Journal of Computer and Communications, 2021, 9, 91-111
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Hybrid K-Means-GRA-SVR Model Based on Feature Selection for Day-Ahead Prediction of Photovoltaic Power Generation

Jiemin Lin, Haiming Li*

School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China
Email: *zjxulhm@163.com

How to cite this paper: Lin, J.M. and Li, H.M. (2021) A Hybrid K-Means-GRA-SVR Model Based on Feature Selection for Day-Ahead Prediction of Photovoltaic Power Generation. *Journal of Computer and Communications*, **9**, 91-111.
https://doi.org/10.4236/jcc.2021.911007

Received: April 13, 2021
Accepted: November 14, 2021
Published: November 17, 2021

Abstract

In order to ensure that the large-scale application of photovoltaic power generation does not affect the stability of the grid, accurate photovoltaic (PV) power generation forecast is essential. A short-term PV power generation forecast method using the combination of K-means++, grey relational analysis (GRA) and support vector regression (SVR) based on feature selection (Hybrid Kmeans-GRA-SVR, HKGSVR) was proposed. The historical power data were clustered through the multi-index K-means++ algorithm and divided into ideal and non-ideal weather. The GRA algorithm was used to match the similar day and the nearest neighbor similar day of the prediction day. And selected appropriate input features for different weather types to train the SVR model. Under ideal weather, the average values of MAE, RMSE and R^2 were 0.8101, 0.9608 kW and 99.66%, respectively. And this method reduced the average training time by 77.27% compared with the standard SVR model. Under non-ideal weather conditions, the average values of MAE, RMSE and R^2 were 1.8337, 2.1379 kW and 98.47%, respectively. And this method reduced the average training time of the standard SVR model by 98.07%. The experimental results show that the prediction accuracy of the proposed model is significantly improved compared to the other five models, which verify the effectiveness of the method.

Keywords

Feature Selection, Grey Relational Analysis, K-Means++, Nearest Neighbor Similar Day, Photovoltaic Power, Support Vector Regression

1. Introduction

In the face of limited fossil energy and the need to adjust the energy structure, the exploration of renewable energy power generation technology is of great significance [1]. A study shows that the earth receives about 1.8 × 10^{11} MW of power per second from solar radiation [2]. Photovoltaic power generation is one of the most promising solar power technologies [3]. Photovoltaic energy has the advantages of cleanliness, wide distribution and abundant reserves, and has become the best substitute for industrial and residential power generation [4]. According to the 2020 report of the International Renewable Energy Agency, in the past 8 years, the global photovoltaic power generation cost has dropped by more than 70%, and the global installed capacity has reached 578.553 GW [5].

However, due to the chaotic nature of the weather system, the production of photovoltaic energy is highly random, volatile and intermittent, which may lead to grid power and voltage imbalances, and also greatly increase the difficulty of large-scale photovoltaic energy applications [6] [7]. In order to improve the power system's ability to consume photovoltaic energy, many solutions have been proposed, including energy storage optimization [8], demand response strategy [9] [10], power flow optimization [11], stand-alone microgrid [12], and PV power forecasting [13]. Considering economy and feasibility comprehensively, photovoltaic power generation forecast is one of the most promising solutions to the impact of large-scale photovoltaic energy application on the grid [14] [15].

The current photovoltaic power generation forecasting technologies have three main directions: physical methods, time series statistical methods and ensemble methods [14]. [16] proposed a partial function linear regression model to forecast the day-ahead photovoltaic power generation. The regression method has a low amount of calculation, but the prediction accuracy is relatively low. [17] proposed an ANN model based on an extreme learning machine algorithm to predict photovoltaic power generation. Artificial neural network can handle nonlinear problems and has excellent self-learning ability, so it has high prediction accuracy. However, the ANN multi-layer network structure greatly increases the complexity of the model, which makes training and optimizing the model consume a lot of computing resources and longer training time. In [18] [19] [20], the support vector machine (SVM) is used for short-term photovoltaic output forecasting. SVM can also handle non-linear problems, has excellent learning ability and does not rely heavily on prior knowledge. The training speed is fast and has the ability to prevent overfitting, with good generalization.

The ensemble method solves the limitations of a single model by mixing different models with unique functions, thereby improving the prediction performance [21]. For the prediction of photovoltaic power generation, the ensemble method that mixes various effective methods is more effective and accurate [22]. For example, the hybrid GA-SVM model [20], which performed better than the SVM model. In [23], a hybrid Kmeans-GRA-Elman model was proposed, the performance of Kmeans-GRA-Elman was better than BP neural network, Elman, GRA-BPNN and GRA-Elman.

Photovoltaic power generation has obvious seasonal and weather characteristics [24]. Weather conditions can be roughly divided into two types: the ideal weather type (sunny day), and non-ideal weather types [25]. For ideal weather, the prediction accuracy of many prediction methods is high enough [26]. It can be seen from [27] [28] that the prediction accuracy of these methods for non-ideal weather was much lower than that of ideal weather. In order to improve the prediction performance under non-ideal weather, similar algorithms have been used in many studies to extract output features under similar weather. For example, [29] proposed a prediction method based on similar days and improved BP neural network. The similarity algorithm can effectively extract the output characteristics of different weather types. Moreover, compared to directly using a large amount of historical data to train the model, the use of similar days not only saves a lot of computing resources, but also improves the prediction accuracy of the model. However, if the time interval between the similar day and the forecast day is too long, the characteristics of the photovoltaic array (surface cleanliness, module aging, conversion efficiency, etc.) have changed a lot, which will cause a large error between the predicted result and the actual value [25].

A short-term photovoltaic power generation forecast method using the combination of K-means++, grey relational analysis (GRA) and support vector regression (SVR) based on feature selection is proposed in this paper. The proposed HKGSVR (hybrid Kmeans-GRA-Support Vector Regression) forecasting model is compared with SVR, HKGLSTM (hybrid Kmeans-GRA-LSTM), HKGBP (hybrid Kmeans-GRA-Back Propagation Neural Network), HKGLR (hybrid Kmeans-GRA-Linear Regression), HKGARIMA (hybrid Kmeans-GRA-Autoregressive Integrated Moving Average), respectively, to demonstrate its superiority in predictive performance. The main contributions of this paper include:

1) A novel day-ahead PV power forecasting method utilizes SVR, clustering and similarity algorithms is proposed.

2) Clustering historical power data through multi-index K-means++ to obtain power generation modes of different weather types. Overcome the limitation of directly categorizing according to weather tags. According to the average power of each cluster, it is divided into ideal weather cluster and non-ideal weather cluster.

3) The GRA algorithm is used to match the nearest neighbor similar day, and the error caused by the long time interval between the similar day and the forecast day is reduced by using the information of the nearest neighbor similar day.

4) By analyzing the correlation between photovoltaic output power and various meteorological factors, 10 feature combinations are proposed. Select appropriate input features for ideal and non-ideal weather to further improve the prediction accuracy of photovoltaic power generation.

