Energy 187 (2019) 115940

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Random forest solar power forecast based on classification optimization

Da Liu [a, c], Kun Sun [a, b, *]

[a] Economics and Management School, North China Electric Power University, Changping District, Beijing, 102206, China
[b] Beijing Key Laboratory of New Energy and Low-Carbon Development (North China Electric Power University), Changping Beijing, 102206, China
[c] Institute of Smart Energy, North China Electric Power University, Changping District, Beijing, 102206, China

ARTICLE INFO

Article history:
Received 24 April 2019
Received in revised form
15 July 2019
Accepted 12 August 2019
Available online 12 August 2019

Keywords:
Principal component analysis
Random forest
K-means clustering
Differential evolution grey wolf optimize

ABSTRACT

With the rapid development of the photovoltaic industry, the share of photovoltaic power generation in the power trading market is growing. The intermittent and uncontrollable characteristics of photovoltaic power generation have a huge impact on the stability of the power system. To reduce the occurrence of such conditions, it is necessary to improve the prediction accuracy of photovoltaic power generation. However, in the traditional modeling process, the accuracy of the model is often poor due to excessive noise in the original data or improper parameter adjustment. In this paper, Principal Component Analysis and K-means clustering algorithm combined with random forest algorithm optimized by Differential Evolution Grey Wolf Optimizer are used to model the photovoltaic power generation in three regions. Principal Component Analysis and K-means clustering are used to obtain the hourly point features similar to the predicted time points, and then the input data is filtered to reduce the noise data interference. At the same time, the popular optimization algorithm quickly selects the Random Forest parameters, which greatly avoids the artificial filtering factors and causes the error to be generated. Through the establishment of comparative experiments, it is found that the recommended model has higher prediction accuracy and robustness.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

At present, the world's energy structure is still based on non-renewable fossil energy, followed by increasingly prominent worldwide energy depletion and environmental pollution. Under this circumstance, optimizing the global energy structure and increasing the proportion of non-fossils have become an important step in building a resource-saving and environment-friendly society [1]. In order to alleviate the energy crisis and respond to global climate change, countries around the world actively carry out emission reduction actions and establish non-fossil energy development targets, which has laid the foundation for the rapid development of photovoltaics in recent years [2–4]. As a large number of renewable-energy sources are connected to the grid, its instability gradually emerges, which has a huge impact on grid reliability, including voltage fluctuations, local power quality fluctuations and stability issues [5,6]. These problems require power generation companies to predict accurately their own power generation to improve the stability of the power grid and reduce the occurrence of light [7–9].

Compared with other time-scale PV predictions, hourly forecasts are used as the basis for real-time grid scheduling due to shorter and closer prediction times. Therefore, in terms of accuracy, the ability to predict the fluctuation of photovoltaic power generation in hours under different weather conditions is more demanding. According to the input data source, the current stage of photovoltaic power generation prediction can be divided into two categories, one is the use of current or delayed sequence of photovoltaic power generation output, and the other is the use of local measurements, ground-based cloud maps, satellite imagery and numerical forecast data, etc [10]. Based on machine learning theory, using a large amount of historical data, combined with different domain methods, scholars around the world have established many hybrid models for mining the volatility of photovoltaic data. For example, when performing numerical analysis, most of the similar day data are selected for analysis, and next combined with the corresponding model for modeling and prediction [11,12].

* Corresponding author. Economics and Management School, North China Electric Power University, Changping District, Beijing, 102206, China.
E-mail address: 1172206255@ncepu.edu.cn (K. Sun).

https://doi.org/10.1016/j.energy.2019.115940
0360-5442/© 2019 Elsevier Ltd. All rights reserved.

Abbreviations

PCA	Principal Component Analysis
RF	Random Forest
DE	Differential Evolution
GWO	Grey Wolf Optimizer
HGWO	Differential Evolution Grey Wolf Optimize
SVM	Support Vector Machine
ARIMA	Autoregressive Integrated Moving Average model
ELM	Extreme Learning Machines
RBNN	Radial Basis Neural Networks
WNN	Wavelet Neural Networks
PSO	Particle Swarm Optimization
FA	Firefly Algorithm
ACA	Ant Colony Algorithm
ANN	Artificial Neural Network
NWP	Numerical Weather Prediction

Otherwise, K-means clustering and grey correlation analysis are used to analyze the input variables, and then Elman model is used for prediction [13]. After using the self-organizing map to divide the whole input space into different regions, the Particle Swarm Optimization (PSO) algorithm is used to optimize the Support Vector Machine (SVM) model to predict the solar irradiance [14,15]. Other literatures include the establishment of Autoregressive Integrated Moving Average model (ARIMA) [16,17], Extreme Learning Machines (ELM) [18,19], Radial Basis Neural Networks (RBNN) [20], Wavelet Neural Networks (WNN) [21], Gaussian process regression [22] and other models [23,24] for photovoltaic power generation prediction. When the model parameters are involved in the modeling process, the literature generally uses genetic algorithms such as genetic algorithm, PSO algorithm and Firefly Algorithm (FA) to optimize the model parameters, and achieves good prediction results in the field of photovoltaic power prediction [25]. For instance, the literature [26] uses the Firefly Algorithm to optimize the best trees and leaf trees in Random Forest (RF), and has made breakthroughs in prediction accuracy and speed. Based on Random Forest, a multi-stage model optimized by Ant Colony Algorithm (ACA) was used to predict the monthly solar radiation in three locations in Australia, and the feasibility of the method was verified [27,28].

However, mixed models have their own limitations more or less. These limitations boil down to two points, how to make model parameters and how to choose input variables [29]. There is a traditional misunderstanding in using the optimization algorithm, that is, it is easy to fall into the local best advantage and then cannot jump out of the vicinity of the local best advantage, resulting in unsatisfactory results [30–32]. In comparison, the Grey Wolf Optimizer (GWO) algorithm is faster and more optimized than similar algorithms [33]. Introducing the Differential Evolution algorithm (DE) in the GWO algorithm can effectively help the model [34]. The differential evolution Grey Wolf Optimizer (HGWO) is used in the literature to solve the economic scheduling problems in nonlinear, non-convex and discontinuous. The experiment proves that the effect is higher than other similar algorithms [35].

In addition to parameter selection, the selection of input samples also affects the accuracy of the final model [36,37]. A good choice of input variables determines the lower limit of the modeling effect, and parameter selection determines the upper limit of the model accuracy [38]. In the literature [39,40], Principal Component Analysis (PCA) is used to reduce the dimension of the original input to solve the problem of slow convergence of the algorithm, and the feasibility of the theory is proved. The literature [41,42] fusion PCA and Artificial Neural Network (ANN) predict the photovoltaic power generation, and obtain more accurate experimental results and provide practical proof of relevant theories. The literature combines PCA and Numerical Weather Prediction (NWP) models to analyze the solar irradiance of tropical islands in Singapore for one year and finds the advantages of its recommended algorithm [43].

