-valley difference, mean load and other 15 indices. These indices can capture reflect the characteristics and power consumption behaviors of consumer. Reference [17] selects serval the most relevant indices which contain information about the load profile of each consumer. In this paper, we have selected 6 load shape indices, including load factor, maximum utilization rate, the peak-valley difference ratio, peak load factor, flat load factor and valley load factor, which can comprehensively reflect the power consumption behaviors of various consumers, as shown in Table 2.

| Parameter | Definition | Period of definition |
|-----------|------------|---------------------|
| Load factor | \( I_1 = P_{av}/P_{max} \) | a whole day |
| Maximum utilization rate | \( I_2 = P_{ave}/(24P_{max}) \) | a whole day |
| The peak-valley difference ratio | \( I_3 = (P_{max} - P_{min})/P_{max} \) | a whole day |
| Peak load factor | \( I_4 = P_{av}/P_{peak} \) | 08:00—11:00, 18:00—21:00 |
| Flat load factor | \( I_5 = P_{ave}/P_{av} \) | 06:00—08:00, 11:00—18:00, 21:00—22:00 |
| Valley load factor | \( I_6 = P_{ave}/P_{val} \) | 22:00—24:00, 00:00—06:00 |

where \( P \) is the load of the consumer, the subscripts sum, av, max, min represent the total amount, mean value, maximum value, and minimum value respectively, and peak, val, and sh represent peak, valley, and flat periods respectively.

2) FEATURE SELECTION OF LOAD PROFILE BASED ON TIME SERIES

Although the above load shape indices can capture the shape feature of load profile, these features cannot reflect the complexity and volatility of the load profile. Therefore, in this paper we introduce 4 load profile features base on time series data mining, including binned entropy, complexity-invariant distance, nonlinear metrics and mean absolute change.

a: BINNED ENTROPY(BE) [23]

By dividing the value of the load profile \( X \) into serval equidistant bins, the interval \([\min(X), \max(X)]\) can be divided equally into \( k \) bins, then the value of the load profile will be distributed among the \( k \) bins. According to this equidistant bin, the entropy of this probability distribution can
be calculated.

$$BE(X) = - \sum_{k=0}^{\min(max\ bins,\ len(X))} p_k \log(p_k) \cdot 1_{(p_k>0)} \tag{4}$$

where $p_k$ is the probability that the value of the load profile falls in the $k$-th bin, $max\ bins$ is the number of bins, which represent the length of the load profile.

**b: COMPLEXITY-INVARIANT DISTANCE (CID) [24]**

The value of CID can measure the complexity of the load profile. A larger CID value indicates that the load profile is more complex and volatile, with more peaks and valleys.

$$CID = \sqrt{\sum_{i=0}^{n-2\ lag} (x_i - x_{i+1})^2} \tag{5}$$

where $x_i$ is the consumer load profile at the $i$-th time, $n$ is the number of load sampling points in a day, and $lag$ is the lag order.

**c: NONLINEAR METRICS (NM) [25]**

The value of NM measures the degree of non-linearity of the load profile, which can capture the fluctuation of the load profile.

$$NM = \frac{1}{n - 2\ lag} \sum_{i=0}^{n-2\ lag} x_{i+2\ lag}^2 \cdot x_{i+lag} \cdot x_i \tag{6}$$

where $x_i$ is the consumer load profile at the $i$-th moment, $n$ is the number of load sampling points in a day, and $lag$ is the lag order.

**d: MEAN ABSOLUTE CHANGE (MAC)**

The value of MAC refers to the arithmetic mean value of the absolute value of the difference between the load value at the time before and after.

$$MAC = \frac{1}{n} \sum_{i=1}^{n-1} |x_{i+1} - x_i| \tag{7}$$

where $x_i$ is the consumer load profile at the $i$-th moment, $n$ is the number of load sampling points in a day.

