Prediction of Short-term PV output power Based on PCA-Stacking under different weather conditions

Zhu Xiaofei
Nanjing University of Posts and Telecommunications, Nanjing 210023, China

1870559155@qq.com

Abstract. With the continuous expansion of photovoltaic scale, the accurate prediction of photovoltaic power generation is increasingly important for grid dispatching and grid optimization operations. In this paper, the photovoltaic power generation mainly uses meteorological factors and historical data as the input and output of the neural network. The input quantity is large, the data is redundant, and the network is difficult to converge, which always has a great adverse effect on the accuracy of photovoltaic output prediction. Firstly, different weather types are classified according to the trend graphs of different weather types. Principal components analysis (PCA) is used to analyze less comprehensive features from multiple meteorological factors and reduce the input of predictive models. At the same time, aiming at the problem that the prediction accuracy of a single prediction model such as the existing neural network and wavelet analysis method is limited, the idea and method of integrated learning are introduced, and a short-term prediction method based on Stacking method combined with SVM and Xgboost is proposed. Compared with the single model of SVM and Xgboost, the results show that the proposed method has a significant improvement compared with the accuracy of a single prediction model.

1. Introduction
In recent years, as China attaches great importance to energy conservation and environmental protection, the demand for renewable energy is increasing. Due to its sufficient cleanliness and safety, solar energy makes solar energy one of the fastest growing renewable energy sources in China. Photovoltaic power generation has become a research hotspot in the energy field and has developed rapidly. Meanwhile, with the persistent progress of solar power generation technology and the rapid growth of global installed capacity of photovoltaic power generation, grid-connected solar power generation technology has become an important development direction of solar photo-voltaic power generation [1]. However, photovoltaic power generation exhibits time-varying, volatility and randomness due to objective factors such as solar irradiance and weather, and its uncertainty brings many difficulties to the power grid on safety and economic operation of the grid [2,3]. Accurate prediction of photovoltaic power generation is of great significance for grid dispatching and optimization operations. In term of different classification criteria, the prediction methods of the photovoltaic output power are different [4]. According to the classification of prediction process, the PV power prediction method can be
divided into direct prediction and indirect prediction; according to the modeling method, it can be divided into physical method prediction and statistical method prediction. Currently, neural networks [5-7], such as Elman neural network [8], support vector machines (SVM) [9], autoregressive moving average models (ARMA) [10], Markov chains and others are the most widely used statistical learning methods. According to different prediction time scales [11], classify photovoltaic power prediction methods. They can be divided into ultra-short-term (0 ~ 6 h), short term (6h-1d) [12], long term (one month-one year). Ultra-short-term PV power prediction can provide power transient information; short-term prediction can be used in scheduling, load tracking and forecasting, power market, etc. Medium and long-term PV power forecasting can be used in optical resource assessment, new photovoltaic power plant planning and other fields.

At present, most of the researches on photovoltaic power generation mainly focus on solar energy monitoring system, short-term prediction of the photovoltaic power generation and performance prediction of grid-connected photovoltaic systems. Voyant et al. used exogenous meteorological data as the optimized multi-layer perceptron (MLP) inputs to forecast daily global horizontal irradiance. Through comparing with different forecasting models containing a persistence model, ARIMA and ANN, they claimed the advantages of the proposed methodology [13]. Chen et al. presented an application of SVM to forecast daily solar irradiation using sunshine duration in Liaoning province of China. The results illustrated that the SVM method could outperform the traditional forecasting methods [14]. Azimi et al. presented a hybrid solar irradiance forecasting framework combined with a novel cluster selection algorithm, a time-series analysis and MLP neural network for different time horizons [15]. Baser and Demirhan developed an approach using fuzzy regression functions with support vector machine (FRFSVM) to estimate yearly mean daily global horizontal solar irradiance in Turkey [16].

In summary, the existing models have the defects of data processing difficulties, complex model structure, and poor dynamic system identification. The prediction error of machine learning method is about 8% [17], the cloud error is about 26.20%, and the rainy day error is about 43.05% [18]. In the above study, the weather forecast method is limited by the accuracy of the weather forecast system, and the forecasted weather. When the parameter error is large, the prediction accuracy will be seriously affected. The machine learning method uses BP neural network to face the problems of local optimization, slow iterative convergence, BP and SVM and its combination model. It is essentially based on single model prediction, and its prediction accuracy is limited, and there is a big improvement space.

Aiming at the existing problems, this paper proposes a short-term output combination forecasting method based on PCA-stacking photovoltaic power station. Considering the non-stationarity of the hourly output power distribution of photovoltaic power plants in the case of sudden weather changes, the data set is first divided by the trend graph analysis of power generation and different weather, and the data is processed by principal component analysis. Feature extracts feature components, compresses data dimensions, reduces data redundancy, obtains principal component components, and introduces the idea of integrated learning. Using Stacking method to combine SVM and Xgboost, use multiple primary learners SVM to predict the predicted samples. Then "generate" a new data set for training the secondary learner Xgbboost. In this new data set, the output of the primary learner is used as a sample input feature, and finally the secondary training obtained by the secondary Xgbboost for the SVM. Set training to get the final prediction result.

