A Modified K-means Clustering Algorithm Based on FMF-GS-DD

Peng Wang, Yiwei Ma, Zhiqi Ling, Genhong Luo

***Abstract*—K-means is one of the most popular clustering algorithms. Its simplicity and efficiency enable it to be used in various industrial fields such as period partition of time-of-use (TOU) tariff. To solve the problems including insufficient determination of cluster number and initial cluster centers in K-means, we propose a modified K-means clustering algorithm based on fuzzy membership functions (FMF), gap statistic (GS), and data density (DD). First, the improved initial cluster centers are modeled to complete initialization through the first calculation based on FMF. Second, the cluster number of the dataset is determined by GS evaluation and further calculation based on DD is performed if necessary. Next, the efficiency and quality of the proposed clustering algorithm are examined by using three evaluation indexes: standard deviation, silhouette coefficient index, and Davies Boldin index. Finally, the effectiveness of the proposed method is verified under a typical case study of TOU period partition, and the results show that the proposed method can significantly shorten convergence time and improve clustering quality, compared to the traditional K-means clustering algorithm.**

***Index Terms*—K-means clustering, fuzzy membership functions, gap statistic, data density, time-of-use tariff, period partition**

I. Introduction

With the advent of the big data era, the development of machine learning technology is also getting faster. Clustering analysis is one of the most representative methods in data analysis [1, 2]. And K-means clustering is one of the classical partition algorithms [3-5]. It is easy to explain, convenient to implement, and has advantages like fast convergence and good robustness. It is widely used in the electric power industry, such as TOU period partition [6-13], as it can divide load curves into different periods.

As we all know, the traditional K-means clustering algorithm (TKM) has two noticeable drawbacks [14,15]: (i) cluster number k is pre-set and its value is often difficult to estimate; (ii) clustering effect is pretty sensitive to initial cluster centers which are usually randomly selected according to convention or experience. How to improve TKM has attracted widespread attention. Many studies on improving clustering algorithm initialization have been carried out [16-19]. Specifically, the initialization can be described as two steps: the determination of k and the value of each center.

In the studies on cluster number k, Literature [20] proposes a Gaussian mixture clustering algorithm that combines the elbow method and expectation maximization to improve the user experience of power system customers. The typical scenarios are obtained in Literature [21] by using the elbow method to determine k. Literature [22] combines with the elbow method and GS to determine the optimal number of clusters, then determines initial cluster centers. Literature [23] gives a GS method based on weighted martingale distance to determine the optimal number k. Literature [24] proposes an approach based on GS to measure clustering performance. The practice of determining clustering number k by statistical principles has been commonly used, and in this paper, we adopt GS to determine k.

In terms of calculating the values of initial centers, the application of intelligence optimization algorithms has solved the over-dependence of TKM on the initial centers [25,26]. Literature [27] makes use of an improved particle swarm optimization based on support vector machine regression to find initial cluster centers. Literature [28] also employs particle swarm optimization to improve the initialization. Other algorithms like firefly metaheuristic optimization algorithm and genetic optimization algorithm have been still commonly combined with TKM [29,30]. But the methods based on intelligence optimization algorithms still have a certain randomness, because the initial swarm and update of them are random.

However, considering the stability, Literature [31] employs a maximum distance method to calculate the values. Literature [32] analyzes the characteristics of initial centers and selects the centers by making the similarity lowest. Based on the feasibility of determining k firstly and calculating the initial centers secondly is verified by Literature [22], we then use FMF and DD to calculate the values of initial centers to avoid randomness and make the most of data distribution.

The main work of this paper is summarized as follows: Section II presents theoretical models of traditional K-means clustering (TKM), gap statistic (GS), fuzzy member functions (FMF), and data density (DD). Section III proposes a modified K-means clustering algorithm (MKM), which consists of improved initial cluster centers based on FMF and DD, and optimal cluster number k based on GS. Section IV presents three different evaluation indexes (standard

Manuscript received April 30, 2023; revised September 4, 2023.

This work was supported by the National Natural Science Foundation of China (No.61703068), and the Chongqing Postgraduate Research and Innovation Project (No. CYS22485). (Corresponding author: Yiwei Ma)

Peng Wang is a postgraduate student of Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: 17623600676@163.com).

Yiwei Ma is an associate professor of Department of Electrical Engineering, School of Automation and Industrial Internet, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: mayw@cqupt.edu.cn).

Zhiqi Ling is a postgraduate student of Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: lzq13535793794@163.com).

Genhong Luo is a postgraduate student of Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: 1468483560@qq.com).

deviation, silhouette coefficient index, and Davies Boldin index) to evaluate the proposed algorithm. Section V analyzes the proposed method which is verified under a typical case study of TOU period partition. Finally, Section VI summarizes this work.

II. Related Work

A. Traditional K-means Clustering Algorithm

TKM is a classical clustering algorithm as it is efficient and simple. The specific steps of the algorithm are shown as follows [33-35].

Step 1: k is set as the number of clusters, and k initial centers are randomly selected: v_1, v_2, …, v_k.

Step 2: For each sample x_t, the distances from it to different initial cluster centers (v_1, v_2, …, v_k) are calculated respectively. Then it is divided into the cluster c_j with the shortest distance. It can be expressed in Eq. (1).

$$x_t \in c_j,\ if\ \ d(x_t, v_j) < d(x_t, v_i) \tag{1}$$

Where, $d(x_t,v_j)$ indicates the distance between x_t and v_j.

