Hour-ahead photovoltaic power forecast using a hybrid GRA-LSTM model based on multivariate meteorological factors and historical power datasets

Biaowei Chen¹², Peijie Lin¹², Yaohai Lin³, Yunfeng Lai¹², Shuying Cheng¹², Zhicong Chen¹², and Lijun Wu¹²

1 School of Physics and Information Engineering, and Institute of Micro-Nano Devices and Solar Cells, Fuzhou University, Fuzhou, China
2 Jiangsu Collaborative innovation Center of Photovoltaic Science and Engineering, Changzhou, China
3 College of Computer and Information Sciences, Fujian Agriculture and Forest University, Fuzhou, China
E-mail: linpeijie@fzu.edu.cn; yunfeng.lai@fzu.edu.cn

Abstract. Owing to the clean, inexhaustible and pollution-free, solar energy has become a powerful means to solve energy and environmental problems. However, photovoltaic (PV) power generation varies randomly and intermittently with respect to the weather, which bring the challenge to the dispatching of PV electrical power. Thus, power forecasting for PV power generation has become one of the key basic technologies to overcome this challenge. The paper presents a grey relational analysis (GRA) and long short-term memory recurrent neural network (LSTM RNN) (GRA-LSTM) model-based power short-term forecasting of PV power plants approach. The GRA algorithm is adopted to select the similar hours from history dataset, and then the LSTM NN maps the nonlinear relationship between the multivariate meteorological factors and power data. The proposed model is verified by using the dataset of the PV systems from the Desert Knowledge Australia Solar Center (DKASC). The prediction results of the method are contrasted with those obtained by LSTM, grey relational analysis-back propagation neural network (GRA-BPNN), grey relational analysis-radial basis function neural network (GRA-RBFNN) and grey relational analysis-Elman neural network (GRA-Elman), respectively. Results show an acceptable and robust performance of the proposed model.

1. Introduction

The consumption of energy is growing dramatically since the rapidly rhythm of human social activities. With the continuous consumption of fossil fuels such as coal, oil and natural gas, their extensive use has induced a significant global climate change. This environmental problem has become the focus of attention worldwide.

Over the past decades, in an increasing number of countries, such as the United States, Japan and China, the proportion of renewable energy in electrical market is increased dramatically. Among these renewable energies, solar photovoltaic (PV) power is one of the largest renewable energy resources on our planet, due to its benefits of being clean, inexhaustible and green. According to the REN21’s 2019 report, in 2018, solar PV capacity additions were more than 100 GW for the first time, and the cumulative capacity reached 505 GW, an increase of 25% from 2017 [1]. However, like many other renewable energy sources, PV power depends highly on weather conditions. The instability of
weather condition makes the output of PV power have strong randomness, fluctuations and intermittence. It not only brings big challenge to PV plant in electrical power dispatching and maintenance management, but also increases the cost of power generation. Therefore, these challenges should be urgently solved to enhance the inflexibility of PV plant.

Enhancing the prediction accuracy of PV power is a good way to improve the inflexibility of PV system [2]. Up to now, a good number of researches has been conducted to develop appropriate models in forecasting PV power generation with the targets of higher accuracy and minimum complexity with computational cost [3]. The time horizon of PV power forecast can be categorized as long-term (LT), short-term (ST), or very short-term (VST) [4]. The VST horizon is defined as several minutes to several hours. This kind of forecasting method would provide reference to unit commitment, scheduling, and dispatching of electrical power. In addition, it also enhances the security of grid operation [5]. The radical increase in the computer’s power makes learning networks (LNs) based forecasting models always provide a more promising performance than physical methods and statistical approaches due to their potential abilities for data-mining and feature-extracting. Typical LNs applied in the studies of PV power forecast includes back propagation neural network (BPNN), radial basis function neural network (RBFNN), Elman neural network (Elman NN) and et al [6]. In addition, the numerical weather prediction (NWP) based prediction method has become the most popular research directions due to the higher accuracy and resolution of NWP [7]. In [8], based on multivariate meteorological factors and historical power datasets, the model applies grey relational analysis (GRA) algorithm and NWP meteorological information to select similar days as training-dataset, and the result shows the accuracy of day-ahead PV power forecasting has been improved. It is clear that the forecasting accuracy of model as well as the cost of training-time can be improved by selecting datasets reasonably [9]. And GRA is an efficient algorithm to select the similarity time periods. In addition, the long short-term memory recurrent neural network (LSTM) also applied to handle time series forecasting task. In [10], the model applies LSTM to forecasting day-ahead solar irradiance, the merit of LSTM in time series task has been proved. However, the huge dataset dramatically increases the training time cost, which makes it unsuitable for VST PV power prediction. Thus, the GRA-LSTM hour-ahead PV forecasting method applies GRA to select the similar and optimal time periods of forecasting hour, and applies LSTM as LN to predict hour-ahead PV power generation output.