2. Prediction model algorithm

2.1. Principal component analysis model

The principal component analysis method is one of the most commonly used dimensionality reduction methods. The linear transformation transforms the original data into a set of linearly independent
representations of each dimension, which can be used to extract the main feature quantities of the data, which is often used for dimensionality reduction of high-dimensional data. Keep the feature of the worst contribution of the data set to the greatest extent. The method is mainly to decompose the covariance matrix to obtain the principal components (feature vectors) of the data and their weights.

The principal component expression is:

\[
g_1 \Delta 77/g_2 \Delta 6 \approx \frac{g_1 \Delta 57}{g_2 \Delta 6} \approx \frac{g_2 \Delta 70}{g_1 \Delta 76/g_2 \Delta 70} + \frac{g_1 \Delta 57}{g_2 \Delta 71/g_1 \Delta 76/g_2 \Delta 71} + \cdots + \frac{g_1 \Delta 57}{g_303 \Delta 70} + \frac{g_1 \Delta 57}{g_2 \Delta 70} \frac{g_3040}{g_1 \Delta 76/g_3040} \\
\]

The paper is the k-dimensional eigenvector corresponding to the i-th eigenvalue of the correlation matrix of the original variable, X is the initial input variable of the k-dimensional and corresponding feature vector \(e_1, e_2, \ldots, e_k\).

2.2. Stacking algorithm

When there is a lot of training data, a more powerful combination strategy is to use the "learning method", which is a combination of another learning device. Stacking is a typical representative of learning method. Here we call the individual learner the primary learning device. The learner used for the combination is called a secondary learner or a meta-learner.

Stacking first trains the primary learner from the initial dataset and then "generates" a new dataset for training the secondary learner. In this new dataset, the output of the primary learner is used as a sample input feature, and the initial The mark of the sample is still treated as a sample mark. Its specific process is shown in Figure 1.

![Figure 1. Stacking forecasting process](image)

2.3. Support vector machine

Support vector machine (SVM): The derivation process of the support vector machine is: given training sample \(D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, y_i \in \mathbb{R}, i=1,2,\ldots,n\). Through a linear regression function:

\[
f(x) = w^T x + b
\]

To fit the sample, use the \(\varepsilon\)-insensitive loss function:

\[
|y - f(x)|_\varepsilon = \begin{cases} 0, & |y - f(x)| \leq \varepsilon \\ |y - f(x) - \varepsilon|, & |y - f(x)| > \varepsilon \end{cases}
\]

Introducing the regularization constant and the relaxation variable and , the optimization problem is:
Introduce the Lagrange multiplier, convert it into a dual form to get the optimal solution. The regression function is:

\[ f(x) = \sum_{i=1}^{m} \left( \hat{\alpha}_i - \alpha_i \right) \kappa(x, x_i) + b \] (5)

Where \( \kappa(x, x_j) = \emptyset(x_j)^T \emptyset(x_j) \) is a kernel function.

### 2.4 Xgboost

Xgboost has received a lot of attention due to its excellent learning effect and effective training speed. The derivation process of Xgboost is: for a given training set \( T = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\} \), define an objective function:

\[ \text{Obj}(t) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + f_t(x_i) + \Omega(f_t) + \text{constant} \] (6)

Constant is a constant, and the regular term \( \Omega(f_t) \) is as follows:

\[ \Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{r} w_j^2 \] (7)

Where \( T \) represents the number of leaf nodes and \( w_j \) represents the weight of the \( j \)-th leaf node. Use the Taylor expansion \( f(x + \Delta x) \approx f(x) + f'(x) \Delta x + \frac{1}{2} f''(x) \Delta x^2 \) to expand equation (6):

\[ \text{Obj}(t) \approx \sum_{i=1}^{n} \left[ L(y_i, \hat{y}_i) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{constant} \] (8)

Where \( g_i \) denotes the first derivative of \( L(y_i, \hat{y}_i) \) for \( \hat{y}_i = 1 \), \( h_i \) denotes \( L(y_i, \hat{y}_i) \). The second derivative of \( \hat{y}_i - L(y_i, \hat{y}_i) \) is known as the real value and the residual calculated by the previous function. So the final result is:

\[ \sum_{i=1}^{n} \left[ \sum_{j=1}^{T} G_j w_j + \frac{1}{2} H_j \lambda w_j^2 \right] + \gamma T + C \] (9)