Step 3: The mean value of the cluster c_j is taken as the new center. It is calculated by Eq. (2).

$$v_j = \frac{1}{|c_j|} \sum_{x_t \in c_j} x_t \tag{2}$$

Step 4: Repeat **Step 2**-**Step 3** until the cluster centers no longer change or the objective function in Eq. (3) is met.

$$J(\boldsymbol{x}, \boldsymbol{v}) = \min \sum_{j=1}^{k} \sum_{x_t \in c_j} \left| x_t - v_j \right|^2 \tag{3}$$

When Eq. (3) is satisfied, the algorithm process stops and the clustering is completed. It is obvious that the random initial centers badly affect the efficiency and quality of the algorithm performance.

B. Gap Statistic

The principle of GS is based on comparing the dispersion degree between the dataset and the data generated from the reference distribution. The detailed steps of the model are shown as follows.

Step 1: The samples in $\boldsymbol{x}$ are divided into k clusters by TKM: c_1, c_2, …, c_k, and the sum of squared distances between samples within each cluster c_r is calculated. The calculation is given in Eq. (4).

$$D_r = \sum_{x_i \in c_r} \sum_{x_j \in c_r} \left| x_i - x_j \right|^2 \tag{4}$$

Step 2: The compactness W_k is obtained by normalizing and summing D_r. It is calculated by Eq. (5).

$$W_k = \sum_{r=1}^{k} \frac{1}{2|c_r|} D_r \tag{5}$$

Step 3: Construct statistic $Gap\ (k)$ by Eq. (6).

$$Gap(k) = E^*\{\log(W_k)\} - \log(W_k) \tag{6}$$

Where, $E^*\{\log(W_k)\}$ is the expectation of $\log(W_k)$. It is estimated by using the logarithmic mean value of $W^*_{k,b} \sim U(\min(\boldsymbol{x}), \max(\boldsymbol{x}))$:

$$E^*\{\log(W_k)\} = \frac{1}{B} \sum_{b=1}^{B} \log W^*_{k,b} \tag{7}$$

Where, B is the number of reference datasets. According to the laws of large numbers, when B is large enough, the error caused by TKM randomness is negligible.

Step 4: Calculate standard deviation $sd(k)$.

$$sd(k) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(\log W^*_{k,b} - E^*\{\log(W_k)\}\right)^2} \tag{8}$$

Step 5: Choose k_{min}, which meets Eqs. (9) and (10), to determine k.

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{9}$$

$$s_k = \sqrt{\frac{1+B}{B}} sd(k) \tag{10}$$

After determining the cluster number k by GS evaluation, only the values of initial cluster centers need to be calculated to complete the initialization.

C. Fuzzy Membership Functions

FMF belongs to fuzzy evaluation methods. It is a comprehensive evaluation method of objects in the form of fuzzy sets [36]. It describes the correlation of each sample with clusters without the specific distribution of the initial cluster centers. For each x_t in the domain $\boldsymbol{x}$, if $u_A(x_t) \in [0,1]$, A is called a fuzzy set on the domain $\boldsymbol{x}$, and u_A is the fuzzy map (FMF) of $\boldsymbol{x}$ to A. The closer $u_A(x_t)$ is to 1, the higher the degree of x_t belongs to A. It is expressed in Eq. (11).

$$\boldsymbol{x} \xrightarrow{u_A} [0,1] \Rightarrow A = \left\{ x_t, u_A(x_t) \mid x_t \in \boldsymbol{x} \right\} \tag{11}$$

When there are multiple fuzzy sets on $\boldsymbol{x}$, we have the following representation in Eq. (12).

$$\mathbf{U} = \begin{bmatrix} u_A(\boldsymbol{x}) \\ u_B(\boldsymbol{x}) \\ \vdots \end{bmatrix} \tag{12}$$

Where, $u_A(\boldsymbol{x})$ is the row vector of the membership degree.

In the fuzzy theory, FMF is diverse and in general can be divided into three types: big-scale, small-scale, and middle-scale [37-39]. These three types of membership degrees can help us calculate general cluster centers.

D. Data Density

DD is the number of samples contained in a given range. It can help us to analyze the distribution of the dataset. To get the density, it is necessary to specify the clustering radius Υ

and calculate the number of contained samples based on the distances d between the centers and other samples. If $d \le \Upsilon$, the samples belong to the range of the center; If $d > \Upsilon$, the samples are out of the range of the center. This is beneficial for us to calculate the density of each sample and further filter out the most compact one.

III. Modified K-Means Clustering Algorithm

The TKM has the drawbacks including insufficient determination of cluster number and initial cluster centers. The modified K-means clustering algorithm (MKM) proposed in this paper aims to find appropriate initial cluster centers and optimum cluster number k to achieve better clustering performance. The basic MKM framework is presented in Fig. 1.

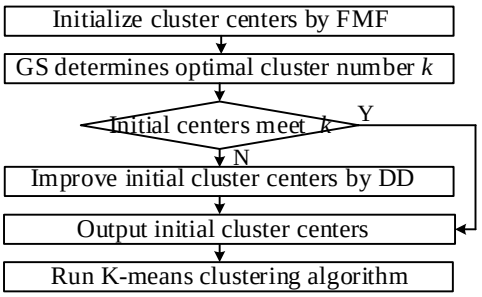


Fig. 1. Basic framework of MKM.