In this paper, a method of clustering input data based on PCA and K-means algorithm is proposed, and the random forest is optimized by HGWO algorithm. In view of the high degree of data fluctuation in solar-energy prediction, the goal of denoising and dividing the original data is realized, and the fitting of the training set and the prediction of the verification set or the test set are completed as much as possible. The structure of the paper is as follows. A brief introduction to the research method and the recommended model is given in Section 2. Section 3 conducts empirical analysis, including raw data analysis, modeling analysis, results analysis, and discussion. Section 4 gives the conclusion of the article.

2. Research methods

Cluster analysis of samples maximizes the homogeneity of elements in the same class and maximizes the heterogeneity of elements between classes. Cluster analysis has been successfully applied to many fields such as graph clustering, computer vision, and clustering of non-convex spherical data, and has achieved good application results. The basis of swarm intelligence is a simple creature that follows ordinary rules. These seemingly uncomplicated individuals can often show amazing creativity when they appear in groups. Many optimization problems in the engineering field are solved by the application of swarm intelligence theory. Random Forests have been widely used in many forecasting fields due to their high tolerance to poor information and superior fitting ability. This paper combines the advantages of the three algorithms to apply these algorithm combinations to hourly PV prediction.

2.1. Principal component analysis

PCA is a widely used method of data dimensionality reduction. The main idea is to map the n-dimensional features of the original data into the k dimensional space ($n > k$), and the reconstructed k-dimensional features are orthogonal to each other, so that the primary data features are maximally preserved. The PCA algorithm flow is as follows:

1) Calculate the mean X_{mean} of the X column vector of the data set, so that $X_{new} = X - X_{mean}$.
2) Solve the covariance matrix Cov of the matrix X_{new} and the corresponding eigenvalues and eigenvectors.
3) sorting according to the size of the feature value, selecting the feature vector corresponding to the maximum k feature values of the feature value as the column vector to form the feature vector matrix W_{nxk}.

The reconstructed k-dimensional features are sorted by importance, and the most important parts are selected according to specific needs, and data compression is implemented assuming maximally maintaining the original data information. In the modeling process, the model complexity is often increased due to too many input variables of the model, and the prediction accuracy is reduced. Using PCA can effectively avoid this problem.

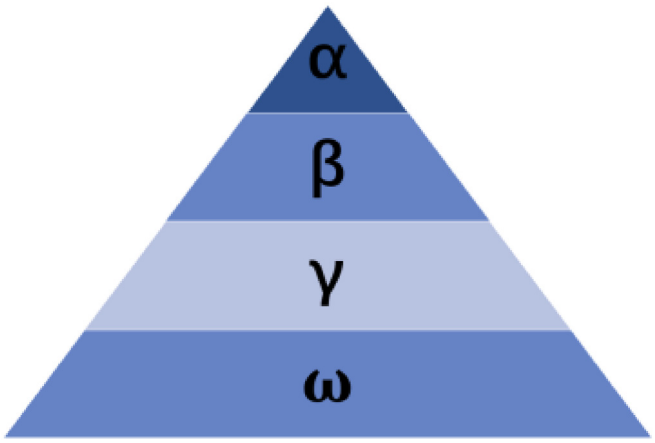

Fig. 1. Schematic diagram of grey wolf hierarchy.

2.2. K-means clustering principles

K-means is an unsupervised clustering algorithm. Although the algorithm is simple to operate, the clustering effect is outstanding. At the same time, for large data sets, K-means also has good scalability and high efficiency, so it is widely used in many clustering fields. The specific steps of the algorithm are as follows:

(1) k samples are randomly selected from the data set D as an initial center point.
(2) Traverse all the data points in all data sets D, and divide each data point into the classification of the nearest central point.
(3) Calculate the average of each cluster set and use it as a new center point for the category.
(4) Repeat steps (2) and (3) until the k center points no longer change, or execute to a certain number of iterations.

2.3. Differential evolution Grey Wolf Optimizer principle

The Grey Wolf algorithm is a heuristic algorithm that simulates the hunting behavior of grey wolf populations. In the algorithm process, the highest level of the α wolf is responsible for making decisions to lead the wolves for hunting. The next level of β and γ wolves complete the assisting work, and the ω wolf follows the high-level wolves for group activities. The grey wolf rank relationship in the algorithm is shown in Fig. 1.

Specific algorithm steps include enclosing, hunting, and attack. The process of enveloping is the process of the wolf group moving from the initial vector position $\vec{X}$ to the target value $\vec{X}^p$ in the random vector of the [0,1] interval. In formula (1) and (2), t represents the number of iterations, $\vec{r}_1$ represents a random vector with a length in the range [0, 1], and ⊗ denotes $\vec{C}$ and $\vec{X}^p(t)$ for component multiplication.

$$\vec{D} = \left|\vec{C} \otimes \vec{X}^p(t) - \vec{X}(t)\right| \tag{1}$$

$$\vec{C} = 2\vec{r}_1 \tag{2}$$

The hunting process is guided by α, β and γ wolves. Other ω wolves update their position according to the best search position of the high-grade wolf, and continue to reach a better position through continuous iteration. In formula (3) and (4), a is gradually reduced to 0 during the iterative process, and $\vec{r}_2$ and $\vec{r}_1$ are random vectors.

$$\vec{X}(t+1) = \vec{X}^p(t) - \vec{A} \otimes \vec{D} \tag{3}$$

$$\vec{A} = a(2\ \vec{r}_2 - 1) \tag{4}$$

The α, β, and γ wolves calculate the distance from the target value by formula (1), and obtain the new position $\vec{X}(t+1)$ by formula (3), and thus repeat. The attack phase is the process in which the wolves disperse around the target value, and finally finds the optimal solution near the target value.

The hybrid grey wolf algorithm better inherits the advantages of the grey wolf algorithm, and continuously updates the wolf position by introducing a differential algorithm to the cross-mutation process. To avoid the algorithm falling into the local optimal misunderstanding. It can also greatly shorten the algorithm calculation time.

The initial position of the wolves is optimized by formula (5), where $X_{k,p}(0)$ is the p-dimensional value of the k th individual in the initial population, and $rand(0,1)$ represents between [0,1] The random number, X_p^{up} and X_p^{low}, represent the upper and lower limits of the p th dimension.

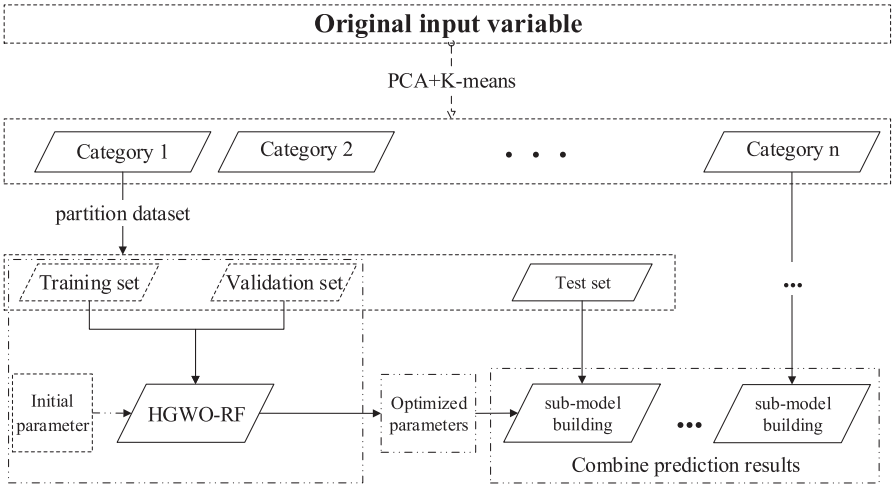

Fig. 2. Flow chart of model implementation.

