Pol. J. Environ. Stud. Vol. 28, No. 3 (2019), 1999-2008
DOI: *10.15244/pjoes/89539* ONLINE PUBLICATION DATE: 2018-12-12

Original Research

Elman-Based Forecaster Integrated by Adaboost Algorithm in 15 min and 24 h ahead Power Output Prediction Using PM 2.5 Values, PV Module Temperature, Hours of Sunshine, and Meteorological Data

Jianguo Zhou[1], Wei Li[1]*, Xuechao Yu[1], Xiaolei Xu[1], Xiaolei Yuan[2], Jiashuai Wang[3]

[1]Department of Economics and Management, North China Electric Power University, Baoding, China
[2]Dezhou Power Supply Company, Dezhou City, Shandong Province, China
[3]Spic Ningjin Thermoelectricity Co., Ning Jin County, China

Received: 4 January 2018
Accepted: 26 March 2018

Abstract

Nowadays, with the depletion of fossil energy and deterioration of environmental quality, solar energy is perceived to be a renewable and clean energy. While developing rapidly all over the world, solar energy is also faced with many challenges resulting from its inherent properties. In order to reduce the impact on the grid and facilitate scheduling, it is a growing problem to build a feasible model to forecast PV power with high precision. Therefore, this paper proposes an Elman-based forecaster integrated by Adaboost algorithm, namely Adaboost + Elman. Before forecasting, input variables containing PM 2.5 values, temperature of the PV module, sunshine hours, and meteorological data are made using correlation, clustering, and discriminate analysis to avoid information redundancy and improve the generalization ability of the model. To verify the developed model's application to short-term PV forecasting in two different time scales, data of Huangsi in 2016 are used for model construction and verification. An additional 7 models are introduced to make comparison. Experimental results prove that the proposed model is effective and practicable for two different scales of short-term PV power prediction.

Keywords: Adaboost + Elman, PM 2.5, temperature of the PV module, meteorological data, 15 min and 24 h ahead power prediction

*e-mail: liweixiaolian@163.com

Introduction

Nowadays, the world has encountered a series of prominent and insurmountable difficulties caused by combustion of fossil fuel that have helped give rise to climate change. The depletion of fossil energy and deterioration of environmental quality do not conform to the construction of harmonious society, which even seriously threatens the survival and sustainable development of mankind. It is unwise to give much reliance on non-renewable resources to satisfy the energy demand of the world's expanding population and economy, because an increasingly tight and critical energy supply may resulted [1, 2]. On the one hand, different initiatives commit themselves to reduce the emission of greenhouse gases into the atmosphere, for example, the Chinese government promised to reduce CO2 emissions by 40~45% by 2020 compared with 2005 [3]. On the other hand, research and development in many countries has been initiated to spare no effort to green transition as alternate energy sources, such as photovoltaic (PV) power generation, hydroelectric power generation, nuclear power, and wind power generation.

PV power's development is perceived to be a fresh growth point following wind power generation all over the world. In a Chinese government work report of 2017, "poverty alleviation" has been explicitly prioritized. PV power generation plants and distributed residential PV power are the top priority. Additionally, China has abundant solar energy resources. The amount in China is equivalent to the United States, while it is much better than Europe and Japan. Actually, PV has much potential in China, and indeed in many countries elsewhere in the world, such as the U.S., Brazil, and so on [4-7]. Therefore, photovoltaic penetration is overwhelming, especially as a distributed PV system.

It cannot be denied that PV penetration is faced with multiple and huge challenges in a photovoltaic grid connected system. Challenges in form of but not limited to: power demand and supply fluctuations, adverse impact on the power grid, uncertain meteorological conditions, and infrastructure challenges.

For the sake of addressing or alleviating several particular challenges, this research was carried out to forecast multi-scale PV power output in the short term. Long-term forecasting is easily obtained, which doesn't require high accuracy. However, as for all types of photovoltaic applications, short-term PV power prediction is closely bound with the safety and stabilization of the power system, which could meet the requirements of different time scales, such as 15 min and 24 h. Forecasts of 15 minutes and intra-hour ahead are crucial for supervising and scheduling purposes, especially for maximum power point controlling. Predictions of 24 h ahead of schedule are vital for planning, operating the reserve capacity economically, and dispatching the management of the grid.

As a sort of typical intermittent power supply, PV power generation systems are affected by numerous uncertain factors seriously, such as irradiance, humidity, temperature, and wind speed. Uncertain factors would result in high volatility and randomness in a PV power generation system, which would have a serious impact on the security and stability of the power grid [8]. PV power information is supposed to be acquired by managers ahead of time to schedule power generation and manage the allocation of reserve capacities. Therefore, PV power forecasting is significant to the reliability and stability of the PV power industry [9, 10].

Many recent international research has been conducted to forecast the PV power output. Generally, different approaches are required based on the different time horizons. Few approaches are applied to exactly forecast power based on the different time horizons simultaneously. The very short-term prediction refers to a scale that is 15 min or subter-hour (T_F<1 h) primarily utilize sky imaging techniques or time series analysis, which has limitations. When the horizons increase (1 h<T_F<6 h), accurate power prediction relies mainly on the use of satellites. Longer time scales (T_F>6 h) heavily depend on advanced numerical weather prediction (NWP) [11]. In China, research in regard to short-term forecasts based on sky imaging techniques are just at the early stage, the results of which are poor in accuracy [12]. In addition, total sky imager is considered too costly for widespread use. Compared with sky imaging, the data of NWP can be available. Time series analysis requires historical power data as inputs, which is propitious to less climate change.