We first use FMF to obtain the dataset membership degree, then calculate the initial cluster centers. Next, the optimal cluster number k is determined by GS evaluation. If the previously initialized centers do not meet GS-based k, then they will be optimized by DD. Finally, the improved centers are used for K-means clustering.

A. Modified Initial Cluster Centers

1) Initial cluster centers based on FMF

The clusters of any dataset satisfy the constraint $k \ge 2$. The fewer the clusters, the easier the distribution of cluster centers can be estimated, but selecting a reasonable method to determine the initial cluster centers is still significant. For any datasets that satisfy $2 \le k \le 3$, FMF can appropriately describe general data distribution. In this paper, p-trapezoidal FMF is used to describe three types of membership degree of the dataset. They are shown as follows.

The middle-scale FMF is shown in Fig. 2.

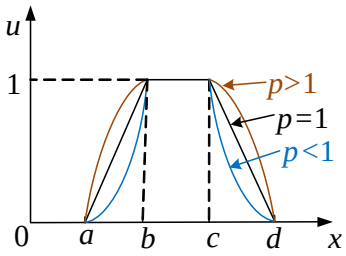


Fig. 2. Middle-scale FMF.

As can be seen from Fig. 2, the membership degree of the data belonging to [b, c] is always 1. When p>1, the membership degree outside [b, c] tends to be convex outward. While p<1, it is concave inward. The middle-scale FMF is calculated as Eq. (13).

$$u(x_t)=\begin{cases}0, & x_t \le a\\ \left(\dfrac{x_t-a}{b-a}\right)^p, & a < x_t \le b\\ 1, & b < x_t \le c\\ \left(\dfrac{d-x_t}{d-c}\right)^p, & c < x_t \le d\\ 0, & x_t > d\end{cases} \tag{13}$$

The big-scale FMF and small-scale FMF are shown in Fig. 3 and Fig. 4, respectively. Similar to the middle-scale FMF, the membership degree tends to be outwardly convex when p>1, and conversely inwardly concave when p<1. The calculation is uniformly expressed by Eq. (14).

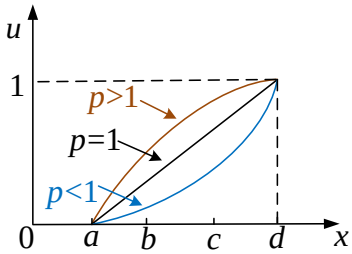


Fig. 3. Big-scale FMF.

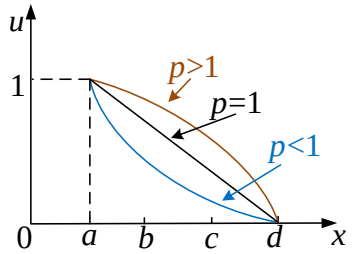


Fig. 4. Small-scale FMF.

$$u(x_t)=\begin{cases}0, & x_t \le a\\ \left(\dfrac{(x_t-a)-\varpi(2x_t-a-d)}{d-a}\right)^p, & a < x_t \le d\\ 0, & x_t > d\end{cases} \tag{14}$$

Where, ϖ is the control parameter. When ϖ=0, Eq. (14) is big-scale; when ϖ=1, it is small-scale.

For comparison, the normalization constraint has to be satisfied in Eq. (15).

$$\text{s.t.}\sum_{j=1}^{k} u_{j,t}=1, \quad 0 \le u_{j,t} \le 1 \tag{15}$$

It can be seen that, when p=1, the function curves are linear and data are neutral to different FMF. While $p \ne 1$, the function curves are all nonlinear: when p<1, the FMF are conservative, and when p>1, the FMF show a positive trend.

TKM is hard and its cluster centers are determined only by the samples within the clusters where they are located. This limits the available information for cluster center initialization. However, combined with FMF, the initialization is softened and more information can be utilized.

The FMF-based initialization is expressed by Eq. (16).

$$v_j = \frac{\sum_{t=1}^{T} (u_{j,t})^m x_t}{\sum_{t=1}^{T} (u_{j,t})^m} \tag{16}$$

Where, m is the fuzzy exponent.

Unlike random initialization, the centers determined by Eq. (16) are global and stable, considering the membership degree of all samples to different clusters. It can definitely improve the robustness of K-means.

To select proper p, we then investigate the attitude change of initial cluster centers to cluster types at different p values. So, the constraint in Eq. (17) must be met.

$$\frac{\left| v_{j,p+\Delta p} - v_{j,p} \right|}{v_{j,p}} \le \varepsilon,\ j = 1,...,k \tag{17}$$

Where, Δp is the step length; ε is the threshold.

It is easy to understand that the smaller ε is, the centers are more sensitive; the larger ε is, the centers are more likely to be found. However, each type of FMF can only calculate one initial cluster center. In other words, the above FMF can only calculate three general centers. For more initial cluster centers, we have the following models.

2) Improved initial cluster centers based on GS-DD

When GS determines k>3, it is necessary to further explore new initial cluster centers based on dataset distribution. But, the method in Eq. (17) cannot be used multiple times to find new centers because we are not always able to find an objective threshold ε, and overuse of the above method will bring great subjectivity to the selection of the centers, so the threshold ε is not valid here.

According to the characteristics of clustering algorithms, the higher the density of the cluster is, the more conducive they are to converge. Here, the distribution of dataset and general initial centers are available, so we propose a method based on DD to improve initial centers [40]. When k>3, we make the following process.

Step 1: Define the maximum distance between center $v_{j,p}$ and other samples in the dataset as $d_{\max}(v_{j,p})$, the minimum distance as $d_{\min}(v_{j,p})$.