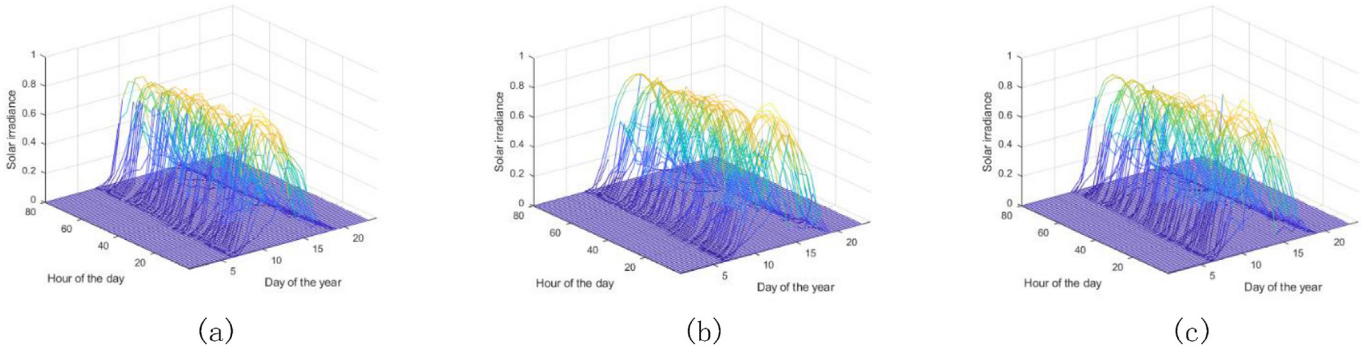

Fig. 3. PV power changes per hour in three regions.

$$X_{k,p}(0) = X_p^{low} + rand(0,1) \times \left(X_p^{up} - X_p^{low}\right) \quad (5)$$

The process of mutation, crossover and selection mainly generates the mutant individuals through the classical difference strategy, and uses the control crossover probability to obtain different crossover results. Finally, the greedy algorithm standard is used to measure whether the mutant individuals meet the actual needs.

2.4. Random forest principle

Random forest is a collection method that performs regression or classification by establishing multiple unrelated decision trees. Random forest mainly uses the idea of Bagging, and adopts random bootstrap method in sample selection. The weighting aspect adopts the method of uniform sampling, and all the prediction function weights are equal, which supports the parallel calculation of each prediction function. Through randomized forests, the random forest is not easy to overfit, extremely strong anti-noise ability and excessively fast calculation speed. Suppose the initial sample size is N. The sample feature dimension is M, and the number of decision trees in the artificially designated random forest is k. The specific modeling steps are as follows:

(1) Constructing k decision trees from the original samples by means of bootstrap.
(2) Select m features in the M dimension as training for different decision trees, and $m < M$.
(3) The decision trees are not pruned and grow as much as possible.
(4) Random forest results are averaged from the results of each decision tree.

2.5. Modeling process

The specific modeling process is shown in Fig. 2. The principal component analysis and K-means are used to cluster the original data, and next each clustering result is divided into data sets according to the proportion, and the training set, verification set and test set are obtained. The training set and the verification set are later modeled using HGWO-RF, and the model parameters are adjusted in the fitness function of the verification set. The optimized model parameters are then applied to the final model to obtain test set prediction results. Combine different types of prediction results to obtain overall prediction results.

3. Empirical research

3.1. Data description and data processing

This paper uses the PV power forecast data from the 2014 Global Energy Forecasting Competition (GEFCom2014). The dataset describes hourly photovoltaic power and environmental data for the three regions from April 1, 2012 to June 29, 2012. Fig. 3 shows the photovoltaic power variation over the three regions during this period. The environmental data includes 13 items, as shown in

Table 1
Table of environmental variables.

No	Variable	No	Variable
1	Time	8	The vertical component of the wind at 10 m
2	Total column of liquid water	9	Temperature at 2 m
3	Total columnar ice water	10	Surface solar radiation
4	Surface pressure	11	The surface heat drop
5	Relative humidity	12	Top solar radiation
6	Total cloud cover	13	The total rainfall
7	Horizontal component of wind speed at 10 m		

Table 2
Statistical characteristics of photovoltaic power generation data sets.

	Mean	Median	Standard deviation	Kurtosis	Skewness	Minimum	Maximum
Region 1	0.2586	0.1211	0.287	1.9416	0.6943	0.0001	0.9162
Region 2	0.2288	0.1025	0.2664	2.15	0.80	0.00	0.9022
Region 3	0.2379	0.1165	0.2917	2.02	0.73	0.00	0.958

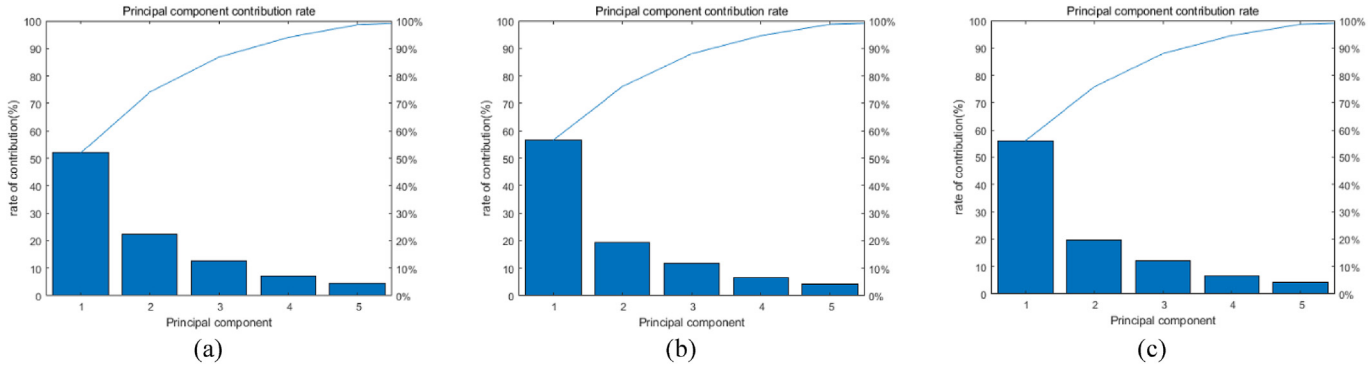

Fig. 4. Contribution rates of principal components of input variables in three regions.

Table 1.

Since photovoltaic power relies on sunlight to illuminate solar panels, there is no daylight exposure at night, so the power generation is zero. Make a preliminary selection of the data set and select the photovoltaic power as an output from 5 to 20 h routine. The environmental factor is used as the original input 4 h before the forecast time point, that is, the daily 1–19 h data is taken as input, and the rolling prediction of the hourly power generation is performed.

