A Novel Clustering Method for PV Power Curve Patterns based on Multidimensional Feature, Entropy Weight, and K-means

Xingzhen Li, Yiwei Ma, Hao Zhong, Miao Huang

***Abstract*—The clustering of photovoltaic (PV) power is a challenging issue, as it is subject to various natural meteorological factors. To address this issue, this paper proposes a novel clustering method for PV power curve patterns based on multidimensional feature, entropy weight method, and K-means algorithm. First, a multidimensional feature model is proposed to better reveal the PV power characteristics, which integrates three major power fluctuation features and four major meteorological features based on factor analysis (FA). Second, the entropy weight method (EWM) is adopted to calculate the weight for each feature, which is used to modify the Euclidean distance in the K-means algorithm for high-quality performance of PV power curve clustering. The experimental results show that this method is more effective than traditional methods in terms of clustering indicators and clustering quality, as it achieves the best results in SC, DBI, and CHI clustering validity indices of 0.4463, 1.0981, and 393.9127, respectively.**

***Index Terms*—Entropy weight, Feature extraction, K-means, Pattern clustering, PV power curve.**

I. Introduction

With the global emphasis on renewable energy, photovoltaic (PV) power generation has developed rapidly and has become one of the main options for replacing fossil fuels. Therefore, PV power generation has become an important means to address carbon emissions and energy crises [1], [2]. However, the intermittency and uncertainty of PV power generation pose significant challenges to the power system. PV clustering effectively describes uncertain PV output features [3], [4]. How to find similarities between uncertain PV power generation outputs and generating representative types of PV outputs is currently the most pressing problem to be solved. At present, research on PV power clustering mainly involves two aspects: clustering feature extraction [5],[6] and clustering algorithm [7], [8].

In the research of clustering feature extraction, traditional power curves usually use curve contour features [9], [10], power features [11], user behavior features [12], and other features for clustering analysis. The various features of PV power curves are closely related to meteorological factors, so the meteorological features of PV power curves must be considered in the process of clustering feature extraction. Ref. [13] employs multi-scale fluctuation feature extraction techniques to capture the fluctuations in PV power. Ref. [14] extracted the global and peak features of PV power curves. Ref. [15] extracts and distinguishes the commonalities and differences of PV power time series features, thereby improving the accuracy and rationality of PV power clustering. Ref. [16] employs the Symbolic Sequence Histogram (SSH) to illustrate the fluctuation features in PV power. Ref. [17] uses the FCM algorithm to perform cluster analysis on multiple meteorological features and comprehensively analyzes the impact of various meteorological features on PV power generation. Ref. [18] uses the Kendall rank correlation coefficient to evaluate the main meteorological features. Ref. [19] considers the randomness of meteorological features and uses weights to enhance the correlation between meteorological features and PV power generation data. Refs. [13]-[16] primarily concentrates on the features of photovoltaic power curves, with little attention paid to the impact of meteorological features. Refs. [17]-[19] fully consider the importance of meteorological features but overlook the inherent features of PV power generation. Therefore, this paper proposes a comprehensive method that combines meteorological and power features.

In the research of clustering algorithm, Ref. [20] proposes a modified K-means clustering algorithm based on fuzzy membership functions (FMF), gap statistic (GS), and data density (DD). Ref. [21] establishes a fusion pattern recognition model based on the K-means algorithm, dividing PV data into three modes. Ref. [22] divides historical PV power generation data into four scenarios representing different output and fluctuation features based on the K-means algorithm. Ref. [23] uses the elbow method to determine the optimal clustering number and uses the

Manuscript received Sep. 12, 2024; revised Jan. 31, 2025.

This work was supported in part by the National Natural Science Foundation of China under Grant 61703068, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN202304206, and in part by the Chongqing Postgraduate Research and Innovation Project under Grant CYS23468 & CYS23469.

Xingzhen Li is a postgraduate student of Electrical Engineering Department, School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (e-mail: S220303007@stu.cqupt.edu.cn).

Yiwei Ma is an associate professor at the School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (Corresponding author to provide e-mail: mayw@cqupt.edu.cn).

Hao Zhong is an associate professor at the Hubei Provincial Key Laboratory for Operation and Control of Cascaded Hydropower Station, China Three Gorges University, Yichang 443002, China. (e-mail: zhonghao022@163.com).

Miao Huang is a senior engineer at the School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (e-mail: huangmiao@cqupt.edu.cn).

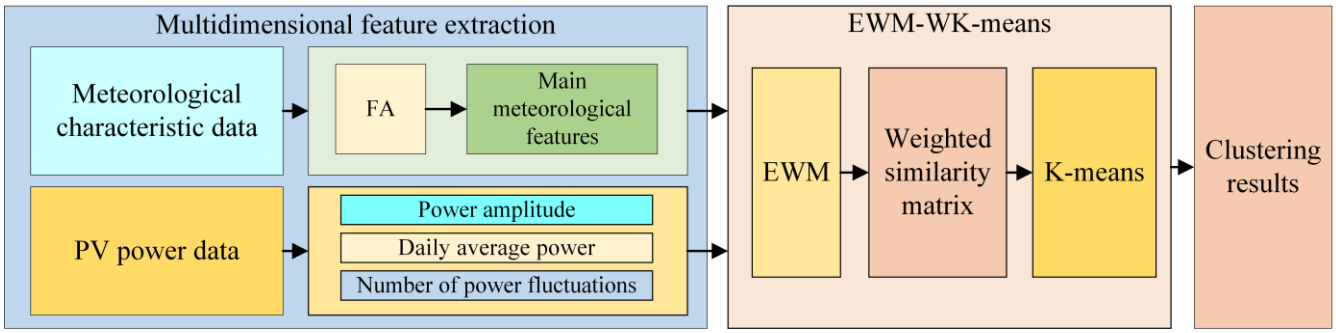

Fig. 1. The framework diagram of PV power curve clustering based on multidimensional features and EWM-WK-means

K-means algorithm to divide the wind-solar power curves into five scenarios. Ref. [24] uses the FCM algorithm to cluster the three fluctuation features of historical data, thereby improving the accuracy and stability of subsequent predictions.

The most of existing references adopt the K-means algorithm for PV data clustering, mainly because this algorithm has fast computation speed and high efficiency when clustering large datasets with multiple features [25]-[27]. However, none of these methods consider the different importance of different features. Therefore, this paper proposes an EWM-K-means (EWK-means) algorithm based on multidimensional feature extraction (MFE), namely the MFE-EWK-means method. The method uses the EWM algorithm to calculate the entropy weights of each feature and then uses the entropy weights to improve the similarity distance of the K-means algorithm. The framework diagram of the MFE-EWK-means method is given in Fig. 1. This method makes the clustering and analysis of PV data more reasonable. The main contributions of this paper are as follows:

1. A multidimensional feature extraction (MFE) method was proposed by studying the impact of various meteorological features and power features on PV clusters.
2. An EWK-means algorithm is proposed as a feature-based weighted clustering algorithm. It uses EWM to calculate the entropy weight of each feature and optimizes the distance calculation method of the K-means algorithm.
3. The MFE-EWK-means method divides photovoltaic data into eight groups, including weather change patterns such as sunny-cloudy, cloudy-sunny, overcast-rainy, and rainy-overcast.

The main work of this article is summarized as follows: Section II presents a novel dimensional feature extraction (MFE) method, Section III gives a feature-based EWK-means clustering algorithm. Section IV shows the comparative experimental results to verify the effectiveness and superiority of the method. Finally, Section V summarizes the work of this article.

II. Feature Construction

The relationship between PV power data and various meteorological factors has been confirmed. The power of PV power generation is affected by a series of meteorological factors, including solar radiation, temperature, humidity, and cloud cover [28]. The power characteristics of different types of PV power curves also have great differences. Therefore, this article extracts the meteorological and power features from PV data.

A Meteorological Feature Extraction

Among all meteorological factors, their impact on PV power generation has a certain repeatability. Therefore, this paper will use the FA algorithm to select the features of all meteorological factors and screen out meteorological features with significant impact.