Step 2: The mean radius $\Upsilon_{j,p}$ is defined by Eq. (18).

$$\Upsilon_{j,p} = \frac{\sum_{p \in \mathbf{P}} (d_{\max}(v_{j,p}) + d_{\min}(v_{j,p}))}{2 \times |\mathbf{P}|} \tag{18}$$

Step 3: Obtain the density $\rho_{j,p}$ of the region with $v_{j,p}$ as the center and $\Upsilon_{j,p}$ as the radius, the number of samples contained in the region is calculated by Eqs. (19) and (20).

$$\rho_{j,p} = \sum_{t=1}^{T} \varphi(\Upsilon_{j,p} - d(v_{j,p}, x_t)) \tag{19}$$

$$\varphi(y) = \begin{cases} 1, & y \ge 0 \\ 0, & y < 0 \end{cases} \tag{20}$$

Step 4: $v_{j,p}$ with the max $\rho_{j,p}$ is taken as the new center. Similarly, the second or third max one is also applicable to optimize more initial centers, until GS-based k is satisfied.

Based on the above analysis, all initial cluster centers are confirmed and the dataset can be clustered directly using K-Means clustering algorithm.

B. Modified K-Means Clustering Algorithm

We provide a detailed solution process of the MKM clustering algorithm as follows.

We first set the threshold ε and the exponent p of FMF, and save the exponent p to **P** before calculating initial centers. If the change of centers under p does not satisfy Eq. (17), we update p by step length Δp and calculate the initial centers again until the constraint in Eq. (17) is satisfied. Subsequently, we calculate k using GS and check the previous centers: if GS-based k is satisfied, clustering is performed directly; if not, new centers are improved based on DD and finally clustering is performed [41]. MKM considers the characteristics of the clustering algorithm, and focuses on the stability of the initial solution and distribution of the dataset. The detailed steps of MKM are shown in Fig. 5.

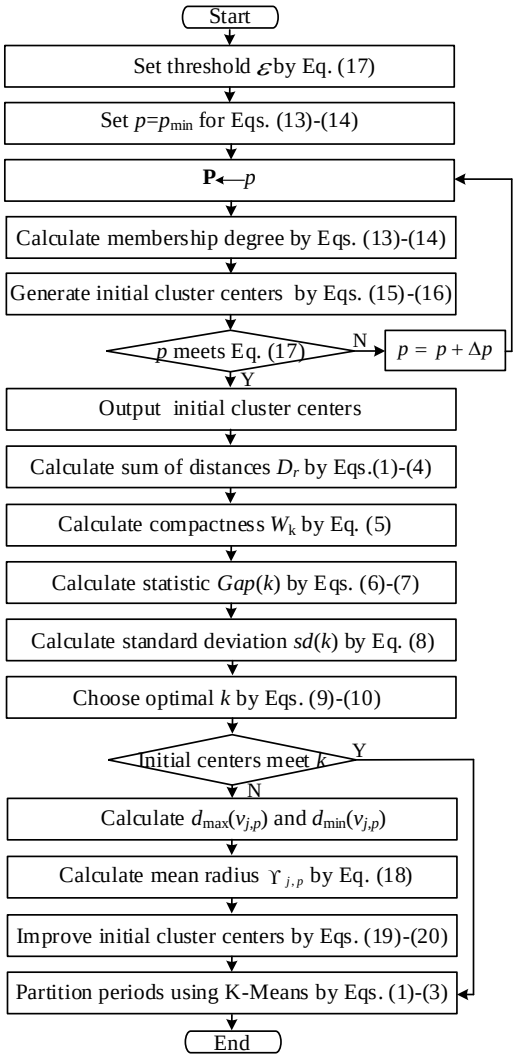


Fig. 5. Solution flowchart of MKM clustering algorithm.

IV. Evaluation Indexes for Modified K-means

A good clustering algorithm is efficient and able to produce high-quality clusters: high intra-cluster similarity

and low inter-cluster similarity. In this paper, the clustering efficiency and quality of the proposed MKM are evaluated by the following indexes.

A. Evaluation Index for Clustering Efficiency

The excellent efficiency of clustering algorithms requires a short convergence time and low error of its objective function value J. For the error, we use standard deviation for efficiency evaluation. It is calculated by Eq. (21).

$$\sigma = \sqrt{\frac{\sum_{t=1}^{T}(J_t - \bar{J})^2}{T}} \tag{21}$$

B. Evaluation Indexes for Clustering Quality

1) Silhouette coefficient evaluation index

Silhouette coefficient index (SC) combines both intra-cluster dissimilarity and inter-cluster dissimilarity to evaluate clustering effect [35]. It is expressed as Eqs. (22)-(23).

$$\mathrm{SC}(k) = \frac{\sum_{t=1}^{T}\frac{b(t)-a(t)}{\max\{a(t),b(t)\}}}{T} \tag{22}$$

$$b(t) = \min\{b_{j,t},\ldots,b_{k,t}\} \tag{23}$$

Where, $a(t)$ is the average distance from sample x_t to other samples in the same cluster; $b_{j,t}$ is the average distance from x_t to all samples in different cluster c_j. The larger the SC, the better the clustering effect.