2. Architecture of the proposed model

2.1. Correlation between PV power output and input variables

The generation power of grid-connected PV power station is mainly related to the following factors: the meteorological factors, the working characteristics of PV module, the type and installation of PV cell. Due to the complexity of the inherent properties of grid-connected PV power plants, it is difficult to fully consider all the performance parameters in many practical engineering applications. However, in many studies [2,3,6,11], the importance of meteorological factors for PV generation have been proved. Thus, the $\rho_{x,y}$ is applied to analyze the influence of meteorological factors on the output power of PV power generation.

The $\rho_{x,y}$ is defined as follows:

$$
\rho_{x,y} = \frac{N \sum X Y - \sum X \sum Y}{(N \sum X^2 - (\sum X)^2)^{1/2} (N \sum Y^2 - (\sum Y)^2)^{1/2}} \quad (1)
$$

The Table 1 below shows the significance of different correlation values. And the Figure 1 represents the correlation of hourly PV power mean value ($P_{\text{mean}}$) with hourly typical meteorological mean value in different month, includes hourly mean value of air temperature ($T_{\text{mean}}$), hourly mean value of relative humidity ($H_{\text{mean}}$), hourly mean value of global horizontal radiation ($G_{\text{mean}}$) and hourly mean value of diffuse horizontal radiation ($D_{\text{mean}}$). In general, it can be seen that the correlation between different impact factors and PV output power change smoothly in different months. According to this, the proposed method applies these four factors as four vectors of the input eigenvector of training
dataset. In addition, considering the ρ between $G$ and $P$ maintains in the highest level, the proposed model also selects $G$ at every time point as the other vector of the input eigenvector.

Table 1. The significance of different correlation values

| $\rho_{x,y}$ value | Significance     |
|---------------------|------------------|
| 0.8-1.0             | extremely strong correlation |
| 0.6-0.8             | strong correlation  |
| 0.4-0.6             | medium correlation |
| 0.2-0.4             | weak correlation   |
| 0-0.2               | very weak correlation |

Figure 1. Correlation between hourly mean value of $P$ and hourly mean value of different meteorological

2.2. Grey relational analysis

Cleaning and optimizing the training dataset of LNs can improve the accuracy as well as decrease the time-cost of forecasting models, the GRA is applied to select similarity and optimal similarity hours from history dataset. The basic idea of GRA algorithm is obtaining the correlation degree between the curve family of the original evaluation sequence set and the reference series according to the geometric similarity degree between them[8]. This analysis method seeks the numerical relation among the subsystems in a system, which provides a quantization for the changing situation of the system. The steps for comprehensive evaluation using GRA algorithm are as follows:

Step1: Obtain the analysis vector

Obtain the comparison sequence and the reference sequence. The reference sequence $y$ and the comparison sequence $x_i$ are defined as follows:

$$y = \{y(k) | k = 1,2,...,n\} \quad (2)$$

$$x_i = \{x_i(k) | k = 1,2,...,n\}, i = 1,2,...m \quad (3)$$

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

Step2: Nondimensionalize variables

Because it is hard to compare and obtain the accurate result when the dimensions of the system variables are different. Therefore, the data are generally dimensionless processing, and given by:
where, $A_j(k)$ represents the combination matrix of the reference sequence and the comparison sequence, $A_{\min}(k)$ and $A_{\max}(k)$ are the minimum and maximum values of each column of the matrix, $j$ represents the sum of the reference sequence and the comparison sequence quantities.

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

The $\xi(k)$ between $x_o(k)$ and $x_i(k)$ is given by:

$$\xi(k) = \frac{\min_{k} |y(k)-x_o(k)| + \rho \max_{k} |y(k)-x_i(k)|}{|y(k)-x_i(k)| + \rho \max_{k} |y(k)-x_i(k)|}$$

(5)

where, $y(k)$ and $x_i(k)$ represent the reference sequence and the comparison vector. $\rho$ represents the resolution coefficient, here, $\rho$ is 0.5.