Table 2 shows the statistical characteristics of photovoltaic power generation in the three regions. The remaining 1440 environmental and power generation data remaining after the initial selection are used as the overall data set of the model. The data set is then segmented according to the ratio of 6:2:1 and divided into training set, verification set and test set. The training set is used for original model training, and the verification set provides data for the fitness calculation when the model parameters are selected. The test set is the data set for verifying the validity of the model.

3.2. Data preprocessing and experimental evaluation indexes

There is a dimensional difference between the input data. To eliminate this difference and speed up the model operation, the data is normalized, and the normalization method is as shown in formula (6).

$$x_i^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{6}$$

where x_i is the input raw data, x_{max} and x_{min} are the maximum and minimum values of the original data of the column, respectively, and $x_i^{'}$ is the normalized data.

In this paper, the hourly power prediction of photovoltaic power generation is carried out, and Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are selected as evaluation indexes, MAE and RMSE expressions. The formula is as follows:

$$E_{\text{mae}} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_{(i)} - y_{(i)}\right| \tag{7}$$

$$E_{\text{rmse}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_{(i)} - y_{(i)}\right|^2} \tag{8}$$

where n is the predicted data length, $y_{(i)}$ is the original data, and $\hat{y}_{(i)}$ is the predicted data. All evaluation indicators are used to measure the deviation between the observed value and the true value. These two evaluation indicators are used as a measure of the accuracy of the prediction of the machine learning model. The larger the value, the lower the performance of the representative model, and the greater the deviation between the predicted value and the true value. There is a certain difference between the two, that is, the MAE is more robust to the outliers and more responsive to the average level of predictive performance. The RMSE pays more attention to the fitting ability of the abnormal point data, and the two complement each other and complement each other.

3.3. Model parameter selection model parameter selection

3.3.1. PCA + K-means

K-means has good adaptability to high-dimensional data. However, the environmental variables used in the modeling process were 4 h ago. The corresponding input variable values may be

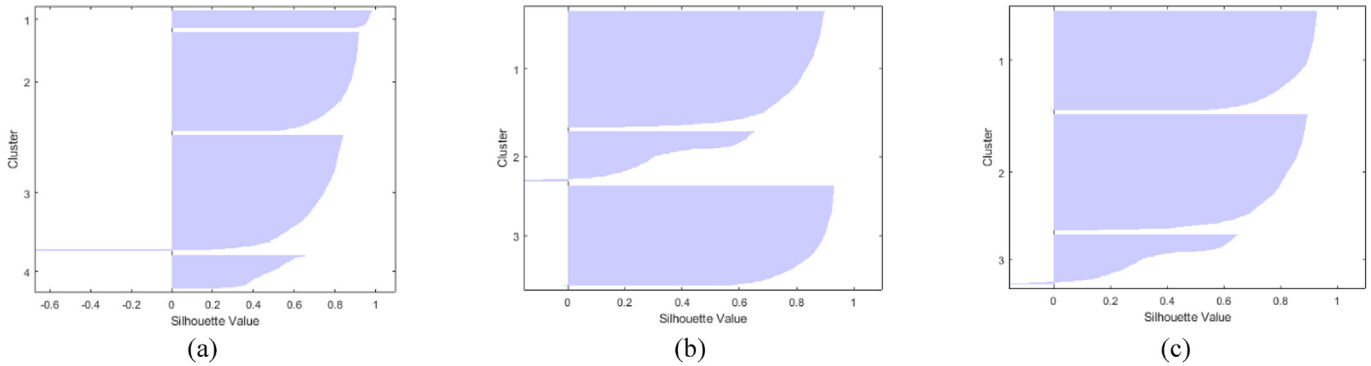

Fig. 5. Contour values of simple clustering in three regions.

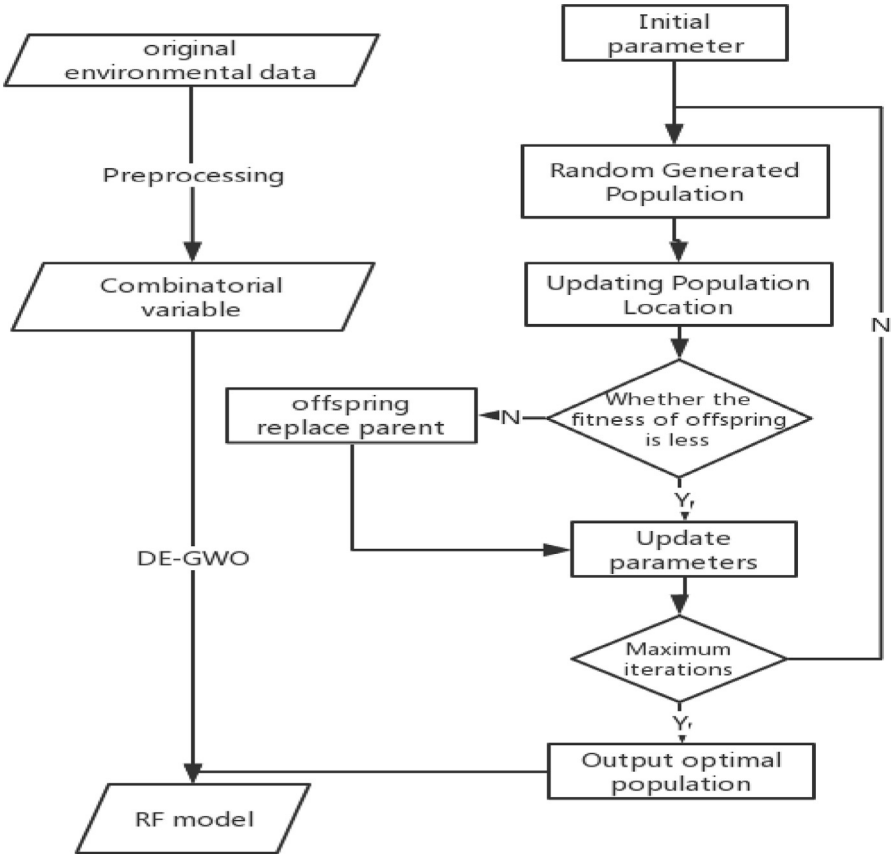

Fig. 6. HGWO-RF modeling flow chart.

related to different hours, which will affect the accuracy of the clustering effect. The initial variable (52 dimensions) is first mapped to a low-dimensional subspace using PCA to form a new linearly independent data set. It can be seen from the principal component contribution rate of the input variables in Fig. 4 that in the low-dimensional space, five variables can express the difference between the original initial variables, that is, the cumulative principal component contribution rate is close to 100%.

In the next K-means clustering, the following analysis is performed using the five principal components of the PCA dimensionality reduction. K-means needs to choose the calculation method when measuring the distance between sample points in low-dimensional space. In this paper, we use 'city block' to calculate the position of the centroid cluster. When calculating the distance d_{ij} between the two points i and j, as shown in formula (9). This formula calculates the sum of absolute differences, also known as the L1 distance. The centroid selection for each class is obtained by calculating the mean of each point in the category.

