Electric Load Clustering in Smart Grid: Methodologies, Applications, and Future Trends

Caomingzhe Si, Shenglan Xu, Can Wan, *Member*, *IEEE*, Dawei Chen, Wenkang Cui, and Junhua Zhao, *Senior Member*, *IEEE*

***Abstract*—With the increasingly widespread of advanced metering infrastructure, electric load clustering is becoming more essential for its great potential in analytics of consumers' energy consumption patterns and preference through data mining. Moreover, a variety of electric load clustering techniques have been put into practice to obtain the distribution of load data, observe the characteristics of load clusters, and classify the components of the total load. This can give rise to the development of related techniques and research in the smart grid, such as demand-side response. This paper summarizes the basic concepts and the general process in electric load clustering. Several similarity measurements and five major categories in electric load clustering are then comprehensively summarized along with their advantages and disadvantages. Afterwards, eight indices widely used to evaluate the validity of electric load clustering are described. Finally, vital applications are discussed thoroughly along with future trends including the tariff design, anomaly detection, load forecasting, data security and big data, etc.**

***Index Terms*—Electric load clustering, similarity measurement, clustering technique, cluster validity indicator, smart grid.**

I. Introduction

With the development of the smart grid, advanced metering infrastructure (AMI) has been gradually popularized. By 2020, the cosmopolitan smart meter installation is expected to reach 780 million [1]. AMI can achieve multi-dimension data measurement of millions of customers with fine-grained data. Therefore, it makes data in the smart grid developed to be real-time and diverse with a higher resolution and a larger volume, which provides an ideal environment for the research of electric load clustering based on data mining techniques. By analyzing the load data and related influencing factors, electric load clustering can extract the power consumption patterns as well as characteristics, and realize consumer classification, thereby supporting progress in other smart grid fields such as load forecasting and demand-side response (DSR). Prior to the widespread distribution of AMI, through statistics, utilities can only obtain the monthly consumption and grid-connected information of households, e.g., voltage levels and nominal demand. By accessing and analyzing load data from AMI, electric load clustering can help obtain the usage habits of households, and even assess the impacts of various variables on consumption patterns, which can be cast into residential characteristics, demographic and socio-economic factors, attitudes toward energy usage (e.g., attention towards energy conservation), power consumption knowledge, and energy efficiency goals [2], [3]. Based on the valuable knowledge mined, electric load clustering helps utilities better implement energy policy and infrastructure planning strategies.

The large-scale integration of renewable energies and the development of electric vehicles also make electric load clustering more meaningful and necessary. On the one hand, the increasing integration of renewable energies [4], [5] has long been a vision of the smart grid, however, the inherent fluctuations, intermittence, and uncertainty generally pose new challenges to modern power systems [6], [7]. In line with international climate goals and national sustainable development plans, renewable energy will occupy more than two-thirds of the global generation by 2040, with solar photovoltaic (PV) and wind energy occupying 40% of the total generation [8]. On the other hand, the commissioning of plug-in electric vehicles (PEVs) and hybrid electric vehicles is expected to increase considerably [9]-[11]. The low controllability of electric vehicles and the stronger coupling between the supply and demand sides further enhance the complexity of the safe and stable operation of the power grid. Demand-side management (DSM) and storage [12] have already aroused great interest in addressing challenges in predictability and controllability of future power supply and system stability. Clustering approaches [13]-[15] have been used to discover power usage patterns based on various measures of power usage data [16]-[20]. The excavation of consumption patterns can help improve the accuracy of load forecasting and support DSR.

Manuscript received: July 16, 2020; accepted: October 12, 2020. Date of CrossCheck: October 12, 2020. Date of online publication: March 9, 2021.

This work was supported in part by the National Natural Science Foundation of China (No. 51877189), National Natural Science Foundation of China Joint Program on Smart Grid (No. U2066601), and Young Elite Scientists Sponsorship Program by China Association of Science and Technology (No. 2018QNRC001).

C. Si is with the School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China (e-mail: 219019004@link.cuhk.edu.cn).

S. Xu, C. Wan (corresponding author), D. Chen, and W. Cui are with the College of Electrical Engineering, Zhejiang University, Hangzhou, China (e-mail: shenglanxu@zju. edu. cn; canwan@zju. edu. cn; dwchen@zju. edu. cn; cuiwenkang@zju.edu.cn).

J. Zhao is with The Chinese University of Hong Kong (Shenzhen), Shenzhen, China, and he is also with Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China (e-mail: zhaojunhua@cuhk.edu.cn)

DOI: 10.35833/MPCE.2020.000472

This paper aims at giving a global view of electric load clustering techniques in the smart grid scenario. After expounding the concept of electric load clustering and the basic process, we provide a systematic summary of five major category clustering algorithms, including partition-based, hierarchical, density-based, grid-based and model-based algorithms. The advantages and disadvantages of each algorithm are compared and analyzed from the perspective of practical problems encountered in electric load clustering researches and aspects of improving directions. Similarity measurements that lay the basis of electric load clustering and evaluation indices commonly used in electric load clustering are also summarized in detail. After the summary of methodologies, a discussion is carried out on the applications and further development in electric load clustering brought by combinations of new progress which includes data privacy, artificial intelligence (AI), big data, renewable energy, and development of Energy Internet. The contributions in this paper can be summarized as follows.

1) Typical algorithms for electric load clustering are classified and summarized thoroughly with the analysis of advantages and disadvantages.

2) Similarity measurements and evaluation indices of electric load clustering algorithms are summarized in detail.

3) A view on the applications of electric load clustering in the smart grid is presented and its future trends are illustrated.

The remainder of this paper is organized as follows. Section II introduces the basic concept of electric load clustering and illustrates the five key parts in electric load clustering. Section III makes an analytic on several common similarity measurements of load data and gives a systematical summary on different electric load clustering algorithms. Section IV holds a mathematical discussion on evaluation indices of electric load clustering performance. Section V puts forward current applications and future trends of electric load clustering. Eventually, the conclusion is drawn in Section VI.

II. Concept and Process of Electric Load Clustering

A. Definition and Mathematical Interpretation

Clustering is a kind of popular unsupervised data mining technique. In the context of electric load, clustering can be used to mine and analyze load datasets provided by massive power consumers. Electric load clustering algorithms can divide massive load profiles into as many groups (clusters) as possible based on similarity evaluations of samples and discover potential patterns among consumers. Generally, the objective of clustering algorithms is to make instances belonging to the same cluster more similar than those belonging to different clusters, that is, high intra-cluster similarity and low inter-cluster similarity should be fulfilled. Mathematically, dataset D which contains n samples can be divided into K disjoint clusters C_1, C_2, …, C_K and the union of all clusters constitute the complete D.

$$D = \bigcup_{i=1}^{K} C_i \tag{1}$$

Especially, the concept of membership is introduced in fuzzy clustering, that is, each sample belongs to each cluster with a membership degree less than one numerically, and the sum of all membership degrees equals to one.

B. Basic Process for Electric Load Clustering

The process of electric load clustering is split into five phases: data preprocessing, data size reduction and feature selection, primary clustering stage, assessment of clustering performance, and formation/selection of customer clusters. It is visually interpreted in Fig. 1.

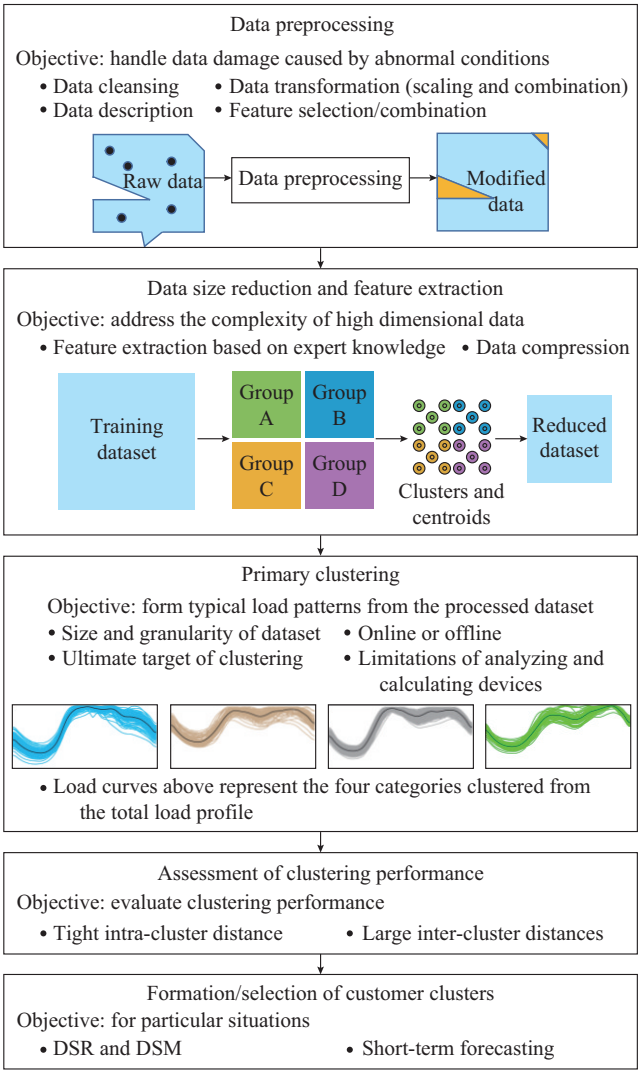

Fig. 1. Basic process of electric load clustering.

1) Data Preprocessing

In practical applications, "dirty data" are often encountered [21], and the data preprocessing mainly aims at handling data damage caused by abnormal conditions such as noise, extreme weather events, and faults. The negative effects on the later generalization can be evaded by discovering and processing lost data and outliers in load datasets. Raw data may share negative characteristics such as incom-

pleteness (empty attribute value), noise, inconsistency, redundancy (quantity of data or number of attributes exceeds the need), imbalance (quantities of data vary stupendously among different categories), outliers (data far from the majority in terms of numerical), and duplicate (data occurs multiple times). Data preprocessing stages fundamentally involve procedures which are data cleansing [22], data transformation such as scaling and normalization [15], data description, feature selection or feature combination, etc.

2) Data Size Reduction and Feature Extraction

This stage processes smart meter data before the primary clustering to reduce the size of input data or define more meaningful characteristics for the later clustering stage. The objective is to address the puzzle of high complexity caused by high dimensional data in subsequent algorithms and to better classify power consumers through various methods such as dimension reduction. The techniques mentioned in literature can be roughly categorized into feature definition that contains feature extraction based on expert knowledge and data compression. Particularly, in data compression, lossy compression and feature extraction are closely linked. An effective feature extraction mechanism is required to mitigate the data storage and processing burden. Methods such as symbolic aggregate approximation (SAX) [23], principle component analysis (PCA) [24], singular value decomposition (SVD) [25], and discrete wavelet transformation (DWT) [26] are the most prevailing for purpose of feature extraction. In addition, in [24], a convolutional autoencoder was utilized to reduce the dimensionality of original input data.