By deriving equal to 0. Can get:

\[ w_j = -\frac{G_j}{H_j + \lambda} \] (10)

The simplified formula for bringing \( w_j \) into the objective function is as follows:

\[ \text{Obj}(t) = -\frac{1}{2} \sum_{j=1}^{T} G_j^2 + \gamma T + C \] (11)

After the objective function is simplified, you can see that the objective function of Xgboost is customizable, and only the first and second derivatives of it are used in the calculation. After obtaining the simplified formula, the next step is to select the appropriate splitting feature for the gain of the selected feature.

\[ \text{gain}(\Phi) = \text{gain(before)} - \text{gain(after)} = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \] (12)
3. Predictive model construction analysis
First, the training data is divided into four data sets, such as sunny, cloudy, cloudy and rainy days, according to the characteristics of photovoltaic power generation, and then input into the principal component analysis model (PCA) for dimensionality reduction, and the dimensionality-reduced data is used as the prediction sample. Enter 5 primary learning SVMs and use the secondary learner Xgboost to combine the predicted outputs of the primary SVM to arrive at the final prediction. The specific process is shown in Figure 2.

The prediction process of the Stacking learner is shown in Figure 1. The prediction results are input into n primary SVM, and the primary SVM is combined using the secondary learner Xgboost to obtain the final prediction result.

3.1. Trends in photovoltaic power generation analysis and training sample composition
Photovoltaic power generation is affected by weather, environment and other factors with randomness and volatility, especially under cloudy and rainy weather conditions. The power generation power curve fluctuates drastically, and multiple peaks are generated, which has a great influence on the prediction accuracy of photovoltaic power generation. The sunny day is relatively stable, the curve is gentle, and it is easy to obtain better prediction accuracy, and the power generation on sunny days is much larger than the extreme weather. In cloudy weather, the difference from the sunny curve fluctuated significantly in different time periods, but we finally found that in cloudy weather, the power generated is very close to that of sunny days.

![Figure 2. Overall workflow](image-url)
3.2. PCA analysis of meteorological factors

Taking the sunny photovoltaic power generation prediction sub-model as an example, the principal component analysis is carried out on eight original meteorological factors (ambient temperature, environmental humidity, irradiance, wind speed, wind direction, plate temperature, temperature difference, and air pressure) that affect photovoltaic power generation. The contribution rate and cumulative contribution rate of the principal components are shown in Table 1. As can be seen from Table 1, when the principal component is 3, a significant inflection point appears, but the cumulative contribution rate at this time is less than 85%, so four of them are selected. The main component, and a new input variable is calculated using equation (1). Input variables are entered into the Stacking learner for prediction.

| Serial number | Initial eigenvalue | Contribution rate% | Grand total% |
|---------------|-------------------|--------------------|--------------|
| 1             | 5.34              | 35.20              | 35.20        |
| 2             | 3.25              | 26.13              | 61.33        |
| 3             | 1.68              | 15.37              | 76.70        |
| 4             | 1.27              | 13.12              | 89.82        |
| 5             | 0.46              | 5.15               | 94.97        |
| 6             | 0.27              | 3.43               | 98.40        |
| 7             | 0.05              | 0.84               | 99.24        |
| 8             | 0.03              | 0.76               | 100%         |

3.3. Data normalization and model evaluation indicators

In this paper, the meteorological data and irradiance data of a PV power plant in Nanjing are used as input samples. The data collection interval is 15 minutes. The normalization method used in this paper will be used to transform the data into intervals. The normalization function is shown in equation (13):

\[
\bar{x}_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} x_i
\]

In this paper, MAE (mean absolute error) is used to estimate the accuracy of the model. The expression is shown in equation (14):

\[
e_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^{N} |y_{\text{model},i} - y_{\text{actual},i}|
\]

Where: \( y_{\text{model},i} \) is the predicted value of the model, which \( y_{\text{actual},i} \) is the measured value.

4. Case analysis

On a sunny day, the amount of clouds is small, and the meteorological conditions such as irradiance and temperature are relatively stable, so the photovoltaic power generation fluctuation is small. Figure 3 shows the comparison of prediction results based on PCA-Stacking model and PCA-SVM, PCA-Xgboost, Xgboost and other models on sunny days. The prediction results are shown in Fig. 3. The power generation of the model in this model is very close to the real value, and it is not much different from the PCA-SVM, Xgboost, PCA-Xgboost and other models.
Under cloudy conditions, illumination has become an important factor affecting photovoltaic power generation. Due to the influence of wind power, the thickness and position of clouds, fluctuations in photovoltaic power generation have caused fluctuations. Figure 4 shows the comparison between the model and PCA-SVM, Xgboost, and PCA-Xgboost in cloudy forecast. It can be seen that there are multiple peaks in the real value, and each model has a large deviation at the peak. While other times performed better. Overcast and rainy days have less exposure time, so the amount of electricity generated is less than sunny and cloudy. Figures 5 and 6 show the predictions of different models of the two weathers. It can be seen that the predicted values of the model in this model are more realistic. In rainy and bad weather conditions, the solar radiation is very small, so the amplitude of the photovoltaic output power is very small, too.