The remainder of this paper is organized as follows. Section 2 describes the hybrid Kmeans-GRA-SVR model. Section 3 illustrates clustering and model evaluation metrics. Section 4 introduces the experiments and result analysis. Finally, conclusions are given in Section 5.

2. Hybrid K-Means-GRA-SVR Model

2.1. K-Means++ Clustering Algorithm

K-means++ clustering algorithm is an improved version of K-means algorithm. This algorithm separates the K initial cluster centers more from each other. In this work, which is selected as the classifier due to its higher efficiency and improved robustness compared with others (e.g., standard K-means, K-medoids, Gaussian mixture models, etc.) [30]. The running process of K-means++ is as follows:

Step 1: Randomly select a sample as the first cluster center C_I;

Step 2: Calculate the probability of each sample being selected as the next cluster center:

$$\frac{D(x)^2}{\sum_{x \in X} D(x)^2} \tag{1}$$

where, $D(x)$ represents the distance between the sample and the nearest cluster center.

Then use the roulette method to select the next cluster center;

Step 3: Repeat step 2 until K cluster centers are selected;

Step 4: For each sample x_i in the datasets, calculate its distance to K cluster centers, and then put it into the class corresponding to the smallest distance cluster center;

Step 5: For each cluster, recalculate its cluster center C_i:

$$c_i = \frac{1}{c_i} \sum_{x \in c_i} x \tag{2}$$

Step 6: Repeat steps 4 and 5 until the position of the cluster center does not change.

In this part, the historical power data is directly clustered by season to obtain different power generation modes due to the diversity of weather. Moreover, the aging of the equipment itself and its own parameters will be different under different weather, it is difficult for us to accurately measure these changes. The characteristics of historical power data will integrate these changes into it. After clustering the historical power data, the centroid value of each cluster is calculated by the minimum, average and maximum of global horizontal irradiance (GHI), diffuse horizontal irradiance (DHI), relative humidity (RH) and temperature (T)(12 meteorological factor eigenvalues).

2.2. Grey Relational Analysis Algorithm

The basic idea of the grey relational analysis algorithm is to judge the correlation degree by comparing the geometric similarity between the reference sequence and several data columns. Generally, the more consistent the change tendency of the reference sequence and the comparison sequence, the higher the degree of correlation between the two variables. The flow of the GRA algorithm is as follows:

Step 1: Determine the reference sequence y and the comparison sequence *x*x_i:

$$y = \{ y(k) \mid k = 1, 2, \cdots, n \} \tag{3}$$

$$x_i = \{ x_i(k) \mid k = 1, 2, \cdots, n \}, i = 1, 2, \cdots, m \tag{4}$$

where, n and m represent the dimension of the eigenvalues and the number of comparison sequence, respectively.

Step 2: Non-dimensionalization of variables:

$$d_j^*(k) = \frac{D_j(k) - D_{av}(k)}{D_{\max}(k) - D_{\min}(k)}, k = 1, 2, \cdots, n; i = 0, 1, 2, \cdots, m; j = 1, 2, \cdots, m+1 \tag{5}$$

where, $D_j(k)$ contains reference sequence and comparison sequence, $D_{av}(k)$, $D_{min}(k)$ and $D_{max}(k)$ are the average, minimum and maximum values of each column, j represents sum of the number of reference sequence and comparison sequence.

Non-dimensionalization is used to solve the problem that the columns cannot be compared due to the different dimensions.

Step 3: Calculate correlation coefficient $\xi_i(k)$:

$$\xi_i(k) = \frac{\min\limits_i \min\limits_k \left| y(k) - x_i(k) \right| + \rho \max\limits_i \max\limits_k \left| y(k) - x_i(k) \right|}{\left| y(k) - x_i(k) \right| + \rho \max\limits_i \max\limits_k \left| y(k) - x_i(k) \right|} \tag{6}$$

where, ρ is called the resolution coefficient, here, ρ is 0.5.

Step 4: Calculate correlation degree.

Calculate the average value of the correlation coefficient at each moment (that is, each point in the curve) r_i:

$$r_i = \frac{1}{n} \sum_{k=1}^{n} \xi_i(k), k = 1, 2, \cdots, n \tag{7}$$

Step 5: Sort correlation degree.

After determining the cluster to which the prediction day belongs, the correlation between the prediction day and each sample in the cluster is calculated by GRA based on 12 meteorological factor eigenvalues, and the date with the correlation degree greater than the threshold (an appropriate correlation value that takes into account the similarity and the number of samples) is regarded as the similar days. Based on GRA global matching results: for the ideal weather, the sample with the highest correlation in the 7 days before the forecast date is set as the nearest neighbor similar day; for non-ideal weather, the sample with the highest correlation in the 30 days before the prediction is set as the nearest neighbor similar day.

2.3. Support Vector Regression

Based on the structural risk minimization theory, the support vector machine constructs a hyperplane in the feature space, thereby overcoming the local optimal problem and requiring fewer training samples [30]. When the data type is complex, support vector regression is used. For a set of data $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$,

X_i is the input variable of the sample, and Y_i is the target value. The support vector machine equation based on Vapnik theory is as follows [31]:

$$f(x) = \omega^{\mathrm{T}} \phi(x) + b \tag{8}$$

where, ω is a vector of weight coefficients, $\Phi(x)$ is the nonlinear mapping function and b denotes a bias constant.

ω and b can be obtained by the following formula:

$$\text{minimize}: \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \xi_i + \xi_i^* \tag{9}$$

$$\text{subject to} \begin{cases} y_i - \langle \omega, \phi(x_i) \rangle - b \le \varepsilon + \xi_i \\ \langle \omega, \phi(x_i) \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases} \tag{10}$$

where, ξ_i and ξ_i^* are slack variables, and C denotes the penalty variable, ε is the insensitive loss function.

By introducing Lagrangian multipliers and optimal constraints, (8) can be transformed into:

$$f(x, a_i a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*) K(x, x_i) + b \tag{11}$$

where, $K(x, x_i) = \Phi(x_i)\Phi(x_j)$ is the kernel function.

In this paper, the radial basis function (RBF) kernel is applied to construct the SVR model. The RBF kernel is presented as:

$$K(x_i, x_i) = \exp\left(-\gamma \|x_i - x_i\|^2\right) \tag{12}$$

where, γ is the kernel parameter.

2.4. The HKGSVR Model Workflow

The flow chart of the hybrid K means-GRA-LSTM model is shown in **Figure 1**, and the workflow is as follows:

Step 1: Obtain historical photovoltaic output power and meteorological factor data, and deal with the missing and abnormal data in the data set.

Step 2: Use the multi-index K-means++ algorithm to cluster historical photovoltaic power data by season, and calculate the 12 meteorological factor eigenvalues as the central value of each cluster. According to the average power of each cluster, it is divided into ideal weather cluster and non-ideal weather cluster.

Step 3: The Euclidean distance, Pearson correlation coefficient and GRA correlation between the 12 meteorological eigenvalues of the forecast day and the centroid value of each cluster are calculated to determine the cluster to which the forecast day belongs.

Step 4: Calculate the correlation between the predicted day and each sample in the matched cluster through GRA to obtain the similar days (as the training set) and the nearest neighbor similar day (as the validation set) and normalize them.