2) Davies Boldin evaluation index

Davies Boldin index (DB) is defined by Davis and Bouldin [35]. It is shown in Eqs. (24)-(26).

$$\mathrm{DB}(k) = \frac{\sum_{j=1}^{k}\max_{j,j\neq i}\left\{\frac{S_i+S_j}{d_{ij}}\right\}}{k} \tag{24}$$

$$S_i = \frac{1}{|c_i|}\sum_{x_t\in c_i}|x_t - z_i| \tag{25}$$

$$d_{ij} = |z_i - z_j| \tag{26}$$

Where, z_j is the mean value of the cluster c_j; S_i is the average distance between the data and the center in the cluster c_j; d_{ij} measures the dispersion degree between the cluster c_i and cluster c_j. The smaller the value of DB(k), the better the quality of clustering.

V. Case Study Based on TOU Period Partition

A. Experimental Case

The peak-valley period partition of TOU tariff is a typical case to verify the performance of partition clustering algorithms [12, 27], therefore we choose it to discuss the clustering performance of MKM. To make experimental case more comprehensive and obtain a more accurate period partition, this paper uses a daily load curve with a 15-minutes time interval, as shown in Fig. 6.

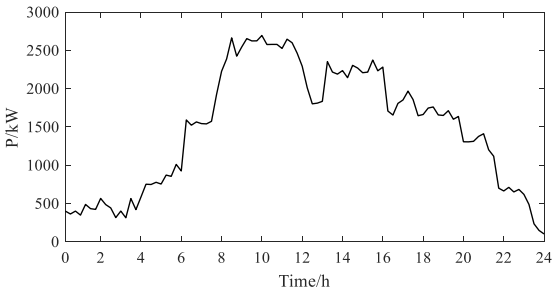


Fig. 6. Daily load curve.

According to Fig.2 and Eq. (13), parameters a and d are the minimum and maximum values of the load data, respectively. The rest parameters b and c are the mean values of the load. The remaining parameters are set as shown in Table I.

TABLE I
Parameter Settings

Parameters	Values
B	1000
$p_{\min}$	0.1
Δp	0.01
m	2
ε	0.01%

B. Results and Discuss

The general initial cluster centers under different p are shown in Fig. 7.

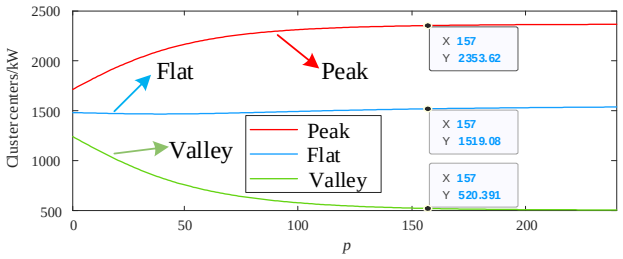


Fig. 7. General cluster centers under different p.

As p increases, the peak period cluster center increases and the valley period cluster center decreases, while the flat period cluster center is almost constant. As marked in the figure, when p=1.57, the change of centers is no longer obvious and meets the constraint in Eq. (17). So, peak, flat and valley centers are determined.

The GS-based k is shown in Fig. 8.

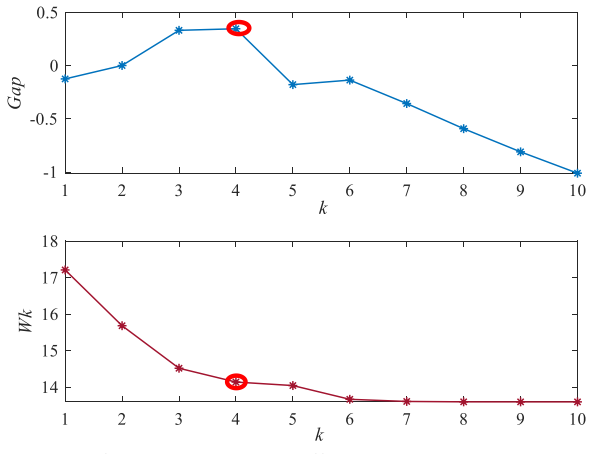


Fig. 8. Change of $Gap(k)$ and W_k under different k.

It is easy to find that W_k decreases gradually. However, from k=4, the decrease of W_k slows down significantly. The change of $Gap(k)$ is not monotonic, but only when k=4, the constraint in Eq. (9) is met. So, the optimal number k is 4.

In TOU period partition studies, the most common is to divide the load curve into three periods or four periods: peak, shoulder, flat and valley. In this paper, simulation scenarios are established as shown in Table II.

TABLE II
SIMULATION SCENARIOS

Scenarios	Clustering algorithms	Period types
Scenario 1	MKM	peak, **shoulder**, flat, valley
Scenario 2	MKM	peak, flat, **leg**, valley
Scenario 3	MKM	peak, flat, valley
Scenario 4	TKM	peak, flat, valley

Because of the special structure of middle-scale FMF, the flat center remains almost linear at the same level as p increases (Fig. 7). Therefore, the initial centers improved by Eqs. (18)-(20) are only considered to exist in valley and peak periods, without considering the flat period. All determined initial cluster centers are shown in Table III.

TABLE III
INITIAL CLUSTER CENTERS

Period types	Values (kW)
Peak	2353.62
Shoulder	1968.42
Flat	1519.08
Leg	863.16
Valley	520.39

Partition TOU periods according to the centers in Table III, we have the iteration process which are shown in Fig. 9.