$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \tag{9}$$

The number of cluster categories will affect the final prediction effect. Too large or too small will lead to a reduction in prediction accuracy. Too many categories will make each type of data sample too small, and the sub-model cannot learn the data distribution law, resulting in under-fitting. Conversely, too few categories can cause data samples in the same category to be too different. The number of clusters is determined according to the value of the Silhouette and the result of the validation set fitting. The contour values depict the similarity of the sample to the similar sample compared to other categories. The value range of the contour value belongs to (−1,1). The larger the value, the higher the matching

Table 3
List of forecast errors (10^{-2}).

Region 1	h+1		h+2		h+3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
HGWO + RF	4.94	9.24	5.26	9.84	6.31	11.45
HGWO + SVM	5.69	9.89	6.03	11.08	7.06	12.40
ANN	6.44	9.95	7.15	12.71	8.50	13.37
Decision tree	6.02	12.08	5.55	11.16	7.86	14.40
Gaussian regression	6.04	10.81	6.83	12.16	7.19	11.49
PCA + K-means + HGWO + RF	**4.76**	**8.88**	**4.89**	**9.36**	**6.26**	**11.34**
PCA + K-means + HGWO + SVM	5.83	10.26	5.96	10.45	6.74	11.58
Region 2	h+1		h+2		h+3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
HGWO + RF	5.35	8.32	5.46	8.68	5.77	**8.79**
HGWO + SVM	5.97	8.92	6.58	9.48	6.78	9.90
ANN	6.93	9.54	7.33	10.33	7.45	11.07
Decision tree	6.01	10.45	6.50	11.99	6.40	11.61
Gaussian regression	5.81	8.95	6.36	9.76	7.02	10.29
PCA + K-means + HGWO + RF	**4.93**	8.04	**4.95**	**8.02**	**5.62**	8.95
PCA + K-means + HGWO + SVM	5.07	**7.95**	5.91	9.27	6.52	10.50
Region 3	h+1		h+2		h+3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
HGWO + RF	5.69	10.02	5.54	9.58	5.85	9.84
HGWO + SVM	6.29	10.60	7.05	11.12	7.27	11.75
ANN	7.74	11.71	7.99	12.07	8.14	11.76
Decision tree	6.79	13.57	5.84	12.03	6.26	11.42
Gaussian regression	6.46	11.07	7.05	11.34	7.85	12.17
PCA + K-means + HGWO + RF	**5.50**	**9.96**	**5.22**	**9.25**	**5.80**	**9.82**
PCA + K-means + HGWO + SVM	6.18	10.75	6.73	10.39	6.52	9.91

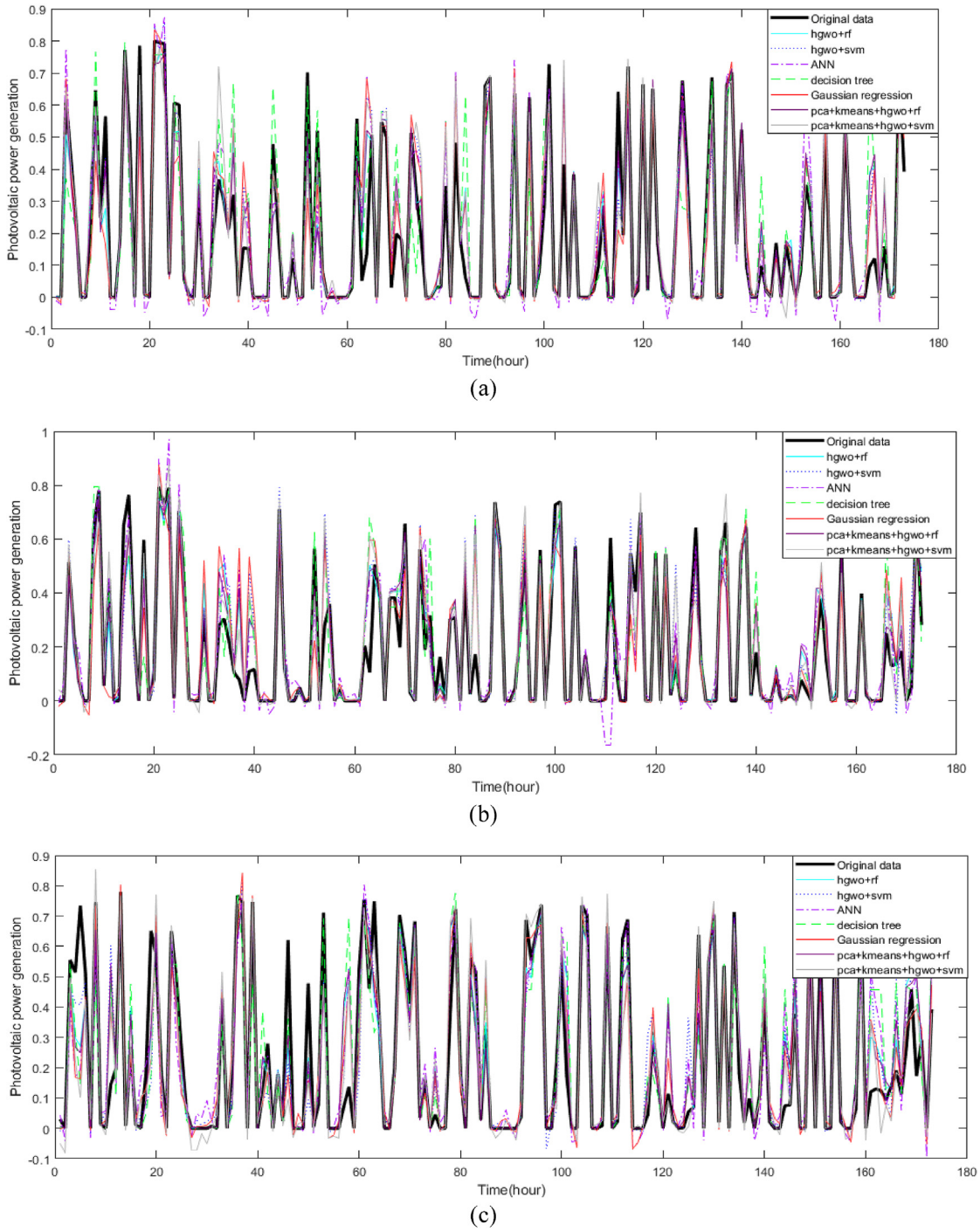

Fig. 7. Region1 forecast map of different advance periods.

degree of the same category of samples, and the lower the similarity with the different categories of samples, the clustering result is more reasonable. Fig. 5 shows the contour values calculated after simple clustering by PCA and K-means. All the samples have good clustering results except for some samples.