With the increasing penetration of solar power in China, it is vital to improve the precision of forecasting PV power production. Many momentous works pertinent to PV power forecasting have been published by excellent scholars. Their methods could be divided into the following categories: physical modeling methods, statistical models, hybrid methods, and ensemble learning methods. Statistical models include time series models, the single nonlinear prediction model, and hybrid methods.

The physical method is based on the geographical information of the photovoltaic power station, detailed meteorological data, and the operation equation of the PV module [13]. Saint-Drenan et al. [14] developed a physical method that can employ historical PV power data to found the parameters equation of a physical model for power output. On the one hand, the modeling process is relatively complex. On the other, it is difficult to simulate some extremely abnormal weather conditions and the slow change of PV modules' parameters over time. In addition, the model has poor anti-interference ability and robustness. Due to the significant superiority of the statistical methods to physical methods, it is feasible to acquire the power production of a plant before construction in the absence of historical data. Wolff et al. [15] compared SVR for solar power with physical approaches, and the SVR showed obviously promising results, which further validated the previous research conclusions.

The statistical model, as a data-driven approach, is able to abstract a relationship between the past data to predict the future situation of the plant, which requires abundant historical data modeling. Statistical models include linear prediction methods and nonlinear prediction models. Considering the nonlinear relationship, artificial intelligence (AI) techniques – including artificial neural networks (ANNs), radial basic function neutral networks (RBF) [16], support vector machines (SVM), k-nearest neighbors (k-NN), random forests (RF), and others – have been employed in the research subject by a substantial amount of scholars aiming to improve the accuracy of prediction, and comparatively satisfied results have been obtained. Among these techniques, the ANNs have the most extensive application in solar power forecasting. Monterio et al. [17] employed ANNs, SVM, and Kalman filter (KF) to estimate PV power generation and analyzed the performance of these three models. The results indicated that the ANN model distinctly outperformed the SVM-based model, and the ANN model and the SVM-based model were obviously superior to Kalman filter.

Hybrid algorithms have the most widespread use. Despite the simplicity, the single models would omit some information inevitably when employed alone. To overcome this limitation, a growing number of scholars have combined algorithms to foster the models' strengths to improve the precision of prediction, which was defined as hybrid models. Hybrid methods applied to solar power forecasting are either combining a statistical technique with a PV performance model (hybrid-physical), or combining two or more statistical techniques (hybrid-statistical). The former approach was introduced in several works [18]. Dolara et al. [19] proposed the physical hybrid artificial neural network (PHANN) method successively. The latter approach, namely hybrid-statistical, could be classified into several groups. The first group is that nonlinear prediction models were combined with decomposition algorithms [20]. The second group is that optimization algorithms were integrated with the neural network model [21]. The third group refers to decomposition algorithms being used to decompose the PV output series, then an artificial bee colony [22] and other optimization algorithms were proposed to optimize the artificial neural network.

Ensemble learning has its irreplaceable advantages, such as increasing accuracy, improving stability, improving the selection of algorithm parameters, and boosting efficiency of learning, which has been put forward and used to improve the efficiency of PV prediction [23]. Marco Pierro et al. [24] built a model of multi-model ensemble (MME) rooting in the averaging of the best data-driven forecasts, which improved the accuracy of the results of forecasting and raised the skill score from 42% to 46%.

On the basis of the aforementioned analysis, the Elman neutral network integrated by Adaboost is proposed in this paper. In addition, the data is conducted with correlation analysis, cluster analysis, and discriminant analysis, which could avoid information redundancy, improve the generalization ability, and raise the prediction accuracy of the model proposed. Then the applicability of the proposed method to different scales of photovoltaic power prediction in the short term is investigated. The different short-term scales incorporating 15 min and 24 h are within the scope of this research. Consequently, this research is distinguished from previous studies on single-scale prediction. It is hoped that this study could contribute to making up the insufficiency in previous research and providing inspiration for future research to some extent.

Material and Methods

This section focuses on the specific process of the proposed method, which is presented in Fig. 1. The specific process consists of 3 parts: data process, optimized basic model, and ensemble learner algorithm. Specific discussion and elaboration of the proposed model are in the following.

Data Processing

Data processing means the inclusion of the selection of indicators, normalization, correlation analysis, cluster analysis, and discriminant analysis.

Selection of Indicators

Meteorological data, PM 2.5 values, power data, hours of sunshine, and temperature of photovoltaic modules (T2) are introduced in the methods proposed in order to make a prediction. The hours of sunshine refers to the length of time each day when the intensity of radiation exceeds or equals 120 W/m2 on the plane perpendicular to the sun's rays [25]. Meteorological data includes pressure, humidity, temperature of atmosphere (T1), and irradiance. What is different from other studies is that the hours of sunshine, the temperature of photovoltaic modules (T2), and PM 2.5 values are contained in the input variables. Similar studies are relatively few. There are three reasons why these indicators are selected. First of all, these factors affect the PV power generation to a great extent, respectively. Secondly, with the deterioration of air quality, the influence of PM 2.5 on air visibility and irradiance is increasingly prominent. The temperature of photovoltaic modules (T2) directly affects the output of the photovoltaic power [26]. Thirdly, the quantification and availability of input variables should be taken into consideration. While the speed of wind and clouds also have impact on PV power, randomness and uncertainty are so difficult to be quantitative. The ground-based cloud image has been introduced to make more precise