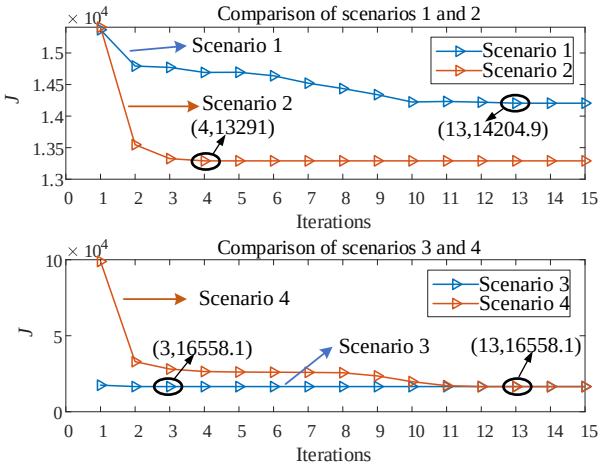
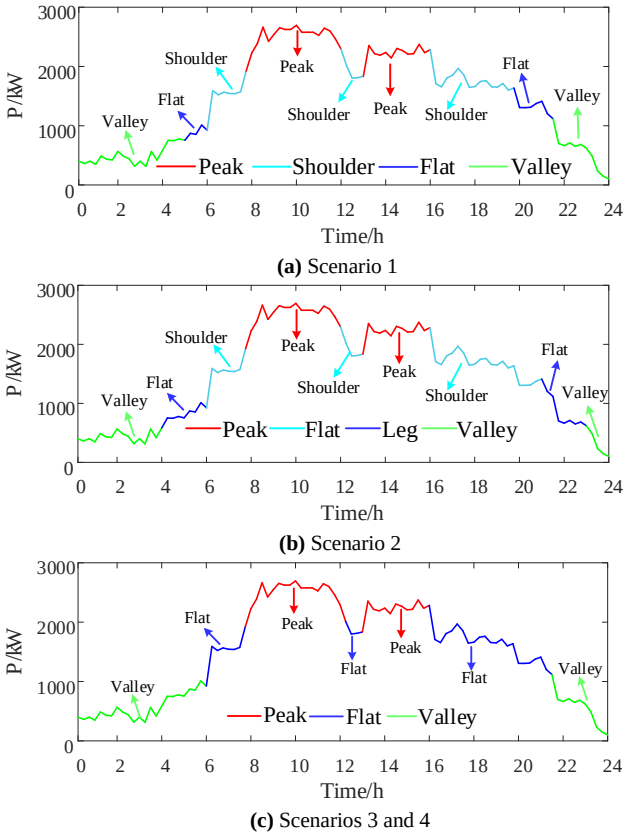


Fig. 9. The iterations of J.

Compare the iterations of scenarios 3 and 4, the MKM converges faster noticeably (3 and 13, respectively) and has a smaller standard deviation of J (242.00 and 20355.70, respectively). Overall, MKM is better than TKM. From the comparison of scenarios 1 and 2, the convergence time of scenario 1 is shorter than scenario 2, and the convergence value of scenario 2 is better. So, for the same algorithm and the same number of clusters, the selection of the initial cluster centers also significantly affects the clustering performance. Since scenarios 3 and 4 have the same convergence values, they divide the load curve into the same periods. The above scenarios can produce three kinds of TOU period partitions. They are shown in Fig. 10.



(a) Scenario 1

(b) Scenario 2

(c) Scenarios 3 and 4

Fig. 10. TOU period partitions in different scenarios.

The specific time points of each period in different scenarios are shown in Table IV.

TABLE IV
SPECIFIC TIME POINTS

Scenarios	Period types	Time points	Total time
1	peak	07:45~12:00 13:00~16:00	7h15min
	shoulder	06:00~07:45 12:00~13:00 16:00~19:45	6h30min
	flat	05:00~06:00 19:45~21:30	2h45min
	valley	00:00~05:00 21:30~00:00	7h30min
2	peak	07:45~12:00 13:00~16:00	7h15min
	flat	06:00~07:45 12:00~13:00 16:00~21:00	7h45min
	leg	04:00~06:00 21:00~23:00	4h
	valley	00:00~04:00 23:00~00:00	5h
3&4	peak	07:45~12:15 13:00~16:00	7h30min
	flat	06:00~07:45 12:15~13:00 16:00~21:30	8h
	valley	00:00~06:00 21:30~00:00	8h30min

Compare scenarios 1 and 2, we can find that the time points and total time of the peak period are the same. The partitions of the peak in scenarios 3 and 4 are also almost the same as scenarios 1 and 2, only with a 15 min difference. However, the other period partitions vary. The total time of the flat in scenario 1 is 5 hours less than scenario 2 and 5h15min than scenario 3, respectively. Besides, the total time of the valley in scenario 2 is 2h30min less than scenario 1 and 3h30min than scenario 3, respectively. This indicates that the peak period of this load curve is significantly different from other periods, which is easily recognized by K-means. The evaluation of each scenario is shown in Table V.

TABLE V
EVALUATION OF CLUSTERING QUALITY

Indexes / Scenarios	SC	DB
Scenario 1	0.7190	0.0188
Scenario 2	0.6960	0.0262
Scenario 3&4	0.5724	0.0423

The evaluation rule is that the larger the SC and the smaller the DB, the better the clustering and the higher the quality of the period partitions. From the above evaluation indexes, the period partition quality of scenarios 1 and 2 is better than that of scenarios 3 and 4, and scenario 1 can better reflect the load feature compared to scenario 2. Therefore, the load curve in this paper is best divided into four periods: peak, shoulder, flat, and valley.