**B. CLUSTERING ANALYSIS OF CONSUMER TYPICAL LOAD PROFILE**

In this paper, we use the K-means algorithm as a demonstration for clustering analysis of TLP of consumers, and studies the similarity of TLP of different consumers. K-means clustering algorithm is a hard clustering algorithm based on partition. The sum of square errors (SSE) is used as the objective function to measure the tightness of clusters with a smaller error being more desirable. The specific steps of K-means algorithm are as follows:

**Step 1** Take the TLPs of 96 points per day of M consumers, $M = \{X_1, \ldots, X_m\}$ and extract the load profile features $M = \{T_1, \ldots, T_m\}$, which is reduced from the original $m \times 96$ dimensions to $m \times 10$ dimensions. Partition of the TLPs into $k$ non-empty subsets, $S = \{S_1, S_2, \ldots, S_k\}$ and randomly select $k$ initial cluster centroids $\mu_i (i = 1, 2, \ldots, k)$.

**Step 2** Calculate the Euclidean distance $d = ||T_j - \mu_i||$ between each load profile feature $T_j (j = 1, 2, \ldots, m)$ and the cluster centroid $\mu_i$, and divide the load profile according to the minimum distance to form a cluster $S_i (i = 1, 2, \ldots, k)$.

**Step 3** Take the mean value of the load profile in each cluster as the updated new cluster centroid, through calculating the equation (8)

$$\mu_i = \frac{1}{|S_i|} \sum_{T_j \in S_i} T_j \tag{8}$$

**Step 4** Repeat the above steps until the cluster centroid no longer changes.

In this paper, we use the “elbow method” [26] to determine the optimal number of clusters $k$ of K-means. Calculating the SSE, drawing a $k$-SSE curve, we can determine the number of clusters $k$ through observing the inflection point in the graph. Besides, we use “K-means++” method to choose initial clustering centers for K-mean clustering to speed up convergence and it can greatly reduce the impact of the initial clustering centers.

**C. CLUSTERING VALIDITY INDICES**

Here we use Silhouette (SIL), Calinski-Harabas (CH) and Davies-Bouldin Index (DBI) to evaluate the clustering quality.

1) SIL INDEX

SIL index is a measure of how similar an object is to its own cluster compared to other clusters. The higher the SIL index value, the better the clustering quality, and the value range is $-1$ to $1$. The SIL index is calculated by the following equation:

$$SIL = \frac{b - a}{\max(a, b)} \tag{9}$$

where $a$ is the mean distance between the sample and all other points in the same cluster, and $b$ is the mean distance between the sample and all other points in the next closest cluster.

2) CH INDEX

The CH index measures the tightness by calculating the sum of the squares of the distance between the sample within the cluster and the cluster center (i.e. the within-class dispersion matrix), and measures the separation by calculating the between-class dispersion matrix. The CH index value is the ratio of the degree of separation to the tightness. The larger the CH index value, the better the clustering effect.

$$CH = \frac{\text{tr}(B_k) \cdot m - k}{\text{tr}(W_k) \cdot k - 1} \tag{10}$$

where $m$ is the number of samples in the training set and $k$ is the number of clusters. $B_k$ is the covariance matrix between
clusters, and $W_k$ is the covariance matrix of data within the clusters. $tr$ is the trace of the matrix.

3) DBI INDEX
The DBI index calculates the ratio of the sum of the distances within the cluster to the distance outside the cluster. The DBI index is calculated with the following equation:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{M_{ij}}$$  \hspace{1cm} (11)$$

where $s_i$ is the mean distance between all samples in the $i$-th class and their cluster centroids; $M_{ij}$ is the distance between the $i$-th class and the $j$-th cluster centroids. The smaller the DBI, the better the clustering quality.

IV. EXPERIMENTS AND RESULTS
The overall process of the electrical load profile analysis method based on time series data mining proposed in this paper is shown in Figure 4.

