


Article

Method for Clustering Daily Load Curve Based on SVD-KICIC

Yikun Zhang ¹, Jing Zhang ^{1,*} , Gang Yao ², Xiao Xu ¹ and Kewen Wei ¹

¹ School of Electrical Engineering, Guizhou University, Guiyang 550025, China; gs.ykzhang18@gzu.edu.cn (Y.Z.); gs.xux18@gzu.edu.cn (X.X.); gs.kwwwei18@gzu.edu.cn (K.W.)

² Guizhou Power Grid Company, Guiyang 550001, China; yaogang0319@gz.csg.cn

* Correspondence: zhangjing@gzu.edu.cn

Received: 1 July 2020; Accepted: 28 August 2020; Published: 31 August 2020



Abstract: Clustering electric load curves is an important part of the load data mining process. In this paper, we propose a clustering algorithm by combining singular value decomposition and KICIC clustering algorithm (SVD-KICIC) for analyzing the characteristics of daily load curves to mitigate some of the traditional clustering algorithm problems, such as only considering intra-class distance and low computational efficiency when dealing with massive load data. Our method identifies effective daily load curve characteristics using the singular value decomposition technique to improve dimensionality reduction, which improves low computational efficiency by reducing the number of dimensions inherent in big data. Additionally, the method performs SVD on the load data to obtain singular values for determination of weight of the KICIC algorithm, which leverages intra-class and inter-class distances of the load data and further improves the computational efficiency of the algorithm. Finally, we perform a series of simulations of actual load curves from a certain city to validate that the algorithm proposed in this paper has a short operation time, high clustering quality, and solid robustness that improves the clustering performance of the load curves.

Keywords: classification of load curves; singular value decomposition; dimensionality reduction; inter-class distance; weighted Euclidean distance; clustering validity

1. Introduction

In recent years, with growing demand for electricity and the popularized use of smart electricity meters, electric power systems have accumulated increasingly massive load data [1,2]. However, the load demand response has many characteristics such as complexity, randomness and high non-linearity. By using load data mining and identifying electricity consumption modes, users' electricity consumption characteristics can be obtained and used to provide an important reference for the reliable operation of the power grids, to refine user partitioning and to personalize interactions between the electric power systems and users [3–5].

Clustering algorithms are an effective approach to mining the electricity consumption characteristics of users [6,7]. Common clustering algorithms are primarily hierarchical, partition-based, grid-based or model-based [8]. In existing research, most of the clustering algorithms specific to load curves use distance similarity for classification. In [9], the major characteristics of the daily load curves are extracted for clustering by combining the principal component analysis method and the K-means algorithm. In [10], a load classification method based on the Gaussian mixture model and multi-dimensional analysis is developed. In this method, the load data are subject to multi-dimensional analysis and dimensionality reduction; then, the resulting data are used with the Gaussian mixture model clustering algorithm for the classification of large-scale load data sets. In [11], an integrated clustering algorithm is developed by comparing the advantages and disadvantages of

multiple clustering algorithms. The resulting algorithm effectively combines the K-means algorithm and the hierarchical clustering algorithm. In [12], major characteristics are obtained by singular value decomposition (SVD) for dimensionality reduction and to perform weighted K-means clustering of the load curves. In [13], six daily load characteristic indexes are selected as dimensionality reduction indexes to express the original load curves before clustering, thus improving clustering efficiency. However, these methods only consider minimizing intra-class distance for improving the intra-class compactness and ignore the effects of the inter-class distance on the clustering results when they use distance as the daily load curve similarity measure for clustering. The methods' clustering results tend to have a blurred boundary, and the load curves at various types of boundaries might be categorized incorrectly, thus leading to lower clustering quality. In addition, the existence of blurred samples might result in an increase in the number of iterations of the algorithm and a decrease in computational efficiency.

In response to the decrease in clustering quality due to inter-class blurred samples, in [14], the ESSC (enhanced soft subspace clustering) algorithm is proposed, which expands the inter-class distance by maximizing the distance between various centers and the global center. However, this method makes the shift of the class center unsatisfactory. Especially when inter-class distribution in clustering samples is nonuniform, maximizing the distance between the global center and the various class centers might cause several closer class centers to be more compact, which contradicts the original purpose of maximizing the inter-class distance and weakens the clustering effect. In [15], to further improve the clustering quality, the weighted K-means algorithm in combination with the intra-cluster and inter-cluster distances (KICIC) is proposed. The algorithm adds inter-class distance to the traditional weighted K-means algorithm and maximizes the distance between the class center and samples of other classes, reducing the effects of boundary samples on the clustering results. However, the method requires iteratively updating the weights of various dimensions, which occupies large computational resources. In particular, its computational efficiency for massive data has to be improved.

Comprehensively considering all the reviewed clustering methods, we propose a clustering algorithm based on SVD-KICIC to deal with the problems of low clustering quality and poor efficiency caused by blurred boundary samples in traditional clustering techniques for daily load curves. Our method performs dimensionality reduction of the original data by applying a singular value decomposition to the data then conducting a clustering analysis of the daily load curves by maximizing inter-class distance with a KICIC algorithm. Our method considers data obtained by sampling at each time point of the load curve as a dimension and extracts the dimensionality weighted via SVD dimensionality reduction, thus forming samples with weights of various dimensions to be clustered. The method establishes an objective function in combination with intra-class and inter-class distance, guaranteeing the minimum intra-class distance and the maximum inter-class distance, achieving effective and accurate clustering of load curves. The measured daily load data from a city are used as a sample here. They are compared against three clustering algorithms that do not consider inter-class distance, traditional K-means, SVD-weighted K-means, and KICIC algorithms, to verify the effectiveness and accuracy of the algorithm proposed in this paper.

2. Basic Principles

2.1. Theory of Singular Value Decomposition

We assume $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m]$ is a $m \times n$ -order real matrix containing m load curves, where the i th load curve is denoted as $\mathbf{X}_i = [x_{i1}, x_{i2}, \dots, x_{in}]$ and n is the number of sampling points of the load curve. The number of load curves for clustering analysis, m , is generally larger than n , the number of the sampling points. Thus, we let $m > n$ in this work.