3.3.2. HGWO-RF modeling

According to previous literature analysis, the accuracy of random forest model prediction depends mainly on whether the number of *ntrees* and leaf *mtry* in each decision tree is successful. In the process of using HGWO-RF optimization. The population size is selected 30. The iterative parameter is selected 30. The scaling factor boundary is selected as [0.2, 0.8], and the crossover probability is selected as 0.1. In the modeling process, the original environmental variables are 13 items. The variable selection idea is to use all the environmental factors within 4 h before the predicted time point as the input variables of the random forest model. Because random forests are greatly adaptable when dealing with high-dimensional data and are highly insensitive to redundant information. If you do variable screening, it is likely to cause certain information to be missing. The random forest first trains according to the initial parameter setting through the training set, predicts

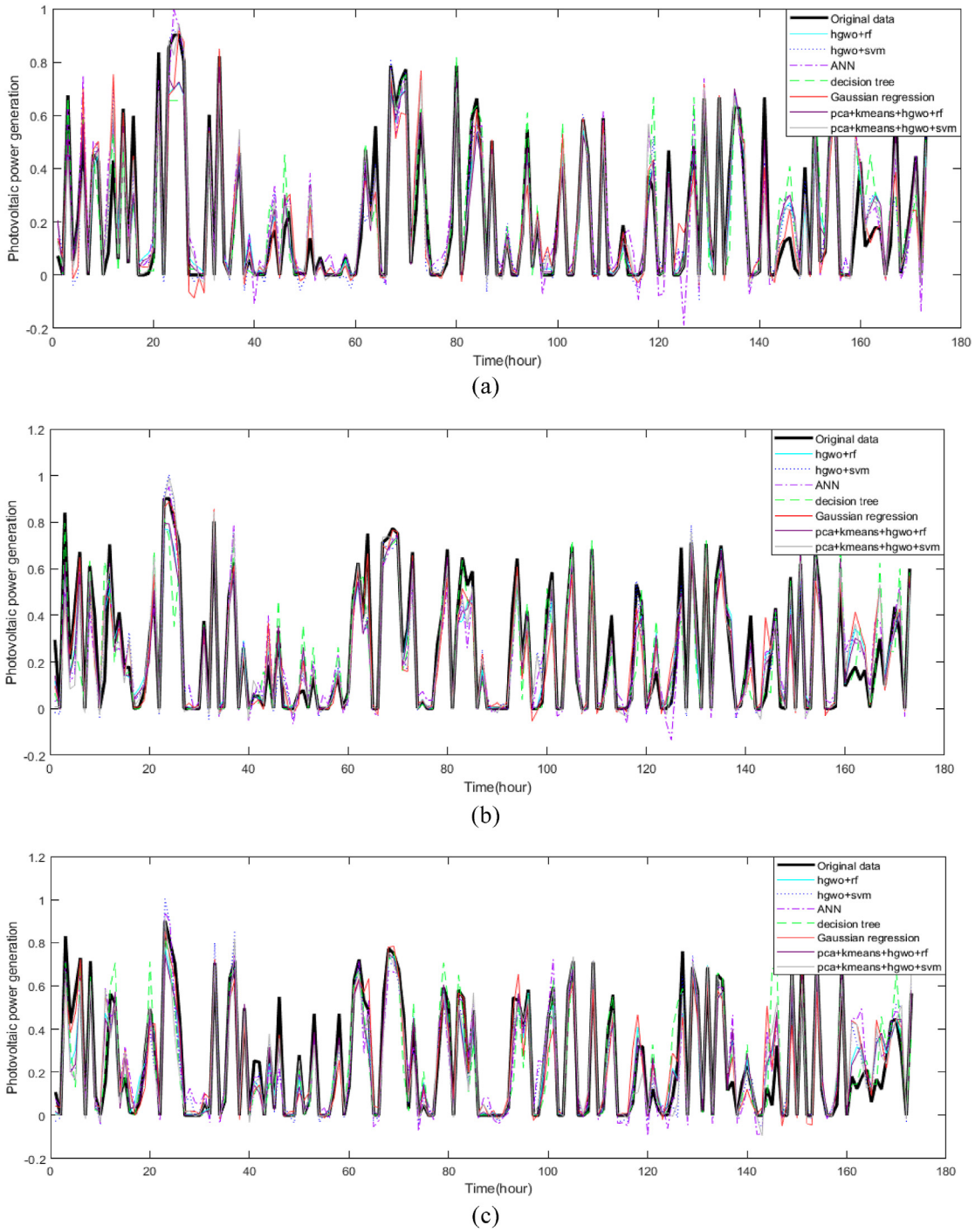

Fig. 8. Region 2 forecast map of different advance periods.

the trained model on the verification set, and calculates E_{rmse} as the fitness function to evaluate the performance of the parameters. In the HGWO algorithm, the fitness function of the parent and the child is continuously calculated, and the position where the particle is located when seeking the desirable value of the fitness function is sought. The calculation is stopped until the maximum number of iterations is reached, and the optimal model parameters are output. The modeling process is shown in Fig. 6.

3.4. Result analysis and discussion

This paper selects Support Vector Machine [44], Artificial Neural Network [37], Decision tree [45] and Gaussian regression model [10] which have been recognized by many experts and scholars in recent years for comparative analysis. Contrast models were created using MATLAB 2018 software. Refer to the existing literature to add stop or optimization conditions to each model using its own optimization function. At the same time, the PCA-K-means and HGWO algorithms are also introduced into the establishment of the