In this paper, we use a UCI Electricity Load Diagrams 2011–2014 dataset [27] as an example for experiment. The dataset includes the electricity consumption data of 370 consumers in Portugal from the start time of 2011 to the end of 2014 with a sampling period of 15min, that is 96 sampling points per day, a total of 140256 sample points. Taking into account the different start times of data collection for different consumers, we pick the consumer load data in 2012 for analysis, and exclude consumers who have not completed records in 2012, leaving 317 consumers.

A. TYPICAL LOAD PROFILE EXTRACTION OF A SINGLE CONSUMER
One consumer MT_166 among 317 consumers is selected as an example to extract its TLP. All daily load profiles of the consumer in 2012 are shown in Figure 2, and we can notice that there are abnormal values and infrequent patterns of load profile.

For all daily load profile of the consumer MT_166, we firstly reduce the dimension of the daily load profile through PAA, and then use SAX method to transform each load profile into SAX word. Here we set the dimensions of PAA representation $w = 4$ and the SAX alphabet size $\alpha = 3$. Table 3 shows the numbers of the SAX words of all daily load profiles of the consumer MT_166.

| SAX word | numbers | SAX word | numbers |
|----------|---------|----------|---------|
| abcc     | 181     | bbcc     | 3       |
| abcb     | 88      | aaaa     | 2       |
| aabb     | 66      | aabb     | 2       |
| accc     | 12      | abca     | 2       |
| bccc     | 9       | aacb     | 1       |

Figure 6 shows part of the load profiles of the discords, whose SAX words are “bbcc”, “aaaa”, “aabb”, “abca”, “aacb”. We can observe that these load profiles may not be the typical power consumption behavior of the consumer, so these load profiles must be eliminated when extracting the TLP of the consumer.
It can be seen from Table 3 that the SAX words of the load profiles with the most occurrences of the consumer MT_166 is “abcc”, and the load profiles represented by the SAX word “abcc” is shown in Figure 7. Compared with Figure 5, Figure 7 has eliminated the consumer’s abnormal and infrequent load profiles, and then we can obtain the consumer’s typical electricity consumption behavior, from equation (3), which is considered as the consumer’s TLP.

**B. CLUSTER ANALYSIS OF TYPICAL LOAD PROFILES OF DIFFERENT CONSUMERS**

Using the same method in Section A, the TLPs of all 317 consumers can be extracted, as shown in Figure 8.

Based on the load profile feature selection proposed in Section III, the TLP features of all consumers are extracted, and then cluster analysis is performed on all consumers based on the K-means algorithm. Use the “elbow method” to determine the optimal number of clusters for K-means, gradually increase the value of $k$ from small to large, observe the changes in SSE, and draw a $k$-SSE line chart, as shown in Figure 9.

Observing the $k$-SSE line chart in Figure 9, we can see that the number of clusters $k = 4$ is the most appropriate. In this paper, we firstly choose the following two methods as the contrast method: Contrast method 1 is clustering directly based on the original load data, Contrast method 2 is clustering based on conventional load shape indices. The clustering results of different methods are shown in Figure 10-12.

The clustering results of the method in this paper are shown in Figure 10. The characteristics of the load profiles of various consumers can be clearly observed: the electricity consumption behaviors of the Type 1 and Type 3 consumers are relatively similar, and both of them are unimodal consumers. The difference is the power consumption peak period of the first type consumers is longer, mainly concentrated in 8:00-21:00, while third type consumers is mainly concentrated in 11:00-20:00. Our approach can distinguish these two type consumers much better. Type 2 Consumers are bimodal, and the peak power consumption period is mainly concentrated...
at 9:00-12:00 in the morning and 18:00-21:00 in the evening; the load profile fluctuations of the fourth type consumers are tiny, and the load profiles are almost flat. The clustering results of the contrast method 1 are shown in Figure 11. The clustering results are similar to the proposed method, but the classification effect of the Type 1 and Type 3 consumers is not good, and the second type of consumers clustering results are even more poor. The clustering result of the contrast method 2 is shown in Figure 11. The clustering quality is the worst, and the Type 4 consumers cannot be clearly distinguished, and the electricity consumption behaviors of all types of consumers cannot be reflected efficiently.
Therefore, a comprehensive observation of the cluster analysis results in Figures 10-12 shows that the cluster analysis directly based on the original load data can roughly distinguish the difference between different type of consumer load profiles, but it is not able to classify each consumer precisely and accurately; The result of clustering analysis based on the conventional load shape indices is far from satisfying, indicating that the conventional load shape indices have lost a lot of detailed information of the load profile and cannot effectively reflect the difference in the shape of the load profile. The clustering quality of the method proposed in this article is much better than the other two methods, and our approach can reflect the difference in load profile shape and distinguish the electricity consumption behaviors of each consumer accurately.