The main task of the FA algorithm is to extract meteorological features from the overlapping information in the original meteorological factors, ultimately reducing the number of meteorological factors [29]. The basic idea is to reduce multiple meteorological factors with complex internal connections into a few unrelated meteorological features through the correlation matrix between each meteorological factor. The FA model is illustrated in Eq. (1).

$$X=\begin{bmatrix}X_1\\X_2\\\vdots\\X_p\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1q}\\a_{21}&a_{22}&\cdots&a_{2q}\\\vdots&\vdots&\ddots&\vdots\\a_{p1}&a_{p2}&\cdots&a_{pq}\end{bmatrix}\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_q\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_p\end{bmatrix}=A\mu+\varepsilon \quad (1)$$

Where, $X=\left(X_1,X_2,\ldots,X_p\right)^T$ denotes the initial meteorological matrix; $A=\left(a_{ij}\right)_{pq}$ denotes the factor loading matrix and the element a_{ij} denotes the load of the i-th variable x_i on the j-th dimensional common factor μ_j; $\mu=\left(\mu_1,\mu_2,\ldots,\mu_q\right)^T$ denotes the matrix of common factor vectors; and $\varepsilon=\left(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_q\right)^T$ denotes the special factor vector which is part of the initial meteorological matrix.

The specific steps of the FA algorithm are as follows:

Step 1. Calculate the Pearson correlation coefficient matrix R according to Eq. (2).

$$R=\left(r_{ef}\right)_{q\times q}=\frac{1}{p-1}\sum_{\kappa=1}^{p}\left(x_{\kappa e}^{'}-\bar{x}_e^{'}\right)\left(x_{\kappa f}^{'}-\bar{x}_f^{'}\right) \quad (2)$$

Where, r_{ef} is the correlation coefficient between standardized indicators e and f and satisfies $r_{ef}=1$, $r_{ef}=r_{fe}$; q is the total number of common factors, and κ is the total number of sample variables.

Step 2. Solve the eigenvalues of the matrix R according to Eq. (3).

$$\left|\lambda R-E\right|=0 \quad (3)$$

Where, λ denotes the eigenvalues of the matrix R, and $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_q\geq 0$; E denotes the identity matrix.

Step 3. Calculate the cumulative factor contribution rate according to Eq. (4) and determine the number of extracted features.

$$g_\tau = \sum_{\tau=1}^{\upsilon}\left(\lambda_\tau \Big/ \sum_{j=1}^{q}\lambda_j\right) \tag{4}$$

Where, g_τ denotes the cumulative contribution of the top τ factors; when it is greater than or equal to the threshold ε_g, record the number υ.

Step 4. Calculate the feature loading matrix A based on the S eigenvalues as shown in Eq. (5).

$$A = \left[\sqrt{\lambda_1}l_1, \sqrt{\lambda_2}l_2, ..., \sqrt{\lambda_\upsilon}l_\upsilon\right] \tag{5}$$

Where, $l_1, l_2, \cdots, l_\upsilon$ are the eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_\upsilon$.

Step 5. Calculate the feature scores matrix G according to Eq. (6).

$$G = A'^T R^{-1} X \tag{6}$$

Where, G is the feature score matrix, which is computed using the least squares method. A' is the rotated characteristic load matrix. It is obtained by orthogonal rotation of the feature load matrix so that the variance value of each column is maximized.

The raw 16-dimensional features comprise the following variables: dry-bulb temperature (DBT), azimuth, cloud transparency (CT), dew-point temperature (DPT), solar scattered irradiance (SSI), solar direct irradiance (SDI), solar horizontal irradiance (SHI), fixed inclination irradiance (FII), tracking inclination irradiance (TII), atmospheric precipitation (AP), relative humidity (RH), snowfall depth (SD), ground pressure (GP), wind direction (WD), wind speed (WS), and zenith angle (ZA).

After extracting the main meteorological features through the FA algorithm, the meteorological features with a cumulative contribution rate exceeding 80% are determined as the main meteorological features and standardized into a meteorological feature matrix M, as shown in Eq. (7).

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1\nu} \\ M_{21} & M_{22} & \cdots & M_{2\upsilon} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{n\upsilon} \end{bmatrix} \tag{7}$$

Where, M_n denotes the normalized meteorological feature vector of the n-th power curve; $M_{n\upsilon}$ denotes the υ-th dimension normalized meteorological feature of the n-th power curve.

B Power Feature Extraction

The PV power curves have randomness and volatility, so relevant features can be extracted to describe the amplitude, fluctuation, and distribution of the curve. These features have a significant impact on the quality of PV power generation. Accurately determining power features is an important foundation for clustering PV power curves.

Taking the β-th PV power curve P_β shown in Eq. (8) as an example, extract its power features.

$$P_\beta = \left\{p_{\beta\tau}, \tau = 1, 2, \ldots, T\right\} \tag{8}$$

Where, $p_{\beta\tau}$ denotes the power at the τ-th sampling point of the β-th PV power curve; T denotes the number of sampling nodes of the PV power curve.

Below are the meanings and expressions of the three major power features [30].

1) Power Amplitude (PA)

The PA reflects the maximum value of PV power in a day, which is affected by a variety of factors such as light intensity and temperature. It usually occurs when light intensity is max. The PA is a good representation of the maximum power level of the power curve. The calculation method of PA is shown in Eq. (8).

$$I_{PA\beta} = \max\left(p_{\beta\tau}\right) \tag{9}$$

Where, $I_{PA\beta}$ denotes the power amplitude of the β-th PV power curve.

2) Daily Average Power (DAP)

The DAP represents the average value of the PV power system's output power for the day, which evaluates the overall power level of the PV power curve for a day. The calculation method of DAP is shown in Eq. (9).

$$I_{DAP\beta} = \frac{1}{T}\sum_{\tau=1}^{T} p_{\beta\tau} \tag{10}$$

Where, $I_{DAP\beta}$ denotes the daily average power of the β-th PV power curve.

3) Number of Power Fluctuations (NPF)

The NPF reflects the fluctuation condition of the power curve. For the sunny day type, the curve is smoother and has fewer fluctuations, while for the cloudy and overcast weather type, the power curve is very zigzag and has more fluctuations. The calculation method of NPF is shown in Eq. (10).

$$I_{NPF\beta} = \sum_{\tau=2}^{T-1}\left[(p_{\beta(\tau+1)} - p_{\beta\tau})\cdot(p_{\beta\tau} - p_{\beta(\tau-1)}) < 0\right] \tag{11}$$

Where, $I_{NPF\beta}$ denotes the number of power fluctuations in the β-th PV power curve; $\left[(p_{\beta(\tau+1)} - p_{\beta\tau})\cdot(p_{\beta\tau} - p_{\beta(\tau-1)}) < 0\right]$ denotes the Iverson parenthesis and takes the value of 1 if the condition in the parenthesis is true and the value of 0 otherwise.

The indicators I_{PA} and I_{DAP} describe the maximum and average values of the PV power curve, reflecting the limit and overall situation of the PV power curve. The indicator I_{NPF} describes the fluctuating feature of the PV power curve, reflecting the local fluctuations of the PV power curve. The larger the value of indicator I_{NPF}, the greater the fluctuation.

The power features are extracted from all the PV power curves and normalized to the PV power feature matrix Z, as illustrated in Eq. (11).

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} I'_{PA1} & I'_{DAP1} & I'_{NPF1} \\ I'_{PA2} & I'_{DAP2} & I'_{NPF2} \\ \vdots & \vdots & \vdots \\ I'_{PAn} & I'_{DAPn} & I'_{NPFn} \end{bmatrix} \tag{12}$$

Where, z_n denotes the normalized power features eigenvector of the n-th power curve; I'_{PAn} denotes the PA of the n-th power curve.

III. PV Power Curve Clustering

To facilitate the classification study of PV power curves,

this paper fully integrates the PV meteorological features and power features extracted in Section II. Considering the varying importance of these features, this paper proposes an EWK-means algorithm for multi-feature clustering analysis of photovoltaic curves. It is divided into feature weight assignment based on EWM and power curve clustering based on EWK-means. Fig. 2 illustrates the flowchart of the EWK-means algorithm.