The SVD technique decomposes the real matrix X into three matrices [16,17]: U , Λ , and V^T

$$\begin{cases} X = U\Lambda V^T \\ \Lambda = \begin{bmatrix} \Lambda_1 \\ 0 \end{bmatrix} \end{cases} \quad (1)$$

where the orthogonal matrix $U = [u_1, u_2, \dots, u_m]$ is an $m \times m$ -order matrix with each column vector being a mutually orthogonal unit vector. Meanwhile, the column vector is the characteristic vector of the matrix XX^T and called the left singular vector. The orthogonal matrix $V = [v_1, v_2, \dots, v_n]$ is an $n \times n$ -order matrix, its column vector is also a mutually orthogonal unit vector. Meanwhile, it is the characteristic vector of the matrix XX^T and called the right singular vector. $\Lambda_1 = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is a diagonal matrix such that its diagonal elements are the singular values of the matrix X , which are decreasing in sequence (i.e., $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$). Equation (1) is expressed as

$$\begin{aligned} X &= U\Lambda V^T \\ &= [u_1 \quad u_2 \quad \dots \quad u_m] \begin{bmatrix} \Lambda_1 \\ 0 \end{bmatrix} [v_1 \quad v_2 \quad \dots \quad v_n]^T = \\ &= [\lambda_1 u_1 \quad \lambda_2 u_2 \quad \dots \quad \lambda_n u_n] \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} = \sum_{j=1}^n \lambda_j u_j v_j^T \end{aligned} \quad (2)$$

We consider a load curve X_i in the matrix X as an example. From Equation (2), we can derive

$$X_i = \begin{bmatrix} \lambda_1 u_{1i} & \lambda_2 u_{2i} & \dots & \lambda_n u_{ni} \\ v_1^T & v_2^T & \dots & v_n^T \end{bmatrix}^T \quad (3)$$

where u_{1i} is the coordinate of the vector u_1 at the first point and u_{2i} is similarly defined.

Thus, the SVD method establishes a new orthogonal coordinate system with the vectors v_1, v_2, \dots, v_n as the coordinate axes. The singular value λ_j denotes the scale from the vector u_j to the coordinate axis v_j ; $\lambda_j u_{j,i}$ is the coordinate value of the load curve X_i on the coordinate axis v_j . In addition, the larger the singular value λ_j , the greater the degree of data dispersion on the coordinate axis v_j , and the greater the variance of the data reflected, the better the coordinate axis can indicate the variation direction of the data. As the singular values obtained from the SVD method are arranged in descending order, the coordinate axes v_1, v_2, \dots, v_q corresponding to the first q singular values $\lambda_1, \lambda_2, \dots, \lambda_q$ are the q major variation directions of the matrix, and can best represent the major characteristics of the original matrix. The matrix X and the load curve X_i can be approximately denoted as

$$\begin{aligned} X &\approx Y \cdot \begin{bmatrix} v_1^T & v_2^T & \dots & v_q^T \end{bmatrix}^T \\ &= \begin{bmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_q u_q \\ v_1^T & v_2^T & \dots & v_q^T \end{bmatrix}^T \end{aligned} \quad (4)$$

$$\begin{aligned} X_i &\approx Y_i \cdot \begin{bmatrix} v_1^T & v_2^T & \dots & v_q^T \end{bmatrix}^T \\ &= \begin{bmatrix} \lambda_1 u_{1i} & \lambda_2 u_{2i} & \dots & \lambda_q u_{qi} \\ v_1^T & v_2^T & \dots & v_q^T \end{bmatrix}^T \end{aligned} \quad (5)$$

According to Equations (4) and (5), the major characteristics of the load curve X_i can be denoted as $Y_i = [y_{i1}, y_{i2}, \dots, y_{iq}]$ in a low dimension coordinate system by reducing inessential coordinate axes.

2.2. Weighting K-Means Algorithm in Combination with the Intra-Class and Inter-Class Distance

The traditional K-means algorithm, and its derivative algorithms, generally use intra-class Euclidean distance as an indicator for judging similarity among daily load curves. However, in actual applications, the daily load curves usually have blurred inter-class samples. As shown in Figure 1, the load curves at the boundaries of various classes might be assigned to other classes, thus leading to low clustering quality. In addition, due to the existence of blurred samples, the number of algorithm iterations increases, thus reducing the computational efficiency.

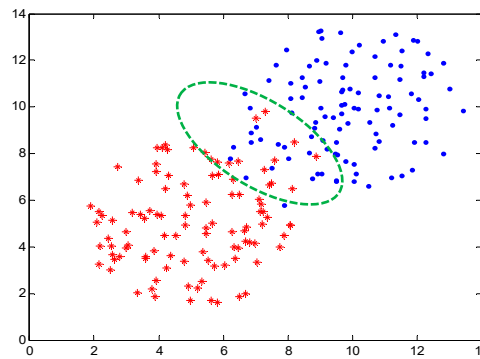


Figure 1. Fuzzy boundary samples.

In response to this problem, [15] proposes the KICIC clustering algorithm. As the traditional algorithm only considers the intra-class distance, this algorithm integrates inter-class distance, and maximizes inter-class distance while minimizing intra-class distance. Therefore, a target function is established:

$$\begin{aligned}
 P &= P(\mathbf{U}, \mathbf{W}, \mathbf{Z}) \\
 &= \sum_{p=1}^k \sum_{i=1}^m u_{ip} \sum_{j=1}^n w_{pj} (x_{ij} - z_{pj})^2 \\
 &\quad - \eta \sum_{p=1}^k \sum_{i=1}^m (1 - u_{ip}) \sum_{j=1}^n w_{pj} (x_{ij} - z_{pj})^2 \\
 &\quad + \gamma \sum_{p=1}^k \sum_{j=1}^n w_{pj} \log w_{pj}
 \end{aligned} \tag{6}$$

and constraint condition:

$$\begin{cases} \sum_{p=1}^k u_{ip} = 1, u_{ip} \in \{0, 1\} \\ \sum_{j=1}^n w_{pj} = 1, 0 \leq w_{pj} \leq 1 \end{cases} \tag{7}$$

where $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_k]$ is the k clustering centers with $\mathbf{Z}_p = [z_{p1}, z_{p2}, \dots, z_{pq}]$ being the p th clustering center. $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_k]$ is the k weight vectors, with $\mathbf{W}_p = [w_{p1}, w_{p2}, \dots, w_{pn}]$ being the weight of each sampling point in the p th class. When the i th load curve is assigned to the p th class, let $u_{ip} = 1$, otherwise $u_{ip} = 0$. Therefore, the $m \times k$ assignment matrix \mathbf{U} is constituted.

Minimizing target function (6) is the optimization principle, and the analysis is conducted on the basis of three items. The purpose of the first item is to minimize the intra-class distance of the sample in its class; the role of the second item is to maximize the inter-class distance; the third item adjusts the characteristic weight distribution via the entropy of weight. However, in actual applications, the KICIC algorithm occupies large computational resources when solving for the weight matrix \mathbf{W} [15]. Thus, the curse of dimensionality arises from this high-dimensionality data and results in low computational efficiency. In this work, we use the SVD technique for dimensionality reduction to improve the computational efficiency of the KICIC algorithm.