When using piecewise aggregate approximation (PAA) to reduce the data size for the load dataset, the time axis would be divided into several intervals, and the amplitudes within each one would be replaced with their respective averages [27]. Based on PAA, SAX divides the axis that represents the power consumption amplitude into intervals and uses appropriate symbols for the explicit representations of each interval. The SAX representation of the time-domain load profile is drawn according to the time intervals divided lastly. However, due to the segmentation properties of SAX, some partition-based clustering algorithms, e. g., *K*-means, cannot be applied for further processing after all transformations of load profiles, while hierarchical clustering and density-based spatial clustering of applications with noise (DBSCAN) are suitable for subsequent clustering processes.

The principle of applying PCA to load data size reduction is to select the dimension with large information entropy numerically in a smart meter dataset and remove the dimension with the small information entropy. Load data are transformed from the original coordinate system to a new one that consists of a few new orthogonal axes with the largest variance of raw data. Large variances represent large differences between different data, which contain a large amount of distinguishable information.

3) Primary Clustering

In the primary stage of electric load clustering, it is significant to select appropriate clustering algorithms for specific situations and set parameters reasonably. The choice highly depends on factors such as the size and granularity of the given dataset, the ultimate target of clustering, online or offline [28], limitations of analyzing and calculating devices. From the perspective of data dimension and size, extensive scales of consumers and attributes will increase the computational burden of the clustering algorithm that requires distance matrices, e.g., hierarchical clustering [29].

In some studies, integrated electric load clustering algorithms [30], [31] can speed up the procedure and achieve better results.

4) Assessment of Clustering Performance

Unlike classification tasks with certain optimal goals and learning processes, clustering tasks are unsupervised without certain objectives and uniform criterion to assess the validity [32]. Intuitively, since the clustering process does not involve labels, the quality is generally evaluated by internal evaluation indicators, which is on account of the calculation of closeness within each cluster and cluster separation. An ideal clustering performance results in tight intra-cluster and large inter-cluster distances [33]. On the other hand, electric load clustering can also be assessed by external evaluation indicators in the case that reference labels are given based on actual application requirements. Various clustering validity indices (CVIs) have been proposed to assess the results, e. g., the F-measure, Davies-Bouldin index (DBI) [34], mean square error (MSE) [35], silhouette coefficient (SC) [36], mean index adequacy (MIA) [37], ratio of within-cluster sum of squares to between cluster variation (WCBCR) [38], Dunn index, silhouette width criterion (SWC) [39], purity [40], and Rand index (RI) [41]. Moreover, evaluations based on fuzzy partitioning are known as Xie-Beni index and non-fuzzy index [42], [43].

5) Formation/selection of Customer Clusters

In view of specific electric load clustering scenarios, subsequent processing of formed clusters is required. Clusters should be chosen to meet the needs of particular applications. Herein, two classical scenes are implemented.

Electric load clustering can realize the classification of DSR resources and support fine implementations of DSM strategies. Supposing that electric load clustering is applied in the formation of DSM strategies, the ultimate number of clusters is not expected to exceed the predetermined sum restricted by practical application constraints, and the level of segmentation may be specified by the needs of distribution network operators (DNOs) and retailers. In this case, some clusters with similar patterns can be merged [44].

It has also aroused the interest of researchers about how to perform clustering on load datasets to improve the load forecasting accuracy. In general, consumers are merged by clustering based on similarities in consumption behaviors and then load forecasting is conducted for each colony. Especially, the dynamic characteristics of load are considered by researchers in order to support load forecasting preferably [19]. Since the target is to improve the result of load forecasting, evaluation indicators such as mean absolute percentage error (MAPE) common to load forecasting are used as the criterion to determine the formation of clusters [45].

III. Clustering Algorithm

This section begins with a summary of common measures in electric load clustering stage, followed by an overview of some classical electric load clustering algorithms. They are specifically designed due to different practical constraints such as the structure of provided load data, scale of dataset, dimensionality of data (sampling frequency), number of outliers, complexity requirement, dependency on the input order, and dependency on each consumer's preset or priori knowledge. To facilitate an intuitive understanding, we classify electric load clustering algorithms into five major categories which include partition-based, hierarchical, density-based, grid-based, and model-based clustering algorithms. Figure 2 depicts the electric load clustering algorithms mentioned in this paper, where the dark blue indicates the major categories, the light blue indicates the typical algorithms, the dark green indicates the subdivision of general algorithms, the green indicates the improvement of algorithms, and the light green indicates the specific algorithms.

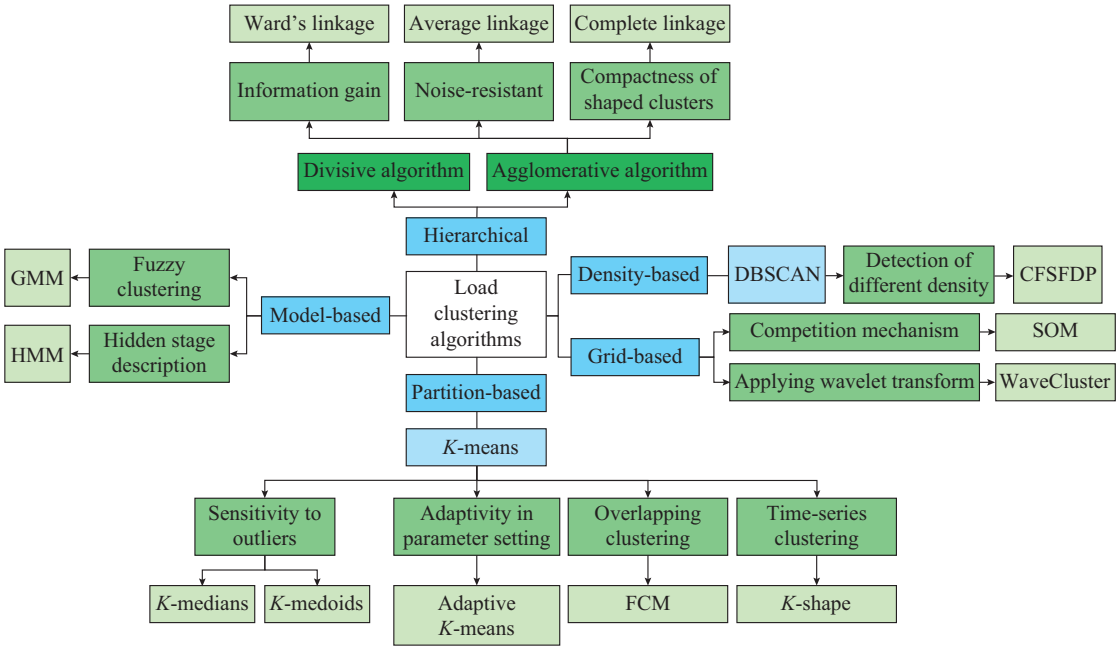

Fig. 2. Classification of electric load clustering algorithms.

A. Measurements in Electric Load Clustering

The concept of using similarity or dissimilarity metrics to construct clusters is prevalent in various types of clustering algorithms. The task is to group instances in a mathematical expression that represents the degree of similarity or dissimilarity between samples [14]. The selection of metrics depends on the data type, the significance of data magnitude, and the data sparsity. After calculating the similarity or dissimilarity among samples, load data are transformed into a similarity or dissimilarity space in the form of matrix analysis. Different kinds of measurements are summarized in [46] from a technical point of view.

1) Minkowski Distance

Minkowski distance (L_p-norm distance) is one of the most popular measures in various literature.

$$L_p = \left(\sum_{i=1}^{dim} \left| x_{di} - y_{di} \right|^p \right)^{\frac{1}{p}} \tag{2}$$

where dim is the feature dimension of samples; x_{di} and y_{di} are the samples of i^{th} dimension; and p is a constant.

This formula calculates the distance value based on the difference between the features of two objects. When $p=1$, the Minkowski distance is noted as Manhattan distance. Most commonly, when $p=2$, the Minkowski distance is known as Euclidean distance [47]. When $p \to +\infty$, the Minkowski distance is called Chebyshev distance. Among the distances listed above, Euclidean distance is most widely used for clustering smart meter data to discover residential load patterns [48], while the defect is that it is unable to recognize the relationship between different shapes [27]. The principle drawback of Minkowski distance is that it does not treat scales of individual components differently and does not consider differences in distributions of the individual components such as expectations and variances. The min-max method is employed for the sensitivity of Euclidean distance to the difference of load data amplitude [27].

2) Gower's Distance

Gower's distance is purposed to calculate the degree of similarity between instances i and j based on the attribute k, and a value s_{ijk} is assigned in the process based on the degree of similarity between the instances, and the possibility index δ_{ijk} of comparing i and j is put forward. When $\delta_{ijk}=1$, it means that the attribute k can be compared between i and j, otherwise the value equals 0. When $\delta_{ijk}=1$, s_{ijk} is unknown and is set to be 0 normally, and the similarity between i and j is defined as the average of all the possible comparisons (attributes may not exist or cannot be compared for the lack of information or in the case of dichotomous variables).

Gower's distance normalizes the variables in [0,1] and uses weighted linear combinations to calculate ultimate distance matrix. The advantage is that Gower's distance is easy to compute, but its drawback is the susceptibility to outliers of unstandardized continuous variables. Therefore, the data transformation process is essential, which will consume a large amount of memory.

In the agglomerative algorithm proposed in [49], Gower's distance was used to measure the similarity of both quantitative and qualitative information.

3) Canberra Distance

Canberra distance can be viewed as a weighted version of Manhattan distance. The measurement is very sensitive to nonnegative changes close to 0. Similar to Mahalanobis distance, the measurement is insensitive to the scale of data. However, Canberra distance assumes that variables are independent of each other, without considering the correlation between them. There is no significant difference in the computation cost between Canberra distance and Euclidean distance on the same dataset, but the correlation obtained by Canberra distance is always higher than that obtained by Euclidean distance [50].

A *K*-means-based load estimation method was proposed in [51] for the smart meter data of households, and the effects of adopting Canberra distance, Manhattan distance, Euclidean distance, and Pearson correlation coefficient were studied in the case.