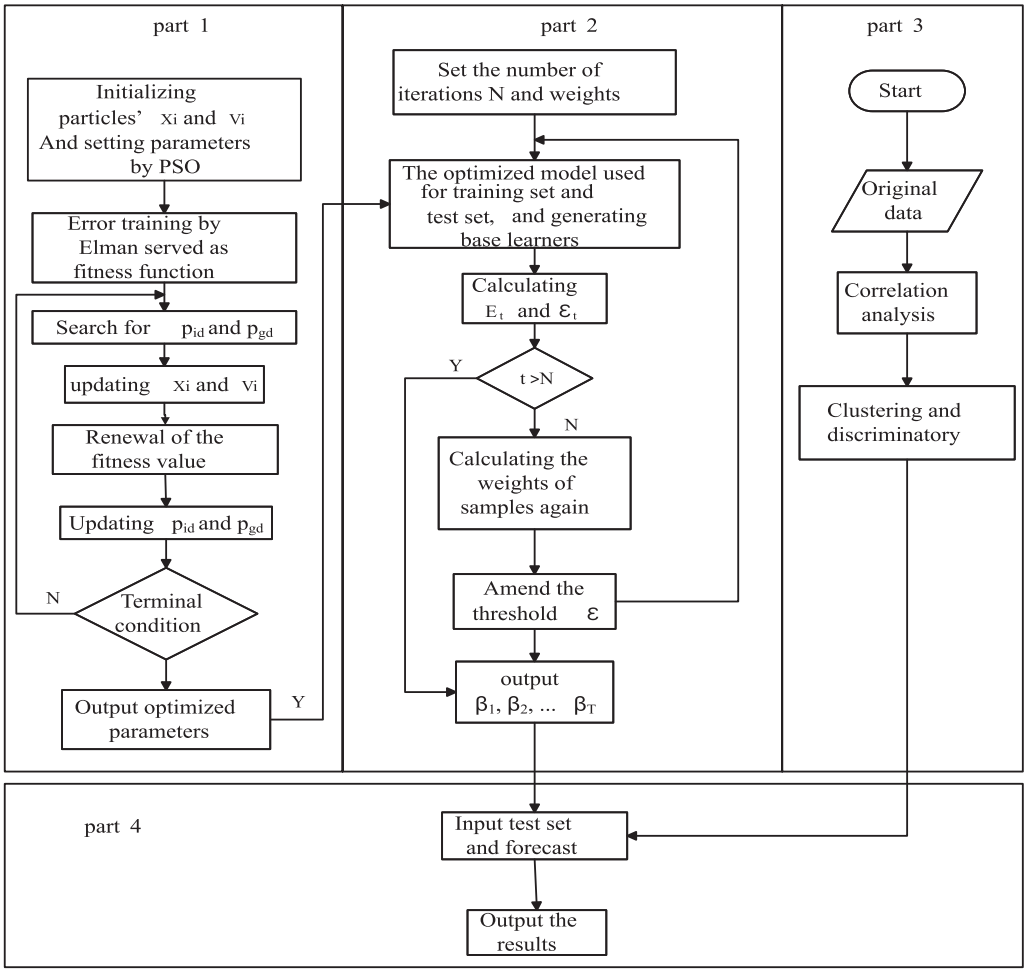

Fig. 1. Flowchart of the proposed model.

predictions, while the cost of the equipment is so high and the accuracy is low. The wind has high randomness, uncertainty, and fast changes. So the amount of cloud and wind speed are not quantified as input variables.

More specifically, the temperature of photovoltaic modules (T2), PM 2.5 values, pressure, humidity, temperature (T1), and irradiance can be available as input variables 15 min ahead forecasting. And sunshine hours, average irradiance, average temperature of atmosphere (average T1), and PM 2.5 values are included in the independent variables in 24-h forecasting.

Normalization

In the interest of eliminating the influence of dimension and improving the training speed and regression effect, the data is normalized to the interval [0, 1] according to Formula (1), where x_i is original data, x'_i is normalized data, x_{max} is the maximum value of original data, and is the minimum value of initial data:

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Correlation Analysis

If variables of high correlation coefficients are placed in the same sample set, great redundancy of information would be caused and the contribution of certain input information would be ignored [27]. To prevent this condition from happening, it is crucial to make correlation analysis. So far, only a few studies have taken into account the correlation between variables.

The correlation coefficients of characteristic attributes are determined as follows: the inputs of training samples are $X = \{x_1, x_2,..., x_n\}$, $x_k \in R^m$, $k = 1,2,...n$, $x_k = \{x_{k1}, x_{k2},..., x_{km}\}$ and the correlation coefficient between i dimension attribute and the j dimension attribute is defined as Formula (2):

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{\sqrt{\left[\sum_{k=1}^{n}(x_{ki} - \overline{x_i})^2\right]\left[\sum_{k=1}^{n}(x_{kj} - \overline{x_j})^2\right]}} \tag{2}$$

The correlation coefficients are presented in Tables 1 and 2, from which correlation coefficients between any

Table 1. Correlation coefficients of input variables in 15 min ahead forecast.

	Humidity	Irradiance	PM 2.5	Pressure	T 2	T 1
Humidity	1	-0.328	0.384	0.210	-0.381	-0.384
Irradiance	-0.328	1	-0.055	-0.050	0.872	0.427
PM 2.5	0.384	-0.055	1	-0.426	0.123	0.319
Pressure	0.210	-0.050	-0.426	1	-0.330	-0.605
T 2	-0.381	0.872	0.123	-0.330	1	0.765
T 1	-0.384	0.427	0.319	-0.605	0.765	1

Table 2. Correlation coefficients of input variables in 24 h ahead forecast.

	Sunshine hours	Average irradiance	PM 2.5	Average T 1
Sunshine hours	1	0.603	-0.561	0.880
Average irradiance	0.603	1	-0.576	0.583
PM 2.5	-0.561	-0.576	1	-0.478
Average T 1	0.880	0.583	-0.478	1

two variables are less than 0.9. Thus it does not matter that all feature variables are placed in the same sample set.