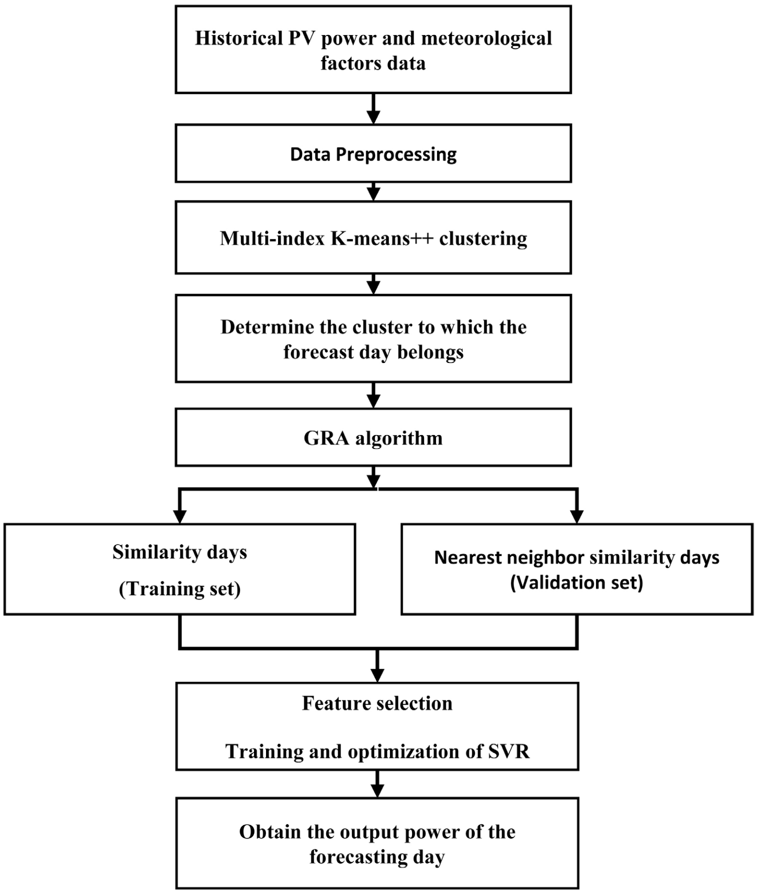
Historical PV power and meteorological factors data
Data Preprocessing
Multi-index K-means++ clustering
Determine the cluster to which the forecast day belongs
GRA algorithm
Similarity days (Training set)
Nearest neighbor similarity days (Validation set)
Feature selection
Training and optimization of SVR
Obtain the output power of the forecasting day

Figure 1. The working process of the hybrid K-means-GRA-SVR model.

Step 5: Select appropriate input features and similar days are used to train SVR. Determine the C and γ of SVR through grid search and cross-validation, and use the nearest neighbor similar day to test.

Step 6: Use the trained model to predict the prediction day.

3. Evaluation Metrics

3.1. Clustering Evaluation Metrics

If the ground truth labels are not known, evaluation must be performed using the model itself. The Silhouette Coefficient is an example of such an evaluation, the score is higher when clusters are dense and well separated. Silhouette Coefficient $S(i)$ is defined as follows [32]:

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{13}$$

where, $a(i)$ is the mean distance between a sample and all other points in the same cluster, $b(i)$ is the mean distance between a sample and all other points in the next nearest cluster. Average the Silhouette Coefficient of all points, which is the total Silhouette Coefficient of the clustering result.

Davies-Bouldin index is defined as follows [33]:

$$\text{DBI} = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left(\frac{\overline{S_i} + \overline{S_j}}{\left\|\omega_i - \omega_j\right\|_2}\right) \tag{14}$$

where, $\overline{S_i}$ is the average distance from the points in the cluster to the cluster centroid, $\left\|\omega_i - \omega_j\right\|_2$ is the distance between the centroid of cluster *i* and *j*.

The Davies-Bouldin index is lower if the model clusters have better separation.

SSE is also an effective metric, that is, the sum of squared errors of the distance between the centroid of each cluster and the points in the cluster. SSE is defined as follows:

$$\text{SSE} = \sum_{i=1}^{K}\sum dist\left(x, c_i\right)^2 \tag{15}$$

3.2. Metrics of Photovoltaic Power Forecasting Techniques

In order to evaluate the performance of the proposed method HKGSVR for photovoltaic power generation forecasting, the root mean square error (RMSE), average absolute error (MAE) and coefficient of determination (R^2) indicators were calculated. The mean absolute error can better reflect the difference between the predicted value and the true value. RMSE is used to measure the deviation between the predicted value and the actual value, so it is more sensitive to outliers (that is, if the predicted value of a point is very different from the true value, the RMSE of the curve will be very large). R^2 is used to test the fit of the predicted value to the true value, and is generally used to evaluate the prediction performance of the model. They are defined as follows [14].

1) The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} P_{fi} - P_{ai}} \tag{16}$$

where, P_{ai} and P_{fi} are the actual and predicted value at *i* hour. *N* refers to the number of hours a sample contains.

2) The MAE is expressed as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|P_{fi} - P_{ai}\right| \tag{17}$$

3) The R^2 is given as:

$$\text{R}^2 = \frac{\left(N\sum_{i=1}^{N} P_{fi}P_{ai} - \sum_{i=1}^{N} P_{fi}\sum_{i=1}^{N} P_{ai}\right)^2}{\left(N\sum_{i=1}^{N} P_{fi}^2 - \left(\sum_{i=1}^{N} P_{fi}\right)^2\right)\left(N\sum_{i=1}^{N} P_{ai}^2 - \left(\sum_{i=1}^{N} P_{ai}\right)^2\right)} \tag{18}$$

4. Experimental Analysis

4.1. Data

In this paper, the general datasets on the DKASC (Desert Knowledge Australia

Solar Center) website are used for related experiments. The photovoltaic array is composed of 22 polycrystalline silicon photovoltaic panels with a rated power of 265 W, whose total rated power is 5.83 kW. The photovoltaic array is located at the Desert Knowledge Precinct in Alice Springs, a town in the Northern Territory that enjoys one of the country's highest solar resources in an arid desert environment. The configuration information of the photovoltaic array is shown in Table 1. Meteorology (global horizontal irradiance, diffuse horizontal irradiance, relative humidity and temperature) and historical power data of PV arrays from March 1, 2018 to February 29, 2020 were used in the experiment. The experiment uses data with an interval of 1 hour from 7:00 to 18:00 every day.

4.2. Number of Clusters and Weather Division

In order to obtain the appropriate number of clusters for each season, SSE, DBI and Silhouette Coefficient (S) are used for evaluation. Taking autumn as an example, the experimental results are shown in Figure 2.