4) Cosine Similarity

The similarity between two vectors can be evaluated by measuring the cosine of the angle between them. The result is independent of the length and only relevant to the direction of vectors. In data mining fields, cosine similarity is used to measure the degree of aggregation within clusters.

The similarity of load profiles in [52] was described by the angle cosine values between two vectors, where the elements were the hourly consumption data in the proposed statistic-fuzzy technique.

5) Pearson Correlation Coefficient

Pearson correlation coefficient is used to measure the correlation between sets X_i and X_j for linearly correlated continuous data of bivariate normal distribution within the range of $[-1, 1]$. The correlation between X_i and X_j enhances with the increase of the absolute value of the correlation coefficient, and a positive or negative coefficient indicates that X_i is positively or negatively correlated with X_j, respectively.

The extreme value in the dataset may result in a serious deviation of coefficient, so it needs to be processed before the calculation of Pearson correlation coefficient. In addition, Pearson correlation coefficient cannot detect more complex associations such as quadratic, cubic, and time varying relationship [53].

6) Kullback-Leibler (KL) Divergence

KL divergence based on information theory is to describe the difference between two probability distributions X and Y obtained from two load datasets, respectively.

For discrete random variables, we can obtain

$$D_{KL}(X \| Y) = \sum_i X(i) \ln \frac{X(i)}{Y(i)} \tag{3}$$

where $D_{KL}(X \| Y)$ is the difference between the two probability distributions X and Y; and $X(i)$ and $Y(i)$ are the probabilities of element i in the distributions X and Y, respectively.

For continuous random variables, we can obtain

$$D_{KL}(X \| Y) = \int_i X(i) \ln \frac{X(i)}{Y(i)} \mathrm{d}i \tag{4}$$

A symmetric generalized KL divergence was used in [54] as the measurement of distributions, where typical load patterns (TLPs) were extracted by hierarchical clustering.

7) Dynamic Time Warping (DTW)

DTW which applies dynamic programming is served for time-warping calculations. To evaluate the similarity of two time series, one can use DTW to find the optimal alignment between two time-correlated sequences. In many cases of electric load clustering, the linear scaling or stretching of load curves cannot take into account the variability in the duration of the segments in load curves. The practical application of DTW is to perform nonlinear processing on the time axis to make shapes of sequences as similar as possible. DTW is a frequently used measurement for comparing time-series datasets and has the highest alignment performance compared with other algorithms [55].

To better classify and predict consumption behaviors at the household level, a DTW-based approach was proposed in [19]. The approach has been optimized to better estimate which electrical devices will be used in which hours.

B. Partition-based Clustering Algorithm

The ideology of partition-based clustering algorithm is to achieve the effect that clusters are sufficiently distant from each other and samples in each cluster are sufficiently close to each other [56]. When applied to electric load clustering, the partition-based clustering algorithm creates K partitions on the original consumption dataset obtained by AMI. It selects K cluster centroids according to a predefined limit and then performs iterative relocation based on the predefined heuristic algorithm until the desired objective has been achieved. In fact, partition-based clustering needs to run many times in different initial states to get better results. *K*-means is widely used among partition-based clustering algorithms [56] in addition to fuzzy c-means (FCM) based on fuzzy theory, which can be considered as another typical example [57].

A number of improvements based on *K*-means have been proposed to deal with different limitations of prototype algorithms in different applications of electric load clustering. Algorithms that contains *K*-medoids [58] and *K*-medians [59] are applied to address the shortcoming that *K*-means is sensitive to noisy data and outliers occurred in the process of acquiring load data. Moreover, *K*-shape is proposed to address the shortcoming that *K*-means can only handle numerical samples and is not suitable for shape-like load samples [60].

1) K-means

K-means is the most popular and simplest electric load clustering algorithm for its outstanding computational efficiency on the large-scale and high-dimension load datasets [45]. In *K*-means, cluster centroids are computed as the average of cluster members. The ideal dissimilarity measurement

summarized from various literature is Euclidean distance generally. Although the algorithm has advantages of easy implementation and high efficiency, it has the drawback of not being applicable to the case with discrete characteristics, and cluster centroids may not be similar to any other instances when dealing with load data containing outliers or asymmetrically distributed load data. When the dataset is considerably large, there is a considerable probability that *K*-means will fall into a local optimal solution, so it is necessary to handle the non-convex problem under this circumstance.

Based on the original *K*-means using Euclidean distance as a similarity measure, a shape clustering was proposed in [35] based on the segmental slope of load curves, which showed the feasibility in improving the accuracy and efficiency of clustering by capturing shape features of load curves. A hierarchical *K*-means was developed in [61] to improve the clustering performance under big data problems in distributed AMI circumstances considering the shortcomings of local optimal solutions for *K*-means. To decompose daily usage patterns into total daily usage and standardized daily load patterns, a two-stage approach that consists of the adaptive *K*-means and hierarchical algorithms was used in [62].

2) K-medians

Since outliers may exist in load profiles in practical cases, *K*-medians, which can avoid the effect of outliers in smart meter data on the overall cluster generation, is introduced in electric load clustering research. Cluster centroids for *K*-medians are the median of all cluster members, and the optimal measurement of coherence is Manhattan distance (vector 1-norm). The advantage of choosing the cluster centroid as the median is that the effect on the median value is even negligible when there are particularly influential noisy data or outliers in load datasets obtained by AMI. Without loss of generality, compared with *K*-means, *K*-medians is more robust to asymmetric distributed data and outliers. Since cluster centroids may be different from any sample, the computation cost is higher.

A comparison of six algorithms was carried out in [59] for down-scaling annual load consumption data to a composite typical load demand day for the energy system. It took seasonal and monthly classification of electrical load into consideration along with complex electric load clustering algorithms. *K*-medians and *K*-medoids show the best performance when analyzing the design, operation, and cost structure of the resulting energy systems.

3) K-medoids

In electric load clustering scenarios, *K*-means requires that all load consumption samples to be in a Euclidean space and can be extremely error-prone to noise. For non-numerical samples, it is not possible to calculate real variables such as the mean. Different from *K*-means whose cluster centroids are distributed in continuous space, *K*-medoids can only take load consumption samples as its cluster centroids. The iterative steps are roughly the same for both algorithms.

K-medoids shares similar advantages with *K*-medians in terms of the robustness to noise and outliers in smart meter data, and guarantees convergence. The disadvantage is that *K*-medoids is computationally expensive, which makes it difficult to be adapted to the clustering of large-scale and high-dimension load datasets.

K-means and *K*-medoids were examined in [63]. Among several electric load clustering algorithms, the best-performing one was selected according to different indices so that they could be divided into groups based on different household daily load patterns. Various clustering indices of *K*-means, *K*-medoids, and hierarchical clustering were evaluated for electric load clustering in [64].

4) Adaptive K-means

In electric load clustering studies, predetermining the amounts of clusters is often a tough issue because a small number of clusters will make the electric load clustering results less persuasive, while a huge number of clusters make applications of electric load clustering rather complex. In contrast to the original *K*-means that the amount of clusters K is needed to be predetermined, adaptive *K*-means is able to determine the ultimate number of clusters during the cluster formation process without trying alternative K. Adaptive *K*-means starts with an initial optimal guess for K, and allows the number to be changed at any time in course of formation.

Adaptive *K*-means differentiates from the classical *K*-means in the following aspects.

1) The cluster number in adaptive *K*-means can be dynamically adjusted depending on whether a distortion threshold condition is satisfied.

2) Adaptive *K*-means retains complete information on isolated load samples by dividing them into individual clusters.

3) Adaptive *K*-means independently applies K'-means to the offending cluster (K' is predetermined) and thus has the potential for parallel computation.

An advanced method that integrated adaptive *K*-means with hierarchical clustering was proposed for decomposing daily power consumption patterns into total daily usage and standardizing daily load shapes in [62].

Time domain along with fluctuation characteristics was considered in [65] in the local modeling, where adaptive *K*-means was used to cluster the load curves.

5) FCM

Considering the drawbacks of exclusive electric load clustering, that is, load curves in one cluster may be also similar to those in other clusters, the introduction of fuzzy theory can solve the problem by defining the membership degree of a load curve to different clusters. In *K*-center algorithms applying fuzzy theory, FCM is a well-known algorithm. Different from *K*-means, in FCM, each load profile belongs to each cluster with a membership degree [57]. Its concept characterizes the uncertainty that load profiles belonging to two or more clusters. Fuzzy theory makes the result of electric load clustering more reasonable, but its specific application significance needs to be explored. The membership matrix defines the membership degree of each load curve to each cluster, requiring that the total sum of membership degrees of load curves in all clusters is 1. Moreover, any load curve should be a member of at least one cluster. With the

introduction of a membership degree in FCM, the iteration of the centroids is easier to achieve global optimum.

A hesitation index was introduced in the membership and non-membership functions to deal with the uncertainty of electric load in [66], and FCM was applied to deal with the load data representation problem of fuzzy rules.

A technique, which combines FCM and artificial neural network (ANN) based pattern recognition techniques, was introduced for initial screening of electric load shapes using load studies and monthly energy usage data of customers in [67].

6) K-shape

K-shape is a novel time-series algorithm [60] that can be applied in electric load clustering studies as load curves share the time-series characteristics. Unlike the original *K*-means that treats time-series load data as independent attributes, *K*-shape creates a consistent set that is also well-separated in the process. To effectively compare time-series load data, cluster centroids are calculated based on shape-based distance (SBD), which are used recursively to capture shared characteristics of underlying data and update the allocation of load curves. Time-series data clusters are divided based on the similarity of load profiles without considering differences in load magnitudes and phases. *K*-shape can be cast into three parts intuitively, including the selection of SBD, the extraction of time-series load profiles, and shape-based electric load clustering analysis. Since the obtained load profiles exist in the form of time series, *K*-shape applies SBD with curve characteristics taken into account and thus has its advantages.

Different algorithms that include *K*-means, *K*-medoids, hierarchical, DTW Barycenter averaging and *K*-shape were compared in [68]. The conclusion showed that *K*-shape has better performance than traditional algorithms in the specific problems using intra-daily variability.

Fundamental characteristics of the mentioned partition-based clustering algorithms are clearly illustrated in Table I.