Table 2 shows the MSE (mean absolute error) values of the four models of PCA-Stacking, PCA-Xgboost, PCA-SVM, and xgboost in four different weather conditions.
Table 2. Four model MSE values

| Model       | Sunny | Cloudly | Overcast | Rainy |
|-------------|-------|---------|----------|-------|
| Pca-stacking | 0.220 | 0.679   | 0.231    | 0.205 |
| Pca-xgboost  | 0.282 | 0.683   | 0.323    | 0.216 |
| Pca-Svm      | 0.339 | 0.709   | 0.357    | 0.269 |
| Xgboost      | 0.429 | 0.759   | 0.442    | 0.337 |

From the comparison of Table 2, it can be seen that the PCA-Stacking proposed in this paper has higher precision on sunny days and cloudy days. In the rainy days, it is similar to the prediction performance of other models. At the same time, from the performance of the Xgboost model, the prediction error is worse than other models due to the lack of dimensionality reduction.

Based on the above, the model is superior to other models in the prediction of four weather conditions, and has good prediction accuracy for the extreme weather such as overcast and rainy.

References
[1] Tao Ma, Hongxing Yang, Lin Lu, Jinqing Peng, Pumped storage-based standalone photovoltaic power generation system: Modeling and techno-economic optimization, Applied Energy, Volume 137, 2015, Pages 649-659.
[2] D.O. Akinyele, R.K. Rayudu, N.K.C. Nair, Development of photovoltaic power plant for remote residential applications: The socio-technical and economic perspectives, Applied Energy, Volume 155, 2015, Pages 131-149.
[3] Arnau González, Jordi-Roger Riba, Antoni Rius, Rita Puig, Optimal sizing of a hybrid grid-connected photovoltaic and wind power system, Applied Energy, Volume 154, 2015, Pages 752-762, S.I. Sulaiman.
[4] Wang Fei, Mi Zengqiang, Yang Qi-xun, et al. Power forecasting approach of photovoltaic plant based on ANN and relevant data. J Sol Ener 201233(7):1172
[5] T.K.A. Rahman, I. Musirin, and S. Shaari, “Performance Analysis of Evolutionary ANN for Output Prediction of a Grid-Connected Photovoltaic System,” World Academy of Science, Engineering and Technology, pp. 1023-1029.
[6] Lee, D., & Kim, K. (2019). Recurrent Neural Network-Based Hourly Prediction of Photovoltaic Power Output Using Meteorological Information. Energies, 12(2), 215.
[7] Wang Fei, Mi Zengqiang, Yang Qi-xun, et al. Power forecasting approach of photovoltaic plant based on ANN and relevant data. J Sol Ener 201233(7):1127–7.
[8] Zhong Chunxia. Output power prediction of a photovoltaic system based on similar day algorithm and Elman neural network. J Nanjing Inst Technol (Nat Sci Ed) 2016;14(1):42–7
[9] Fei Wang, Zhao Zhen, Zengqiang Mi, Hongbin Sun, Shi Su, Guang Yang, Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting, Energy and Buildings, Volume 86, 2015, Pages 427-438.
[10] Y. Sun et al., "Short-term prediction model of module temperature for photovoltaic power forecasting based on support vector machine," International Conference on Renewable Power Generation (RPG 2015), Beijing, 2015, pp. 1-6.
[11] Shuang Han. Study on short-term forecasting method of wind power in wind farm. North China Electric Power University; 2008.
[12] Stanley KH, Chow, Lee Eric WM, Li Danny HW. Short-term prediction of photovoltaic energy generation by intelligent approach. Energy Build 2012;55:660–7.
[13] Voyant C, Muselli M, Paoli C, Nivet ML. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy 2011;36(1):348–59
[14] Chen JL, Li GS, Wu SJ. Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy Convers Manage 2013;75:311–8.

[15] Azimi R, Ghayekhloo M, Ghofrani M. A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. Energy Convers Manage 2016;118:331–44.

[16] Lima FJL, Martins FR, Pereira EB, Lorenz E, Heinemann D. Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks. Renew Energy 2016;87:807–18.

[17] LEVA S, DOLARA A, GRIMACCIA F, et al. Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power[J]. Mathematics and Computers in Simulation, 2017, 131: 88-100.

[18] T. Yamazaki, S. Wakao, Y. Fujimoto and Y. Hayashi, "Improvement of prediction interval estimation algorithm with just-in-time modeling for PV system operation," 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), New Orleans, LA, 2015, pp. 1-6.