VI. CONCLUSION

To improve the clustering quality and solving efficiency of the traditional K-means clustering algorithm, we proposed an MKM clustering algorithm that improves the initial cluster centers and number of clusters. Due to the significant impact of initial cluster centers on K-means clustering performance, we first used fuzzy membership functions to obtain the general initial cluster centers, rather than randomly selecting based on experience. Then, we used DD to further improve the previous initial cluster centers and optimized the number of clusters through gap statistics. In addition, the silhouette coefficient index and Davies Boldin index were used to evaluate the performance of the proposed MKM algorithm. Finally, the comparative analysis results based on different scenarios indicate that the proposed MKM clustering algorithm has significant performance advantages in solving efficiency and clustering quality compared to the TMK algorithm.

REFERENCES

[1] T. Xie, R. H. Liu, Z. Y. Wei, "Improvement of the fast clustering algorithm improved by K-means in the big data," *Applied Mathematics and Nonlinear Sciences*, vol. 05, no. 01, pp. 1-10, 2020.

[2] Jain A K, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 08, pp. 651-666, 2010.

[3] L. L. Zha, P. F. Niu, L. F. Chang, X. C. Zhang, "Research on boiler thermal efficiency based on optimized K-means clustering algorithm," *Control Engineering of China*, vol. 28, no. 01, pp. 29-34, 2021.

[4] C. W. Chen, J. B. Luo, K. J. Parker, "Image segmentation via adaptive K-means clustering and knowledge-based morphological operations with biomedical application," *IEEE Transactions on Image Processing*, vol. 07, no. 12, pp. 1673-1683, 1998.

[5] YK. Lam, PWM. Tsang, CS. Leung, "PSO-based K-means clustering with enhanced cluster matching for gene expression data," *Neural Computing and Applications*, vol. 22, pp. 1349-1355, 2013.

[6] J. P. Huang, H. Y. Chen, Z. J. Lin, J. Y. Zhong, "A summary of time-of-use research and practice in a demand response environment," *Power System Protection and Control*, vol. 49, no. 09, pp. 178-187, 2021.

[7] Y. J. He, Z. Ye, W. Wei, "Research on time division of peak-valley TOU tariff considering price elasticity of user demand," *Prices Monthly*, vol. 547, no. 12, pp. 1-9, 2022.

[8] L. Zhang, G. Li, Z. Bie, X. Li, Y. Ling and H. Tian, "FCM based demand response baseline load estimation using smart meter data," *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*, pp. 1-5, 2021.

[9] F. L. Xu, R. J. Zhou, W. L. Zhang, Y. L. Cheng, J. B. Cao, Y. Z. Wang, "Peak-to-valley heat price decision model considering user's satisfaction degree with thermal methods under the contradiction between thermoelectric supply and demand," *Proceedings of the CSU-EPSA*, vol. 31, no. 11, pp. 16-22, 2019.

[10] D. N. Liu, E. F. Xu, M. G. Liu, B. Z. Zhou, Y. H. Ying, Q. B. Yu, "TOU pricing method for park considering local consumption of distributed generator," *Automation of Electric Power Systems*, vol. 44, no. 20, pp. 19-28, 2020.

[11] H. J. Yang, R. T. Shi, Y. H. Ma, J. Ma, Y. M. Shen, "Scheduling strategy of electric energy storage system considering multiple time-of-use electricity prices and potential benefit," *Electric Power Automation Equipment*, vol. 41, no. 10, pp. 130-137, 2021.

[12] B. Jiang, G. R. Li, Z. M. Sun, Z. Q. Pang, "A residential peak and valley time division model based on long short-term memory and improved K-means clustering algorithm," *Modern Electric Power*, vol. 38, no. 06, pp. 620-629, 2021.

[13] Z. Y. Song, J. Zhang, D. X. Niu, X. L. Xiao, "Green time of use electricity price optimization based on fuzzy density K-medoids algorithm and NSGA-II," *Smart Power*, vol. 47, no. 03, pp. 38-45, 2019.

[14] J. Wang, S. T. Wang, Z. H. Deng, "Some problems in cluster analysis," *Control and Decision*, vol. 27, no. 03, pp. 321-328, 2012.

[15] W. Lu, "Improved K-means clustering algorithm for big data mining under hadoop parallel framework," *Journal of Grid Computing*, vol. 18, pp. 239-250, 2020.

[16] Y. Shen, D. H. Yu, H. L. Wang, "Improvement of K-means based on particle swarm optimization clustering algorithm," *Computer Engineering and Applications*, vol. 50, no. 21, pp. 125-128, 2014.

[17] T. Bezdan, C. Stoean, AA. Naamany, et al, "Hybrid fruit-fly optimization algorithm with K-means for text document clustering," *Mathematics*, vol. 09, no. 16, pp. 1929, 2021.

[18] Aggarwal S, Singh P. Cuckoo, "Bat and krill herd based K-means++ clustering algorithms," *Cluster Computing*, vol. 22, no. 06, pp. 14169-14180, 2019.

[19] Kuo R J, Mei C H, Zulvia F E, et al, "An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation," *Neurocomputing*, vol. 205, pp. 116-129, 2016.

[20] Y. Chen, B. J. Tian, Y. Z. Peng, Y. Liao, "Gaussian mixture clustering algorithm combining elbow method and expectation-maximization for power system customer segmentation," *Journal of Computer Applications*, vol. 40, no. 11, pp. 3217-3223, 2020.