TABLE I
CHARACTERISTICS OF PARTITION-BASED CLUSTERING ALGORITHMS

Algorithm	Calculation of centroids	Best measure	Advantage	Disadvantage
K-means	Mean of members	Euclidean	1) Operational ease 2) Quick convergence 3) Highly explanatory 4) Preferable load clustering result	1) Sensitive to outliers 2) Easy to fall into a local optimum
K-medians	Median of members	Manhattan	Robust to noisy data in load dataset	1) Dissimilarity between cluster centroids and instances 2) More costly to calculate
K-medoids	The least dissimilar member to others generally	Various measurements	1) Noise and outlier resistant 2) Stability of convergence	More expensive computationally than *K*-means and *K*-medians
Adaptive *K*-means	1) Mean of clusters 2) Final cluster number is determined during the cluster formation	Euclidean	1) Cluster number adjusted adaptively 2) Potential of parallel computation	1) Sensitive to outliers 2) Easy to fall into a local optimum
FCM	1) Mean of members 2) An instance may not be classified into a sole cluster	Euclidean	1) Uncertainty modeling 2) Easier to reach global optimum 3) Robust to outliers	Higher computational complexity over *K*-means
K-shape	On account of the cross-correlation characteristics	SBD	Able to explore sequential features	1) Hard to be extended to large datasets 2) Limits of validity to specific datasets

C. Hierarchical Clustering Algorithm

In hierarchical clustering algorithm [69], load curves to be classified need to be numerically measured on a set of attributes, and rows of the constructed array are analyzed. Rows of matrices can be considered as a vector where the number of dimensions is the number of attributes in a multi-dimensional space. Hierarchical clustering algorithm is more flexible and explicit compared with *K*-center families, so it also arouses the interest of electric load clustering researchers.

Hierarchical clustering algorithm has significant merits such as easy to define, which does not require a predetermined number of clusters, and has fewer restrictions on the similarity of distances and rules. The hierarchical clustering algorithm is also highly exploitable in terms of hierarchical relationships and has a variety of cluster shapes. On the other hand, the disadvantages are the higher complexity, being more sensitive to singular values, and the higher possibility to aggregate into chains eventually compared with *K*-means. Adopting hierarchical clustering algorithms to cluster load curves can help intercept the desired number of clusters from any hierarchy according to the specific scenario applied by the electric load clustering process.

1) Agglomerative Algorithm

The agglomerative algorithm is considered as greedy, and the hierarchy is constructed through a series of irreversible steps [70]. Its clustering is based on a similarity measure. First, distance $d(i,j) \in \boldsymbol{D}$ ($\boldsymbol{D}$ denotes the similarity matrix) between load profiles (observation) i and j is established using distance criterion. Based on matrix $\boldsymbol{D}$, electric load instances are grouped by linkage criteria using evaluation functions, where the linkage criteria indicate the best candidates for merging and have an essential influence on the results of electric load clustering. Therefore, at each level, the closest pair of clusters in $\boldsymbol{D}$ is merged until a single cluster contain-

ing all load profiles is obtained.

Agglomerative algorithms can be divided into two broad categories based on whether a particular cluster centroid is selected. The first category is the linkage methods that can be represented using graphs that include single linkage [71], complete linkage [72], [73], and average linkage [46]. The second category is the algorithms that select the cluster centroids particularly (e.g., select the mean value or weighted mean of samples as the cluster centroid), where the ward's linkage [74], [75] is classical. Algorithms in the second category can specify cluster centroids individually or according to the coordinates and dissimilarities of the centroids.

In the linkage methods that can be presented using graphs, two common algorithms are chosen as samples, where the first one is complete linkage, and the other is average linkage. Firstly, the complete linkage allows for more compactly shaped clusters and is sensitive to outliers in load consumption data. The similarity between two load clusters is based on the similarity of the least similar load profile attributed to each of them. Secondly, the average linkage is expensive in computation for massive smart metering load data, but it is noise-resistant and is a compromise between single linkage and complete linkage. The average linkage criterion considers the similarities between all pairs of load curves in two clusters.

A novel computational method based on complete linkage was applied to analyze hourly power usage data in [76] and it was capable of handling large amounts of data.

Applications of several electric load clustering algorithms and indicators for classifying power consumers were described in [77]. The results showed that an improved version based on average linkage was the most appropriate.

Among methods that select cluster centroids particularly, the ward's linkage method is the most popular one in the electric load scenario, which is in conjunction with the notion of information gain and considers the minimization of the intra-cluster membership of all clusters with the cluster sum of squared differences between centroids of mass. For each of the two clusters C_j and C_j, the ward's linkage method merges them in $C_i \bigcup C_j$, and the resulting value of clusters' error sum of squares (ESS) increases. In practical applications, the ward's linkage method avoids these negative effects as single linkage ends up with a small number of large clusters, while the complete linkage can end up with too many clusters [78].

Considering the number of clusters in different groups, and based on the average daily net electric load model, a hierarchical clustering algorithm with the ward's linkage was proposed in [79] to obtain the clustering labels for each customer.

2) Divisive Algorithm

In two general classifications of hierarchical electric load clustering algorithms, divisive algorithm adopts an opposite strategy from the agglomerative one, which can be regarded as an inverse algorithm. From the other side, by combining with other electric load clustering algorithms, divisive algorithm is not only less complex than the agglomerative one, but also easier to obtain higher clustering robustness, since divisive algorithm starts from the whole and thus allows for a better analysis of noise and outliers in power consumption datasets. However, the divisive algorithm is more technically challenging to be implemented, mainly in terms of ensuring the correct load dataset splitting and setting up split termination conditions suitable for electric load clustering.

A double-level electric load clustering algorithm that integrates *K*-means and divisive algorithm was developed in [80], which combined the speed of *K*-means and divisive algorithm used in hierarchical clustering to cluster the smart metering data.

The mathematical descriptions and fundamental features of agglomerative algorithms mentioned for electric load clustering are illustrated in Table II in detail, where c_i and c_j are the centroids of C_i and C_j, respectively; and n_i and n_j are the numbers of data points belonging to the i^{th} and j^{th} cluster, respectively.

TABLE II
CHARACTERISTICS OF AGGLOMERATIVE ALGORITHMS

Linkage criterion	Description	Feature
Complete linkage	$\max_{x \in C_i, y \in C_j} d(x,y)$	1) More compact shaped clusters 2) Sensitive to outliers
Average linkage	$\frac{1}{n_i n_j} \sum_{x \in C_i, y \in C_j} d(x,y)$	1) Compromise between single and complete linkages 2) Computationally expensive, especially for large datasets 3) Noise resistant
Ward's linkage	$\sqrt{\frac{n_i n_j}{n_i + n_j}}\, d(c_i, c_j)$	Based on the objective function instead of similarities between data points of the two clusters

D. Density-based Clustering Algorithm

Due to the similarity and dissimilarity of load profiles obtained in actual practice, there are both high-density and low-density areas in the vector space consisting of load data, so the density-based clustering algorithm has its rationality. The density-based algorithm is a nonparametric algorithm that does not take the number of clusters as an input parameter, and does not need to assume the potential density $p(x)$ and the potential intra-cluster variance in the load consumption dataset. This category of algorithm can effectively process noise by only scanning the dataset once, so it can support load anomaly detection. A significant difference from other algorithms is that the dense point areas are separated by sparse point areas, where clusters are considered as dense areas of density $p(x)$. Therefore, the shape of density-based clusters and the distribution of points in the cluster are not necessarily convex. As measured by various distance functions, they can fit any shape in the data space, which makes density-based methods capable to realize electric load clustering reasonably in non-convex feature space.

1) DBSCAN

DBSCAN adopts a set of parameters on the concept of "neighborhood" to describe the compactness of the load pattern distribution, divides areas with sufficient density into clusters, and discovers clusters of any shape under noisy conditions [81] obtained in the processing of AMI. The core of

DBSCAN can be summarized as expanding from an initially selected core load pattern to a region of patterns with reachable density, resulting in a maximum region containing core points and boundary points, where any two load patterns (points) are connected taking density as the criterion.

In DBSCAN, it does not need to specify the number of clusters in advance and clusters of arbitrary shape can be discovered. Moreover, it is insensitive to load anomalies and is able to automatically identify them during the process [81].

Apart from fine properties mentioned above, DBSCAN has the disadvantages as follows: the quality of clustering becomes worse for datasets with non-uniform density or with large differences in distribution between clusters; it takes longer to converge when the dataset is large; the parameters *eps* and *minSam* are considered at the same time, so the parameter adjustment is complicated [81].

The pricing of power retail was studied in [82], in which DBSCAN was used for load pattern analysis to mine inherent power consumption patterns from historical load data. In order to better capture the load patterns of terminal consumers, the historical load consumption was analyzed statistically.

2) Clustering by Fast Search and Find of Density Peaks (CFSFDP)

An improved version of DBSCAN namely clustering by CFSFDP [83] adopts visualization methods to help find different clusters of different densities. This algorithm is particularly useful for analyzing massive high-dimension load profiles because it requires no repeated iterations and is highly efficient.

CFSFDP is able to find load clusters in different densities, while the drawback is that each load cluster must have a maximum density point as the cluster centroid. If the density of a cluster is uniform or if there are multiple high-density points, some load clusters will be separated into several subclusters. The number of clusters should be specified in CFSFDP.

CFSFDP was incorporated in process with KL distance to evaluate the differences between the two load patterns and obtain the typical dynamics of consumer behavior [84].

The characteristics of the density-based clustering algorithms are shown in Table III.

TABLE III
CHARACTERISTICS OF DENSITY-BASED CLUSTERING ALGORITHMS

Algorithm	Advantage	Disadvantage
DBSCAN	1) No need to specify the cluster number in advance 2) Capable to discover clusters of arbitrary shape 3) Insensitive to anomalies 4) Anomalies identification automatically	1) Worse clustering quality for load datasets with non-uniform density or with large differences in distribution between clusters 2) High time complexity 3) Complicated parameter adjustment
CFSFDP	Able to find clusters of different densities	1) Needs for a maximum density point as the cluster centroid in each cluster 2) Needs to specify the number of clusters

E. Grid-based Clustering Algorithm

Grid-based clustering algorithms make use of multi-dimension grid structures of feature spaces and divide the feature space of load into a finite number of units [85]. The advantage is that the processing time is independent of the number of load curves and the order of input load data. Therefore, it can handle arbitrary types of load consumption data. In the meantime, the disadvantage is that the processing time is limited by the number of units divided in each dimensional space.

1) Self-organizing Map (SOM)

In load characterization, one of the most used algorithms is SOM. It visually projects input load patterns into a reduced output space and keeps the topology unchanged at the same time. Then, the results are grouped via visual inspection. Iterative learning of the input pattern allows the weighted vector space to be consistent with the probability distribution region of the input pattern.

The advantage of SOM is that clusters of mutual nearest neighbors are more correlated than the others, which facilitates the interpretability and visual representation of electric load clustering results. The drawbacks include the need to select parameters, neighborhood function, grid type, centroid number, and the lack of a specific objective function. The clusters in SOM do not often correspond to natural clusters for the possible merging and splitting of natural clusters. Moreover, SOM does not guarantee convergence.