3. SVD-KICIC Algorithm

3.1. Data Preprocessing

3.1.1. Identification and Correction of Abnormal or Missing Data

As a series of problems may occur during the process of load data acquisition, such as communication interruption, measurement equipment failures, interference of environmental factors, and so on, the load data could be abnormal or even missing. When the amount of missing or abnormal data exceeds 10% (included) of the number of samples, the invalid curves need to be rejected. To improve the data quality, the load data need to be preprocessed prior to SVD. The load variation rate $\delta_{i,j}$ of a certain load curve $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,n}]$ is set as the basis. The abnormal data are evaluated using

$$\delta_{i,j} = \frac{x_{i,j+1} - x_{i,j}}{x_{i,j}} \quad (8)$$

where $\delta_{i,j}$ is the load variation rate of the load curve \mathbf{X}_i at the j th sampling point. When the number of sampling points $n = 48$, the threshold value ε generally ranges from 0.6 to 0.9. In other words, when $\delta_{i,j} \geq \varepsilon$, the data at the sampling point are considered abnormal.

For abnormal data points, a smooth correction equation is used for correction and substitution:

$$x_{i,j}^* = \frac{\sum_{g=1}^{g_1} x_{i,j-g} + \sum_{h=1}^{h_1} x_{i,j+h}}{g_1 + h_1} \quad (9)$$

where $x_{i,j}^*$ is the corrected value of the abnormal data point $x_{i,j}$, g denotes a forward value assignment, h denotes a backward value assignment, and values are assigned to g_1, h_1 according to the number of sampling points in the actual case, generally 4~7.

3.1.2. Load Curve Normalization

The amplitude values of the load data collected from different users may differ substantially. Direct clustering lacks objective accuracy, and the clustering results are unreliable if the load data does have the same order of magnitude before clustering. In this research, we process load data using the maximum value normalization principle. The processing method can be characterized by

$$x'_{i,j} = \frac{x_{i,j}}{\max(\mathbf{X}_i)} \quad (10)$$

where $x'_{i,j}$ is the normalized data at the sampling point j of the i th load curve. Then a new normalized matrix \mathbf{X}' is obtained with element $x'_{i,j}$.

3.2. SVD-KICIC and Its Implementation

A traditional KICIC algorithm uses the complete information of the samples as its input. Because of this, the computation is complex for larger sample sizes. In this work, we calculate the characteristic information matrix \mathbf{Y} using an SVD dimensionality reduction technique and use it as the input of the KICIC algorithm. Meanwhile, based on the singular values obtained by the SVD method, we redefine the weights of the KICIC algorithm and improve the overall clustering performance. The specific method is described in Section 3.2.

3.2.1. Improving the Target Function

The weight matrix $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_k]$ in Equation (11) of the KICIC algorithm independently assigns weight vectors to the load data of each class. The value w_{pj} denotes the weight value of the

j th sampling point in the p th class. The algorithm needs to solve the assignment matrix \mathbf{U} , clustering center matrix \mathbf{Z} , and the weight matrix \mathbf{W} , which consume large computational resources.

To improve the overall performance of the KICIC algorithm, we use the characteristic information matrix \mathbf{Y} obtained from SVD as the input, and use the weight vector $\mathbf{W}' = [w'_1, w'_2, \dots, w'_q]$ obtained from normalization with the sum of the first q singular values being 1 as the dimensionality weight of the information matrix, then propose the target function of the SVD-KICIC algorithm.

$$\begin{aligned} P &= P(\mathbf{U}, \mathbf{Z}) \\ &= \sum_{p=1}^k \sum_{i=1}^m u_{ip} \sum_{j=1}^q w'_j (y_{ij} - z_{pj})^2 \\ &\quad - \eta \sum_{p=1}^k \sum_{i=1}^m (1 - u_{ip}) \sum_{j=1}^q w'_j (y_{ij} - z_{pj})^2 \end{aligned} \quad (11)$$

constraint condition:

$$\sum_{p=1}^k u_{ip} = 1, u_{ip} \in \lim_{x \rightarrow \infty} \{0, 1\} \quad (12)$$

Compared with target function (6) of KICIC, target function (11) of the SVD-KICIC algorithm proposed in this work only needs to solve the data object assignment matrix \mathbf{U} and the class center matrix \mathbf{Z} via Equations (14) and (15), as the weight vector \mathbf{W}' is known. As a result, the computational complexity is reduced. In addition, the SVD-KICIC algorithm uses the characteristic information matrix \mathbf{Y} as the input instead of complete data information, thus improving the capacity of the algorithm to analyze massive data.

3.2.2. Determination of the Dimensionality of the Characteristic Information Matrix

The dimensionality of the characteristic information matrix \mathbf{Y} has an important impact on the effectiveness of the method proposed in this paper. To determine the dimensionality, the sum of squares of singular values is defined to represent the information contained in the matrix. The amount of information contained in matrix \mathbf{X}' is defined as $F = \lambda_1^2 + \lambda_2^2 + \dots + \lambda_n^2$; the amount of information contained in matrix \mathbf{Y} after dimensionality reduction is $F_1 = \lambda_1^2 + \lambda_2^2 + \dots + \lambda_q^2$. The ratio of the characteristic information matrix \mathbf{Y} to the amount of information of the original matrix is $A = \frac{F_1}{F}$. In this research, a large number of experiments show that when $A > 0.9$, the characteristic information matrix \mathbf{Y} can effectively express the information contained by the matrix \mathbf{X}' . The value of q is the dimensionality of the characteristic information matrix \mathbf{Y} .

3.2.3. Solving the Assignment Matrix \mathbf{U}

When solving \mathbf{U} , Equation (11) can be simplified to

$$P(\mathbf{U}, \mathbf{Z}) = \sum_{p=1}^k \sum_{i=1}^m u_{ip} \sum_{j=1}^q w'_j (y_{ij} - z_{pj})^2 \quad (13)$$

The target function $P(\mathbf{U}, \mathbf{Z})$ can be minimized if and only if

$$u_{ip} = \begin{cases} 1, & \text{if } \sum_{j=1}^q w'_j (y_{ij} - z_{pj})^2 \leq \sum_{j=1}^q w'_j (y_{ij} - z_{p'j})^2 \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where $u_{ip} = 1$ denotes that the i th load curve is assigned to the p th class. Therefore, we can find the assignment matrix \mathbf{U} . See [15] for specific demonstration details.