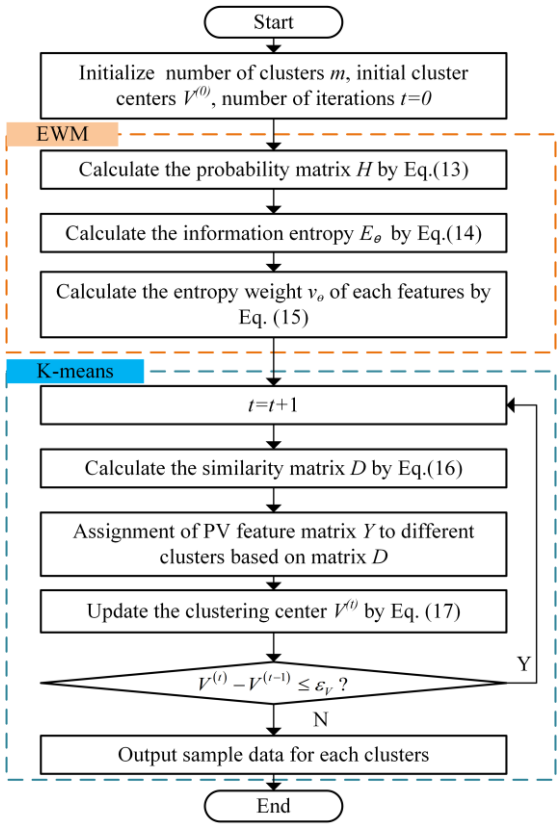

Fig. 2. Flowchart of the EWM-WK-means algorithm

A Feature Weight Assignment based on EWM

The features of meteorological and PV power impact the magnitude and shape of the PV power curve. However, their impact on the curve is not entirely consistent. Therefore, in this paper, the EWM method is used to determine the relative importance of each feature in the standardized feature matrix $Y=[M,Z]$.

The specific steps of the EWM are as follows [31]:

Step 1. Compute the probability matrix H, where the element $h_{\alpha\theta}$ is obtained as in Eq. (13).

$$h_{\alpha\theta}=\frac{y_{\alpha\theta}}{\sum_{\sigma=1}^{n} y_{\sigma\theta}} \tag{13}$$

Where, $h_{\alpha\theta}$ denotes the probability of the θ-th feature in the α-th PV feature data; $y_{\sigma\theta}$ denotes the normalized data for the θ-th feature in the σ-th PV feature data; n denotes the total number of power curves.

Step 2. Calculate the information entropy E_θ of each feature according to Eq. (14).

$$E_\theta=-\frac{1}{\ln n}\left(\sum_{\alpha=1}^{n} h_{\alpha\theta}\ln h_{\alpha\theta}\right) \tag{14}$$

Step 3. Calculate the entropy weight w_θ for each feature according to Eq. (15).

$$w_\theta=\frac{1-E_\theta}{\eta-\sum_{\theta=1}^{\eta}E_\theta} \tag{15}$$

Where, η denotes the total number of the PV features.

B Power Curve Clustering based on EWK-means

In the context of multidimensional feature clustering of PV power curves, it is essential to recognize that the impact of each dimensional feature on the clustering effect varies considerably. The conventional K-means algorithm is merely a straightforward summation of similarity measures between disparate features. However, this approach cannot distinguish the relative importance of various features, so it is unsuitable for multi-feature clustering. Therefore, this paper proposes the entropy weights to improve the K-means algorithm.

The basic idea of EWK-means algorithm is to improve the calculation of distance between features through entropy weight. The objective function is to minimize the weighted sum of squared Euclidean distances between each pattern and its cluster center, as expressed in Eq. (16).

$$\min J(Y,C)=\sum_{\theta=1}^{\eta}\sum_{\alpha=1}^{n}\sum_{\delta=1}^{m} w_\theta\left(y_{\alpha\theta}-c_{\delta\theta}\right)^2 \tag{16}$$

Where, Y denotes the feature matrix; C denotes the clustering centers; w_θ denotes the entropy weight of the θ-th feature; $y_{\alpha\theta}$ denotes the θ-th feature of the α-th feature data; $c_{\delta\theta}$ denotes the θ-th feature of the δ-th clustering center.

The following are the specific steps of the EWK-means algorithm.

Step 1. Set the number of clusters m and initialize the clustering centers $C^{(0)}=\left(C_1^{(0)};C_2^{(0)};\ldots;C_m^{(0)}\right)$ of the feature matrix $Y=\left(Y_1;Y_2;\ldots;Y_n\right)$.

Step 2. The sample set is assigned to the nearest cluster according to the principle of minimum distance. The distance between the feature data and the clustering center is calculated following Eq. (17).

$$d_{\alpha\delta}=\sqrt{\sum_{\alpha=1}^{n}\sum_{\delta=1}^{m}\sum_{\theta=1}^{\eta} w_\theta\cdot\left(Y_\alpha-V_\delta^{(t)}\right)^2} \tag{17}$$

Where, $d_{\alpha\delta}$ denotes the similarity between the α-th power feature data to the δ-th clustering center; Y_α denotes the α-th clustering center; $V_\delta^{(t)}$ denotes the δ-th clustering center after iteration t times.

Step 3. Update the cluster center with the mean of the feature data for each cluster according to Eq. (18).

$$C_\delta^{(t)}=\frac{1}{\gamma_\delta}\sum_{k=1}^{\gamma_\delta} Y_k \tag{18}$$

Where, t denotes the number of iterations; $V_\delta^{(t)}$ denotes the δ-th clustering center after iteration t times; γ_δ denotes the total number of feature data in the category W_δ; Y_k denotes the k-th PV feature data in the category W_δ.

Step 4. Repeat Step 2 and Step 3 until Eq. (19) is satisfied. At this juncture, the clustering centers remain unchanged,

and the PV feature data is classified into different clusters.

$$C^{(t)} - C^{(t-1)} \le \varepsilon_C \tag{19}$$

Where, ε_C denotes the threshold for determining whether the clustering centers have changed.

All the feature data is assigned to different clusters $S = [S_1, S_2, ..., S_m]$ according to the above steps and then all the clusters are output and indexed to the original PV data.

It can be observed that, compared to the K-means algorithm, the EWK-means algorithm simply adds a weight parameter w_θ to the calculation of the objective function and the Euclidean distance. Its function is to adjust the impact of each feature on the clustering results by using different weights to calculate the weighted distance sum of each feature while minimizing the distance within the cluster.

IV. Results and Discussions

To determine the effectiveness of the MFE-EWK-means method, this paper selected the data from a PV power station in Ningxia from August 1, 2019 to July 31, 2020, for experimental verification. The PV generation farm is located at east longitude 102° to 106° and north latitude 37.3° to 40°, with a rated power of 120MW and a maximum power of 150MW, as shown in Fig. 3. The data collection interval is 15 minutes.

Fig. 3. PV plant experimental prototype.

A Feature Extraction

The original photovoltaic dataset includes 16-dimensional meteorological factor data and power curve data. The main stages are as follows. Firstly, downscale the meteorological factor data and extract key meteorological features. Secondly, extract power features from the initial power data.

1) Meteorological feature extraction

The original meteorological factors were downscaled using the FA algorithm, and the top four main meteorological features with a contribution rate greater than 80% were extracted based on Fig. 4. These four main meteorological features contain 81.43% of the initial meteorological factor information. The correlation between these features and the initial meteorological factors is shown in Fig. 5.