To obtain the SOM graph, two different demand data processing methods in frequency and time domain were tested and their respective advantages were evaluated, and the ability of SOM to classify new customers in different clusters was investigated finally in [86].

2) Other Grid-based Clustering Algorithms

Among other grid-based clustering algorithms, WaveCluster treats multi-dimension load usage as multi-dimension signals [85]. The algorithm is proposed to deal with the high dimensionality within load datasets with its high computing power. It first divides the data space and forms the structure of a grid. Subsequently, the wavelet transform is used to transform load data space into a frequency domain. After performing convolution calculations with a kernel function in the frequency space, the clustering properties of the load data can be well represented.

To reduce the number of features that represent each load pattern relative to the time domain data, a discrete wavelet transform was used to extract some spectra features [66]. A fused load curve clustering on account of wavelet transform (FCCWT) was presented in [39], which aimed to obtain the daily load patterns of power consumers. Since the clustering based on wavelet transformation always detects edges, its performance may not be excellent for cases without significant edges between clusters.

The characteristics of the above-mentioned grid-based clustering algorithms are illustrated in Table IV.

TABLE IV
CHARACTERISTICS OF GRID-BASED CLUSTERING ALGORITHMS

Algorithm	Brief description	Advantage	Disadvantage
SOM	Unsupervised neural network	1) Capable to identify the most significant characteristics with self-stability 2) Strong ability of anti-noise	Prone to be affected by factors such as weights of network connection, etc.
WaveCluster	Grid-based algorithm with wavelet transformation principle	1) Used for large-scale high-dimension data containing a large number of isolated outliers 2) Robust to noise and insensitive to input sequence 3) Capable to find clusters of complex structure with different precisions without assuming any particular shape 4) Independent of the prior knowledge of cluster number	Slightly less effective when there are no obvious edges between clusters

F. Model-based Clustering Algorithm

Model-based clustering algorithm fits probability distributions for different load clusters and assumes that the load data conform to some distribution functions. This subsection depicts two algorithms of model-based clustering commonly seen in electric load scenarios, which are Gaussian mixed model (GMM) and hidden Markov model (HMM).

1) GMM

The most common Gaussian distribution is also often used in electric load clustering because load data themselves conform to characteristics of a probability distribution and incorporate the advantages of the aforementioned soft clustering. Accordingly, GMM can be considered as a type of soft clustering in model-based clustering algorithms which gives the probability of electric load samples being assigned to each cluster [87]. The problem is to estimate the parameters for each cluster and determine the cluster that produces each observation [88]. In practice, all components in the model are from the same kind of distribution. By adding models, a mixed model can be used to approximate any continuous probability distribution. GMM is one of the most popular representation, in which components are Gaussian distributions with different mean and variance. From the perspective of probability distributions, GMM gives the probability that load curves belong to each cluster, which makes up for the defect of exclusive electric load clustering.

In [89], Bayesian information criterion (BIC) and Akaike's information criterion (AIC) are vital guideline to select the optimal number of components in GMM. A multi-resolution clustering was used to extract the spectral features of load curves from the multi-resolution smart meter data in the first phase, and then GMM is used to cluster the spectral features of the load curves in the second phase [26].

2) Other Model-based Clustering Algorithms

Due to the time-varying feature of load profiles, HMM can be seen as a kind of time-series probability model that describes a sequence of hidden states randomly generated by a hidden Markov chain [90]. Random observations generated in each hidden state form an observable random sequence. HMM provides a good perspective to model the dynamic characteristics of the consumption behavior of power customers. It can well identify power consumption modes at different levels.

HMM can infer occupancy status from the time-series load dataset of power consumers [90]. Characteristics of occupancy including magnitude, duration, and variability were contained in the model proposed by [90].

The characteristics of the above-mentioned model-based clustering algorithms are illustrated in Table V.

TABLE V
CHARACTERISTICS OF MODEL-BASED CLUSTERING ALGORITHMS

Algorithm	Brief description	Advantage	Disadvantage
GMM	Parametric model that can be viewed as combination of K single Gaussian models	Capable to model the probability of electric load samples being assigned to each cluster	1) Computationally intensive iteration 2) Possibility of falling into a local optimum
HMM	Probability model of time series	Memory-free	Assuming that current state is only related to the previous state

IV. EVALUATION OF ELECTRIC LOAD CLUSTERING

Due to the difference of input data, algorithm, and the initialization of parameters, the electric load clustering validity will be different. In practical research, CVIs for electric load clustering can be divided into two categories, i. e., internal and external evaluation indicators.

Internal evaluation indicators can evaluate electric load clustering validity by calculating the compactness within each load cluster and the separation among load clusters. External evaluation indicators can assess the validity of electric load clustering by comparing the distribution of labels obtained by clustering with the reference labels.

CVIs can be used for different purposes in process of clustering power consumers: identifying an appropriate number of power consumer groups [91], [92], comparing the performance of different electric load clustering algorithms [46], [93], studying the effects of parameters on electric load clustering results [61], [94], and evaluating the performance of clustering when changing (adding or removing) certain attributes [95], [96].

In conjunction with the content mentioned previously, this section focuses on metrics that are commonly used in electric load clustering literature. Assuming that a dataset contains N electric load instances, K is the number of load clusters, C_k is the k^{th} cluster, c_k is the centroid of the k^{th} cluster, and $C_{k'}$ is the cluster other than cluster C_k.

A. Internal Evaluation Indicators for Electric Load Clustering

1) MSE calculates the sum of square Euclidean distances within each load cluster and takes their average value to measure the cluster compactness.

$$MSE = \frac{1}{N}\sum_{k=1}^{K}\sum_{x_i \in C_k} d^2(x_i, c_k) \tag{5}$$

where x_i is the i^{th} load instance that belongs to cluster C_k.

2) SC is an evaluation index of the density and dispersion of load clusters. It ranges from −1 to 1. The more closer to 1 it is, the more compact the load clusters are; on the contrary, the more closer to −1 it is, the looser the load clusters are. Compact clusters indicate an ideal clustering performance.

$$SC = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\left(\sum_{x_j \in C_i}\frac{\psi(j)-\varphi(j)}{\max\left(\psi(j),\varphi(j)\right)}\right) \tag{6}$$

$$\varphi(j) = \frac{1}{n_k - 1}\sum_{x_i \in C_k, x_i \neq x_j} d(x_i, x_j) \tag{7}$$

$$\psi(j) = \min_{k' \neq k}\left(\frac{1}{n_{k'}}\sum_{x_i \in C_{k'}} d(x_i, x_j)\right) \tag{8}$$

where $\varphi(j)$ is the within-cluster mean distance; and $\psi(j)$ is the smallest one of all mean distances to other clusters.

3) DBI is an evaluation index that utilizes the inherent quantities and characteristics in a partitioned cluster to verify the effectiveness of methods. It measures the mean of the maximum similarity for each load cluster. The smaller the DBI is, the better the clustering performance will be, due to the lower degree of dispersion.

$$DBI = \frac{1}{K}\sum_{i=1}^{K}\max_{j \neq i}\frac{\frac{1}{n_i}\sum_{x \in C_i} d(x, c_i) + \frac{1}{n_j}\sum_{x \in C_j} d(x, c_j)}{d(c_i, c_j)} \tag{9}$$

4) MIA gives a value that depends on the compactness of clusters, indicating the distance between load curves of the same type. It represents the average distance between each cluster centroid and all other samples in the corresponding cluster.

$$MIA = \left(\frac{1}{K}\sum_{k=1}^{K} d_{C_k}^2\right)^{\frac{1}{2}} \tag{10}$$

where d_{C_k} is the distance between the cluster centroid c_k and the member of k^{th} cluster, $d_{C_k} = \sqrt{\frac{1}{n_k}\sum_{x_i \in C_k} d^2(x_i, c_k)}$.

5) The value of WCBCR depends on the sum of squares of distances between each input set and its representative set, as well as the similarity of cluster centroids. The smaller the value is, the better the performance of electric load clustering will be.

$$WCBCR = \frac{\sum_{k=1}^{K}\sum_{x_i \in C_k} d^2(x_i, c_k)}{\sum_{k=1, k \neq i}^{K} d^2(c_i, c_k)} \tag{11}$$

6) Dunn index I_{Dunn} is used to evaluate the clustering performance, the higher the value is, the better the performance will be. Dunn index assumes that ideal clusters are compact and well separated from others.

$$I_{Dunn} = \frac{\min_{i \neq j} d(i,j)}{\max_i D_i} \tag{12}$$

where D_i is the largest distance between two instances that both belong to cluster i.

B. External Evaluation Indicators for Electric Load Clustering

1) The F-measure of clustering result $F_{1,score}$ can be viewed as the weighted harmonic mean of precision indicator I_P and recall indicator I_R. Precision and recall are often mutually constrained in massive smart metering datasets. An ideal performance expects both indicators to be high, but in general, recall is low when precision is high and vice versa.

$$F_{1,score} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{13}$$

$$I_P = \frac{TP}{TP + FP} \tag{14}$$

$$I_R = \frac{TP}{TP + FN} \tag{15}$$

where the value of true positive TP predicts a positive class as a positive class; the value of true negative TN predicts a negative class as a negative class; the value of false positive FP predicts a negative class as a positive class; and the value of false negative FN predicts a positive class as a negative class.

2) RI is used to compare the clustering results with the true classification. The principle of RI is to enumerate all pairs in load instances and then observe the number of pairs that are consistent in the clustering algorithm and in the real circumstance.

$$RI = \frac{a+b}{C_N^2} \tag{16}$$

where a is the number of sample pairs that belong to the same cluster both in clustering and reference partitions; b is the number of sample pairs that do not belong to the same cluster either in clustering or reference partition; and C_N^2 is the number of possible sample pairs.

Apart from the classical CVIs shown above, a series of improved versions have been proposed. For example, TSI was applied to assess clustering stability in [27], the smaller the TSI is, the better the stability will be. A stability index for selecting the most appropriate clustering algorithm was developed in [56], and a priority index was proposed to determine the priority of clusters based on the stability index.

In recent years, different CVIs have been involved in the improvement of algorithms in different power scenarios. For example, in [97], the clustering and the qualitative validation process were combined, and the optimal results were extracted based on validity indices considering both the compactness and separateness.