Cluster Analysis

Due to the training samples' similarity exerting a tremendous influence on prediction accuracy, it could effectively improve the generalization ability and prediction accuracy to select data of high similarity. Thus it is obviously crucial for conducting the clustering and discriminatory analysis of input variables and selection of samples. The algorithm of K-means clustering is introduced to divide data into k categories.

1) For the n samples x_i, $i = 1,2,...n$ in a given dataset, determine k centroids primarily: $x_1^{(1)}$, $x_2^{(1)}$,..., $x_k^{(1)}$, among which the observations of the st centroid is ($x_{j1}^{(1)}$, $x_{j2}^{(1)}$,..., $x_{jp}^{(1)}$). And k centroids should be craftily located as different positions lead to disparate outputs.

2) For every point x_i, calculate distance between x_i and every centroid according to the following calculated Formula (3). If d_{ij} is minimum value among (d_{i1}, d_{i2},..., d_{ik}), x_i should be connected with the centroid j. When no point is left, preliminary grouping is accomplished.

$$d_{ij} = \sqrt{\sum_{q=1}^{p} (x_{i,q} - x_{j,q})} \tag{3}$$

Then k new centroids $x_1^{(2)}$, $x_2^{(2)}$,..., $x_k^{(2)}$ must be recalculated. Repeat the above steps until the criterion function converges. That is, this iteration continues until no more alteration of centroids' locations appear and centroids do not move any further. In general, the square error criterion is used, which is defined as follows: For a given set of observations (x_1, x_2,..., x_n), where each observation is a p-dimensional real vector. The ideal situation in k-means clustering is to divide the n observations into k (k≤n) sets $S = \{S_1, S_2,..., S_k\}$, in which the within-cluster sum of squares is minimized. As shown in Formula (4), where μ_i is the mean of points of S_i [28]:

$$\arg\min_{S} \sum_{j=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{4}$$

Discriminant Analysis

Discriminant analysis is a statistical analysis method used to judge the category of sample data. Fisher discriminant model is applied in this paper to determine which category the data belongs to. Suppose there are k categories of m-dimensional space, G_1, G_2,..., G_k, the mean vectors are μ_1, μ_2,..., μ_k and covariance matrices are Σ_1, Σ_2,..., Σ_k (Σ_i>0) For the sake of obtaining Fisher linear discriminant function $u(y) = u^{*T}y$, it is necessary to compute the eigenvectors u^* corresponding to eigenvalue λ^* of $E^{-1}B$ computed by formulas (5-7). The distance between sample x and categories G_i is calculated according to Formula (8). If $L(x, G_e) = \min_{1 \le i \le k} L(x, G_i)$, $x \in G_e$.

$$\bar{\mu} = \frac{1}{k} \sum_{j=1}^{k} \mu_j \tag{5}$$

$$B = \sum_{i=1}^{k} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \tag{6}$$

$$E = \sum_{i=1}^{k} \Sigma_i \tag{7}$$

$$L(x, G_i) = \left| u^{*T} x - u^{*T} \mu_i \right| \tag{8}$$

Optimized Basic Model

The neural network, first introduced in 1943, [29] is the simulation of the structure and function of biological neurons. It has the characteristics of parallel processing and can realize the nonlinear mapping of dimension input space to dimension output space [30].

Elman networks are typically multi-layer feed-forward and dynamic recurrent neural networks. An Elman neural network comprises 4 layers: input, hidden, context, and output. The structure of an Elman network is shown in Fig. 2. The network has additional units compared with BP neutral networks, which are called context units. The additional units are used to memorize the output value of hidden layer unit and make a recurrent connection from the output of the hidden layer to its input. By storing the internal state, the system can adapt to the change of the input variables over time and express the time delay between input and output [31].

The context layer feeds back to the hidden layer value last time, which could equip the networks' dynamic learning ability and improve the prediction accuracy of the system. Therefore, the section of feed-forward network can be used to amend the correction. The connection between output layer, hidden layer, and input layer is similar to the feed-forward network.

Although the Elman neutral network has satisfactory performance in most cases, it still has some limitations of reducing accuracy. Although in Elman's learning process the parameters are trained by gradient descent method, this method has inherent characteristics of slow learning speed and can easily fall into local minimum. In addition, self-feedback gain coefficient is usually acquired by trial, which leads to low efficiency of learning. These limitations would debase Elman's performance and lead to unstable predictive results.

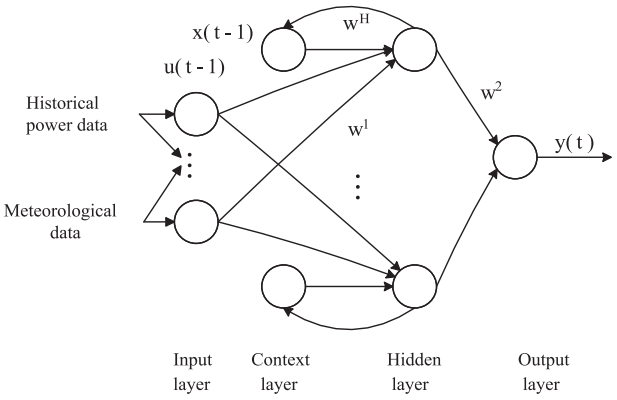

Fig. 2. Structure of Elman neutral network.

Aiming at the above problems, this paper proposed that Elman be optimized by the particle swarm optimization (PSO) algorithm, which taps PSO into training weights and self feedback gain factors in Elman. PSO was proposed by Kennedy and Eberhart in 1995 [32]. The optimization procedure of specific flow chart is shown in Fig. 1.