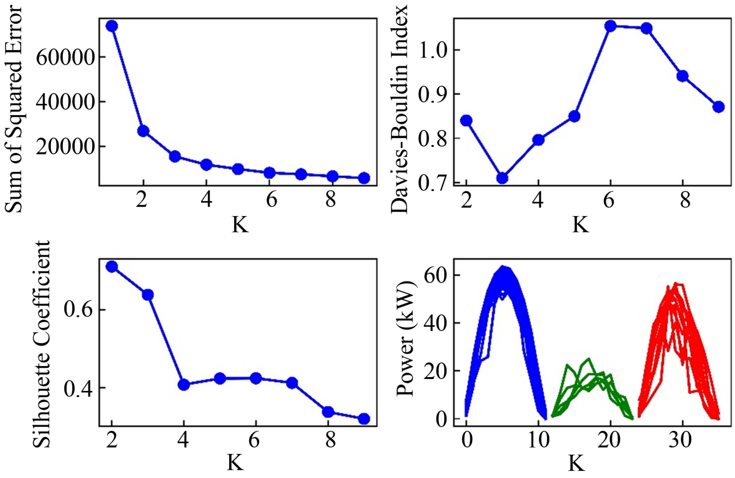

Figure 2. Evaluation of clustering results of historical power data in autumn.

Table 1. The configuration information of photovoltaic array.

Item	Information
Array Rating	5.83 kW
Panel Rating	265 W
Number Of Panels	22
Panel Type	HSL 60 S
Array Area	36.74
Inverter Size/Type	SMA SMC 6000A
Array Tilt/Azimuth	Tilt = 20, Azimuth = 0 (Solar North)
Nominal working temperature	45 ± 3 Celsius
Temperature coefficient of power	−0.41%/Celsius

It can be seen from Figure 2 that SSE decreases as the number of clusters increases. When the number of clusters is 3, the downward trend begins to slow down. DBI has the best performance when the value of K is 3. When the value of K is 2 and 3, the value of S is 0.71 and 0.64, respectively. Then, as the value of K increases, the value of S drops sharply. So the value of K is chosen between 2 and 3. When K = 2, the blue cluster and the red cluster merge into one cluster. However, the blue clusters are mostly smooth arcs, while the red clusters are mostly polylines. Therefore, the value of K is chosen to be 3. The blue cluster is selected as the ideal weather cluster (most of the curves are smooth and the average power is larger in the cluster), and the green and red clusters are non-ideal weather clusters (most of them are broken lines in the clusters, and the average power is small, the average power of the green cluster is 123.09 kW, and the average power of the red cluster is 301.90 kW).

The evaluation of clustering results in each season is shown in Table 2. In order to prevent local optima or other abnormal situations, 100 rounds of experiments were carried out. Considering all indicators and clustering results comprehensively, the number of clusters in spring is 3, the number of clusters in summer is 2, the number of clusters in autumn is 3, and the number of clusters in winter is 3.

The clustering results of each season are divided into ideal weather clusters and non-ideal weather clusters by comparing the average power and geometric shape (arc and polyline) of each cluster. The average power of each cluster in the four seasons is shown in Table 3. The ideal weather clusters are mostly smooth

Table 2. Cluster evaluation metrics for each season.

Metrics	Spring	Summer	Autumn	Winter
SSE K = 2	44,578.41	20,278.67	26,813.26	4218.80
K = 3	32,045.12	14,899.18	15,383.50	2942.09
K = 4	24,354.12	11,971.90	11,657.92	2473.00
DBI K = 2	0.6723	0.9740	0.8402	0.7100
K = 3	0.7846	0.9440	0.7098	0.6375
K = 4	1.0187	1.0247	0.7962	0.9049
S K = 2	0.6383	0.6156	0.7119	0.4861
K = 3	0.5893	0.5866	0.6390	0.5074
K = 4	0.5231	0.5705	0.4073	0.5018

Table 3. Average power of each cluster in four seasons (kW).

Cluster number	Spring	Summer	Autumn	Winter
Cluster1	437.41	440.65	423.34	339.07
Cluster2	281.24	331.46	123.09	376.46
Cluster3	111.08		301.90	419.81

arcs, and the average power is relatively large. The non-ideal weather clusters are mostly broken lines, and the average power is small. Therefore, spring cluster 1, summer cluster 1, autumn cluster 1 and winter cluster 2 and 3 are divided into ideal weather clusters, and the rest are non-ideal weather clusters.

4.3. Selection of Similar Day Threshold and Nearest Similar Day

The similar days are obtained by calculating the GRA correlation between the predicted days and the samples in the matching clusters. In order to improve the prediction accuracy while reducing the computational cost and speeding up the training speed of the model, it is necessary to select an appropriate correlation threshold. A higher correlation threshold can improve the prediction accuracy, but too few training samples may cause overfitting. After comprehensive consideration, the similar day correlation threshold of each forecast day, the nearest neighbor similar day and its correlation degree are shown in Table 4 and Table 5. It can be seen that the nearest neighbor similar days of ideal weather are mostly adjacent days, while the time intervals of nearest neighbor similar days of non-ideal weather are relatively long.

4.4. Design of SVR Model

This part is mainly to explore the optimal C and γ of SVR, which are usually related to the characteristics of power generation in different seasons. Grid search and cross-validation are used to find the optimal number of C and γ for SVR. This experiment uses PyCharm (python3.6) to train and optimize the SVR model on a Win 7 System personal computer with Intel core i5-3230CPU, 2.60 GHz processor and 4 GB RAM.

Table 4. Forecasting day, similar day correlation threshold, nearest neighbor similarity day and correlation under ideal weather.

Item	Spring	Summer	Autumn	Winter
Forecasting day	09/15/2019	02/09/2020	04/18/2019	08/17/2019
similar days threshold	0.90	0.88	0.92	0.86
nearest neighbor similarity day	09/14/2019	02/08/2020	04/15/2019	08/16/2019
Correlation degree	0.9295	0.9452	0.9913	0.9579

Table 5. Forecasting day, similar day correlation threshold, nearest neighbor similarity day and correlation under non-ideal weather.

Item	Spring	Summer	Autumn	Winter
Forecasting day	11/02/2019	02/25/2020	04/28/2019	08/07/2019
similar days threshold	0.92	0.87	0.91	0.91
nearest neighbor similarity day	10/18/2019	02/05/2020	04/17/2019	07/12/2019
Correlation degree	0.9295	0.9452	0.9913	0.9579

It can be seen from Table 6 and Table 7 that the optimal training time of the ideal weather model for each season is 2.0342, 1.9506, 2.3272, 0.6826 s, and the average time is 1.74865 s. And the optimal training time of the model for each season of non-ideal weather is 0.2490, 0.2400, 0.2240, 0.2060 s, and the average time is 0.22975 s. The number of similar days matched has a greater impact on the training optimization time. Comparing Table 6 and Table 7, it can be found that because the data complexity of non-ideal weather is higher than that of ideal weather, the C of non-ideal weather is generally larger than that of ideal weather, and the γ of non-ideal weather is generally smaller than that of ideal weather.