V. Applications and Future Trends

For the present, with the popularity of AMI and the large-scale integration of renewable energy to the power grid [98], [99], a variety of electric load clustering algorithms have been applied to the field of electric load clustering with the feasibility on different power consumption datasets verified. Load clustering research has promising application prospects in areas such as consumer segmentation and enactment of tariff policy [100], [101], load anomaly detection, load forecasting, and DSR and DSM programs [27], [102]. For the future, the data security concern, the big data problems, the organic integration with AI, connections of new energies, and the technology expansion for Energy Internet would be hot spots for later researchers to run studies.

A. Applications of Electric Load Clustering

1) Consumer Segmentation and Enactment of Tariff Policy

Most researches in user segmentation focus on consumer preference for power and marketing management [103]. Almost all the segmentation of these areas depends on the cognitive psychological survey of consumers' self-worth, attitude, knowledge, and behavior. Consumer segmentation which based on electric load clustering can help obtain the number of consumption categories without a prior knowledge. Therefore, it is considered as a fine tool for early exploration when there lacks sufficient understanding of consumers. However, current methods based on electric load clustering still have their own defects. Firstly, they cannot automatically recognize the importance of different features. Secondly, it is difficult to distinguish the influence of correlations between power features on the clustering results. Finally, high-dimension features may make electric load clustering results lack practical significance, which is also the future development direction of research in this field.

With the gradual opening of the power market, consumer segmentation based on electric load clustering will usher in a series of new developments, providing decision information for customizing price design.

2) Load Anomaly Detection

Electric load clustering has already been widely used in detecting abnormal power consumption. Globally, unexpected events such as power theft, fraud, etc. cause utilities $96 billion in non-technical losses yearly [104]. These losses may have serious consequences such as pushing up the electricity price of paying customers and preventing the utilities from obtaining resources needed for future capital investment [105]. Due to customer fraud, power suppliers in the United States suffer an economic loss up to $6 billion [106]. Electric load clustering can help solve this problem by mining load patterns. Abnormal phenomena from conventional curves can be distinguished through the organic combination of electric load clustering and non-intrusive load monitoring (NILM).

3) Load Forecasting

Due to the time-varying characteristics of the load and the susceptibility to social, meteorological, and other factors, utilities and operators have the requirement of short-term load forecasting for power generation scheduling and daily decision making [107], [108]. By means of load pattern extraction based on the analysis of energy consumption characteristics of users, electric load clustering can provide prior distribution knowledge of load characteristics for load forecasting. Besides, electric load clustering can also provide users' data with similar patterns for load forecasting to improve forecasting accuracy in the absence of samples.

4) DSR and DSM Programs

In general, consumers with low volatility and high usage are suitable for incentive-based DSR programs (e.g., direct load control) for their predictability. However, customers with high volatility and high usage are suitable for price-based DSR programs (e.g., time of use pricing) due to their flexibility in electricity usage planning. For decision makers, the effectiveness of the program can be evaluated based on consumption data [109]. Both programs can be further designed to ultimately achieve orderly electricity utilization and operational optimization of power grids. Since different types of households have their own consumption preferences, it is necessary for utilities to use clustering to gain insight into the electricity consumption of different types of households. The data-supported decision allows flexible and personalized DSR plans and increases satisfaction at the same time [110].

Table VI provides the corresponding references and the aims for each application of electric load clustering.

TABLE VI
Corresponding References and Aims for Applications of Electric Load Clustering

Application	Reference	Aim
Consumer segmentation	[29], [46], [47], [56], [57], [77]	Manage power and market
Enactment of tariff policy	[78], [82], [95], [100], [102]	Offer differentiate services and conduct electricity price packages
Load anomaly detection	[23], [53], [111]	Ensure safe operation of power grid and reduce non-technical loss
Load forecasting	[19], [20], [62], [81]	Improve the accuracy of load forecasting
DSR and DSM programs	[2], [12], [27], [54], [62], [89]	Alleviate voltage shortage and reduce electricity costs

B. Future Trend of Electric Load Clustering

1) Data Security

DNOs and retailers utilize clustering to learn consumers' behavior centrally. The traditional electric load clustering process requires an access to all smart meter data, which can lead to privacy issues for both electricity consumers and retailers. Therefore, it is essential to propose a distributed framework that can protect the privacy. For instance, with

the aid of accelerated mean consistency with convergence for load analysis, the privacy-preserving accelerated average consensus algorithm proposed in [112] performed only through local computation and information sharing between neighboring data without sacrificing privacy. Furthermore, information from multiple industries often needs to be taken into account in electric load clustering to improve accuracy, though these data are difficult to communicate due to barriers and leakage concerns. The development of federated learning framework for holistic considerations can address the security challenges to data access. Unlike the former linear models used in privacy protection, the federated learning framework [113] can protect multi-party security dynamically via sharing trained models rather than data to each other.

2) Big Data

In recent years, the scale and complexity of load data have increased geometrically, which generate a big data scenario with respect to electric load clustering. Current electric load clustering algorithms cannot be entirely satisfactory in face of massive power consumers due to the insufficient computing power. For example, to reduce the impacts of under-computing on electric load clustering and the computation pressure of head servers, edge computing [114] characterized by local computing can partition load data by distribution areas and cluster load on much smaller datasets. Then, the clustering results can be uploaded to the cloud computing center for unified data management and model calculation. The process will facilitate hierarchical and zonal control of the power grid. In addition, a large number of links and heterogeneous convergence of edge computing [115] will have broad prospects in data-driven operation and control of the power grid.

3) AI

In recent years, besides clustering, various AI technologies such as fuzzy logic, deep learning, and adversarial learning have made continuous breakthroughs in data resolution, learning, and computing ability, which will have broad application prospects in the research of electric load clustering [116]. In addition, with the increasing load complexity, the organic combination of AI technologies in electric load clustering will matter a lot. It includes intelligent real-time perception combined with physical state, data-driven models combined with simulation and assistant decision-making combined with operation control. It can change the limitation of traditional electric load clustering algorithms and provide more refined information on load profile. Consequently, it can help improve the operation security of power grid and change the operation and service mode.

4) Integration of Renewable Energy

With the development of renewable energy, distributed PV generations that could be installed in households are becoming a hotspot for residential power consumption. Distributed PV generation converts solar energy into power for self-generation and uses, with the remainder fed into the power grid. The integration of PV generation has greatly changed the electric load characteristics of consumers, which transformed the traditional load nodes into the generalized load nodes [117]. The random fluctuation and intermittence of renewable energies increased the uncertainty of generalized load nodes [118]. Therefore, in the background of renewable energy access, it is important to conduct electric load clustering studies for generalized load nodes in power system analysis.

5) Development of Energy Internet

The Energy Internet is proposed and developed as a complex multi-network stream system that takes the internet and other advanced information technologies as the basis, and the distributed renewable energy as the main primary energy source. In the Energy Internet, the electric power system is the core, which is coupled closely with other energy networks and transportation systems. It is an extremely meaningful research area to extend electric load clustering to the clustering of various energy consumption data with coupling characteristics, so as to fully portray the energy portrait of users in the Energy Internet.

At present, in research on the smart grid, controllable loads mainly take localized absorption and control. While in the Energy Internet, due to the massive number of distributed devices, the research focus will shift from localized absorption to wide-area coordination. Different kinds of renewable energies are coupled with each other and the analysis of relations between loads and generations will be more difficult. In the future, as an important distributed device of the power system, the controllable load with its fast response time and wide geographic distribution can be an effective mean to suppress the intermittence of renewable energies and maintain the power balance of the system in the event of failure [119]. It is also a future research trend about how to model the complex characteristics of controllable loads and realize reasonable electric load clustering considering the loads coupled with multi-energy on the basis of the modeled feature space.

VI. Conclusion

The spread of AMI and the access of renewable energy promote the research of electric load clustering. Through data analysis, electric load clustering can help detect different load patterns and provide theoretical support for other researches and applications in smart grid. In this paper, previous clustering algorithms in load scenarios are thoroughly summarized. We first illustrate the process of electric load clustering stage by stage and introduce the common similarity measures in electric load clustering. Several well-known clustering algorithms are then explained separately with their virtues and drawbacks summarized. In addition, eight indices for evaluating the validity of electric load clustering are introduced. Finally, the applications and future trends of electric load clustering are discussed in detail.

References

[1] J. Leiva, A. Palacios, and J. A. Aguado, "Smart metering trends, implications and necessities: a policy review," *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 227-233, Mar. 2016.

[2] W. Labeeuw, J. Stragier, and G. Deconinck, "Potential of active demand reduction with residential wet appliances: a case study for Belgium," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 315-323, Jan. 2015.