Step 4: Calculate correlation degree

Get the average of the correlation coefficients (that is, the points in the curve), as a quantitative representation of the correlation degree between the comparison sequence and the reference sequence. Correlation degree $r_i$ is defined as follows:

$$r_i = \frac{1}{n} \sum_{k=1}^{n} \xi(k)$$

(6)

Step 5: Sort correlation degree

The correlation degree is sorted according to the value, which reflects the relation of each comparison sequence relative to the reference sequence, calculate the correlation degree between the forecasting day and each sample in the same cluster sample sets. The hour with the highest correlation degree is determined as the optimal similarity hour, and the hours when the correlation degree is greater than a threshold $r_{\text{threshold}}$ is determined as similarity hours.

2.3. LSTM recurrent neural network

The Long-short term memory network, as one of the most advanced recurrent neural networks, has shown remarkable result in numerous time series learning tasks [10]. Fundamentally, there are three logic gate structures in every single cell, including forgetting gate, input gate and the output gate. And each operation process mainly includes four sub-operations. The formula corresponding to each part of the operation is as follows [12]:

**Forget gate:**

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

(7)

**Input gate:**

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

(8)

$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

(9)

**Merge process:**

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

(10)

**Output gate:**

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

(11)

$$h_t = o_t \cdot \tanh(C_t)$$

(12)

The cascade structure and cell sample of LSTM is shown in Figure 2.
where, $h_t$ represents the output at time-step $t$, and $x$ represents the input at time-step $t$.

2.4. The implementation procedure of the GRA-LSTM model

The procedure of the GRA-LSTM model is shown in Figure 3. The detailed steps are described as follows:

Step 1: Collect the historical of power datasets and multivariate meteorological factors before forecasting hour. The multivariate meteorological factors include global horizontal radiation, diffuse horizontal radiation, relative humidity and air temperature.

Step 2: Preprocess the data, including abnormal data and normalization.

Step 3: Determine the similarity hours and the optimal similarity hour of the forecasting hour in the historical sample sets with GRA algorithm according to the meteorological characteristic ($F_T$) values of the forecasting hour. The $y$ and $x_i$ represent the eigenvector of forecasting hour and same-period hours of history dataset.

The eigenvector $F_T$ composed of $(T_{start}, T_{mean}, T_{end}, G_{min}, G_{mean}, G_{max})$, these parameters represent start, average, end value of temperature, and minimum, average, maximum value of global radiation, respectively.

Step 4: Collect the training samples, the threshold $r_{threshold}$ is set as 0.85. The training dataset include two main part. Include first 20 same time periods before forecasting day and first 10 samples of similarity hour samples, 30 samples in total. And the highest $r_i$ sample is applied as optimal similarity hour.

Step 5: Determine the neuron numbers in the input, output and hidden layer, and initialize the threshold values and weights of LSTM RNN, respectively. The LSTM recurrent neural network is trained by using training hour samples, and the prediction model is obtained. In this study, $h_t$ of LSTM is $P_t$, it represents the power output at time-step $t$, and $x_i$ is an eigenvector composed of $(G_t, M_t)$, $G_t$ represents global radiation at time-step $t$, $M_t$ composed of $(T_{mean}, H_{mean}, G_{mean}, D_{mean})$ represent 4 meteorological factors in $T$ hour period.

Step 6: Input the forecasting meteorological characteristic values ($M_{forecasting}$) of the forecasting hour and global radiation data ($G_{best}$) of optimal similarity hour into the prediction model to forecast the output power of the forecasting hour.

Figure 2. Cascade structure and cell sample of LSTM
Historical PV power data and multivariate meteorological factors in same periods hours (1 year before forecasting hour) → Data preprocessing → GRA algorithm(according to 6 meteorological data, \( F_T(T_{\text{start}}, T_{\text{mean}}, T_{\text{end}}, G_{\text{min}}, G_{\text{mean}}, G_{\text{max}}) \)) →

First 20 same time samples before forecasting day and first 10 similarity hours(training samples) → Optimal similarity day (test sample input) → Forecasting hour (test sample input) →

13 global radiation data of the hour, \( G \) → 4 meteorological data of the hour, \( M(T_{\text{mean}}, H_{\text{mean}}, G_{\text{mean}}, D_{\text{mean}}) \) →

13 PV output data of the hour, \( P \) → 13 global radiation data of the hour, \( G_{\text{best}} \) →

LSTM neural network initialize → LSTM neural network prediction model →

13 power data of forecasting day, \( P \)

Figure 3. Flowsheet of the proposed GRA-LSTM model

3. Experiment and discussions

3.1. Evaluation metrics

In order to verify the validity of the method. Three different evaluation metrics \( R^2 \), \( RMSE \), and \( MAPE \) are applied to verify the prediction accuracy of the proposed model. These error metrics are defined as follows.