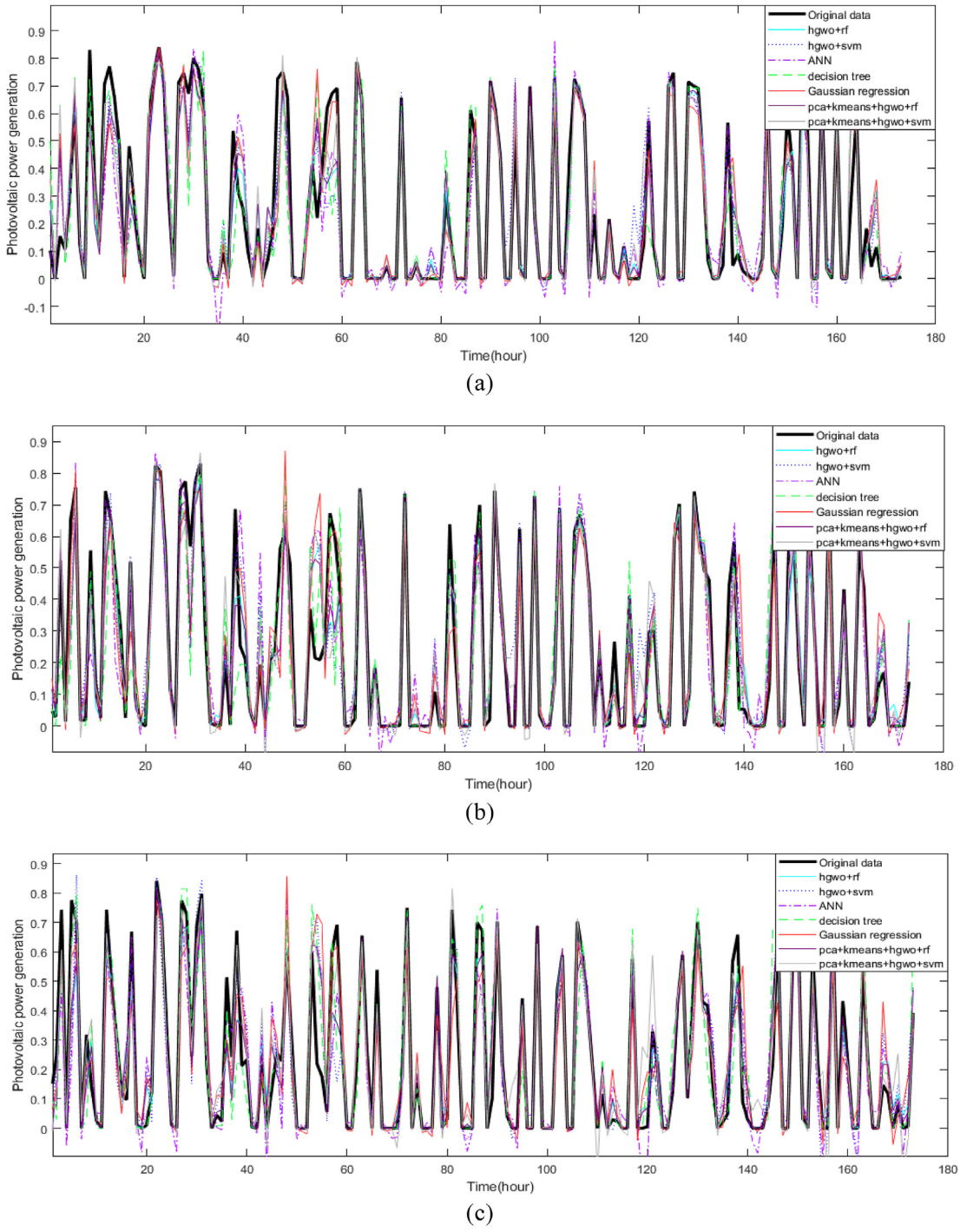

Fig. 9. Region 3 forecast map of different advance periods.

SVM model, to test whether the prediction system is applicable to other models. The distinct types of data in the test set are separately predicted, and the prediction ability of each model under different lead times is compared. Table 3 shows the prediction errors of the seven models in different regions and unlike lead times. The data set described above has been desensitized, so the unit of MAE and RMSE is 1. For the convenience of the reader, both MAE and RMSE in the table propose 10^{-2}.

For region1, the forecast for different lead times is shown in Fig. 7. Subgraphs (a), (b) and (c) show predictions with lead times of 1, 2 and 3 h, respectively. Combined with Fig. 7 and Table 2, it can be found that the random forest optimized by HGWO achieves better prediction ability and extremely higher stability than other original models. The introduction of the PCA and K-means algorithms to HGWO-RF recommended in this paper gives the model a higher precision. In order to verify the feasibility of PCA and K-means in forecasting, observing the HGWO-optimized SVM model, the accuracy is also improved after the introduction of the classification prediction method. As the lead time increases, the prediction accuracy of the same model decreases, but the overall prediction

ability ranks the same as before. PCA and K-means can be found to classify and predict input variables. In the case of extended lead time, the accuracy improvement effect is affected, which is not as obvious as the short lead time.

For region 2, the predictions are shown in Fig. 8, and a conclusion similar to region1 can be obtained in conjunction with Table 2. It can be further found that the stability of random forest prediction is higher than that of a single decision tree. In the process of using a single decision tree, it is easy to overfit. That is to say, the prediction effect is superior on the training set, but when transplanting to the verification set or the test set prediction, there is often a high error. There is a certain amount of noise data in the data. In the process of modeling, the single decision tree may use the noise data as the segmentation criterion, which leads to the failure of the established decision tree to represent the characteristics of the real data, resulting in inaccurate results. In the experiment, it is also found that although the neural network model has a good fitting ability, the characteristics of poor stability in photovoltaic prediction are difficult to ignore. Observing Figs. 7 and 8, it is found that the prediction results of the neural network model are sometimes much higher than the real data, and occasionally greatly lower than the original data. And the neural network model belongs to the technical black box, which can't control or reproduce the experiment, and the good model is difficult to transplant. Compared with other models, the Gaussian regression model has lower accuracy than the random forest.

In the prediction process of region 3 photovoltaic power generation shown in Fig. 9, one point is unlike from the previous two locations. The optimal lead time is not the first hour of forecasting, which is related to the characteristics of the data set. Because the time span of the selected input variables is the 4-h environmental variable before the prediction, due to the different geological characteristics of the region 3, the environmental variables most relevant to the generation of electricity appear within the fifth hour before the prediction. Therefore, the forecast with a lead time of 2 h is more accurate than the forecast of other lead times.

4. Conclusion

This paper proposes a random forest photovoltaic prediction algorithm based on PCA-K-means clustering and Differential Evolution Grey Wolf algorithm. The algorithm considers that the environmental data used in predicting photovoltaic power generation often has low correlation and chaos. Therefore, it is necessary to conduct principal component dimension reduction analysis on multiple hours of original environmental data. Then use K-means to cluster them to form a target cluster similar to the prediction period, and reduce the possibility of irrelevant data interfering model accuracy. In the process of modeling three different regions, six models except the recommended model were selected for comparative analysis, and the characteristics of each model were gradually understood. The HGWO algorithm was used to select quickly the optimal parameters of the model with the verification set as the target. Using the MAE, RMSE and predictive image analysis in the experiment, it is found that the recommended algorithm PCA-K-means-HGWO-RF inherits the insensitivity and good fitting ability of Random Forest to noise data, and the accuracy and stability of the algorithm prediction. The recommended model MAE values in the three regions were lower than the optimal results in the comparison model by 0.18, 0.14 and 0.19 in the h+1 period. This indicator reflects the average level of predictive performance of the recommended model over other models. Due to the limitations of the comparison model algorithm, its stability may be affected by noise data. For example, a single decision tree is prone to over-fitting, neural network parameters are difficult to estimate, and reproducibility is poor. The recommended model is also in the lead position on the RMSE indicator, indicating that the model is recommended to have higher stability in dealing with abnormal points. In addition, the PCA-K-means algorithm also has an improved prediction accuracy for the SVM model.

However, if the lead time is selected too long or the input variable is particularly low in correlation with the result, the effect of the algorithm may be affected. Hybrid algorithm clustering overemphasizes the degree of difference in the same variable, ignoring the information contained in seemingly constant variable changes. At the same time, when using HGWO algorithm for parameter optimization, due to a large number of iterative operations, the calculation time will be higher than that required for direct modeling. How to find effective and fast variable screening and parameter selection methods is the direction of the next research. Perhaps adding automatic feature engineering can help in this regard. Deep learning elements can be added to existing frameworks to automate clustering modeling of input variables. Or integrate the idea of this hybrid model into the deep learning model construction, and use the change of the dropout layer property to construct different model structures.

Funding

This research was funded by the 111 Project (B18021), the 2018 Key Project of Philosophy and Social Sciences Research, Ministry of Education, China (18JZD032) and the Fundamental Research Funds for the Central Universities (2019FR004).