[21] H. S. Gao, Y. M. Zhang, X. Q. Ji, X. Zhang, Y. J. Yu, "Scenario clustering based distributionally robust comprehensive optimization of active distribution network," *Automation of Electric Power Systems*, vol. 44, no. 21, pp. 32-41, 2020.

[22] Sagala N T M, Gunawan A A S, "Discovering the optimal number of crime cluster using elbow, silhouette, gap statistics, and Nbclust methods," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 13, no. 01, pp. 1-10, 2022.

[23] Z. H. Yan, Z. J. Zhang, Y. P. Wang, Y. Z. Jin, T. Yan, "Improved deep embedding clustering algorithm based on weighted Mahalanobis distance," *Journal of Computer Applications*, vol. 39, no. S2, pp. 122-126, 2019.

[24] Amira M. El-Mandouh, Laila A. Abd-Elmegid, Hamdi A. Mahmoud and Mohamed H. Haggag, "Optimized K-means clustering model based on gap statistic" *International Journal of Advanced Computer Science and Applications(ijacsa)*, vol. 10, no. 01, 2019.

[25] Moftah H M, Azar A T, Al-Shammari E T, et al, "Adaptive K-means clustering algorithm for MR breast image segmentation," *Neural Computing and Applications*, vol. 24, pp. 1917-1928, 2014.

[26] M. L. Liu, B. X. Zhang, X. Li, W. D. Tang, G. Q. Zhang, "An optimized K-means algorithm based on information entropy," *The Computer Journal*, vol. 64, no. 07, pp. 1130-1143, 2021.

[27] X. Zhang, X. Li, "Research on peak and valley periods partition and distributed energy storage optimal allocation considering load

characteristics of industrial park," *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Mauritius, Mauritius, pp. 1-6, 2021.
[28] Atabay H A, Sheikhzadeh M J, Torshizi M, "A clustering algorithm based on integration of K-means and PSO," *2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*, pp. 59-63, 2016.
[29] A. Alam and M. Muqeem, "Automatic clustering for selection of optimal number of clusters by K-means integrated with enhanced firefly algorithms," *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, pp. 343-347, 2022.
[30] M. Yuan et al., "Study of driving cycle of city tour bus based on coupled GA-K-means and HMM algorithms: a case study in Beijing," *IEEE Access*, vol. 09, pp. 20331-20345, 2021.
[31] W. Huang et al., "Peak and valley periods partitioning based on improved K-medoids algorithm," *2020 IEEE Sustainable Power and Energy Conference (iSPEC)*, pp. 1860-1866, 2020.
[32] D. Y. Jiang, W. Zheng, X. Q. Lin, "Research on selection of initial center points based on improved K-means algorithm," *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, pp. 1146-1149, 2012.
[33] Ahmad Wahyu Rosyadi, and Nanik Suciati, "Image Segmentation Using Transition Region and K-Means Clustering," *IAENG International Journal of Computer Science*, vol. 47, no.1, pp. 47-55, 2020.
[34] H. Xue, H. Zhu, F. Tian, and T. Chai, "Busy Fishing Area Recognition based on Improved K-means with Random Walking Centroid," *IAENG International Journal of Computer Science*, vol. 49, no.3, pp. 919-925, 2022.
[35] Guan Wang, Cheng Xing, Jie-Sheng Wang, Hong-Yu Wang, and Jia-Xu Liu, "Clustering Validity Evaluation Method Based on Two Typical Clustering Algorithms," *IAENG International Journal of Computer Science*, vol. 49, no.3, pp. 871-879, 2022.
[36] S. Hu, Y. Xiang, J. Y. Liu, R. Wang, "Fuzzy power flow calculation in distribution networks with distributed generation based on parameterized membership matching function," *Proceeding of the CSEE*, vol. 39, no. 18, pp. 5370-5379, 2019.
[37] Y. Zhang, G. Zhao, C. G. Hu, W. B. Gu, "Comprehensive evaluation method of the portable electricity metering devices based on the improved optimization of membership functions," *Electrical Measurement & Instrumentation*, vol. 58, no. 11, pp. 186-193, 2021.
[38] J. H. Yang, S. Ouyang, Y. L. Shi, R. Y. Huang, Z. W. Liu, "Combined membership function and its application on fuzzy evaluation of power quality," *Advanced Technology of Electrical Engineering and Energy*, vol. 33, no. 02, pp. 63-69, 2014.
[39] Y. Li, J. Wang, S. R. Wang, F. Chen, X. Zheng, J. Tang. "Research on evaluation indices of new energy generation characteristics and membership functions," *Power System Protection and Control*, vol. 46, no. 08, pp. 43-49, 2018.
[40] E. Z. Zhu, Z. H. Wang, F. Liu, Z. J. Ma, "Dh-Kmeans: an improved K-means clustering algorithm based on dynamic initial cluster center determination and hierarchical clustering," *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 170-176, 2022.
[41] Xue Deng, Keyao Zheng, and Ye Xiong, "Cluster Analysis Based on Indicator System on the Development of Digital Economy in Guangdong," *IAENG International Journal of Applied Mathematics*, vol. 51, no.3, pp. 728-735, 2021.

Peng Wang is currently working toward the M.S. degree at School of Automation and Industrial Internet, Chongqing University of Posts and Telecommunications. (e-mail: 17623600676@163.com)

Yiwei Ma is currently an Associate Professor of Electrical Engineering, in Chongqing University of Posts and Telecommunications. (e-mail: mayw@cqupt.edu.cn)