The \( R^2 \) is define as:

\[
R^2 = \left( \frac{\sum_{i=1}^{N} P_{f,i} P_{a,i} - \sum_{i=1}^{N} P_{f,i} \sum_{i=1}^{N} P_{a,i}}{\sum_{i=1}^{N} P_{f,i}^2 - \left( \sum_{i=1}^{N} P_{f,i} \right)^2} \right) \left( \frac{\sum_{i=1}^{N} P_{a,i}^2 - \left( \sum_{i=1}^{N} P_{a,i} \right)^2}{\sum_{i=1}^{N} P_{a,i}^2} \right)^2 \]  

(13)

The \( RMSE \) is defined as:

\[
RMSE = \left( \frac{1}{N} \sum_{i=1}^{N} (P_{f,i} - P_{a,i})^2 \right)^{1/2}
\]

(14)

The \( MAPE \) is defined as:

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{P_{f,i} - P_{a,i}}{P_{a,i}} \right) \times 100
\]

(15)

\( \bar{P}_{a,i} = \frac{1}{N} \sum_{i=1}^{N} P_{a,i} \)
Where $P_{f,i}$ and $P_{a,i}$ are the predict and practical PV output power of forecasting hour. $\bar{P}_{a,i}$ is the mean practical PV output power, and $N$ is the sample point numbers in the PV power generation period which equals to 13 in this study.

3.2. Experiment result

The dataset of the PV systems from the Desert Knowledge Australia Solar Center (DKASC) is applied to verify the effect of the proposed forecast method. The multivariate meteorological factors (air temperature, relative humidity, global horizontal radiation and diffuse horizontal radiation) and historical power datasets from January 1, 2017 to November 30, 2018 are applied for the experiment. And the LSTM, grey relational analysis-back propagation neural network (GRA-BPNN), grey relational analysis-radial basis function neural network (GRA-RBFNN) and grey relational analysis-Elman neural network (GRA-Elman) are adopted as the comparison models. The forecasting curves by the GRA-LSTM forecasting model and the other forecasting comparison models are shown in Figure 4.

**Figure 4.** The forecasting curves of GRA-LSTM and comparison models: (a) 9-10 o’clock; (b) 11-12 o’clock; (c) 13-14 o’clock; (d) 15-16 o’clock.

Figure 5 shows the corresponding absolute errors in different time periods. It is clear that the GRA-LSTM model-based prediction method has the best results in most cases.

In addition, to further test the performance of proposed model. The March 18, 2018 (summer in Australia), May 10, 2018 (autumn in Australia) and August 19, 2018 (winter in Australia), December 21, 2018 (spring in Australia) are selected as test days. And the Table 2 shows the prediction accuracy evaluation (includes the $R^2$, RMSE and MAPE) of forecasting result. Considering the proposed GRA-LSTM model-based forecast technology and the other comparative forecast approaches (LSTM, GRA-BPNN, GRA-RBFNN and GRA-Elman). It can be concluded that the proposed model results in better prediction accuracy: the RMSE and MAPE have 5.7766 kW and 2.2784% average values. The average RMSE enhancement of the GRA-LSTM model with respect to the previous comparative models is 66.57%, 42.81%, 71.54%, 47.91%, respectively. And the average MAPE enhancement is 70.42%, 46.01%, 72.73% and 54.24%, respectively. And considering the $R^2$ as given in Table 2, the LSTM has highest 0.9267 average value, it shows the outstanding time series learning ability of LSTM, however, the GRA-LSTM has better performance in robustness. Therefore, compared with comparative approaches. Synthetically, the proposed GRA-LSTM model is a novel and effective hour-ahead PV power generation prediction model.
Figure 5. Absolute error curves in different time periods: (a) 9-10 o’clock; (b) 11-12 o’clock; (c) 13-14 o’clock; (d) 15-16 o’clock.