[3] L. Dethman and D. Thomley. (2009, Aug.). Comparison of segmenta-

tion plans for residential customers. [Online]. Available: https://energytrust.org/library/reports/091231
[4] C. Wan, C. Zhao, and Y. Song, "Chance constrained extreme learning machine for nonparametric prediction intervals of wind power generation," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3869-3884, Sept. 2020.
[5] A. Shahid, "Smart grid integration of renewable energy systems," in *Proceedings of 2018 7th International Conference on Renewable Energy Research and Applications (ICRERA)*, Paris, France, Oct. 2018, pp. 944-948.
[6] C. Wan, Z. Xu, P. Pinson *et al*., "Probabilistic forecasting of wind power generation using extreme learning machine," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1033-1044, May 2014.
[7] C. Wan, Z. Xu, P. Pinson *et al*., "Optimal prediction intervals of wind power generation," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1166-1174, May 2014.
[8] L. Cozzi and T. Goodson. (2020, Jan.). Empowering electricity consumers to lower their carbon footprint. [Online]. Available: https://www.iea.org/commentaries/empowering-electricity-consumers-to-lower-their-carbon-footprint
[9] K. J. Dyke, N. Schofield, and M. Barnes, "The impact of transport electrification on electrical networks," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 12, pp. 3917-3926, Dec. 2010.
[10] H. Turker, S. Bacha, and A. Hably, "Rule-based charging of plug-in electric vehicles (PEVs): impacts on the aging rate of low-voltage transformers," *IEEE Transactions on Power Delivery*, vol. 29, no. 3, pp. 1012-1019, Jun. 2014.
[11] S. Rivera, B. Wu, S. Kouro *et al*., "Electric vehicle charging station using a neutral point clamped converter with bipolar DC bus," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 1999-2009, Apr. 2015.
[12] M. A. Z. Alvarez, K. Agbossou, A. Cardenas *et al*., "Demand response strategy applied to residential electric water heaters using dynamic programming and *K*-means clustering," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 524-533, Jan. 2020.
[13] R. Xu and D. Wunsch, *Clustering*. Hoboken: John Wiley & Sons, 2008.
[14] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Berlin: Springer, 2005.
[15] P. Flach, *Machine Learning: the Art and Science of Algorithms that Make Sense of Data*. New York: Cambridge University Press, 2012.
[16] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870-1891, Dec. 1992.
[17] K. Benmouiza and A. Cheknane, "Forecasting hourly global solar radiation using hybrid *K*-means and nonlinear autoregressive neural network models," *Energy Conversion and Management*, vol. 75, pp. 561-569, Nov. 2013.
[18] R. Granell, C. J. Axon, and D. C. H. Wallom, "Clustering disaggregated load profiles using a Dirichlet process mixture model," *Energy Conversion and Management*, vol. 92, pp. 507-516, Mar. 2015.
[19] T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-based approach to household electric load curve clustering and prediction," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5196-5206, Sept. 2018.
[20] I. P. Panapakidis, "Clustering based day-ahead and hour-ahead bus load forecasting models," *International Journal of Electrical Power & Energy Systems*, vol. 80, pp. 171-178, Sept. 2016.
[21] Y.-X. Cai, L.-J. Cai, and Z. Lu, "Anomaly detection of online monitoring data of power equipment based on association rules and clustering algorithm," in *Proceedings of 2nd Annual International Conference on Electronics*, *Electrical Engineering and Information Science (EEEIS 2016)*, Xi'an, China, Dec. 2016, pp. 289-298.
[22] C. Gao, Y. Wu, J. Tang *et al*., "Daily power load curves analysis based on grey wolf optimization clustering algorithm," in *Proceedings of Purple Mountain Forum 2019 – International Forum on Smart Grid Protection and Control*, Nanjing, China, Aug. 2020, pp. 661-671.
[23] M. Cui, J. Wang, and M. Yue, "Machine learning-based anomaly detection for load forecasting under cyberattacks," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5724-5734, Sept. 2019.
[24] S. Ryu, H. Choi, H. Lee *et al*., "Convolutional autoencoder based feature extraction and clustering for customer load analysis," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1048-1060, Mar. 2020.
[25] J. C. S. D. Souza, T. M. L. Assis, and B. C. Pal, "Data compression in smart distribution systems via singular value decomposition," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 275-284, Jan. 2017.
[26] R. Li, F. Li, and N. D. Smith, "Multi-resolution load profile clustering for smart metering data," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4473-4482, Nov. 2016.
[27] S. Lin, F. Li, E. Tian *et al*., "Clustering load profiles for demand response applications," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1599-1607, Mar. 2019.
[28] G. L. Ray and P. Pinson, "Online adaptive clustering algorithm for load profiling," *Sustainable Energy*, *Grids and Networks*, vol. 17, p. 100181, Mar. 2019.
[29] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1120-1128, Aug. 2007.
[30] K. Mets, F. Depuydt, and C. Develder, "Two-stage load pattern clustering using fast wavelet transformation," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2250-2259, Sept. 2016.
[31] G. Chicco, O. Ionel, and R. Porumb, "Electrical load pattern grouping based on centroid model with ant colony clustering," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1706-1715, May 2013.
[32] P. Cichosz, *Data Mining Algorithms: Explained Using R*. Warsaw: John Wiley & Sons, 2014.
[33] Y. Wang, Q. Chen, C. Kang *et al*., "Load profiling and its application to demand response: a review," *Tsinghua Science and Technology*, vol. 20, no. 2, pp. 117-129, Apr. 2015.
[34] K. Park and S. Son, "A novel load image profile-based electricity load clustering methodology," *IEEE Access*, vol. 7, pp. 59048-59058, May 2019.
[35] Y. Xiang, J. Hong, and Z. Yang, "Slope-based shape cluster method for smart metering load profiles," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1809-1811, Mar. 2020.
[36] S. Hasanvand, M. Nayeripour, S. A. Arefifar *et al*., "Spectral clustering for designing robust and reliable multi-MG smart distribution systems," *IET Generation*, *Transmission & Distribution*, vol. 12, no. 6, pp. 1359-1365, Oct. 2017.
[37] K. Li, X. Ge, X. Lu *et al.*, "Meta-heuristic optimization based two-stage residential load pattern clustering approach considering intracluster compactness and inter-cluster separation," *IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 3375-3384, Jul. 2020.
[38] A. K. Zarabie, S. Lashkarbolooki, S. Das *et al*., "Load profile based electricity consumer clustering using affinity propagation," in *Proceedings of 2019 IEEE International Conference on Electro Information Technology (EIT)*, Brookings, USA, May 2019, pp. 474-478.
[39] Z. Jiang, R. Lin, F. Yang *et al*., "A fused load curve clustering algorithm based on wavelet transform," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 1856-1865, May 2018.
[40] M. A. Masud, J. Z. Huang, M. Zhong *et al*., "Cluster survival model of concept drift in load profile data," *IEEE Access*, vol. 6, pp. 51269-51285, Sept. 2018.
[41] Y. Jin and Z. Bi, "Power load curve clustering algorithm using fast dynamic time warping and affinity propagation," in *Proceedings of 2018 5th International Conference on Systems and Informatics (ICSAI)*, Nanjing, China, Nov. 2018, pp. 1132-1137.
[42] N. Anuar and Z. Zakaria, "Cluster validity analysis for electricity load profiling," in *Proceedings of 2010 IEEE International Conference on Power and Energy*, Kuala Lumpur, Malaysia, Nov. -Dec. 2010, pp. 35-38.
[43] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847, Aug. 1991.
[44] J. Wong and R. Rajagopal. (2012, Jul.). A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting. [Online]. Available: https://www.aceee.org/files/proceedings/2012/data/papers/0193-000182.pdf
[45] F. L. Quilumba, W. Lee, H. Huang *et al*., "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918, Mar. 2015.
[46] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933-940, May 2006.
[47] G. Chicco, R. Napoli, F. Piglione *et al*., "Load pattern-based classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 1232-1239, May 2004.
[48] Z. Yu, "A temperature match based optimization method for daily load prediction considering DLC effect," *IEEE Transactions on Power Systems*, vol. 11, no. 2, pp. 728-733, May 1996.
[49] B. Goehry, Y. Goude, P. Massart *et al*., "Aggregation of multi-scale experts for bottom-up load forecasting," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 1895-1904, May 2020.
[50] A. Bouguettaya, Q. Yu, X. Liu *et al*., "Efficient agglomerative hierar-

chical clustering," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2785-2797, Apr. 2015.
[51] A. Al-Wakeel, J. Wu, and N. Jenkins, "*K*-means based load estimation of domestic smart meter measurements," *Applied Energy*, vol. 194, pp. 333-342, May 2017.
[52] W. Li, J. Zhou, X. Xiong *et al*., "A statistic-fuzzy technique for clustering load curves," *IEEE Transactions on Power Systems*, vol. 22, no. 2, pp. 890-891, May 2007.
[53] K. Zheng, Q. Chen, Y. Wang *et al*., "A novel combined data-driven approach for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1809-1819, Mar. 2019.
[54] H. Hino, H. Shen, N. Murata *et al*., "A versatile clustering method for electricity consumption pattern analysis in households," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1048-1057, Jun. 2013.
[55] X. Wang, A. Mueen, H. Ding *et al*., "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 275-309, Feb. 2012.
[56] T. Zhang, G. Zhang, J. Lu *et al*., "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 153-160, Feb. 2012.
[57] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organising maps to classify electricity customers," in *Proceedings of 2003 IEEE Bologna Power Tech Conference*, Bologna, Italy, Jun. 2003, pp. 7-13.
[58] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for *K*-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336-3341, Mar. 2009.
[59] T. Schütz, M. H. Schraven, M. Fuchs *et al*., "Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis," *Renewable Energy*, vol. 129, pp. 570-582, Dec. 2018.
[60] J. Paparrizos and L. Gravano, "*K*-shape: efficient and accurate clustering of time series," in *Proceedings of the International Conference on Management of Data*, Melbourne, Australia, May 2015, pp. 1855-1870.
[61] T. Xu, H. Chiang, G. Liu *et al*., "Hierarchical *K*-means method for clustering large-scale advanced metering infrastructure data," *IEEE Transactions on Power Delivery*, vol. 32, no. 2, pp. 609-616, Apr. 2017.
[62] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420-430, Jan. 2014.
[63] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied Energy*, vol. 141, pp. 190-199, Mar. 2015.
[64] L. Kotzur, P. Markewitz, M. Robinius *et al*., "Impact of different time series aggregation methods on optimal energy system design," *Renewable Energy*, vol. 117, pp. 474-487, Mar. 2018.
[65] C. Li, W. Cai, C. Yu *et al*., "Electricity consumption behaviour analysis based on adaptive weighted-feature *K*-means-AP clustering," *IET Generation*, *Transmission & Distribution*, vol. 13, no. 12, pp. 2352-2361, Jun. 2019.
[66] M. Charwand, M. Gitizadeh, P. Siano *et al*., "Clustering of electrical load patterns and time periods using uncertainty-based multi-level amplitude thresholding," *International Journal of Electrical Power & Energy Systems*, vol. 117, p. 105624, May 2020.
[67] R. F. Chang and C. N. Lu, "Load profile assignment of low voltage customers for power retail market applications," *IEE Proceedings: Generation*, *Transmission and Distribution*, vol. 150, no. 3, pp. 263-267, May 2003.
[68] H. Teichgraeber and A. R. Brandt, "Clustering methods to find representative periods for the optimization of energy systems: an initial framework and comparison," *Applied Energy*, vol. 239, pp. 1283-1293, Apr. 2019.
[69] P. Nahmmacher, E. Schmid, L. Hirth *et al*., "Carpe diem: a novel approach to select representative days for long-term power system modeling," *Energy*, vol. 112, pp. 430-442, Oct. 2016.
[70] S. S. Tabatabaei, M. Coates, and M. Rabbat, "GANC: greedy agglomerative normalized cut for graph clustering," *Pattern Recognition*, vol. 45, no. 2, pp. 831-843, Feb. 2012.
[71] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, pp. 30-34, Jan. 1973.
[72] D. Defays, "An efficient algorithm for a complete link method," *The Computer Journal*, vol. 20, no. 4, pp. 364-366, Jan. 1977.
[73] I. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis, "Three-stage clustering procedure for deriving the typical load curves of the electricity consumers," in *Proceedings of 2013 IEEE Grenoble Conference*, Grenoble, France, Jun. 2013, pp. 1-6.
[74] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236-244, Mar. 1963.
[75] 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.
[76] T. Räsänen, D. Voukantsis, H. Niska *et al*., "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data," *Applied Energy*, vol. 87, no. 11, pp. 3538-3545, Nov. 2010.
[77] G. Chicco, R. Napoli, F. Piglione *et al*., "Emergent electricity customer classification," *IEE Proceedings: Generation*, *Transmission and Distribution*, vol. 152, no. 2, pp. 164-172, Mar. 2005.
[78] A. Gabaldon, A. Guillamon, M. C. Ruiz *et al*., "Development of a methodology for clustering electricity-price series to improve customer response initiatives," *IET Generation*, *Transmission & Distribution*, vol. 4, no. 6, pp. 706-715, Jun. 2010.
[79] M. Sun, T. Zhang, Y. Wang *et al*., "Using bayesian deep learning to capture uncertainty for residential net load forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 188-201, Jan. 2020.
[80] Z. A. Khan, D. Jayaweera, and M. S. Alvarez-Alvarado, "A novel approach for load profiling in smart power grids using smart meter data," *Electric Power Systems Research*, vol. 165, pp. 191-198, Dec. 2018.
[81] W. Kong, Z. Y. Dong, Y. Jia *et al*., "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019.
[82] J. Yang, J. Zhao, F. Wen *et al*., "A model of customizing electricity retail prices based on load profile clustering analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3374-3386, May 2019.
[83] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492-1496, Jun. 2014.
[84] Y. Wang, Q. Chen, C. Kang *et al*., "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2437-2447, Sept. 2016.
[85] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "Wavecluster: a multi-resolution clustering approach for very large spatial databases," in *Proceedings of the 24th VLDB Conference*, New York, USA, Aug. 1998, pp. 428-439.
[86] S. V. Verdu, M. O. Garcia, C. Senabre *et al*., "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1672-1682, Nov. 2006.
[87] M. T. Bina and D. Ahmadi, "Aggregate domestic demand modelling for the next day direct load control applications," *IET Generation*, *Transmission & Distribution*, vol. 8, no. 7, pp. 1306-1317, Jul. 2014.
[88] J. Tang, S. Alelyani, and H. Liu, "Data classification: algorithms and applications," in *Data Mining and Knowledge Discovery Series*. Boca Raton: CRC Press, pp. 37-64, 2014.
[89] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136-144, Jan. 2016.
[90] A. Albert and R. Rajagopal, "Smart meter driven segmentation: what your consumption says about you," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4019-4030, Nov. 2013.
[91] T. Räsänen, J. Ruuskanen, and M. Kolehmainen, "Reducing energy consumption by using self-organizing maps to create more personalized electricity use information," *Applied Energy*, vol. 85, no. 9, pp. 830-840, Sept. 2008.
[92] V. Figueiredo, F. Rodrigues, Z. Vale *et al*., "An electric energy consumer characterization framework based on data mining techniques," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 596-602, May 2005.
[93] 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.
[94] L. O. Hall, A. M. Bensaid, L. P. Clarke *et al*., "A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 672-682, Sept. 1992.
[95] G. Chicco, R. Napoli, P. Postolache *et al*., "Customer characterization options for improving the tariff offer," *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 381-387, Feb. 2003.
[96] E. Carpaneto, G. Chicco, R. Napoli *et al*., "Electricity customer classi-