Every particle denotes a potential optimal position of the problem. At the outset of the iterations, parameters are initialized, including the position, velocity, and fitness value. In the process of exploring the best position, each particle's position corresponds to a fitness value updated by P_{best} and G_{best}. P_{best} is the optimal position by the particle itself, and G_{best} is the global best position of the entire swarm for the moment. Suppose there are n particles in the D-dimensional search space, for the renewal of speed and position of each particle, certain formulas (9-10) are shown as follows, where d = 1, 2,…, D represents dth dimensional space, i = 1, 2,…, n refers to ith particle, k indicates an iteration count, c_1 is the cognitive scaling parameter and c_2 is the social scaling parameter, r_1 and r_2 are random numbers uniformly distributed between 0 and 1, $X_i = [x_{i1}, x_{i2},..., x_{iD}]^T$ and $V_i = [V_{i1}, V_{i2},..., V_{iD}]^T$ respectively refer to the ith particle's position and velocity, and pbest and gbest are the individual extremum and the global extremum, respectively [33].

$$V_{id}^{k+1} = \omega V_{id}^{k} + c_1 r_1 \left(pbest_{id}^{k} - X_{id}^{k}\right) + c_2 r_2 \left(gbest_{gd}^{k} - X_{id}^{k}\right) \tag{9}$$

$$X_{id}^{k+1} = X_{id}^{k} + V_{id}^{k+1} \tag{10}$$

Ensemble Learner

AdaBoost.RT is an ensemble learning algorithm of wide application in addressing regression problems. The letters R and T stand for regression and threshold, respectively. And the rough process is described next [34].

1) Input: sequence of m samples (x_1, y_1), (x_2, y_2),..., (x_m, y_m); weak learner; integer T as the number of iterations; threshold ϕ ($0 < \phi < 1$) for distinguishing correct and incorrect predictions

2) Initialization: iteration t = 1; the initial weight of each training sample is $D_t(i) = 1/m$; error rate is $\varepsilon_t = 0$

3) Training process ($t \le T$):

Move 1: call weak learners, and offer it distribution D_t

Move 2: establish the regression model, $f_t(x_i) \rightarrow y$

Move 3: calculate the error of each sample by Formula (11)

$$Et = \left| (f_t(x_i) - y_i) / y_i \right| \tag{11}$$

Move 4: calculate the error of base learner by Formula (12)

$$\varepsilon_{\mathrm{t}} = \sum_{i:E_t(i)\rangle\phi} D_t(i) \tag{12}$$

Move 5: by Formula (13), calculate the , where

$$D_{\mathrm{t+1}}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t, E_t(i) \le \phi \\ 1, E_t(i) \rangle \phi \end{cases} \tag{13}$$

Move 6: set, output the final hypotheses by Formula (14)

$$f_{\mathrm{fin}}(x) = \sum_t (\log \frac{1}{\beta}) f_t(x) / \sum_t (\log \frac{1}{\beta_t}) \tag{14}$$

Results and Discussion

In this section, for the purposes of exploring the efficiency and practicability of the proposed model in 15 min and 24 h ahead forecasting, specific data was served to conduct empirical research described in detail below. PM 2.5, the temperature of the PV module (T2), meteorological data, the hours of sunshine, and power data are collected from the PV power plant in Huangsi, HeBei. Data from October 1, 2016 to January 31, 2017 are selected as training samples and test samples. An inland solar plant, Huangsi is located in Xingtai, of approximately 20 km to the northwest. The altitude of the site is 145~257 m. The installed capacity of the photovoltaic power plant is 50 MW, and the annual power generation is 5700 MWh. The plant covers about 94 hectares.

This paper investigates thoroughly the applicability of the integrated model proposed to different short-term time scales by the comparisons of several models. In order to highlight the superior performance of Adaboost+Elman proposed in this paper, RBF and BP neutral network integrated are applied. Fig. 3 shows the curves of the real and the 15 min ahead PV power prediction of three base models integrated by Adaboost, namely Adaboost+Elman, Adaboost+RBF, and Adaboost+BP. Fig. 4 presents the real curves and the 24 h ahead PV power prediction curves of Adaboost+Elman, Adaboost+RBF, and Adaboost+BP. The analysis shows that: A) Adaboost+Elman has the best fitting effect in both time horizons, while Adaboost+BP shows the worst performance of the three models. This is because Elman neutral network has a context layer, which gives Elman a better dynamic performance. And the very nature of BP leads to its low efficiency and local optimum. B) The goodness of fit reaches relatively high levels in two time scales, which is between the predicted values by the three models and the real values. This is accounted for in that the ensemble learning algorithms have good performance in reducing errors and improving accuracy.

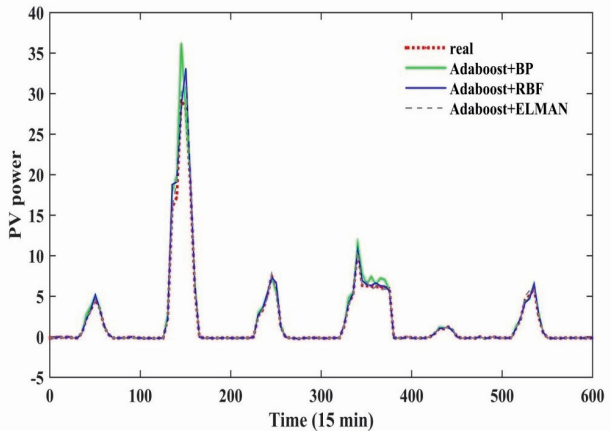

Fig. 3. Curves of the real and the 15 min ahead PV power prediction of three base models integrated by Adaboost.