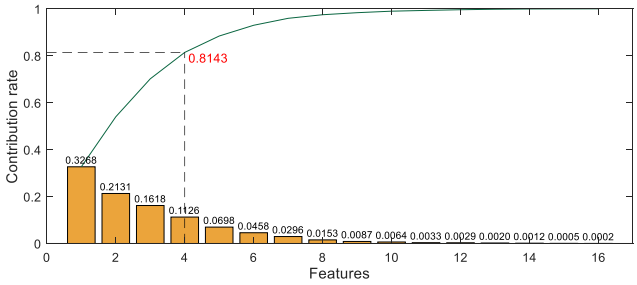

Fig. 4. Contribution of factors and cumulative contribution

Among the main factors I, SSI, SDI, SHI, FII, and TII account for the largest proportion. Therefore, it is designated as the "Radiation Feature (RF)". Regarding the main factor II, DPT accounts for the largest proportion and is hence classified as the "Temperature Feature (TF)". In the main factor III, relative humidity accounts for the largest proportion, so it is designated as the "Humidity Feature (HF)". Similarly, in the main factor IV, CT accounts for the largest proportion, so this indicator is designated as the "Cloud Feature (CF)".

2) Power feature extraction

The raw power curve data is subjected to feature extraction using the methods described in Eqs. (8)-(10), thereby extracting three power features: PA, DAP, and NPF. Table I lists the power feature values of some PV data.

TABLE I
The Power Feature Values

	PA (kW)	DAP (kW)	NPF
1	108.0651	30.1780	9
2	110.9685	32.4213	7
3	99.2486	26.1801	10
4	118.8685	38.4069	1
5	43.7680	7.9978	11
6	89.0027	23.7995	2

B Analysis of clustering results

1) Determination of weights for each feature

The extracted four major meteorological features (RF, TE, HF, CF) and three high-power features (PA, DAP, NPF) were normalized. Subsequently, the entropy weights were obtained by using the EWM to determine the entropy weight magnitude among the features. Table II shows the entropy weights of each feature.

TABLE II
The Entropy Weight Values

Meteorological Feature	Entropy Weight	Power Feature	Entropy Weight
RF	0.1923	PA	0.1669
TF	0.1538	DAP	0.1568
HF	0.0770	NPF	0.1762
CF	0.0770		

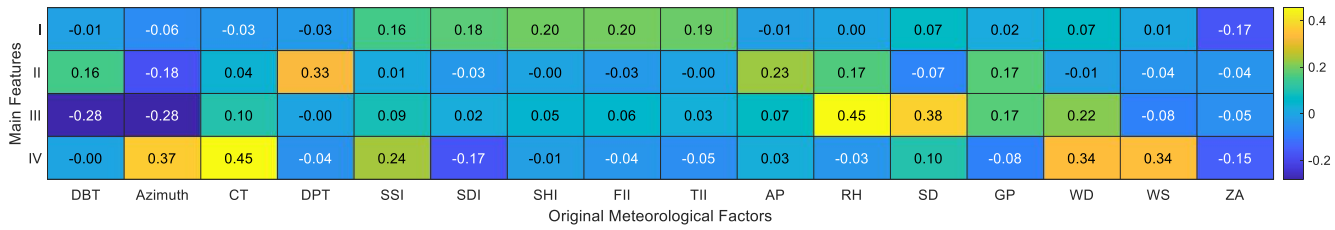

Fig. 5. Pearson correlation coefficients between major factors and original meteorological factors

2) Comparative analysis of clustering results

Different numbers of clusters can lead to different clustering results. This study used three clustering indices, the Silhouette Coefficient (SC), the Davis Bouldin Index (DBI), and the Calinski-Harabasz Index (CHI) [32] for a comprehensive evaluation and selected the optimal number of clusters.

The range of SC values varies from -1 to +1, and is calculated based on the similarity between the sample point and its corresponding cluster, as well as the nearest neighbor cluster. The larger the value, the better the clustering effect. The DBI mainly depends on the density of clustering. The smaller the value, the more compact the clustering and the higher the degree of separation. The CHI is calculated based on the ratio of the variance within a cluster to the variance between clusters. The higher the value, the better the clustering effect. Table III presents the values of each evaluation index for a range of clustering numbers.

TABLE III
NUMBER SELECTION OF CLUSTERS BASED ON CLUSTERING INDICES

Number of Clusters	SC ↑	DBI ↓	CHI ↑
5	0.4040	1.2113	411.3756
6	0.4202	1.2200	401.4596
7	0.4326	1.1468	381.4806
8	**0.4463**	**1.0981**	**393.9127**
9	0.4416	1.1038	298.1806
10	0.4345	1.2692	296.8760
11	0.4387	1.1108	279.1755
12	0.4134	1.1408	292.4205

As illustrated in Table III, As shown in Table III, the clustering effect is optimal when the number of clusters is set to 8. At this point, SC takes the maximum value, DBI takes the minimum value, and CHI takes the larger value. Table IV shows the values of each cluster of PV power data on different features. Fig. 6 shows the clustering results of the MFE-EWK-means method proposed in this article.

The range of meteorological and power features in Table IV varies for different clusters. Fig. 6 shows the effect of the MFE-EWK-means method proposed in this article.

Cluster “a” has significantly higher values in RF and TF compared to other clusters, while HF and CF have the smallest range of values. It has a value range of 96.43-128.03kW in PA and 26.59-44.59kW in DAP, which is the largest among all clusters. However, its value range in NPF is only 1-3. The curves shown in Fig. 6 (a) have large amplitudes and no obvious power fluctuations, so the cluster “a” is named the sunny pattern.

Cluster “b” has a larger range of values and fluctuations in RF and TF, while it has a wider range of values but smaller values in HF and CF. The range of values in PA is 84.78-123.38kW, in DAP it is 19.80-41.08kW, and in NPF it is 5-10, all of which are relatively large. The curves shown in Fig.5 (b) have large amplitudes, but their fluctuation periods occur in the afternoon, so cluster “b” is named the sunny-cloudy pattern.

Cluster “c” has a middle value in RF, a large value and fluctuation range in TF, and a wide range but not large values in HF and CF. The value range in PA is 99.24-119.36kW, in DAP it is 22.66-41.77kW, and in NPF it is 4-10. The curves shown in Fig.5 (c) have large amplitudes, but their fluctuation periods occur in the morning, so cluster “c” is named the sunny-cloudy pattern.

Cluster “d” has a central value in RF, a large and fluctuating range in TF, and a wide and large range of values in HF and CF. The value range in PA is 87.17-120.43kW, in DAP it reaches 18.91-41.05kW, and in NPF it ranges from 6 to 14. The curves in Fig. 6 (d) have a large amplitude, but their fluctuations are obvious and the fluctuation periods are relatively scattered, so cluster “d” is named the cloudy pattern.

Cluster “e” has a small and wide range of values in RF, and is relatively small in TF. However, it has a wide and wide range of values in HF and CF. The value range in PA is 53.98-94.91kW, and in DAP it reaches 12.00-26.28kW, but in NPF it is only 1-3. The curves shown in Fig. 6 (e) are smooth but the amplitude is really small, so cluster “e” is named the overcast pattern.

The value of cluster “f” is relatively small on RF and TF, but it has a wide range of values and the highest value on HF and CF. The value range in PA is 23.27-90.21 kW, in DAP it reaches 4.8-15.33 kW, and in NPF it ranges from 5 to 14. The amplitude of the curves in Fig. 6 (f) is small, the fluctuation of the curves is large, and the fluctuation time is random, so cluster “f” is named the rainy pattern.

Cluster “g” has smaller values in RF and TF, larger and wider ranges in HF, and larger values in CF. The value range in PA is 42.07-75.37 kW, in DAP it reaches 8.54-17.66 kW, and the range in NPF is 4-9. The curves in Fig. 6 (g) have a smaller amplitude, more fluctuations, and are concentrated in the afternoon period, so cluster “g” is named the overcast-rainy pattern.

Cluster “h” has a relatively small value in RF, the smallest value in TF, and a relatively large value in HF and CF. The value range in PA is 52.27-73.70kW, in DAP it reaches 11.53-17.56kW, and in NPF it ranges from 4 to 8. The curves in Fig. 6 (h) have a smaller amplitude, more fluctuations, and are concentrated in the morning period, so cluster “h” is named the overcast-rainy pattern.