fication using frequency-domain load pattern data," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 1, pp. 13-20, Jan. 2006.
[97] M. A. Maniar and A. R. Abhyankar, "Validity index based improvisation in reproducibility of load profiling outcome," *IET Smart Grid*, vol. 2, no. 1, pp. 131-139, Mar. 2019.
[98] C. Wan, J. Lin, J. Wang *et al*., "Direct quantile regression for nonparametric probabilistic forecasting of wind power generation," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2767-2778, Jul. 2017.
[99] C. Zhao, C. Wan, and Y. Song, "An adaptive bilevel programming model for nonparametric prediction intervals of wind power generation," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 424-439, Jan. 2020.
[100] T. Namerikawa, N. Okubo, R. Sato *et al*., "Real-time pricing mechanism for electricity market with built-in incentive for participation," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2714-2724, Nov. 2015.
[101] X. Liang, X. Li, R. Lu *et al*., "UDP: usage-based dynamic pricing with privacy preservation for smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 141-150, Mar. 2013.
[102] Z. Jiang and Q. Ai, "Agent-based simulation for symmetric electricity market considering price-based demand response," *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 5, pp. 810-819, Sept. 2017.
[103] H. Khalkhali and S. H. Hosseinian, "Novel residential energy demand management framework based on clustering approach in energy and performance-based regulation service markets," *Sustainable Cities and Society*, vol. 45, pp. 628-639, Feb. 2019.
[104] LLC. (2017, May). Electricity theft and non-technical losses: global markets, solutions, and vendors. [Online]. Available: http://www.northeast-group. com/reports/Brochure-Electricity% 20Theft% 20&% 20Non-Technical%20Losses%20-%20Northeast%20Group.pdf
[105] C. Wan, Z. Xu, Y. Wang *et al*., "A hybrid approach for probabilistic forecasting of electricity price," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 463-470, Jan. 2014.
[106] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security & Privacy,* vol. 7, no. 3, pp. 75-77, May-Jun. 2009.
[107] Z. Cao, C. Wan, Z. Zhang *et al*., "Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1881-1897, May 2020.
[108] C. Zhao, C. Wan, Y. Song *et al*., "Optimal nonparametric prediction intervals of electricity load," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 2467-2470, May 2020.
[109] A. Nilsson, P. Stoll, and N. Brandt, "Assessing the impact of real-time price visualization on residential electricity consumption, costs, and carbon emissions," *Resources*, *Conservation and Recycling*, vol. 124, pp. 152-161, Sept. 2017.
[110] A. Hatami, H. Seifi, and M. K. Sheikh-El-Eslami, "A stochastic-based decision-making framework for an electricity retailer: time-of-use pricing and electricity portfolio optimization," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 1808-1816, Nov. 2011.
[111] L. A. P. Júniora, C. C.O. Ramosc, D. Rodrigues *et al.*, "Unsupervised non-technical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413-423, Nov. 2016.
[112] M. Jia, Y. Wang, C. Shen *et al.*, "Privacy-preserving distributed clustering for electrical load profiling," *IEEE Transactions on Smart Grid*. doi: 10.1109/TSG.2020.3031007
[113] Q. Yang, Y. Liu, T. Chen *et al.*, "Federated machine learning: concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1-19, Jan. 2019.
[114] W. Shi, J. Cao, Q. Zhang *et al.*, "Edge computing: vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637-646, Oct. 2016.
[115] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30-39, Jan. 2017.
[116] E. Cuevas, E. B. Espejo, and A. C. Enríquez, "Clustering representative electricity load data using a particle swarm optimization algorithm," in *Metaheuristics Algorithms in Power Systems*. Cham: Springer International Publishing, pp. 187-210, 2019.
[117] J. Zhang and Y. Sun, "Generalized load modeling considering wind generators connected to distribution network," *Power System Technology*, vol. 8, no. 35, pp. 41-46, Aug. 2011.
[118] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: application to intra-day household-level load curves," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 411-419, Jan. 2014.
[119] Z. Dong, J. Zhao, F. Wen *et al.*, "From smart grid to Energy Internet: basic concept and research framework," *Automation of Electric Power Systems*, vol. 38, no. 15, pp. 1-11, Aug. 2014.

Caomingzhe Si received the B. Eng. degree in electrical engineering from Chongqing University, Chongqing, China, in 2019. Currently, he is working toward the Mphil. degree with the School of Science and Engineering, Chinese University of Hong Kong (Shenzhen), Shenzhen, China. His research interests include the application of artificial intelligence (AI) in power system, including intelligent dispatching, urban intelligent distribution networks and intelligent buildings, and the evaluation and improvement of power system and grid security.

Shenglan Xu received the B.Eng. degree from the College of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2018. He is currently pursuing the M.S. degree with the College of Electrical Engineering, Zhejiang University. His research interests include electric load clustering and its applications.

Can Wan received the B. Eng. degree from Zhejiang University, Hangzhou, China, in 2008, and the Ph. D. degree from The Hong Kong Polytechnic University, Hong Kong, China, in 2015. He serves as a Professor with the College of Electrical Engineering, Zhejiang University, under the University Hundred Talents Program. He was a Postdoctoral Fellow with the Department of Electrical Engineering, Tsinghua University, Beijing, China, and held research positions at the Technical University of Denmark, Copenhagen, Denmark, The Hong Kong Polytechnic University, and City University of Hong Kong, Hong Kong, China. He was a Visiting Scholar with the Center for Electric Power and Energy, Technical University of Denmark, and Argonne National Laboratory, Lemont, USA. He is an Associate Editor for the IEEE Systems Journal and CSEE Journal of Power and Energy Systems. His research interests include forecasting, renewable energy, active distribution network, integrated energy systems, and machine learning.

Dawei Chen received the B.Eng. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2016. Currently, he is working toward the Ph. D. degree with the College of Electrical Engineering, Zhejiang University. His research interests include the coordinated optimization and control of the multi-energy systems.

Wenkang Cui received the B.Eng. degree from the College of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2018. Currently, he is working toward the Ph. D. degree with the College of Electrical Engineering, Zhejiang University. His research interests include the renewable energy forecasting and its applications.

Junhua Zhao is the External Experts of "Australian National Outlook", the Co-chair of the IEEE Special Interest Group (SiG) on Active Distribution Grids and Microgrids as well as the Secretary of the Asia Pacific Working Group of the IEEE PES SBLC (Smart Building, Load and Customer). He is a member of a Global Smart Grid Federation (GSGF) working group. Besides, he is a member of the "Interfaces of Grid Users/Focus on EV and Local Storage" working group, Smart Grid Australia (SGA) working group, "Cyber Physical Security of the Smart Grid" group and "Critical Infrastructure Program for Modelling and Analysis (CIPMA)" expert group. He is the Vice-chair of the Shenzhen AI Industry Society's Expert Committee. He is the editorial board members of IET Energy Conversion and Economics, Journal of Modern Power Systems and Clean Energy, Electric Power Components and Systems, and Power System Protection and Control. He is the expert reviewer of Australian Research Council (ARC), National Natural Science Foundation of China Reviewer, and Hong Kong Research Grants Committee (RGC). Also, he is the reviewer of IEEE Transactions on Power Systems, IEEE Transactions on Smart Grid, IEEE Transactions on Neural Networks and Learning Systems, Applied Energy, and IET Generation, Transmission & Distribution. His research interests include smart grid, electricity market, energy economic, data mining, and artificial intelligence.