3.2.4. Solving the Clustering Center Matrix \mathbf{Z}

Assuming that the assignment matrix \mathbf{U} obtained in Section 3.2.3 is fixed, the target function $P(\mathbf{U}, \mathbf{Z})$ can be minimized if and only if

$$z_{pj} = \frac{(1 + \eta) \sum_{i=1}^m u_{ip} y_{ij} - \eta \sum_{i=1}^m y_{ij}}{(1 + \eta) \sum_{i=1}^m u_{ip} - \eta m} \quad (15)$$

where z_{pj} denotes the coordinate value of the j th dimension of the p th clustering center. Therefore, we can get the clustering center matrix \mathbf{Z} . See [18,19] for a specific demonstration process.

The algorithm process is shown in Figure 2. The characteristic information matrix \mathbf{Y} and weight matrix \mathbf{W}' are obtained by the SVD of matrix \mathbf{X}' . Then, the object assignment matrix \mathbf{U} and clustering center \mathbf{Z} are iteratively solved by Equations (14) and (15). Iteration is repeated until the value of target function (11) is no longer decreasing.

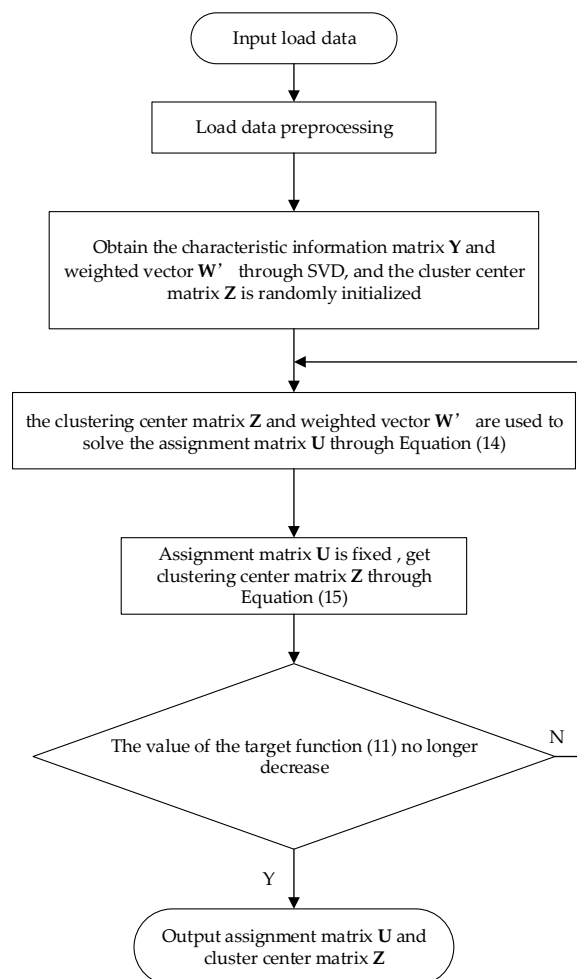


Figure 2. Clustering flow chart based on SVD-KICIC algorithm.

3.3. Clustering Effectiveness Indicator

The clustering effectiveness test is a process that assesses the clustering quality to determine the optimal clustering data set [20–22].

We assume that m load curves are divided into k classes. The Silhouette indicator Ω_{sil} of the i th sample in the p th class is defined as follows:

$$\Omega_{sil}(i) = \frac{d_a(i) - d_b(i)}{\max(d_a(i), d_b(i))} \quad (16)$$

The overall clustering quality could be assessed by the mean value of Silhouette indicator Ω_{silM} of all load curves. A greater value denotes higher clustering quality. The clustering number p , corresponding to the maximum value of Ω_{silM} , is used as the optimal clustering number.

The computational expression of Ω_{silM} is characterized by

$$\Omega_{silM} = \frac{1}{m} \sum_{i=1}^m \Omega_{sil}(i) \quad (17)$$

At the same time, the clustering results are assessed with the Davies–Bouldin index (DBI). DBI refers to the maximum value of the ratio of the sum of the average values of the intra-class distance of any two classes to the distance of their two clustering centers. A smaller value represents a better clustering result. The computational method is defined as follows:

$$I_{DB} = \frac{1}{k} \sum_{a=1}^k \max_{a \neq b} \left(\frac{\bar{C}_a + \bar{C}_b}{M_{ab}} \right) \quad (18)$$

where I_{DB} is the numerical value derived from DBI, \bar{C}_a and \bar{C}_b are the mean values of the sum of the distance from the two class samples to their clustering centers and M_{ij} is the distance between the two clustering centers.

4. Analysis of Examples

The examples in this paper were realized on a PC configured with Inter(R) Core(TM) i5-8300H CPU 2.50 GHz, RAM 16 GB. The operating system used was Windows 10. To verify the accuracy and efficiency of the method proposed in this paper, the K-means algorithm, SVD-weighted K-means algorithm, KICIC algorithm, and SVD-KICIC algorithm were used to independently study the actual load curves of a certain city and the results, of the four algorithms are compared and analyzed.

4.1. Clustering of Actual Daily Load Curves

The experimental data in this paper are derived from the data of 5263 measured daily load curves of a certain city. The sampling interval was 30 min, and each load curve had 48 sampling points. Because some data were missing or abnormal after data preprocessing, 5158 daily load curves were obtained, forming an initial matrix of 5158×48 orders.

After performing an exhaustive number of simulations, the clustering result with the highest accuracy was associated with the parameter $\eta = 0.07$.

Singular value decomposition was performed for the matrix X' to obtain a characteristic value matrix (48×48). As shown in Figure 3, as the dimension of the feature information matrix increases, the information ratio of the matrix Y gradually increases at a lower rate. As shown in Figure 3, when the dimension of the characteristic information matrix is 5, the ratio of information is $A = \frac{F_1}{F} \geq 0.9$. Thus, matrix X' is expressed as a 5158×5 matrix Y . The weight vector corresponding to the five [column] dimensions is $W' = [0.542, 0.225, 0.118, 0.096, 0.019]$.