Next, we compare the effectiveness of the proposed method with the traditional data-driven feature dimension reduction method. Two common data dimension reduction methods are selected to reduce the dimension of original load data, which are principal component analysis (PCA) [28] and locally linear embedding (LLE) [29]. Then cluster analysis is conducted based on the data after reduction features and the clustering results are shown in Figure 13-14. Among them, PCA algorithm is set to extract the principal components whose variance contribution rate reaches 98%, and the algorithm determines the number of principal components is 8 in the end; the parameter setting of LLE algorithm is: the number of nearest neighbors is 30, and the dimension of low dimensional feature space is 8.

Observing the Figures 13-14, we can see that the clustering results of PCA+Kmeans are similar to the proposed method, and the clustering quality is fine; while LLE+Kmeans has a slightly poor clustering quality, which is not good for the Type 4 consumers. However, the method of data-driven feature dimension reduction cannot reflect the physical meaning of each feature dimension. The feature selection of the method proposed in this paper combines the time series analysis method and the load shape indices of load profile. Each feature has a clear meaning, which can better reflect the electrical consumption behaviors of consumer.

The clustering results of the above different methods are compared numerically. Table 4 compares the number of various consumers in the clustering results of different methods, and Table 5 compares the performance of different experiment methods based on clustering validity indices.

It can be seen from Table 4 that the number of clustering profiles of the proposed method is similar to that of the contrast method 1 and PCA+Kmeans method, while the other two methods are quite different. From the perspective of clustering validity indices, compared with other methods, as seen in Table 5, the proposed method has higher SIL value, lower DBI value and higher CH value, which indicates that the clustering quality of the proposed method is much better, and the extracted load profile features can improve the quality of clustering.

V. CONCLUSION

By introducing the Symbol Aggregate approXimation analysis method in time series data mining, we propose a consumer TLP extraction method in this paper, which can better extract the consumer’s electricity consumption behavior, effectively reflect the overall trend of the consumer load profiles and eliminate infrequent profiles. Then combined with the load shape indices and time series characteristics of the TLPs of different users, cluster analysis of the TLPs of different users based on the K-means clustering algorithm. The experiments and results show that the proposed method in this paper can be applied to load profile extraction and cluster analysis to greatly improve the analysis efficiency and clustering quality, which verifies the feasibility and effectiveness of the method.

Studying the characteristics of the load profile of different power consumers and digging out the power consumption rules of consumers can help electric utilities to understand the differences in power consumption characteristics of different consumers and provide guidance to formulate power supply strategies and optimize power supply facilities planning and construction. With the widespread use of power big data technology and the promotion of demand-side response, the next step will focus on how to consider the impact of demand-side
response on consumer load profiles and how to apply the method in this article to actual systems.

REFERENCES

[1] R. R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, “A survey on advanced metering infrastructure,” Int. J. Elect. Power Energy Syst., vol. 63, pp. 473–484, Dec. 2014.

[2] Y. Wang, Q. Chen, T. Hong, and C. Kang, “Review of smart meter data analytics: Applications, methodologies, and challenges,” IEEE Trans. Smart Grid, vol. 10, no. 3, pp. 3125–3148, May 2019, doi: 10.1109/TSG.2018.2818167.