Table 2. Comparison of the prediction error by adopting various prediction methods.

| Model        | Spring | Summer | Autumn | Winter | Average | Standard deviation |
|--------------|--------|--------|--------|--------|---------|--------------------|
| R²           | GRA-LSTM | 0.8891 | 0.9065 | 0.9396 | 0.9011  | 0.9091             | 0.0187             |
|              | LSTM   | 0.8707 | 0.9245 | 0.9683 | 0.9435  | **0.9267**         | 0.0359             |
|              | GRA-BPNN | 0.4616 | 0.9238 | 0.9040 | 0.9220  | 0.8028             | 0.1972             |
|              | GRA-RBFNN | 0.6098 | 0.9139 | 0.9538 | 0.9473  | 0.8562             | 0.1431             |
|              | GRA-Elman | 0.6847 | 0.9890 | 0.9563 | **0.9529** | 0.8957            | 0.1226             |
| RMSE(kW)     | GRA-LSTM | **3.6194** | **6.0800** | 7.3728 | 6.0343  | **5.7766**         | **1.3565**         |
|              | LSTM   | 7.2274 | 16.6588 | 13.250 | 31.9979 | 17.2835           | 9.1419             |
|              | GRA-BPNN | 21.2371 | 8.6991 | **6.6957** | **3.7780** | 10.1025          | 6.6625             |
|              | GRA-RBFNN | 10.6054 | 22.1715 | 17.6918 | 30.7319 | 20.3002           | 7.2992             |
|              | GRA-Elman | 17.4779 | 13.7508 | 6.3992 | 6.7371  | 11.0912           | 4.7127             |
| MAPE (%)     | GRA-LSTM | **1.2109** | **2.3017** | 3.0235 | 2.5776  | **2.2784**         | **0.6680**         |
|              | LSTM   | 2.3637 | 7.2106 | 6.6749 | 14.5652 | 7.7036           | 4.3846             |
|              | GRA-BPNN | 8.4146 | 4.2270 | **2.9075** | **1.3288** | 4.2194          | 2.6304             |
|              | GRA-RBFNN | 3.8040 | 10.6479 | 7.3737 | 11.5985 | 8.3560           | 3.0599             |
|              | GRA-Elman | 7.4808 | 6.4619 | 3.0439 | 3.0102  | 4.9992            | 2.0048             |

4. Conclusion

A GRA-LSTM model-based method is proposed for hour-ahead PV power generation forecasting in this study. Four forecasting meteorological values are designed and are taken as the inputs of the prediction model along with the global horizontal radiation. To optimize the training dataset, the GRA algorithm is applied to select the similarity hours. Then the LSTM RNN is adopted to train the forecasting model. In addition, the datasets on the DKASC website are employed to verify the proposed method by comparing with four other models: LSTM, GRA-BPNN, GRA-RBFNN and GRA-Elman. The experimental results show that the smaller forecast error can be obtained by
applying the proposed model, and the forecast accuracy and robustness of GRA-LSTM is superior to the compared models.

5. References

[1] REN21 2019 Renewables 2019 Global Status Report (Paris: REN21 Secretariat)
[2] Antonanzas J and Osorio N, et al. 2016 Review of photovoltaic power forecasting Solar Energy vol 136 pp 78-111
[3] Wang H and Lei Z, et al. 2019 A review of deep learning for renewable energy forecasting Energy Conversion and Management vol 198 p 111799
[4] Zheng L and Liu Z, et al. 2018 Very short-term maximum Lyapunov exponent forecasting tool for distributed photovoltaic output Applied Energy vol 229 pp 1128-113
[5] Das U K and Tey K S, et al. 2018. Forecasting of photovoltaic power generation and model optimization: A review Renewable and Sustainable Energy Reviews vol 81 pp 912-928
[6] Sobri S and Koohi-Kamali S, et al. 2018 Solar photovoltaic generation forecasting methods: A review Energy Conversion and Management vol 156 pp 459-497
[7] Wang F and Zhang Z, et al. 2019 Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting Energy Conversion and Management vol 181 pp 443-462
[8] Lin P and Peng Z, 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 vol 77 pp 704-717
[9] Zhou H and Zhang Y, et al. 2019 Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism IEEE Access vol 7 pp 78063-78074
[10] Qing X and Niu Y 2018 Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM Energy vol 148 pp 461-468
[11] Daniel K and Frank M, et al. 2019 Short-term and regionalized photovoltaic power forecasting, enhanced by reference systems, on the example of Luxembourg Renewable Energy vol 103 pp455-470
[12] Wang F and Yu Y, et al. 2018 Wavelet Decomposition and Convolutional LSTM Networks Based Improved Deep Learning Model for Solar Irradiance Forecasting applied sciences vol 8 p 1286