References

[1] Sobri S, Koohi-Kamali S, Rahim NA. Solar photovoltaic generation forecasting methods: a review. Energy Convers Manag 2018;156:459–97.
[2] Voyant C, Notton G, Kalogirou S, Nivet M, Paoli C, Motte F, Fouilloy A. Machine learning methods for solar radiation forecasting: a review. Renew Energy 2017;105:569–82.
[3] Gala Y, Fernández Á, Díaz J, Dorronsoro JR. Hybrid machine learning forecasting of solar radiation values. Neurocomputing 2016;176:48–59.
[4] Yuan J, Xu Y, Hu Z, Zhao C, Xiong M, Guo J. Peak energy consumption and CO2 emissions in China. Energy Policy 2014;68:508–23.
[5] Yang D, Kleissl J, Gueymard CA, Pedro HTC, Coimbra CFM. History and trends in solar irradiance and PV power forecasting: a preliminary assessment and review using text mining. Sol Energy 2018;168(SI):60–101.
[6] Painter TH, Skiles SM, Deems JS, Brandt WT, Dozier J. Variation in rising limb of Colorado river snowmelt runoff hydrograph controlled by dust radiative forcing in snow. Geophys Res Lett 2018;45(2):797–808.
[7] Zendehboudi A, Baseer MA, Saidur R. Application of support vector machine models for forecasting solar and wind energy resources: a review. J Clean Prod 2018;199:272–85.
[8] Salcedo-Sanz S, Deo RC, Cornejo-Bueno L, Camacho-Gomez C, Ghimire S. An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia. Appl Energy 2018;209:79–94.
[9] Deo RC, Tiwari MK, Adamowski JF, Quilty JM. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch Environ Res Risk Assess 2017;31(5):1211–40.
[10] Fouilloy A, Voyant C, Notton G, Motte F, Paoli C, Nivet M, Guillot E, Duchaud J. Solar irradiation prediction with machine learning: forecasting models selection method depending on weather variability. Energy 2018;165(A):620–9.
[11] Sun S, Wang S, Zhang G, Zheng J. A decomposition-clustering-ensemble learning approach for solar radiation forecasting. Sol Energy 2018;163:189–99.
[12] Liu D, Niu D, Wang H, Fan L. Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm. Renew Energy 2014;62:592–7.
[13] Lin P, Peng Z, Lai Y, Cheng S, Chen Z, Wu L. 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 Convers Manag 2018;177:704–17.
[14] Eseye AT, Zhang J, Zheng D. Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information. Renew Energy 2018;118:357–67.
[15] Dong Z, Yang D, Reindl T, Walsh WM. A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization

to forecast solar irradiance. Energy 2015;82:570–7.
[16] Wang Y, Wang C, Shi C, Xiao B. Short-term cloud coverage prediction using the ARIMA time series model. Remote Sens. Lett. 2018;9(3):274–83.
[17] Al-Musaylh MS, Deo RC, Adarnowski JF, Li Y. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv Eng Inf 2018;35:1–16.
[18] Bouzgou H, Gueymard CA. Fast short-term global solar irradiance forecasting with wrapper mutual information. Renew Energy 2019;133:1055–65.
[19] Prasad R, Deo RC, Li Y, Maraseni T. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 2018;330:136–61.
[20] Khosravi A, Koury RNN, Machado L, Pabon JJG. Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. J Clean Prod 2018;176:63–75.
[21] Hussain S, AlAlili A. A hybrid solar radiation modeling approach using wavelet multiresolution analysis and artificial neural networks. Appl Energy 2017;208:540–50.
[22] Sheng H, Xiao J, Cheng Y, Ni Q, Wang S. Short-Term solar power forecasting based on weighted Gaussian process regression. IEEE Trans Ind Electron 2018;65(1):300–8.
[23] Benali L, Notton G, Fouilloy A, Voyant C, Dizene R. Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components. Renew Energy 2019;132:871–84.
[24] Wu L, Zhang Z. Grey multivariable convolution model with new information priority accumulation. Appl Math Model 2018;62:595–604.
[25] Song J, Wang J, Lu H. A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting. Appl Energy 2018;215:643–58.
[26] Ibrahim IA, Khatib T. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers Manag 2017;138:413–25.
[27] Prasad R, Ali M, Kwan P, Khan H. Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation. Appl Energy 2019;236:778–92.
[28] Liu D, Sun K, Huang H, Tang P. Monthly load forecasting based on economic data by decomposition integration theory. Sustainability 2018;10(32829).
[29] Wang J, Huang X, Li Q, Ma X. Comparison of seven methods for determining the optimal statistical distribution parameters: a case study of wind energy assessment in the large-scale wind farms of China. Energy 2018;164:432–48.
[30] Fallah SN, Deo RC, Shojafar M, Conti M, Shamshirband S. Computational intelligence approaches for energy load forecasting in smart energy management grids: state of the art, future challenges, and research directions. Energies 2018;11(5963).
[31] Du P, Wang J, Yang W, Niu T. Multi-step ahead forecasting in electrical power system using a hybrid forecasting system. Renew Energy 2018;122:533–50.
[32] Abedinia O, Amjady N, Ghadimi N. Solar energy forecasting based on hybrid neural network and improved metaheuristic algorithm. Comput Intell 2018;34(1):241–60.
[33] Liu H, Wu H, Li Y. Smart wind speed forecasting using EWT decomposition, GWO evolutionary optimization, RELM learning and IEWT reconstruction. Energy Convers Manag 2018;161:266–83.
[34] Niu M, Hu Y, Sun S, Liu Y. A novel hybrid decomposition-ensemble model based on VMD and HGWO for container throughput forecasting. Appl Math Model 2018;57:163–78.
[35] Jayabarathi T, Raghunathan T, Adarsh BR, Suganthan PN. Economic dispatch using hybrid grey wolf optimizer. Energy 2016;111:630–41.
[36] Salcedo-Sanz S, Cornejo-Bueno L, Prieto L, Paredes D, García-Herrera R. Feature selection in machine learning prediction systems for renewable energy applications. Renew Sustain Energy Rev 2018;90:728–41.
[37] Meenal R, Selvakumar AI. Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. Renew Energy 2018;121:324–43.
[38] Verbois H, Huva R, Rusydi A, Walsh W. Solar irradiance forecasting in the tropics using numerical weather prediction and statistical learning. Sol Energy 2018;162:265–77.
[39] Verbois H, Rusydi A, Thiery A. Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting. Sol Energy 2018;173:313–27.
[40] Papaioannou A, Anastasiadis A, Kouloumvakos A, Paassilta M, Vainio R, Valtonen E, Belov A, Eroshenko E, Abunina M, Abunin A. Nowcasting solar energetic particle events using principal component analysis. Sol Phys 2018;293(1007).
[41] Mallika IL, Ratnam DV, Ostuka Y, Sivavaraprasad G, Raman S. Implementation of hybrid ionospheric TEC forecasting algorithm using PCA-NN method. IEEE J Selected Topics Appl Earth Observations Remote Sensing 2019;12(1SI):371–81.
[42] van der Meer DW, Widen J, Munkhammar J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew Sustain Energy Rev 2018;81(1):1484–512.
[43] Bouzgou H, Gueymard CA. Minimum redundancy - maximum relevance with extreme learning machines for global solar radiation forecasting: toward an optimized dimensionality reduction for solar time series. Sol Energy 2017;158:595–609.
[44] Baser F, Demirhan H. A fuzzy regression with support vector machine approach to the estimation of horizontal global solar radiation. Energy 2017;123:229–40.
[45] Ahmad MW, Mourshed M, Rezgui Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. Energy 2018;164:465–74.