[3] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” Energy, vol. 42, no. 1, pp. 68–80, Jun. 2012.

[4] D. Xu, B. Zhou, Q. Wu, C. Y. Chung, C. Li, S. Huang, and S. Chen, “Integrated modelling and enhanced utilization of Power-to-ammonia for high renewable penetrated multi-energy systems,” IEEE Trans. Power Syst., early access, Apr. 22, 2020, doi: 10.1109/TPWRS.2020.2989533.

[5] P. Zhang, Y. Xu, F. Luo, and Z. Y. Dong, “Power big data: New assets of electric power utilities,” J. Energy Eng., vol. 145, no. 3, Jun. 2019, Art. no. 04019009.

[6] J. Yang, J. Zhao, F. Luo, F. Wen, and Z. Y. Dong, “Decision-making for electricity retailers: A brief survey,” IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 4140–4153, Sep. 2018, doi: 10.1109/TSG.2017.2651499.

[7] X. Liu, A. Heller, and P. S. Nielsen, “CITIESData: A smart city data management framework,” Knowl. Inf. Syst., vol. 53, no. 3, pp. 699–722, Dec. 2017.

[8] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, “A review of deep learning for renewable energy forecasting,” Energy Convers. Manage., vol. 198, Oct. 2019, Art. no. 111799.

[9] J. Hu and A. V. Vasilakos, “Energy big data analytics and security: Challenges and opportunities,” IEEE Trans. Smart Grid, vol. 7, no. 5, pp. 2423–2436, Sep. 2016, doi: 10.1109/TSG.2016.2563461.

[10] D. Xu, Q. Wu, B. Zhou, C. Li, L. Bai, and S. Huang, “Distributed multi-energy operation of coupled electricity, heating and natural gas networks,” IEEE Trans. Sustain. Energy, early access, Dec. 23, 2020, doi: 10.1109/TSTE.2019.2961432.

[11] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, “Clustering load profiles for demand response applications,” IEEE Trans. Smart Grid, vol. 10, no. 2, pp. 1599–1607, Mar. 2019.

[12] M. J. Hossain, A. N. Kabir, M. M. Rahman, B. Kabir, and M. R. Islam, “Determination of typical load profile of consumers using fuzzy c-means clustering algorithm,” Int. J. Soft Comput. Eng., vol. 1, no. 5, pp. 169–173, Nov. 2011.

[13] A. Capozzoli, M. S. Piscitelli, and S. Brandi, “Mining typical load profiles in buildings to support energy management in the smart city context,” Energy Procedia, vol. 134, pp. 865–874, Oct. 2017.

[14] R. Granell, C. J. Aixon, and D. C. H. Wallom, “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles,” IEEE Trans. Power Syst., vol. 30, no. 6, pp. 3217–3224, Nov. 2015.

[15] H. Wang, Y. Liu, B. Zhou, C. Li, G. Cao, N. Voropai, and E. Barakhtenko, “Taxonomy research of artificial intelligence for deterministic solar power forecasting,” Energy Convers. Manage., vol. 214, Jun. 2020, Art. no. 112909.

[16] R. Al-Otaibi, N. Jin, T. Wilcox, and P. Flach, “Feature construction and calibration for clustering daily load curves from smart-meter data,” IEEE Trans. Ind. Informat., vol. 12, no. 2, pp. 645–654, Apr. 2016.

[17] V. Figueroa1, 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, pp. 596–602, May 2005.

[18] C. Dinesh, S. Makonin, and I. V. Bajic, “Residential power forecasting based on affinity aggregation spectral clustering,” IEEE Access, vol. 8, pp. 99431–99444, 2020.

[19] M. Emmanuel and J. Giraudel, “Net electricity clustering at different temporal resolutions using a SAX-based method for integrated distribution system planning,” IEEE Access, vol. 7, pp. 123689–123697, 2019.