4.5. Feature Selection

In order to select the appropriate input feature, GRA and Pearson correlation analysis is performed between the power generation and various meteorological factors. The historical power and meteorological data for the year from March 1, 2018 to February 28, 2019 are used for analysis. The result is shown in Figure 3. The definition of Pearson correlation coefficient is as follows:

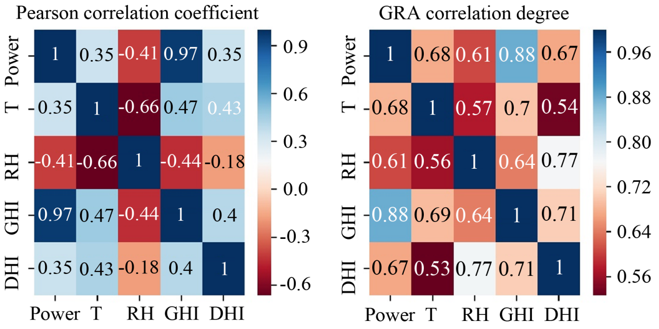

Figure 3. The correlation between photovoltaic power generation and various meteorological factors.

Table 6. The optimal SVR structure for each season under ideal weather.

Structure	Spring	Summer	Autumn	Winter
Training data	(564, 2)	(408, 2)	(612, 2)	(156, 2)
C	100,000.00	100,000.0	100,000.00	1.00
γ	0.0001	0.0001	0.002	1.00
Time (s)	2.0342	1.9506	2.3272	0.6826

Table 7. The optimal SVR structure for each season under non-ideal weather.

Structure	Spring	Summer	Autumn	Winter
Training data	(108, 3)	(180, 3)	(144, 3)	(60, 3)
C	1,000,000.0	10,000.0	1,000,000.0	1,000,000.0
γ	0.000001	0.0001	0.00001	0.000001
Time (s)	0.2490	0.2400	0.2240	0.2060

$$\rho_{X,Y} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \tag{19}$$

where, X and Y are meteorological factors and photovoltaic output power respectively, and N is the number of sampling points per day.

It can be seen from Figure 5 that the Pearson correlation coefficients between photovoltaic power generation and T, RH, GHI, and DHI are 0.35, −0.41, 0.97, and 0.35, respectively. GHI has the greatest impact on photovoltaic output, and there is a negative correlation between relative humidity and photovoltaic power. The GRA correlations between photovoltaic power generation and T, RH, GHI, and DHI are 0.68, 0.60, 0.88 and 0.67, respectively. GHI still has the largest impact on photovoltaic output.

Based on the above analysis, the paper proposes 10 feature combinations. The prefix N represents the nearest neighbor similar day, P, G, and M respectively represent Power, GHI and meteorological factor eigenvalues. For example, NG_MG represents the nearest neighbor day GHI and predicted day meteorological factor eigenvalues and GHI.

For ideal weather, due to its high prediction accuracy, the main consideration for the selection of its input features is to select a feature combination that is easier to obtain and requires less data accuracy while ensuring sufficient prediction accuracy. Therefore, the input features of the ideal weather are power of the nearest neighbor similar day and 12 meteorological factor eigenvalues of the forecast day (NP_M). For non-ideal weather, the main goal of feature selection is to improve the prediction accuracy. Tables 8-10 show the evaluation of 10 feature combinations of non-ideal weather in each season. The best performance of the evaluation indicators in the table is bolded.

It can be seen from Tables 8-10 that the MAE of NG_MG feature combination in each season is 1.3733, 2.0817, 1.6475, and 2.2323 kW, respectively. And

Table 8. MAE (in kW) evaluation of 10 feature combinations.

Feature	Spring	Summer	Autumn	Winter	Average
NG_MG	**1.3733**	2.0817	**1.6475**	2.2323	**1.8337**
NMP_M	6.9802	5.8165	9.2623	3.1488	6.3020
NP_M	5.2198	6.7131	10.2075	2.7698	6.2275
NG_M	3.6054	6.3498	8.3381	3.6163	5.4774
NM_M	12.1144	12.9347	15.7812	16.1616	14.2480
NMPG_MG	2.7202	**2.0359**	3.9729	2.0185	2.6869
NPG_MG	2.9866	2.3195	4.2225	2.1003	2.9072
NP_MG	2.5947	2.2328	4.1497	2.4596	2.8592
NG_G	2.8937	2.2657	3.3403	1.6866	2.5466
MG	1.6100	3.3220	3.7172	**1.6253**	2.5686

Table 9. RMSE (in kW) evaluation of 10 feature combinations.

Feature	Spring	Summer	Autumn	Winter	Average
NG_MG	**1.4699**	2.6625	**1.8700**	2.5492	**2.1379**
NMP_M	7.2800	6.3683	10.2603	3.8242	6.9332
NP_M	6.0625	7.8041	10.6872	3.2925	6.9616
NG_M	4.2990	8.1207	10.2338	4.5935	6.8118
NW_M	12.8652	14.9801	17.5154	18.6947	16.0138
NMPG_MG	2.9276	**2.4981**	4.3010	2.3739	3.0251
NPG_MG	3.1398	2.7112	4.4120	2.4219	3.1712
NP_MG	3.0082	2.6862	4.2862	2.8474	3.2070
NG_G	3.1922	2.7433	3.4749	**1.9734**	2.8459
MG	1.7974	3.7430	3.9093	2.0774	2.8818

Table 10. R^2 evaluation of 10 feature combinations.

Feature	Spring	Summer	Autumn	Winter	Average
NG_MG	**0.9918**	0.9731	**0.9926**	0.9814	**0.9847**
NMP_M	0.7987	0.8464	0.7770	0.9582	0.8451
NP_M	0.8604	0.7693	0.7580	0.9690	0.8392
NG_M	0.9298	0.7502	0.7781	0.9396	0.8494
NM_M	0.3714	0.1499	0.3501	−1e−5	0.2178
NMPG_MG	0.9675	**0.9764**	0.9608	0.9839	0.9721
NPG_MG	0.9626	0.9722	0.9588	0.9832	0.9692
NP_MG	0.9656	0.9727	0.9611	0.9768	0.9690
NG_G	0.9613	0.9715	0.9744	**0.9889**	0.9740
MG	0.9877	0.9469	0.9676	0.9877	0.9725

the average MAE is 1.8337 kW, which is the smallest among all feature combinations. The RMSE of the NG_MG feature combination in each season was 1.4699, 2.6625, 1.8700, 2.5492 kW. And the average RMSE was 2.1379 kW, the best performance among all the feature combinations. The R^2 of each season of NG_MG feature combination is 99.18%, 97.31%, 99.26% and 98.14%. And the average R^2 is 98.47%, which is the highest degree of fit among all feature combinations. From the perspective of comprehensive performance, NG_MG feature combination has higher prediction accuracy and robustness, so NG_MG is selected as the input feature of non-ideal weather.