TABLE IV
FEATURE VALUES FOR DIFFERENT CLUSTER CLASSES

Clusters	Meteorological Features				Power Features		
	RF	TF	HF	CF	PA/kW	DAP/kW	NPF
a	[-0.03, 0.53]	[0.30, 2.59]	[-1.26, 0.25]	[-1.16, 0.53]	[96.43, 128.04]	[26.59, 44.59]	[1, 3]
b	[-0.12, 0.36]	[-1.12, 2.19]	[-1.37, 0.62]	[-0.89, 1.05]	[84.78, 123.38]	[19.80, 41.08]	[5, 10]
c	[-0.32, 0.23]	[-0.13, 2.12]	[-0.87, 0.74]	[-0.56, 0.90]	[99.24, 119.36]	[22.66, 41.77]	[4, 10]
d	[-0.33, 0.20]	[-0.83, 2.55]	[-1.25, 1.45]	[-0.43, 1.53]	[87.17, 120.43]	[18.91, 41.05]	[6, 14]
e	[-0.31, 0.18]	[-1.29, 0.44]	[-0.94, 1.76]	[-0.36, 1.45]	[53.98, 94.91]	[12.00, 26.28]	[1, 3]
f	[-0.53, 0.08]	[-1.39, 0.19]	[-0.49, 3.03]	[-0.48, 2.03]	[23.27, 90.21]	[4.80, 15.33]	[5, 14]
g	[-0.47, 0.06]	[-1.26, 0.17]	[-0.99, 1.75]	[-0.27, 1.81]	[42.07, 75.37]	[8.54, 17.66]	[4, 9]
h	[-0.36, 0.05]	[-1.17, 0.03]	[-0.20, 2.10]	[-0.46, 1.66]	[52.27, 73.70]	[11.53, 17.56]	[4, 8]