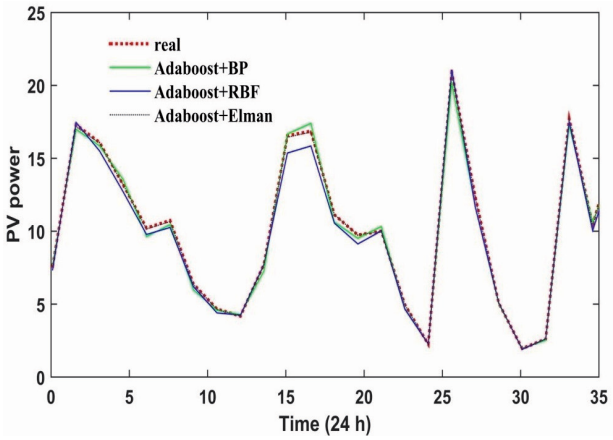

Fig. 4. Curves of the real and the 24 h ahead PV power prediction of three base models integrated by Adaboost.

The following focus on the comparison of the 8 models' prediction results and the measurements in different time horizons of 15 min and 24 h. These 8 models are Adaboost + Elman, Adaboost + RBF, Adaboost + BP, RandomForest, single Elman, single RBF, single BP, and single ARIMA, which are shown in Fig. 5.

To make a more visual and effective comparison of those models, three generally adopted error criteria are presented to measure the accuracy of all involved models, including mean bias error (MBE), mean absolute error (MAE), and root mean square error (RMSE). MBE, MAE, and RMSE are calculated by formulas (15-17), where P_i is the measured power output, namely actual data; $\hat{P}_i$ is the value of forecasting for P_i; and N denotes the number of data points. MBE could measure the bias between the expected value from forecasting and the true value. MAE is utilized to estimate the proximity between the actual value and the predicted value in absolute scale. RMSE introduces the square form to large and severely punish errors.

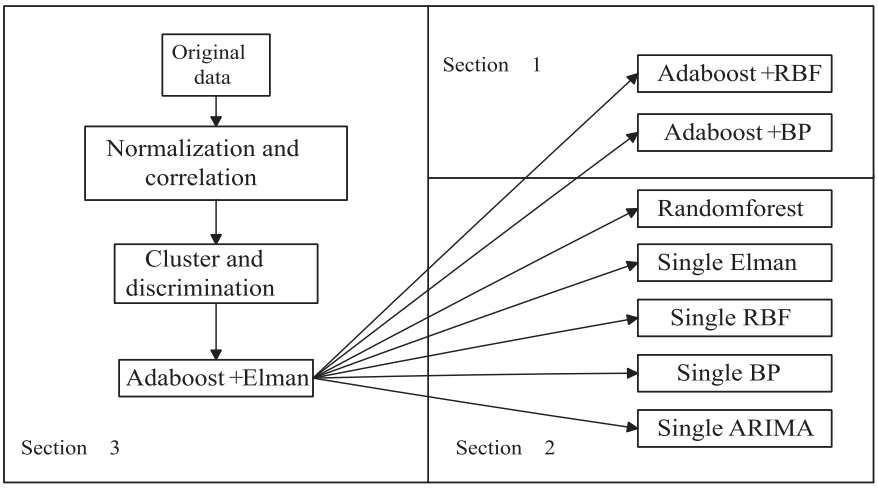

Fig. 5. Framework of the forecasting model comparisons.

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - P_i) \tag{15}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - \hat{P}_i\right| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - \hat{P}_i)^2} \tag{17}$$

From Fig. 6, namely the results of 15 min ahead forecast of PV power, three base models integrated by Adaboost perform best combining with three error indicators. The estimation of errors of three base models integrated are obviously much smaller than the other 5 models. ARIMA ranked fourth in the prediction results, the fifth is the Random Forest, and the remaining three base models have the worst performance. The results of Fig. 7 are roughly the same as in Fig. 6, which strongly indicates that the proposed model of the Elman neutral network integrated by Adaboost is suitable for two different scales of short-term PV power prediction.

From the results of Figs 6-7, the following conclusions can be drawn: A) All the errors of integrated base models are small. This is because "bad" samples with lower accuracy in Adaboost algorithms are paid more attention by models endowed with larger weights. The weights are related to the learning result of the last iteration. And through bootstrap, Random Forest makes unbiased estimation on generalization error, which improves the generalization ability and prediction precision. B) In general, the error of Adaboost is smaller than the Random Forest. The reason for this result mainly is that Adaboost algorithm is more concerned about the larger sample of errors. C) The fitting effect of Elman is better than that of BP and RBF in general, mainly because the Elman neutral network has a context layer, which leads to Elman's better dynamic performance. D) The ARIMA model is more suitable

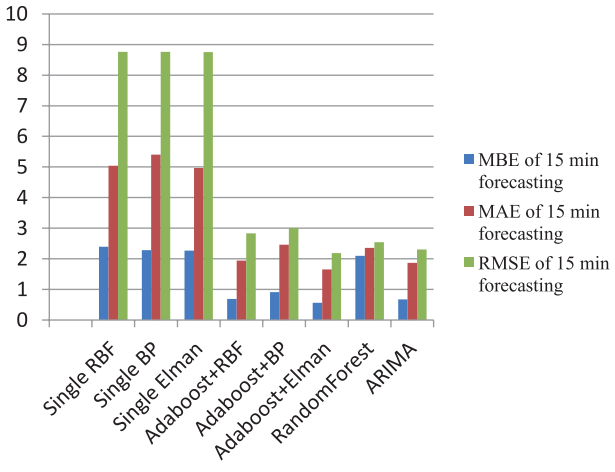

Fig. 6. Analysis of 15 min ahead PV power production forecasting: MBE, MAE, and RMSE of 8 models in 15 min ahead forecasting.