[20] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” in Proc. 8th ACM SIGMOD workshop Res. Issues Data Mining Knowl. Discovery, 2003, pp. 2–11.

[21] J. Lin, E. Keogh, L. Wei, and S. Lonardi, “Experiencing SAX: A novel symbolic representation of time series,” Data Mining Knowl. Discovery, vol. 15, no. 2, pp. 107–144, Aug. 2007.

[22] L. G. B. Ruiz, M. C. Pegalajar, R. Arcucci, and M. Molina-Solana, “A time-series clustering methodology for knowledge extraction in energy consumption data,” Expert Syst. Appl., vol. 160, Dec. 2020, Art. no. 113731.

[23] M. Christ, A. W. Kempa-Liehr, and M. Feindt, “Distributed and parallel time series feature extraction for industrial big data applications,” 2016, arXiv:1610.07717. [Online]. Available: http://arxiv.org/abs/1610.07717

[24] G. E. A. P. A. Batista, E. J. Keogh, O. M. Tataw, and V. M. A. de Souza, “CID: An efficient complexity-invariant distance for time series,” Data Mining Knowl. Discovery, vol. 28, no. 3, pp. 634–669, May 2014.

[25] T. Schreiber and S. Andreas, “Discrimination power of measures for nonlinearity in a time series,” Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 55, no. 5, p. 5443, May 1997.

[26] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in K-means clustering,” Int. J., vol. 1, no. 6, pp. 90–95, 2013.

[27] D. Dua, and C. Graff, UCI Machine Learning Repository, Irvine, CA, USA: University of California, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[28] F. M. Bianchi, E. De Santis, A. Rizzi, and A. Sadeghian, “Short-term electric load forecasting using echo state networks and PCA decomposition,” IEEE Access, vol. 3, pp. 1931–1943, 2015.

[29] A. A. H. Barocio, J. C. Ayon, and H. A. Garcia-Baleon, “Comparison of dimensionality reduction techniques for clustering and visualization of load profiles,” in Proc. IEEE PES Transmiss. Distrib. Conf. Exposition-Latin Amer., Sep. 2016, pp. 1–6.

YING SHI received the B.S. degree in electrical engineering from South China Agricultural University, Guangzhou, China, in 2017. He is currently pursuing the M.S. degree in electrical engineering with the South China University of Technology, Guangzhou. His research interests include optimization and data analytics of power and energy systems.

TAO YU (Member, IEEE) received the B.Eng. degree in electrical power system from Zhejiang University, Hangzhou, China, in 1996, the M.Eng. degree in hydroelectric engineering from Yunnan Polytechnic University, Kunming, China, in 1999, and the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2003. He is currently a Professor of power system with the School of Electric Power, South China University of Technology (SCUT), Guangzhou, China. His research interests include nonlinear and coordinated control theory, artificial intelligence techniques in planning, and operation of power systems.
QIANJIN LIU (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in power plant engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1989 and 1995, respectively. He is currently an Associate Professor with the College of Electric Power, South China University of Technology, Guangzhou, China. His research interests include distribution network operation control and its advanced application analysis and distribution network data fusion model and its information applications.

HANXIN ZHU received the B.S. degree in electrical engineering from the South China University of Technology, Guangzhou, China, in 2018, where he is currently pursuing the M.S. degree in electrical engineering with the College of Electric Power. His research interest includes optimal operation of integrated energy systems.

FUSHENG LI received the B.Eng. degree in electrical engineering from the South China University of Technology, Guangzhou, China, in 2018, where he is currently pursuing the M.S. degree with the School of Electric Power Engineering. His research interests include artificial intelligence techniques and its application in smart grid.

YAXIONG WU received the B.S. and M.S. degrees in electrical engineering from the South China University of Technology, Guangzhou, China, in 2015. He is currently a Senior Engineer with the Grid Planning and Research Center, Guangdong Power Grid Company Ltd., CSG, Guangzhou. His research interest includes distribution network planning.