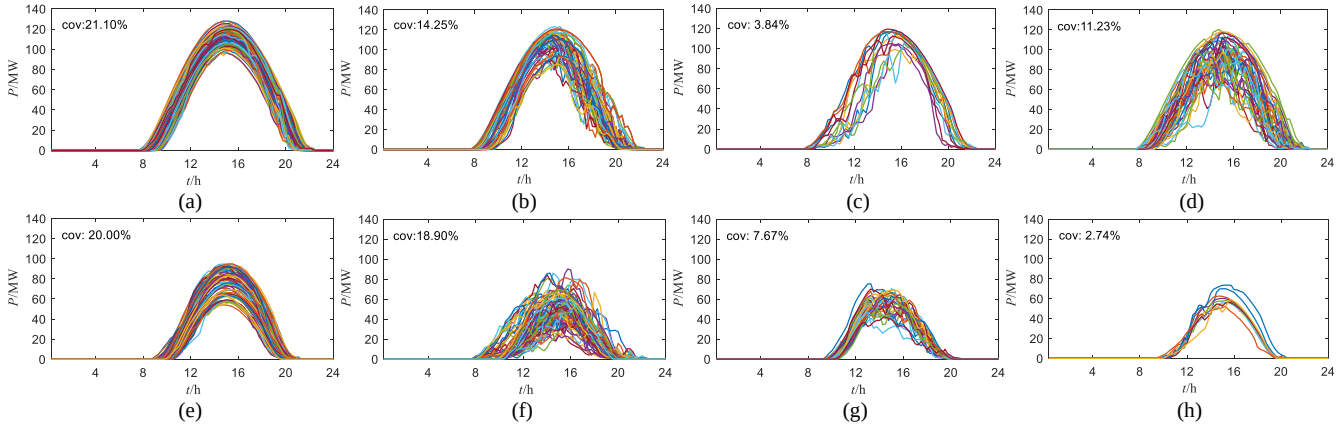

Fig. 6. PV power clustering results based on MFE-EWK-means

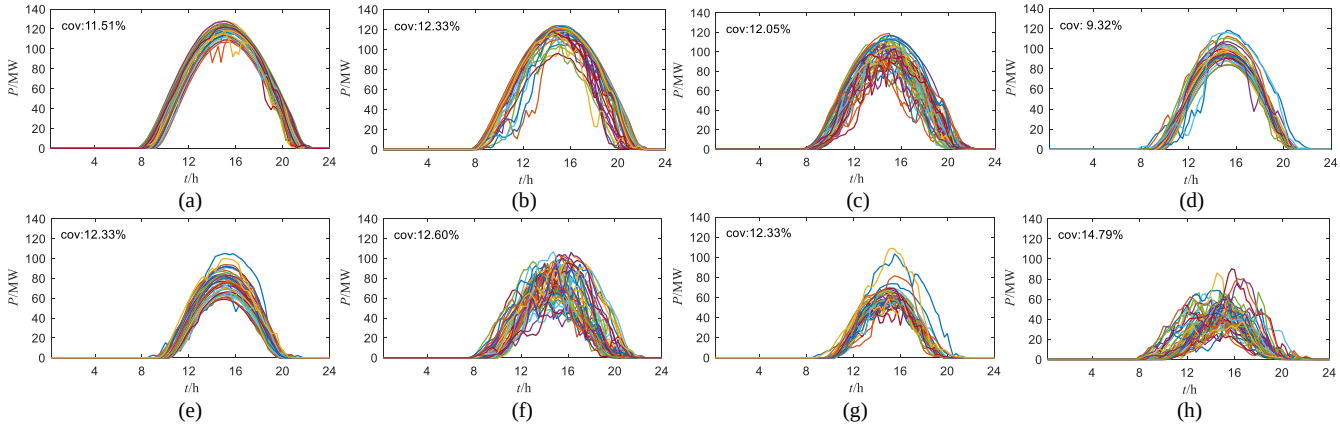

Fig. 7. PV power clustering results based on MFE-EWFCM

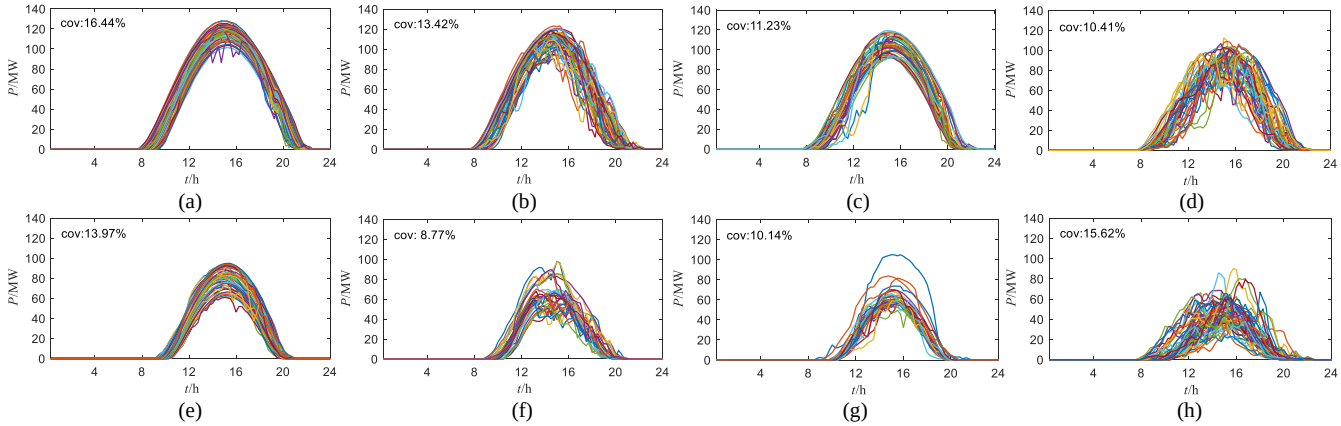

Fig. 8. PV power clustering results based on MFE-K-means

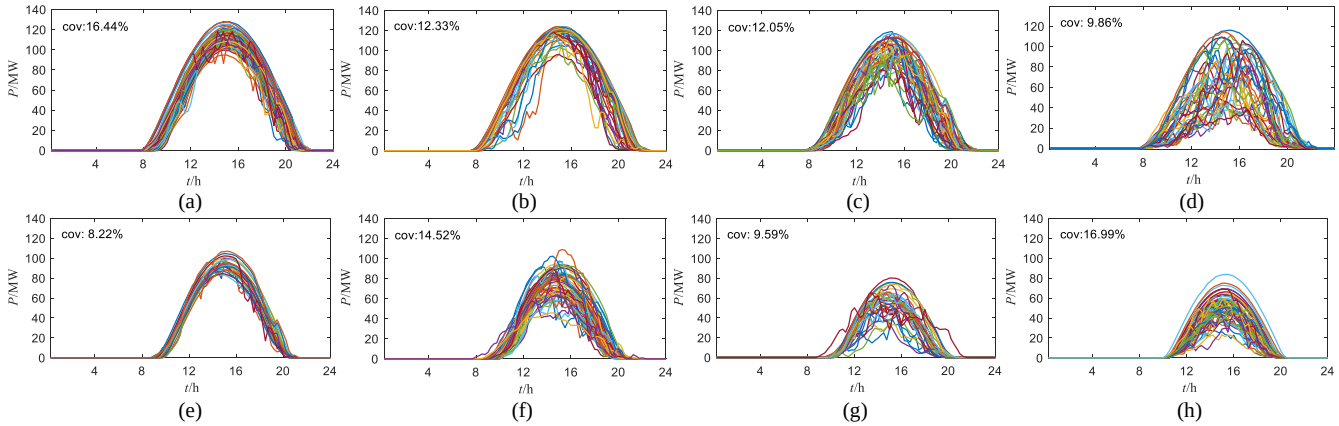

Fig. 9. PV power clustering results based on MFE-FCM

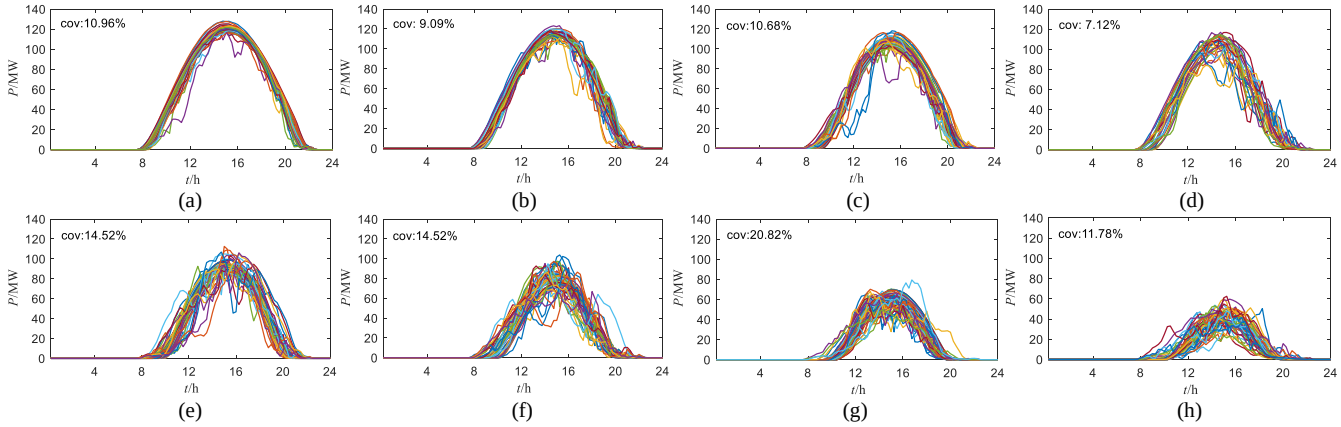

Fig. 10. PV power clustering results based on K-means

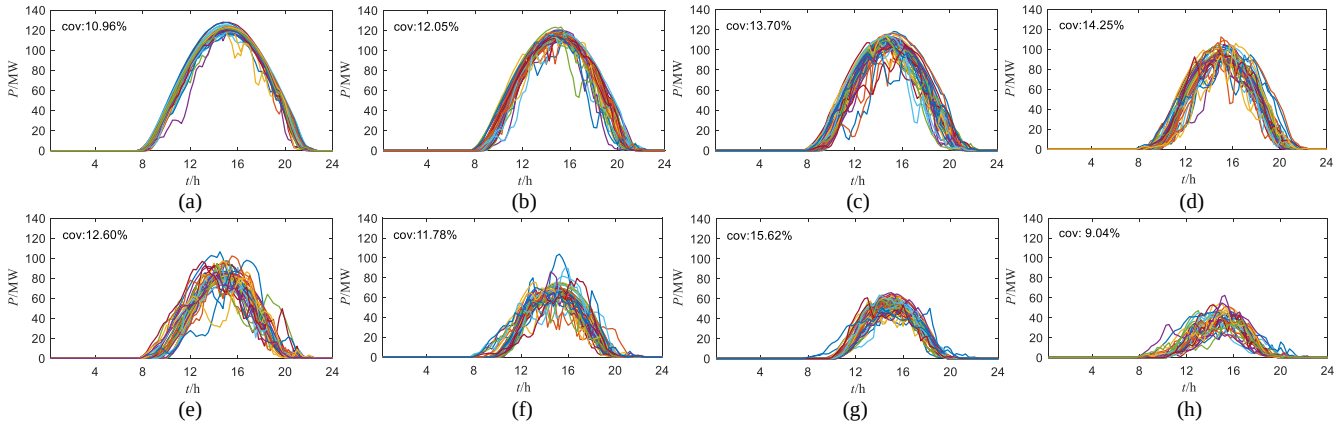

Fig. 11. PV power clustering results based on FCM

In order to verify the superiority of the proposed method, we compared it with other clustering methods, such as MFE-EWFCM, MFE-K-means, MFE-FCM, K-means, and FCM, and the results are indicated in Figs. 6-11.

It can be seen that the clustering results of the FCM method shown in Fig. 11 and the K-means method shown in Fig. 10 are both divided according to the power amplitude, but the curve shapes of the same cluster are not similar. Many smooth curves and fluctuating curves are mixed. From the clustering results of the MPE-FCM method shown in Fig. 9 and the MFE-K-means method shown in Fig. 8, it can be seen that both MFE-FCM and MPE-K-means methods consider the volatility of curves and distinguish between fluctuating curves and smooth curves. However, the weight values of each feature are the same, which makes it difficult to distinguish between the different types of curves effectively.

As shown in Fig. 7, the MFE-WFCM method indicates that adding entropy weight can enhance the differentiation of the majority of power curves. However, the partitioning effect is less optimal for curves with low power values. As shown in Fig. 6, the MFE-WK-means method separates the fluctuation curve from the smooth curve and divides the fluctuation curve in an orderly manner according to the fluctuation period. These fluctuation curves represent different weather types under meteorological conditions, such as sunny-cloudy pattern, cloudy-sunny pattern, cloudy pattern, rainy pattern, overcast-rainy pattern, and rainy-overcast pattern.

Table V lists the values of clustering indicators for each method, and Fig. 12 compares the clustering indicators of all methods.