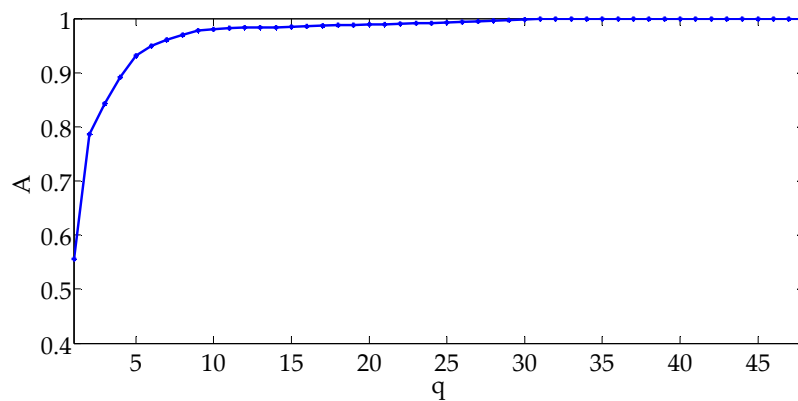


Figure 3. Proportion of information corresponding to dimensionality number.

The traditional K-means clustering, SVD-weighted K-means clustering algorithm, KICIC algorithm, and SVD-KICIC clustering algorithm were compared to perform a clustering analysis of the load data. The results are shown in Figure 4. In the case of setting different numbers of clusters, the validity test demonstrated that when the number of clusters was 5, the average values of the silhouette of the four clustering algorithms were all maximized. Therefore, we selected $k = 5$ as the fixed number of clusters.

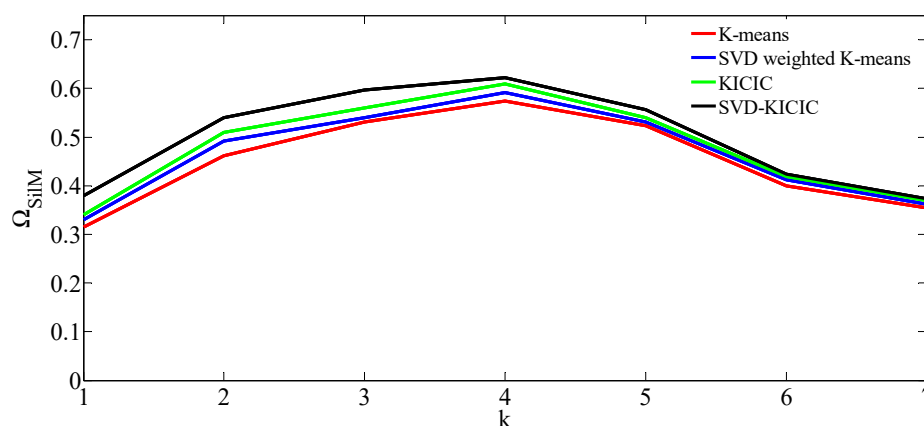


Figure 4. Determining the best clustering number based on validity index.

The clustering result based on our clustering algorithm in this paper is shown in Figure 5. The numbers of various load curves were 1582, 1038, 1269, 845, and 424. The numbers of various load curves corresponding to the KICIC algorithm were 1618, 1022, 1237, 829, and 452. The numbers of various load curves of the SVD-weighted K-means clustering algorithm were 1650, 1015, 1217, 829, and 447. The numbers of various load curves of the traditional K-means clustering algorithm were 1691, 1026, 1258, 807, and 376. The clustering results are shown in Figure 5, and detail a number of load curves for a narrow time period as an example to make different load curves visible.

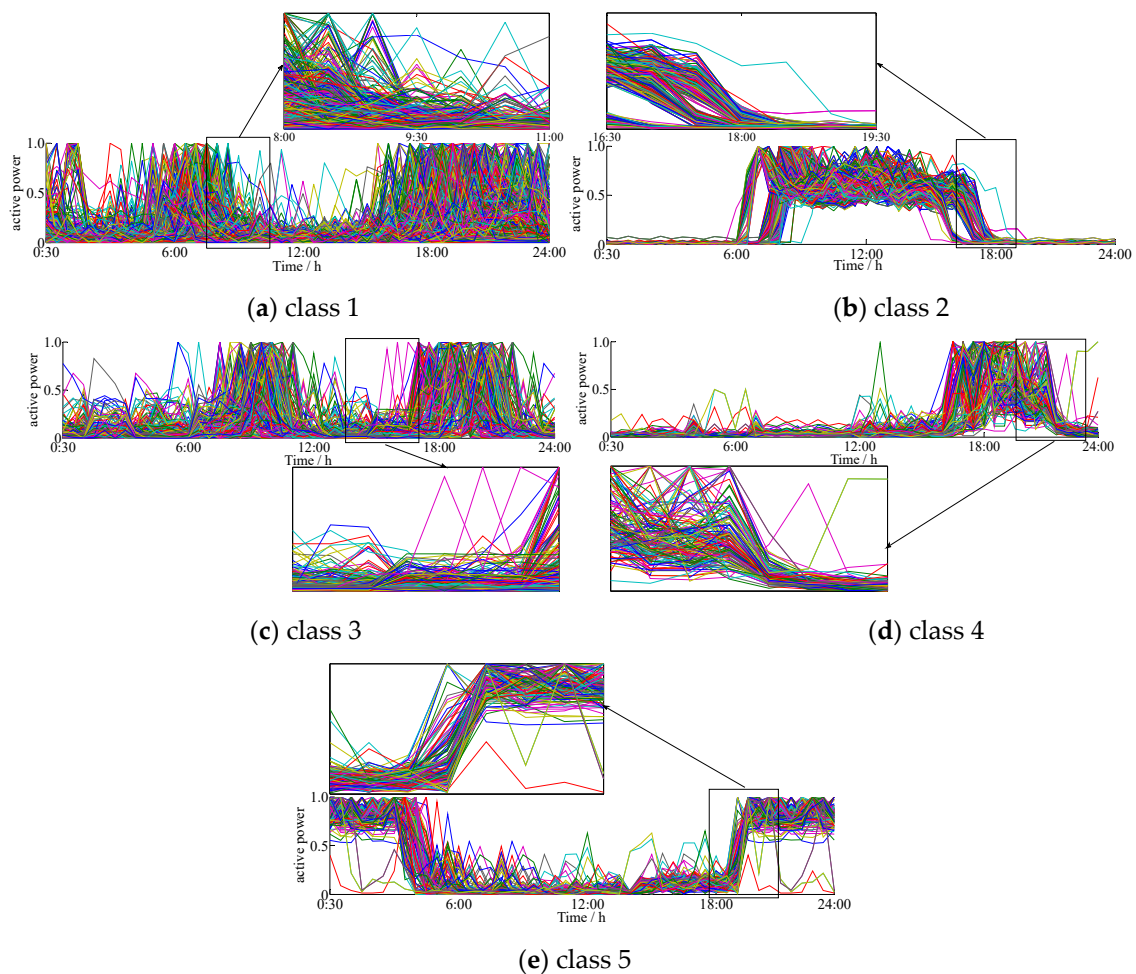


Figure 5. Clustering results of load curves based on the SVD-KICIC algorithm.