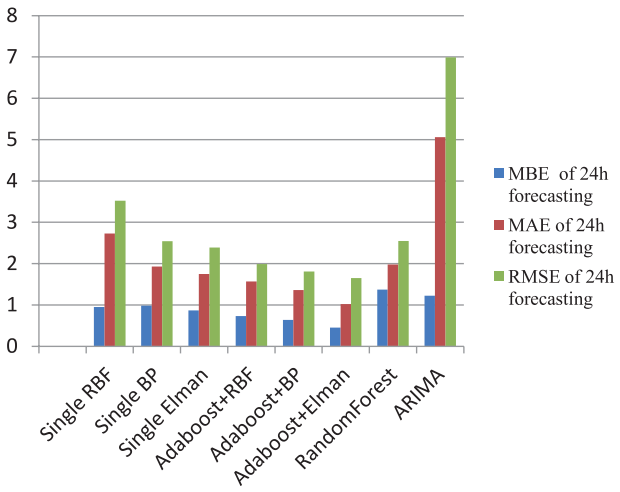

Fig. 7. Analysis of 24 h ahead PV power production forecasting: MBE, MAE, and RMSE of 8 models in 24 h ahead forecasting.

for short-term prediction, and the error becomes larger with the increase of prediction time scale. Moreover, the model considers only the characteristics of the time series itself and does not take other uncertainties into account.

Conclusions

This paper proposed an Elman-based forecaster integrated by Adaboost, namely Adaboost + Elman, for 15 min and 24 h, respectively, ahead of PV power forecasting and the RBF, Elman, BP neutral network, RandomForest, and ARIMA are exploited to make a prediction, as the comparison of the proposed model, Adaboost + Elman, which is the greatest innovation in this article. Obviously, a single base model is of poor generalization, and may easily get trapped in a local optimum. Thus, ensemble algorithms are applied to improving generalization ability and precision. Besides, correlation analysis is conducted to calculate the correlation coefficient of input variables. Once the correlation coefficient of the variable exceeds the threshold, it will not be put into the same feature set, which could effectively reduce the redundancy of information, avoid over fitting, and boost the generalization ability. Moreover, it is crucial for clustering and discriminatory, which enhance the generalization ability and decrease the risk of over-fitting as well.

In addition to the innovation of methods applied to this paper, there is also a shining point that the temperature of photovoltaic modules, PM values, and sunshine hours are taken into account as influential factors, which proves to be an effect. The temperature of photovoltaic panels directly affects the conversion efficiency of photovoltaic cells. Air quality affects power by affecting conspicuity and irradiance. Previous studies have rarely predicted power by incorporating PV module temperature and air quality into the input variables system.

The forecasting results indicate that the 8 designed models are effective and efficient for the 15 min and 24 h ahead PV power prediction. Based on the results of prediction in this research, conclusions could be drawn as below: A) Combining ensemble learners and base models is an innovative application to predict PV power production. B) The Elman model integrated by Adaboost has the best ability of forecasting in 8 models, which better adapts to time-varying change and can be easily put into effect in a photovoltaic plant. C) The proposed model outperforms other methods in both two time horizons of solar power output forecasting, thus greatly expanding the application of the model and meeting the demands of the solar farm.

Although the proposed models have distinct advantages in PV power prediction, other methods with good performance, such as SVM, KELM, and so on, are not adopted in this research. It may be better to set the threshold of coefficients to 0.8, so splitting the variables into different feature sets may make the prediction more perfect. The next study is to incorporate these methods into the prediction and discuss the application to more time scales.

Acknowledgements

The current work is supported by the National Social Science Foundation of China (grant No. 15BGL145), the National Natural Science Foundation of China (grant No. 71471061), the Fundamental Research Funds for the Central Universities (No. 2016MS125), and the Philosophy and Social Science Research Base of Hebei Province.

Conflict of Interest

The authors declare no conflict of interest.

References

1. CERVONE G., CLEMENTE-HARDING L., ALESSANDRINI S., MONACHE L.D. Short-term photovoltaic power forecasting using Artificial Neutral Networks and an Analog Ensemble. Renewable Energy, **108**, 274, **2017**.
2. LEWIS N.S., NOCERA D.G. Powering the planet: chemical challenges in solar energy utilization. Proceedings of the National Academy of Sciences of the United States of America, **103** (43), 15729, **2006**.
3. SUN W., WANG C.F., ZHANG C.C. Factors analysis and forecasting of CO2 emissions in Hebei, using extreme learning machine based on particle swarm optimization. Journal of Cleaner Production, **162**, 1095, **2017**.
4. YI J.W., ZHAO D.Q., HU X.L., CAI G.T. Study on the development of Guangdong's electricity power under CO_2 emission constraints. Journal of University of Science & Technology of China, **41** (5), 452, **2011**.
5. BECKER S., FREW B.A., ANDRESEN G.B., ZEYER T., SCHRAMM S., GREINER M., JACOBSON M.Z. Features of a fully renewable US electricity system: optimized mixes of wind and solar PV and transmission grid extensions. Energy, **72** (7), 443, **2014**.
6. ARENT D., PLESS J., MAI T., WISER R., HAND M., BALDWIN S., HEATH G., MACKNICK J., BAZILIAN M., SCHLOSSER A., DENHOLM P. Implications of high renewable electricity penetration in the us for water use, greenhouse gas emissions, land-use, and materials supply. Applied Energy, **123** (3), 368, **2014**.
7. LIMA F.J.L., MARTINS F.R., PEREIRA E.B., LORENZ E., HEINEMANN D. Forecast for surface solar irradiance at the Brazilian northeastern region using NWP model and artificial neural networks. Renewable. Energy, **87**, 807, **2016**.
8. GONG Y.F., LU Z.X., QIAO Y., WANG Q. An Overview of Photovoltaic Energy System Output Forecasting Technology. Automation of Electric Power System, **40** (4), 140, **2016**.