4.6. Forecasting Results and Discussion

Figure 4 shows the prediction results of the models in each season under ideal

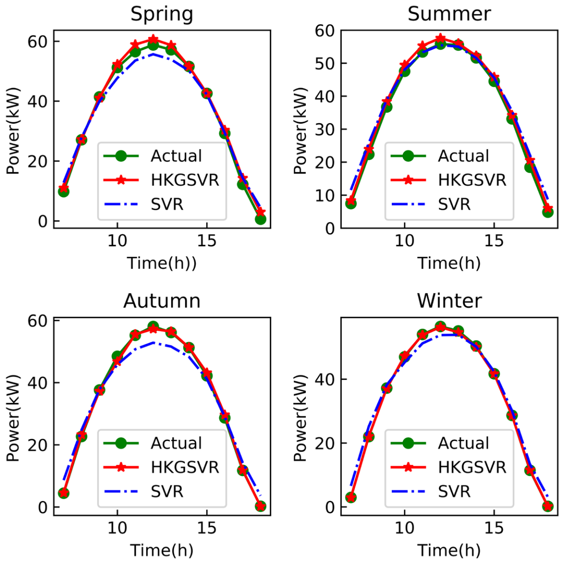

Figure 4. Forecast results of each season under ideal weather.

weather. The feature combination is NP_M. The average value of R^2 is 0.9966. The MAE are 1.4521, 1.4661, 0.7120, and 0.2132 kW respectively. The average value of RMSE is 0.9608 kW. The proposed model's MAE enhancement with respect to the SVR model is 45.51%, 25.37%, 81.21%, 91.72%, respectively. The presented model's RMSE improvement relative to the SVR model is 41.63%, 38.80%, 77.97%, 90.37%, respectively. The average R^2 of the proposed model is also better than the SVR model. And this method reduced the average training time by 77.27% compared with the standard SVR model.

Under non-ideal weather, the forecast results of the HKGSVR model and the other five forecast models for the four seasons are shown in Figures 5-8. It can be seen from the figure that the HKGSVR model has the highest degree of fit in each season.

From the spring forecast results in Figure 5, it can be seen that HKGSVR has the highest degree of fit, HKGLSTM is the second, and the SVR trend is more consistent with the predicted day. From the summer forecast results in Figure 6, it can be seen that each point of HKGSVR has a high degree of fit, and HKGLSTM has a good performance except for one point that has a lower degree of fit. Through the observation of the autumn forecast results in Figure 7, HKGSVR still performs best, and the trends of SVR and HKGARIMA are more consistent with the forecasted day. From the winter forecast results in Figure 8, it is found that both HKGSVR and SVR perform better, and HKGLSTM and HKGBP have poor performance due to less training data.

According to the MAE value of each model given in Figure 9, the HKGSVR's average value of MAE is 1.8337 kW, which is the minimum value of all models. The presented model's average MAE improvement relative to the compared five models (SVR, HKGLSTM, HKGBP, HKGLR, HKGARIMA) are 26.61%, 41.80%, 58.18%, 57.36%, 52.98%, respectively.

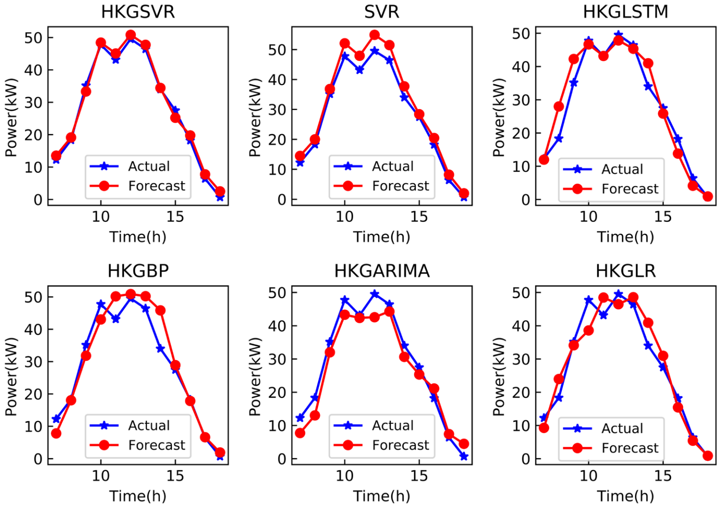

Figure 5. 6 model prediction results for a spring day.

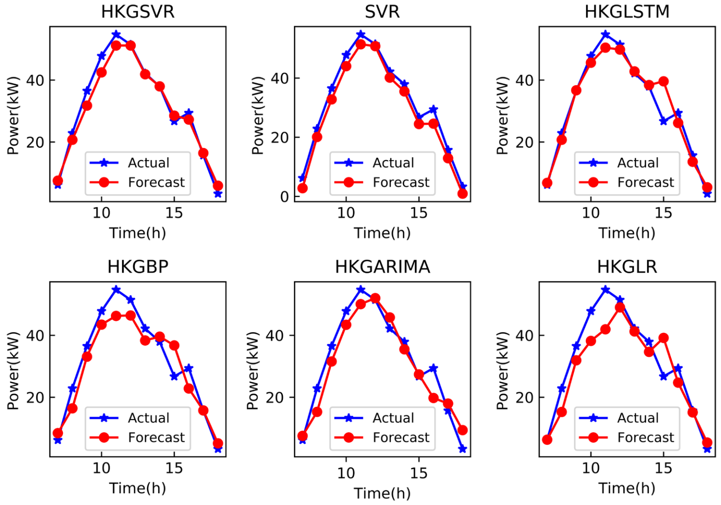

Figure 6. 6 model prediction results for a summer day.

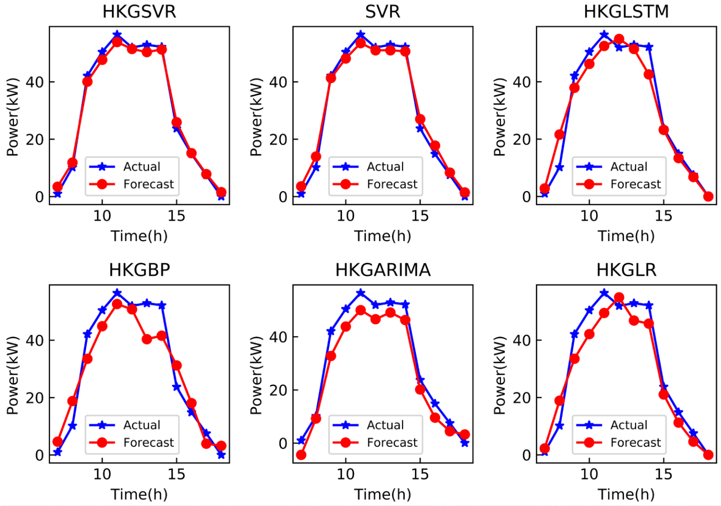

Figure 7. 6 model prediction results for an autumn day.

Observation in Figure 10 finds that the average RMSE of the HKGSVR model is 2.1379 kW, which is the best value among all models. The proposed model's average RMSE enhancement with respect to the compared five models is 24.13%, 52.12%, 59.87%, 61.37%, 52.66%, respectively.

Comparing the R^2 values of the models shown in Table 11 shows that the average R^2 of the proposed model is 0.9847, which is better than other models.