Figure 6 shows that the four clustering algorithms extracted five typical load curves with similar shapes to the load curve. There were four types: double peak, flat peak, single peak, and avoided peak. Class 1 was at the peak of electricity consumption at 8:00 and 19:00, which is typically household electricity, and Class 2 was at the peak of electricity consumption from 6:00 to 18:00, which is typically commercial electricity. Class 3 and 1 were both the double peak load types, the difference being that that Class 3 used less electricity than Class 1 from 23:00 to 6:00 the next morning. Additionally, Class 3 used electricity for small industries, and the electricity consumption time was more regular. Classes 1 and 3 were more likely to be misclassified. The clustering centers obtained by the SVD-KICIC algorithm and the KICIC algorithm were basically consistent. The clustering centers obtained by the K-means algorithm and the SVD-weighted K-means algorithm were relatively close, but the former two were partially different from the latter two in the clustering centers. This is because the SVD-KICIC algorithm and KICIC algorithm minimize the intra-class distance while maximizing the inter-class distance, making the classification results more accurate. Class 4 had the lower power consumption throughout the day but showed increased power consumption at night, which was a typical night load. Class 5 was the peak avoidance load. The clustering results reflect the five actual power load conditions and demonstrate the reliability of the SVD-KICIC algorithm.

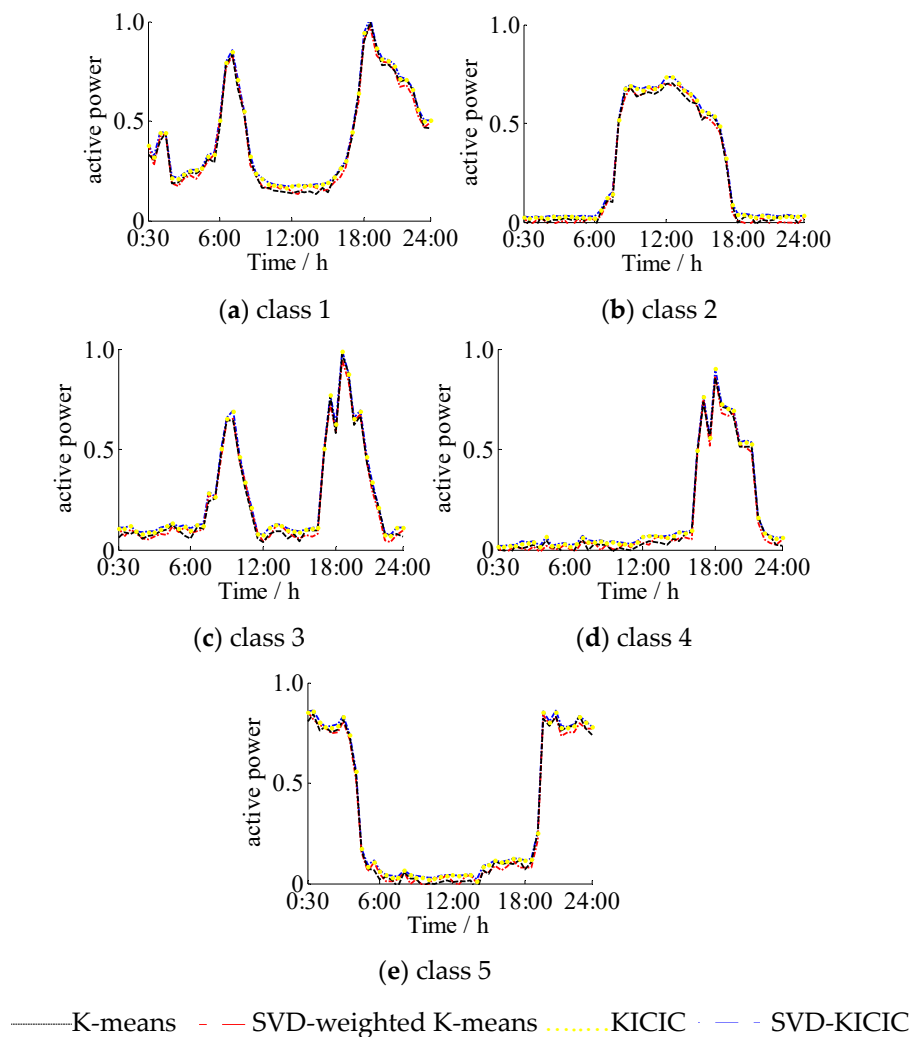


Figure 6. Comparison of typical load profiles between four algorithms.

Based on a comparative analysis of the average values of the clustering indexes obtained from 50 tests listed in Table 1, the SVD-KICIC algorithm can achieve better clustering quality when clustering the load curves. The proposed algorithm fully considered the intra-class and inter-class distance, thus making the intra-class distance of the load smallest and the inter-class distance the largest. The clustering center and the non-class samples were far away from each other, which reduced the impact of the non-class samples on the clustering accuracy and accelerated the clustering iteration process. Therefore, our algorithm runs faster than the traditional K-means algorithm, SVD-weighted clustering algorithm, and KICIC algorithm.

Table 1. Clustering property comparison between four methods.

Algorithm	Best k	Ω_{silM}	DBI	Running Time/s
K-means	5	0.574	1.283	61.32
SVD-weighted K-means	5	0.591	1.206	24.71
KICIC	5	0.609	1.127	45.83
SVD-KICIC	5	0.622	1.007	19.35

In order to further verify the stability of the algorithm, the standard deviation of the number of various load curves in 10 experiments was compared and shown in Figure 7. It can be seen that the

mean value of standard deviation of the SVD-KICIC algorithm was the smallest, and the stability of this algorithm was better.

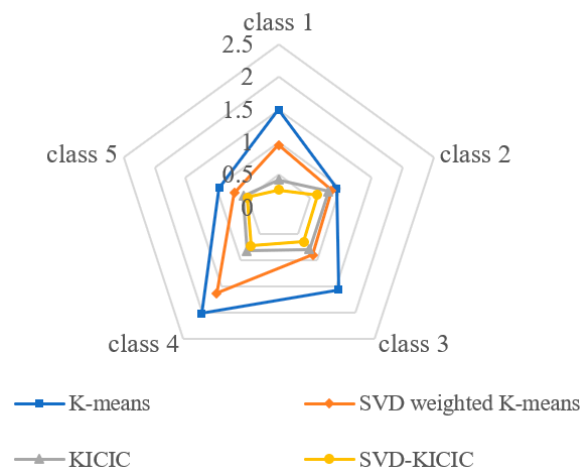


Figure 7. Comparison of stability between four algorithms.