9. BAI J.L., MEI H.W. Improved similarity based fuzzy clustering algorithm and its application in the PV array power short-term forecasting. Power System Protection & Control, **42** (6), 84, **2014**.
10. LI Z.X., RAHMAN S.M., VEGA R., DONG B. A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. Energies, **9** (1), 1, **2016**.
11. MASSIDDA L., MARROCU M. Use of Multilinear Adaptive Regression Splines and numerical weather prediction to forecast the power output of a PV plant in Borkum, Germany. Solar Energy, **146**, 141, **2017**.
12. XIE Y.H., HU X.L., ZHANG H.D. Research on Recognition of Ground-Based Cloud Images Based on Multi-Scale Analysis. Computer Simulation, **31** (11), 212, **2014**.
13. ALMONACID F., PEREZ-HIGUERAS P.J., EDUARDO E.F., HONTORIA L. A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator. Energy Conversion and Management, **85**, 389, **2014**.
14. SAINT-DRENAN Y.M., BOFINGER S., FRITZ R., VOGT S., GOOD G.H., DOBSCHINSKI J. An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation. Solar Energy, **120**, 479, 2015.
15. WOLFF B., KUHNERT J., LORENZ E., KRAMER O., HEINEMANN D. Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction and cloud motion data. Solar Energy, **135**, 197, **2016**.
16. CHEN Z.B., DING J., ZHOU H., CHENG X., ZHU X. A model of very short-term photovoltaic power forecasting based on ground-based cloud images and RBF neural network. Proceedings of the CSEE, **35** (3), 561, **2015**.
17. MONTEIRO R.V.A., GUIMARAES G.C., MOURA F.A.M., ALBERTINI M.R.M.C., ALBERTINI M.K. Estimating photovoltaic power generation: Performance analysis of artificial neutral networks, Support Vector Machine and Kalman filter. Electric Power System Research, **143**, 643, **2017**.
18. ANTONANZAS J., OSORIO N., ESCOBAR R., URRACA R., MARTINEZ-DE-PISON F.J., ANTONANZAS-TORRES F. Review of photovoltaic power forecasting. Solar Energy, **136**, 78, **2016**.
19. DOLARA A., GRIMACCIA F., LEVA S., MUSSETTA M., OGLIARI E., A physical hybrid artificial neural network for short term forecasting of PV plant power output. Energies, **8**, 1138, **2015**.
20. YU Q., PIAO Z.L., HU B. A Hybrid Model for Short-Term Photovoltaic Power Forecasting Based on EEMD-BP Combined Method. Power System and Clean Energy, **32** (7), 132, **2016**.
21. LI Q., ZHOU B.Q., ZHANG J.C., LI J.J. Photovoltaic Power Prediction Based on Adaptive Differential Evolution and BP Neural Network. Shaanxi Electric Power, **42** (2), 23, **2014**.
22. GAO X.M., YANG S.F., PAN S.B. A forecasting model for output power of grid-connected photovoltaic generation system based on EMD and ABC-SVM. Power System Protection and Control, **43** (21), 86, **2015**.
23. ALESSANDRINI S., MONACHE L.D., SPERATI S., CERVONE G. An analog ensemble for short-term probabilistic solar power forecast. Applied Energy, **157** (1), 95, **2015**.
24. PIERRO M., BUCCI F., FELICE M.D., MAGGIONI E., MOSER D., PEROTTO A., SPADA F., CORNARO C. Multi-model ensemble for day ahead prediction of photovoltaic power generation. Solar. Energy, **134**, 132, **2016**.
25. ANGSTROM A. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Quarterly Journal of the Royal Meteorological Society, **50** (210), 121, **1924**.
26. QIAN Z., CAI S.B., GU Y.Q., TONG J.J., BAO G.J. Review of PV power generation prediction. Journal of Mechanical - Electrical Engineering, **32** (5), 651, **2015**.
27. ZHANG W.Q., FU Y.J., YANG H.Z. Multi-model soft-sensor modeling based on improved clustering and weighted bagging. CIESC Journal, **63** (9), 2697, **2012**.
28. BARAK S., ARJMAND A., ORTOBELLI S. Fusion of multiple diverse predictors in stock market. Information Fusion, **36**, 90, **2017**.
29. NOORI R., KARBASSI A.R., ASHRAFI K., ARDESTANI M., MEHRDADI N. Development and application of reduced-order neural network model based on proper orthogonal decomposition for BOD5 monitoring: Active and online prediction. Environmental Progress & Sustainable Energy, **32** (1), 120, **2013**.
30. ZHUANG T., YANG C.J. Silicon content forecasting method for hot metal based on Elman-Adaboost strong predictor. Metallurgical Industry Automation, **41** (4), 1, **2017**.
31. LIU H., TIAN H.Q., LIANG X.F., LI Y.F. Wind speed forecasting approach using secondary decomposition algorithm and Elman neural networks. Applied Energy, **157**, 183, **2015**.
32. XIAO Y., XIAO J., LU F.B., WANG S.Y. Ensemble ANNs-PSO-GA Approach for Day-ahead Stock E-exchange Prices Forecasting. International Journal of Computational Intelligence Systems, **7**, 272, **2014**.
33. CHANG W.Y. Short-Term Wind Power Forecasting Using the Enhanced Particle Swarm Optimization Based Hybrid Method. Energies, **6**, 4879, **2013**.
34. ZHANG P.B., YANG Z.X. A Robust AdaBoost. RT Based Ensemble Extreme Learning Machine. Mathematical Problems in Engineering, **6**, 1, **2015**.