TABLE V
CLUSTERING INDICATOR VALUES FOR DIFFERENT METHODS

Clustering Methods	Clustering Indicators		
	SC ↑	DBI ↓	CHI ↑
FCM	0.3925	1.2589	348.7063
K-means	0.3975	1.2787	351.0337
MFE-FCM	0.3969	1.2440	354.1833
MFE-K-means	0.4059	1.2083	369.8786
MFE-EWFCM	0.3982	1.2418	363.2598
MFE-EWK-means	**0.4463**	**1.0981**	**393.9127**

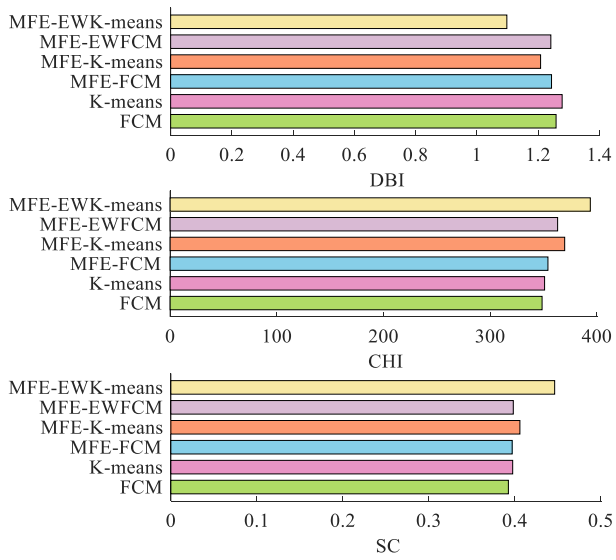

Fig. 12. Comparison of different methods using clustering indicators

From Table V and Fig. 12, it can be seen that the method proposed in this paper is the best among all clustering

indicators. The clustering index of the method using the K-means algorithm is better than that of the method using the FCM algorithm. This indicates that the K-means algorithm has more advantages than the FCM algorithm in processing large datasets with multiple features. Comparing the clustering metrics of the K-means method and the MFE-K-means method, it can be found that after multi-feature extraction, all clustering metrics have improved. This indicates that after extracting features using the MFE method, factors that do not affect the shape of the curve will be removed, greatly increasing the accuracy and precision of clustering. The MFE-EWK-means method adds entropy weight to the MFE-K-means method, which improves all clustering metrics. This indicates that assigning different weights to different features can enhance the influence of important features, thereby increasing clustering accuracy.

Fig. 13 shows the iteration curves of all methods.

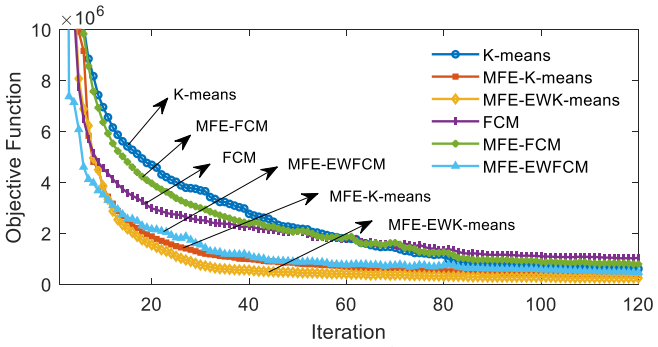

Fig. 13. Iterative curves of all methods

As seen from Fig. 13, the method proposed in this paper is the best in terms of iterative effect. The convergence value of the FCM algorithm is larger than that of the K-means algorithm. This shows that the convergence effect of the K-means algorithm is better, and the early convergence speed will be accelerated after using the MFE method for feature extraction. This is because part of the redundant data is eliminated by the MFE method, and the most critical features are retained. Compared with the MFE-K-means method, the MFE-EWK-means method adds weight value to minimize the objective function and has a better convergence effect.

Table VI lists the clustering index comparison results of the methods used in this paper and those used in Refs. [20]-[24], and Fig. 14 compares the clustering indicators of these methods.

TABLE VI
CLUSTERING INDICATOR VALUES FOR DIFFERENT REFERENCES METHODS

Clustering Methods	Clustering Number	Clustering Indicators		
		SC ↑	DBI ↓	CHI ↑
Ref. [20]	6	0.4202	1.2200	401.4596
Ref. [21]	3	0.3721	1.2782	362.0217
Ref. [22]	4	0.3969	1.2548	369.1856
Ref. [23]	5	0.4040	1.2113	411.3756
Ref. [24]	4	0.3812	1.2679	356.9782
MFE-EWK-means	**8**	**0.4463**	**1.0981**	**393.9127**

It can be seen from Table VI and Fig. 14 that the clustering number and clustering index values obtained by various literature methods are different. The method used in this paper divides the samples into 8 categories, which is more detailed and clearer than other methods. In terms of clustering index, the method used in this paper is superior to other methods in terms of SC and DBI index, and slightly inferior to Ref. [20] and Ref. [23] in terms of CHI index. This is because there are many clustering categories in this paper, which leads to a smaller CHI index.

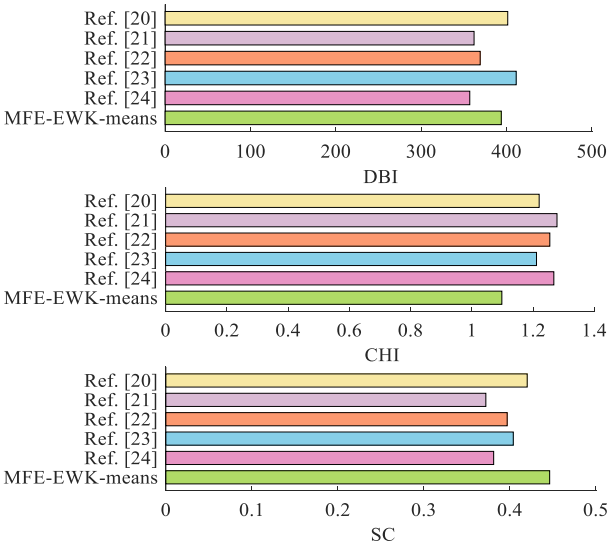

Fig. 14. Comparison of different References using clustering indicators

Therefore, the MFE-EWK-means method proposed in this paper is superior to other methods in clustering performance, clustering index, and convergence. It can effectively identify different fluctuation intervals and divide fluctuation curves.

V. CONCLUSION

To enhance the precision of PV data clustering, this paper proposes an EWM-EWK-means clustering methodology that is founded upon both PV meteorological features and power features. Firstly, the feature extraction method is employed to eliminate some irrelevant factors in the original dataset, thereby obtaining a more optimal feature dataset. Subsequently, the entropy weight calculation is conducted for each feature data set using the EWM method, which assigns distinct weight values to each feature, thereby enlarging the proportion of the features that are more influential and enhancing the closeness of the clustering. Finally, the clustering is conducted using the WK-means algorithm, which is highly effective in addressing the challenges posed by multiple features and multi-influence factors in nonlinear large samples.

The experimental results and comparative analysis demonstrate that this paper's method is capable of differentiating between various types of changing weather conditions, including transitions from sunny-cloudy, cloudy-sunny, cloudy-rainy, rainy-cloudy, and rainy-sunny. This represents a significant advancement in the traditional clustering of weather types, which has typically focused on distinguishing between sunny, cloudy, and rainy conditions. The proposed method offers a more nuanced approach to clustering weather types, enhancing the understanding of their complex relationships. In comparison to alternative methodologies, the approach presented in this paper demonstrates superior efficacy in clustering metrics SC, DBI, and CHI.

REFERENCES

[1] W. Khan, S. Walker, and W. Zeiler, "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach," *Energy*, vol. 240, no. 2, Art. no. 122812, 2022.

[2] Z. Ying, X. Dong, X. Wang, P. Zhang, M. Liu, Y. Zhang, *et al.*, "The relationship between the low-carbon industrial model and human well-being: a case study of the electric power industry," *Energies*, vol. 16, no. 3, pp 1357-1375, 2023.

[3] Z. Song, Y. Huang, H. Xie, and X. Li, "Generation method of multi-regional photovoltaic output scenarios-set using conditional generative adversarial networks," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 13, no. 3, pp 861-870, 2023.

[4] L. Liu, X. Hu, J. Chen, R. Wu, and F. Chen, "Embedded scenario clustering for wind and photovoltaic power, and load based on multi-head self-attention," *Prot. Contr. Mod. Pow.*, vol. 9, no. 1, pp 122-132, 2024.

[5] X. Wu, Z. Zhen, J. Zhang, F. Wang, F. Xu, and H. Ren, et al., "Multidimensional feature extraction based minutely solar irradiance forecasting method using all-sky images," *IEEE Trans. Ind. Appl.*, vol. 60, no. 3, pp 4494-4504, 2024.

[6] Z. Zhang, W. Cui, Y. Tao, and T. Shi, "Road Damage Detection Algorithm Based on Multi-scale Feature Extraction," *Engineering Letters*, vol. 32, no. 1, pp 151-159, 2024.

[7] S. M. Miraftabzadeh, M. Longo, and M. Brenna, "Knowledge extraction from PV power generation with deep learning autoencoder and clustering-based algorithms", *IEEE Access*, vol. 11, no. 7, pp 69227-69240, 2023.