4.2. Simulation Examples

To verify the anti-interference ability and robustness of our algorithm, 5000 simulated load curves with known clustering results were selected and the distribution of the curves was randomly disrupted to perform clustering again. The results contained eight classes of typical load curves with different shapes, each comprising 625 load curves. Noise interference of different degrees within the interval of 5~40% was added to each sampling point of the simulated load curve to produce eight sets of simulated data. The step size 5% was chosen to effectively show the changing trend of the clustering results of the SVD-KICIC and K-means algorithms, so as to compare the clustering performance of the two algorithms. The load data were processed through the proposed algorithm. Simulation of 5000 typical load curves with $r = 30\%$ is shown in Figure 8.

The effects of the noise of different degrees on the experimental results were tested and analyzed with the best cluster number, classification accuracy, and the Silhouette index mean. Table 2 shows the results of the comparison between the SVD-KICIC algorithm and the traditional K-means algorithm in terms of clustering load curves under various noise interferences.

Table 2. Comparison of robustness between two algorithms.

r/%	SVD-KICIC			K-means		
	Best k	Ω_{silM}	Acc/%	Best k	Ω_{silM}	Acc/%
5	8	0.9516	100	8	0.9516	100
10	8	0.9233	100	8	0.9233	100
15	8	0.9085	100	8	0.8536	94.30
20	8	0.8860	100	8	0.8033	91.16
25	8	0.8649	100	7	0.7294	89.56
30	8	0.8327	99.85	7	0.6697	88.34
35	7	0.7638	90.63	7	0.6185	80.16
40	7	0.6079	78.41	6	0.5839	69.83

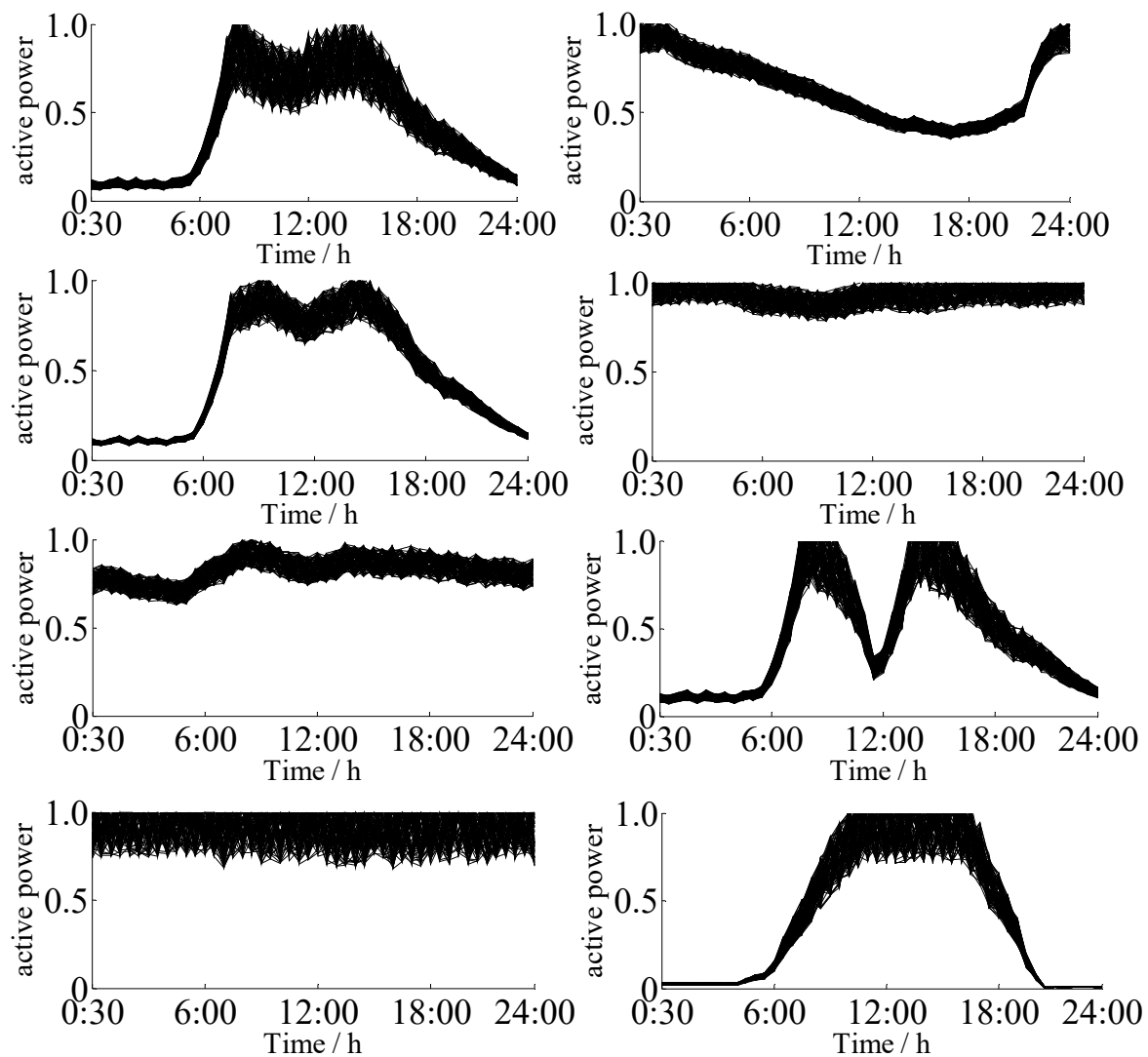


Figure 8. Simulation of 5000 typical load curves ($r = 30\%$).

Based on the comparative analysis of the average value of the clustering indexes obtained from the 50 tests in Table 2, we find that, for the two algorithms, with an increase in the degree of noise interference, the optimal number of clusters deviate. The mean value of Silhouette indexes and classification accuracy show a decreasing trend, indicating that three indexes can be used to test the robustness of the algorithm.