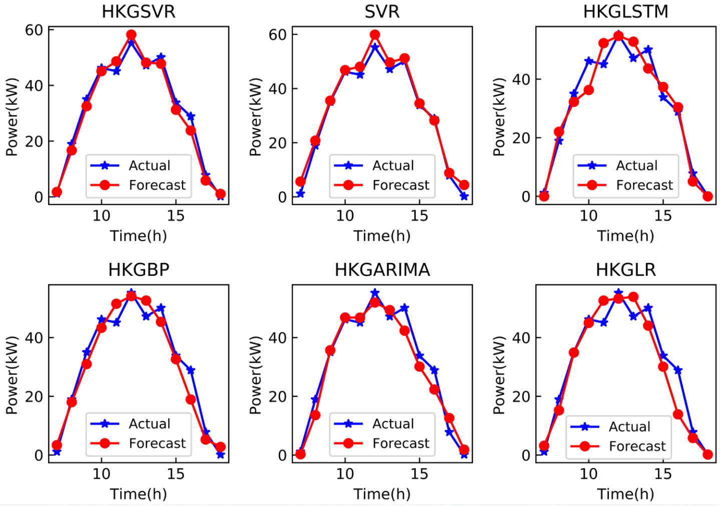

Figure 8. 6 model prediction results for a winter day.

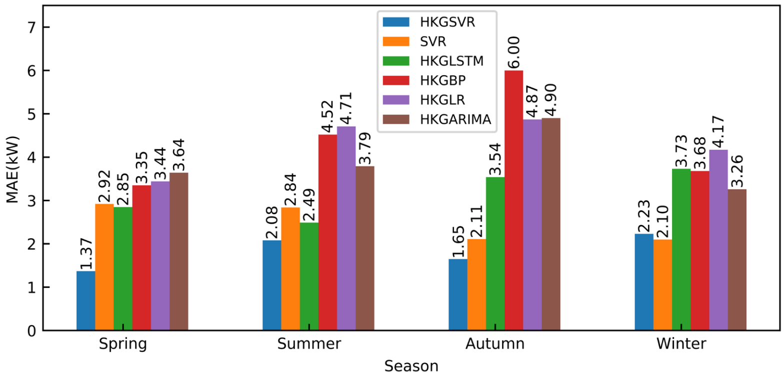

Figure 9. MAE evaluation under non-ideal weather.

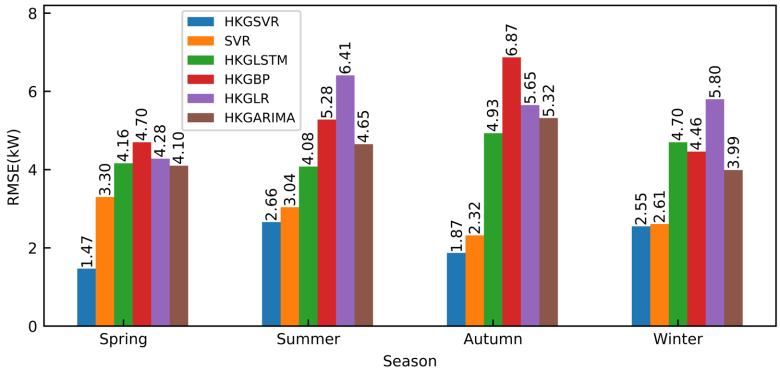

Figure 10. RMSE evaluation under non-ideal weather.

Table 11. Daily R^2 comparison results of 6 models.

Models	Spring	Summer	Autumn	Winter	Average
HKGSVR	**0.9918**	**0.9731**	**0.9926**	**0.9814**	**0.9847**
SVR	0.9586	0.9650	0.9886	0.9804	0.9732
HKGLSTM	0.9406	0.9468	0.9486	0.9369	0.9432
HKGBP	0.9242	0.9109	0.9001	0.9431	0.9196
HKGARIMA	0.9424	0.9308	0.9400	0.9544	0.9419
HKGLR	0.9374	0.8685	0.9323	0.9038	0.9105

Under non-ideal weather, compared to the standard SVR (the average training optimization time is 11.5389 s), the average training optimization time of the HKGSVR model is 0.2225 s, which is 98.07% less than the standard SVR.

5. Conclusion

A hybrid day-ahead photovoltaic power generation prediction model based on K-means++, GRA and SVR is proposed. The historical power data are clustered by multi-index K-means++, and divided into ideal weather clusters and non-ideal weather clusters according to the average power of each cluster. And it chooses the appropriate feature combination for different weather, different feature combination which has a greater impact on the model performance. It also uses GRA to match the similar day and the nearest neighbor similar day of the prediction day to improve the prediction accuracy and reduce the training optimization time of the model. Compared with the standard SVR model under ideal weather, the HKGSVR model not only improves the prediction accuracy but also greatly shortens the training time. Under non-ideal weather, the average MAE, RMSE and R^2 of the proposed model are 1.8337, 2.1379 kW and 98.47%, respectively, which have better performance than the other five models. And the training time is 0.2225 s, which is 98.07% less than the standard SVR. When there are more accurate forecasting weather information and less training data, the HKGSVR model has higher forecast accuracy. In general, HKGSVR has higher accuracy, shorter training time and better generalization performance. Therefore, the model can be used to predict the daily power generation of photovoltaic power plants. However, in terms of model structure optimization, this paper uses grid search, so there is room for improvement in the optimization speed and search range, which will be the direction of further research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Tsai, Y.C., Chan, Y.K., Ko, Y.K., *et al.* (2018) Integrated Operation of Renewable

Energy Sources and Water Resources. *Energy Conversion and Management*, **160**, 439-454. https://doi.org/10.1016/j.enconman.2018.01.062

[2] Shah, A.S.B.M., Yokoyama, H. and Kakimoto, N. (2015) High-Precision Forecasting Model of Solar Irradiance Based on Grid Point Value Data Analysis for an Efficient Photovoltaic System. *IEEE Transactions on Sustainable Energy*, **6**, 474-481. https://doi.org/10.1109/TSTE.2014.2383398

[3] Wang, F., Zhen, Z., Liu, C., *et al.* (2018) Image Phase Shift Invariance Based Cloud Motion Displacement Vector Calculation Method for Ultra-Short-Term Solar PV Power Forecasting. *Energy Conversion and Management*, **157**, 123-135. https://doi.org/10.1016/j.enconman.2017.11.080

[4] Koo, C., Hong, T., Jeong, K., *et al.* (2017) Development of the Smart Photovoltaic System Blind and Its Impact on Net-Zero Energy Solar Buildings Using Technical-Economic-Political Analyses. *Energy*, **124**, 382-396. https://doi.org/10.1016/j.energy.2017.02.088

[5] IRENA (2020) Renewable Energy Statistics 2020. The International Renewable Energy Agency, Abu Dhabi.