[8] X. Wang, T. Wang, H. Xiang, and L. Huang, "A parallel genetic K-means algorithm based on the island model", *Engineering Letters*, vol. 32, no. 8, pp 1632-1643, 2024.

[9] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: application to intra-day household-level load curves", *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp 411-419, 2014.

[10] F. Lu, X. Cui, J. Xing, S. Liu, Z. Lin, and X. Fei, *et al.*, "Electricity load profile characterisation for industrial users based on normal cloud model and iCFSFDP algorithm," *IEEE Trans. Power Syst*, vol. 38, no. 4, pp 3799-3813, 2023.

[11] J. Zhang, Y. Zhang, X. Xu, X. Fu, and Y. He, "Unsupervised and supervised learning combined power load curve classification based on sequential trajectory feature extraction algorithm," *IEEE Access*, vol. 10, no. 8, pp 90312-90320, 2022.

[12] P. Nystrup, H. Madsen, E. M. Blomgren, G. Zotti, "Clustering commercial and industrial load patterns for long-term energy planning," *Smart Energy*, vol. 2, no. 3, Art. no. 100010, 2021.

[13] J. Zhu, M. Li, L. Luo, B. Zhang, M. Cui, and L. Yu, "Short-term PV power forecast methodology based on multi-scale fluctuation characteristics extraction," *Renew Energy*, vol. 208, no. 3, pp 141-151, 2023.

[14] M. Hu, D. Ge, R. Telford, B. Stephen, and D. C. Wallom, "Classification and characterization of intra-day load curves of PV and non-PV households using interpretable feature extraction and feature-based clustering," *Sustain Cities Soc*, vol. 75, no.10, Art. no. 103380, 2021.

[15] Q. Li, X. Zhang, T. Ma, D. Liu, H. Wang, and W. Hu, "A multi-step ahead photovoltaic power forecasting model based on TimeGAN, Soft DTW-based K-medoids clustering, and a CNN-GRU hybrid neural network," *Energy Rep.*, vol. 8, no. 11, pp 10346-10362, 2022.

[16] L. Zheng, R. Su, X. Sun, and S. Guo, "Historical PV-output characteristic extraction based weather-type classification strategy and its forecasting method for the day-ahead prediction of PV output," *Energy*, vol. 271, no. 5, Art. no. 127009, 2023.

[17] N. Li, L. Li, F. Huang, X. Liu, and S. Wang, "Photovoltaic power prediction method for zero energy consumption buildings based on multi-feature fuzzy clustering and MAOA-ESN," *J. Build. Eng.*, vol. 75, no. 3, Art. no. 106922, 2023.

[18] M. Li, W. Wang, Y. He, and Q. Wang, "Deep learning model for short-term photovoltaic power forecasting based on variational mode decomposition and similar day clustering," *Comput. Electr. Eng.*, vol. 115, no. 2, Art. no. 109116, 2024.

[19] F. Sun, L. Li, D. Bian, H. Ji, N. Li, and S. Wang, "Short-term PV power data prediction based on improved FCM with WTEEMD and adaptive weather weights," *J. Build. Eng.*, vol. 90, no. 4, Art. no. 109408, 2024.

[20] P. Wang, Y. Ma, Z. Ling, and G. Luo, "A Modified K-means Clustering Algorithm Based on FMF-GS-DD," *Engineering Letters*, vol. 31, no.4, pp 1518-1525, 2023.

[21] F. Wang, J. Li, Z. Zhen, C. Wang, H. Ren, and H. Ma, *et al.*, "Cloud Feature Extraction and Fluctuation Pattern Recognition Based Ultrashort-Term Regional PV Power Forecasting," *IEEE Trans. Ind. Appl.*, vol. 58, no. 5, pp 6752-6767, 2022.

[22] Z. Wang, C. Wang, L. Cheng, and G. Li, "An approach for day-ahead interval forecasting of photovoltaic power: a novel DCGAN and LSTM based quantile regression modeling method," *Energy Rep.*, vol. 8, no. 11, pp 14020-14033, 2022.

[23] F. Zheng, X. Meng, L. Wang, and N. Zhang, "Operation optimization method of distribution network with wind turbine and photovoltaic considering clustering and energy storage," *Sustainability*, vol. 15, no. 3, Art. no. 2184, 2023.

[24] L. Zhang, Y. He, H. Wu, X. Yang, and M. Ding, "Ultra-short-term multi-step probability interval prediction of photovoltaic power: A framework with time-series-segment feature analysis," *Sol. Energy*, vol. 260, no. 8, pp 71-82, 2023.

[25] F. Han, T. Pu, M. Li, and G. Taylor, "Short-term forecasting of individual residential load based on deep learning and K-means clustering," *CSEE J. Power Energy Syst*, vol. 7, no. 2, pp 261-269, 2021.

[26] J. Yuan and Y. Tian, "Practical privacy-preserving mapreduce based k-means clustering over large-scale dataset," *IEEE Trans. Cloud Comput.*, vol. 7, no. 2, pp 568-579, 2019.

[27] Y. Gu, K. Li, Z. Guo, and Y. Wang, "Semi-supervised k-means ddos detection method using hybrid feature selection algorithm," *IEEE Access*, vol. 7, no. 5, pp 64351-64365, 2019.

[28] G. G. Kim, J. H. Hyun, J. H. Choi, S. -H. ahn, B. G. Bhang and H. -K. Ahn, "Quality analysis of photovoltaic system using descriptive statistics of power performance index," *IEEE Access*, vol. 11, no. 3, pp 28427-28438, 2023.

[29] A. Ahmadian, M. Sedghi, H. Fgaier, B. Mohammadi-ivatloo, M. A. Golkar, and A. Elkamel, *et al.*, "PEVs data mining based on factor analysis method for energy storage and DG planning in active distribution network: introducing S2S effect," *Energy*, vol. 175, no. 5, pp 265-277, 2019.

[30] M. Mohamed, F. E. Mahmood, M. A. Abd, A. Chandra, and B. Singh, "Dynamic forecasting of solar energy microgrid systems using feature engineering," *IEEE Trans. Ind. Appl.*, vol. 58, no. 6, pp 7857-7869, 2022.

[31] H. Zhao and J. Li, "Energy efficiency evaluation and optimization of industrial park customers based on PSR model and improved Grey-TOPSIS method," *IEEE Access*, vol. 9, no. 5, pp 76423-76432, 2021.

[32] L. E. Brito Da Silva, N. M. Melton, and D. C. Wunsch, "incremental cluster validity indices for online learning of hard partitions: extensions and comparative study", *IEEE Access*, vol. 8, no. 1, pp 22025-22047, 2020.

Xingzhen Li received the B.Sc. degree in electrical engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2022. He is currently working toward the M.S. degree at School of Automation, Chongqing University of Posts and Telecommunications. His research interests include demand-side management, and load pattern clustering involved in the fields of vehicle-grid integration, microgrid and new energy power system.

Yiwei Ma received the M.S. degree in control engineering (2007) and the Ph.D. degree in electrical engineering (2015) from South China University of Technology in China. In 2015, She joined Chongqing University of Posts and Telecommunications, where she is currently an Associate Professor with School of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China. Her research interests include optimization design, operation control, and artificial intelligence in the fields of microgrids, smart grids, vehicle-grid integration, and power internet of things.

Hao Zhong received the M.S. and Ph.D. degrees in electrical engineering from Hunan University, Changsha, China, in 2008 and 2011. In 2011, he joined China Three Gorges University, where he is currently an Associate Professor with College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, China. His research interests include power system operation and optimization techniques.

Miao Huang received the M.S. and Ph.D. degrees in electrical engineering from Chongqing University, Chongqing, China, in 2006 and 2011. In 2011, he was employed by State Grid Chongqing Electric Power Co. Electric Power Research Institute in China. In 2015, he joined Chongqing University of Posts and Telecommunications, where he is currently a Senior Engineer with School of Automation and Industrial Internet, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include power system operation and simulation.