When the noise ratio is between 5% and 20%, the optimal number of clusters for both algorithms is 8, and the classification accuracy is equal to or very close to 100%. When the noise ratio is between 25% and 30%, the optimal clustering number of the SVD-KICIC algorithm is always 8, the mean of the Silhouette is greater than 0.85, and the classification accuracy is equal to or very close to 100%. However, the optimal clustering number of the traditional K-means algorithm are changed to 7, and the clustering accuracy and Silhouette mean greatly reduce. When the noise ratio reaches 35~40%, the optimal clustering number of the two types of algorithms is not 8, and the classification accuracy and Silhouette mean of the SVD-KICIC algorithm decrease somewhat, but the fluctuation is small. Therefore, the SVD-KICIC algorithm is more robust than the traditional K-means algorithm.

5. Conclusions

In this work, we propose a clustering method of daily load curves based on SVD-KICIC by combining the advantages of the SVD dimensionality reduction technique and the maximized inter-class

distance of KICIC. This method uses an SVD dimensionality reduction technique to extract the effective characteristics of load data and greatly reduce the data dimension; it also uses the singular value to determine the weight coefficient of KICIC and reduce the number of iterative calculations. Compared with traditional K-means, SVD-weighted K-means, and KICIC clustering methods, the simulation studies show that the method can effectively use the intra- and inter-class distances of load data to improve clustering quality, computational efficiency, and robustness.

Author Contributions: Conceptualization, Y.Z. and J.Z.; methodology, Y.Z. and J.Z.; software, Y.Z.; validation, Y.Z., J.Z., G.Y., and X.X.; formal analysis, Y.Z. and J.Z.; resources, X.X.; data curation, Y.Z. and K.W.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z., J.Z.; supervision, J.Z.; project administration, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Nature Science Foundation of China, grant number 51867005. The Science and Technology Foundation of Guizhou Province, grant number [2016]1036. Guizhou Province Science and Technology Innovation Talent Team Project, grant number [2018]5615. Guizhou Province Reform Foundation for Postgraduate Education, grant number [2016]02, The Science and Technology Foundation of Guizhou Province, grant number [2018]5781.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Lu, S.; Fang, H.; Wei, Y. Distributed Clustering Algorithm for Energy Efficiency and Load-Balance in Large-Scale Multi-Agent Systems. *J. Syst. Complex.* **2018**, *31*, 234–243. [[CrossRef](#)]
- Xu, X.; Chen, W.; Sun, Y. Over-sampling algorithm for imbalanced data classification. *J. Syst. Eng. Electron.* **2019**, *30*, 1182–1191. [[CrossRef](#)]
- Qiang, C.; Xiuli, W.; Weizhou, W. Stagger peak electricity price for heavy energy-consuming enterprises considering improvement of wind power accommodation. *Power Syst. Technol.* **2015**, *39*, 946–952.
- Lin, R.; Wu, B.; Su, Y. An Adaptive Weighted Pearson Similarity Measurement Method for Load Curve Clustering. *Energies* **2018**, *11*, 2466. [[CrossRef](#)]
- Zhang, T.; Gu, M. Overview of Electricity Customer Load Pattern Extraction Technology and Its Application. *Power Syst. Technol.* **2016**, *40*, 804–811.
- Bu, F.; Chen, J.; Zhang, Q.; Tian, S.; Ding, J. A controllable and refined recognition method for load patterns based on two-layer iterative clustering analysis. *Power Syst. Technol.* **2018**, *42*, 903–913.
- Kim, N.; Park, S.; Lee, J.; Choi, J.K. Load Profile Extraction by Mean-Shift Clustering with Sample Pearson Correlation Coefficient Distance. *Energies* **2018**, *11*, 2397. [[CrossRef](#)]
- Chicco, G.; Napoli, R.; Piglion, F. Comparisons among clustering techniques for electricity customer classification. *IEEE Trans. Power Syst.* **2006**, *21*, 933–940. [[CrossRef](#)]
- Koivisto, M.; Heine, P.; Mellin, I.; Lehtonen, M. Clustering of connection point and load modeling in distribution systems. *IEEE Trans. Power Syst.* **2013**, *28*, 1255–1265. [[CrossRef](#)]
- Zhang, M.; Li, L.; Yang, X.; Sun, G.; Cai, Y. A Load Classification Method Based on Gaussian Mixture Model Clustering and Multi-dimensional Scaling Analysis. *Power Syst. Technol.* **2019**. [[CrossRef](#)]
- Bin, Z.; Chijie, Z.; Jun, H.; Shuiming, C.H.; Mingming, Z.H.; Ke, W.; Rong, Z. Ensemble clustering algorithm combined with dimension reduction techniques for power load profiles. *Proc. CSEE* **2015**, *35*, 3741–3749.
- Ye, C.; Hao, W.; Junyi, S. Application of singular value decomposition algorithm to dimension-reduced clustering analysis of daily load profiles. *Autom. Electr. Power Syst.* **2018**, *42*, 105–111.
- Liu, S.; Li, L.; Wu, H.; Sun, W.; Fu, X.; Ye, C.; Huang, M. Cluster analysis of daily load profiles using load pattern indexes to reduce dimensions. *Power Syst. Technol.* **2016**, *40*, 797–803.
- Deng, Z.; Choi, K.S.; Chung, F.L.; Wang, S. Enhanced soft subspace clustering integrating within-cluster and between-cluster information. *Pattern Recognit.* **2010**, *43*, 767–781. [[CrossRef](#)]
- Huang, X.; Wang, C.; Xiong, L.; Zeng, H. A Weighting k-Means Clustering Approach by Integrating Intra-cluster and Inter-cluster Distances. *Chin. J. Comput.* **2019**, *42*, 2836–2848.
- Golub, G.; Loan, C. *Matrix Computation*; Johns Hopkins University Press: Baltimore, MD, USA, 1996.
- Zhang, J.; Xiong, G.; Meng, K. An improved probabilistic load flow simulation method considering correlated stochastic variables. *Int. J. Electr. Power Energy Syst.* **2019**, *111*, 260–268. [[CrossRef](#)]

18. Bezdek, J.C. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **1980**, *1*, 1–8. [[CrossRef](#)]
19. Selim, S.Z.; Ismail, M.A. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 81–87. [[CrossRef](#)]
20. Al-Otaibi, R.; Jin, N.; Wilcox, T.; Flach, T. Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data. *IEEE Trans. Ind. Inf.* **2016**, *12*, 645–654. [[CrossRef](#)]
21. Li, Z.; Yuan, W.; Ren, C.; Huang, C.; Dong, X. Approximate computing method based on cross-layer dynamic precision scaling for the k-means. *J. Xidian Univ.* **2020**, *47*, 1–8.
22. Liu, Y.; Liu, Y.; Xu, L. High-performance back propagation neural network algorithm for mass load data classification. *Autom. Electr. Power Syst.* **2018**, *42*, 131–140.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).