[6] Jiang, H. and Dong, Y. (2017) Forecast of Hourly Global Horizontal Irradiance Based on Structured Kernel Support Vector Machine: A Case Study of Tibet Area in China. *Energy Conversion and Management*, **142**, 307-321. https://doi.org/10.1016/j.enconman.2017.03.054

[7] Köhler, C., Steiner, A., Saint-Drenan, *et al.* (2017) Critical Weather Situations for Renewable Energies—Part B: Low Stratus Risk for Solar Power. *Renewable Energy*, **101**, 794-803. https://doi.org/10.1016/j.renene.2016.09.002

[8] Wang, F., Zhou, L., Ren, H., *et al.* (2018) Multi-Objective Optimization Model of Source-Load-Storage Synergetic Dispatch for a Building Energy Management System Based on TOU Price Demand Response. *IEEE Transactions on Industry Applications*, **54**, 1017-1028. https://doi.org/10.1109/TIA.2017.2781639

[9] Talari, S., Shafie-khah, M., Wang, F., *et al.* (2019) Optimal Scheduling of Demand Response in Pre-Emptive Markets Based on Stochastic Bilevel Programming Method. *IEEE Transactions on Industrial Electronics*, **66**, 1453-1464. https://doi.org/10.1109/TIE.2017.2786288

[10] Chen, Q., Wang, F., Hodge, B.M., *et al.* (2017) Dynamic Price Vector Formation Model-Based Automatic Demand Response Strategy for PV-Assisted EV Charging Stations. *IEEE Transactions on Smart Grid*, **8**, 2903-2915. https://doi.org/10.1109/TSG.2017.2693121

[11] Biswas, P.P., Suganthan, P.N. and Amaratunga, G.A.J. (2017) Optimal Power Flow Solutions Incorporating Stochastic Wind and Solar Power. *Energy Conversion and Management*, **148**, 1194-1207. https://doi.org/10.1016/j.enconman.2017.06.071

[12] Wang, F., Zhou, L., Wang, B., *et al.* (2017) Modified Chaos Particle Swarm Optimization-Based Optimized Operation Model for Stand-Alone CCHP Microgrid. *Applied Sciences*, **7**, 754. https://doi.org/10.3390/app7080754

[13] Zhen, Z., Xuan, Z., Wang, F., *et al.* (2019) Image Phase Shift Invariance Based Multi-Transform-Fusion Method for Cloud Motion Displacement Calculation Using Sky Images. *Energy Conversion and Management*, **197**, 11853. https://doi.org/10.1016/j.enconman.2019.111853

[14] Sobri, S., Koohi-Kamali, S. and Rahim, N.A. (2018) Solar Photovoltaic Generation Forecasting Methods: A Review. *Energy Conversion and Management*, **156**, 459-497. https://doi.org/10.1016/j.enconman.2017.11.019

[15] Wang, F., Li, K., Liu, C., *et al.* (2018) Synchronous Pattern Matching Principle-

Based Residential Demand Response Baseline Estimation: Mechanism Analysis and Approach Description. *IEEE Transactions on Smart Grid*, **9**, 6972-6985. https://doi.org/10.1109/TSG.2018.2824842

[16] Wang, G., Su, Y. and Shu, L. (2016) One-Day-Ahead Daily Power Forecasting of Photovoltaic Systems Based on Partial Functional Linear Regression Models. *Renewable Energy*, **96**, 469-478. https://doi.org/10.1016/j.renene.2016.04.089

[17] Teo, T.T., Logenthiran, T., Woo, W.L., *et al.* (2016) Forecasting of Photovoltaic Power Using Extreme Learning Machine. *IEEE Region* 10 *Conference*, Singapore, 455-458. https://doi.org/10.1109/TENCON.2016.7848040

[18] Felice, M.D., Petitta, M. and Ruti, P.M. (2015) Short Term Predictability of Photovoltaic Production over Italy. *Renewable Energy*, **80**, 197-204. https://doi.org/10.1016/j.renene.2015.02.010

[19] Zeng, J.W. and Qiao, W. (2013) Short-Term Solar Power Prediction Using a Support Vector Machine. *Renewable Energy*, **52**, 118-127. https://doi.org/10.1016/j.renene.2012.10.009

[20] Wang, J., Ran, R., Song, Z. and Sun, J. (2017) Short-Term Photovoltaic Power Generation Forecasting Based on Environmental Factors and GA-SVM. *Journal of Electrical Engineering & Technology*, **12**, 64-71. https://doi.org/10.5370/JEET.2017.12.1.064

[21] Leva, S., Dolara, A., *et al.* (2017) Analysis and Validation of 24 Hours Ahead Neural Network Forecasting of Photovoltaic Output Power. *Mathematics and Computers in Simulation*, **131**, 88-100. https://doi.org/10.1016/j.matcom.2015.05.010

[22] Du, P., Wang, J., Yang, W., *et al.* (2018) Multi-Step Ahead Forecasting in Electrical Power System Using a Hybrid Forecasting System. *Renewable Energy*, **122**, 533-550. https://doi.org/10.1016/j.renene.2018.01.113

[23] Lin, P., Peng, Z., Lai, Y., *et al.* (2018) Short-Term Power Prediction for Photovoltaic Power Plants Using a Hybrid Improved Kmeans-GRA-Elman Model Based on Multivariate Meteorological Factors and Historical Power Datasets. *Energy Conversion and Management*, **177**, 704-717. https://doi.org/10.1016/j.enconman.2018.10.015

[24] Han, Y.T., Wang, N.B., Ma, M., *et al.* (2019) A PV Power Interval Forecasting Based on Seasonal Model and Nonparametric Estimation Algorithm. *Solar Energy*, **184**, 515-526. https://doi.org/10.1016/j.solener.2019.04.025

[25] Gao, M., Li, J., Hong, F. and Long, D. (2019) Day-Ahead Power Forecasting in a Large-Scale Photovoltaic Plant Based on Weather Classification Using LSTM. *Energy*, **187**, Article ID: 115838. https://doi.org/10.1016/j.energy.2019.07.168

[26] Benali, L., Notton, G. and Fouilloy, A. (2019) Solar Radiation Forecasting Using Artificial Neural Network and Random Forest Methods: Application to Normal Beam, Horizontal Diffuse and Global Components. *Renewable Energy*, **132**, 871-884. https://doi.org/10.1016/j.renene.2018.08.044

[27] Ding, M., Wang, L. and Bi, R. (2011) An ANN-Based Approach for Forecasting the Power Output of Photovoltaic System. *Procedia Environmental Sciences*, **11**, 1308-1315. https://doi.org/10.1016/j.proenv.2011.12.196

[28] Shi, J., Lee, W.J., Liu, Y., *et al.* (2015) Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. *IEEE Transactions on Industry Applications*, **48**, 1064-1069. https://doi.org/10.1109/TIA.2012.2190816

[29] Li, P., Zang, C. and Wang, K. (2013) Photovoltaic Generation Prediction Based on Similar Days and Neural Network. *Renewable Energy Resources*, **31**, 1-9.

[30] Luo, X., Zhu, X. and Lim, E.G. (2019) A Parametric Bootstrap Algorithm for Cluster

Number Determination of Load Pattern Categorization. *Energy*, **180**, 50-60. https://doi.org/10.1016/j.energy.2019.04.089

[31] Vapnik, V.N. (1998) Statistical Learning Theory. Wiley-Interscience, New York.

[32] Rousseeuw, P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[33] Davies, D.L. and Bouldin, D.W. (1979) A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224-227. https://doi.org/10.1109/TPAMI.1979.4766909