Application of clustering analysis in the prediction of photovoltaic power generation based on neural network

K Cheng[1,2], L M Guo[1], Y K Wang[1] and M T Zafar[1]

[1]School of Power and Energy, North-western Polytechnical University, Xi'an 710072, Shaanxi province, China

E-mail: cksolar@163.com

Abstract. In order to select effective samples in the large number of data of PV power generation years and improve the accuracy of PV power generation forecasting model, this paper studies the application of clustering analysis in this field and establishes forecasting model based on neural network. Based on three different types of weather on sunny, cloudy and rainy days, this research screens samples of historical data by the clustering analysis method. After screening, it establishes BP neural network prediction models using screened data as training data. Then, compare the six types of photovoltaic power generation prediction models before and after the data screening. Results show that the prediction model combining with clustering analysis and BP neural networks is an effective method to improve the precision of photovoltaic power generation.

1. Introduction

With the expansion of installed capacity of photovoltaic power generation, its proportion in the power grid is also increasing these days. The random fluctuation characteristics of its output will cause the considerable impact on the power grid system, which indirectly affects the operation of the security and stability of the power system. Therefore, it is extremely important to predict the photovoltaic power generation effectively to ensure the security and stability of the power system. At present, some forecasting models for photovoltaic power generation are SVM model [1,2], ARMA model [3], gray theoretical model [4] and neural network model, etc. [5,6]. However, the neural network model generally has better performance among them [7]. Therefore, this study uses this method to predict the photovoltaic power generation.

In the process of BP neural network prediction, the quality of the input samples will directly affect the accuracy of the prediction model. Practically, when the photovoltaic power plant collects and stores the running data, the errors occur which form the abnormal samples. These anomaly samples deviate from the factual basis and the corresponding weather patterns are not correctly described according to its established model. Therefore, it is necessary to screen out these anomalous samples.

Manually, it is very difficult to screen out anomalous samples because of a large amount of data recorded in the past years. Therefore, most of the research papers use raw data collected from the untreated power plants [4-8], and only a few papers deal with filtered data [9]. However, these papers do not focus on the specific screening method. There are some papers, which use the wavelet analysis method, the mean fill method and other methods to deal with the original data [10,11]. However; there is no specified method to verify the accuracy and effectiveness of the power generation forecasting model.

Therefore, based on the statistical data of different kinds of weather, this paper cleans the historical data using the clustering analysis. Then, based on the BP neural network, 6 different kinds of photovoltaic power generation forecasting models are established and verified in the sunny, cloudy and rainy days in MATLAB software [12,13]. The results show the validity and accuracy of the clustering analysis of historical data.

2. Prediction model of photovoltaic power generation

2.1. Prediction model of photovoltaic power generation

In this paper, the clustering analysis and BP neural network are combined to process the raw data collected by the power station. The processed data is used as the input of the neural network to establish a reliable photovoltaic power generation forecasting model. The data in the research process comes from the actual record of the data collection device of Xi'an Jiayang New Energy Co., Ltd. PV power plant. The model is designed based on the photovoltaic system installed on the roof of the building. The total installed capacity of the photovoltaic power station is 1.2 MW. The measured data of the power station in the time period of 2012.01-2015.01 is selected as the original sample. The selected samples consist of total 868 data samples.

The main process of the model is shown in figure 1.

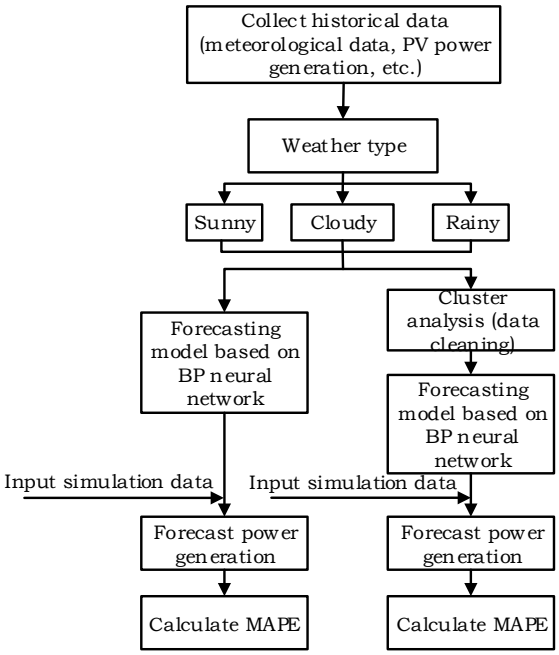

Figure 1. The framework of photovoltaic power generation forecasting model.

- Firstly, 3 typical weather types of sunny, cloudy and rainy days are taken as examples to analyze the theoretical radiation, actual radiation and power generation. The outliers are filtered out according to the cluster pedigree.
- Secondly, the historical data of 2012.01~2014.01 is used as the training sample. The theoretical radiation and the highest and the lowest temperatures are used as the input of BP neural network. The historical data collected by the data collector of PV power plant are used as the output of BP neural network. The neural network model is trained to establish the prediction models of 6 kinds of photovoltaic power generation
- Thirdly, the simulation data of 2014.02 ~ 2015.01 is put into the prediction model of photovoltaic power generation. The purpose is to make use of the BP neural network's

nonlinear learning ability to predict the daily power generation under different weather types

- Finally, in order to verify the need of data screening, the MAPE (the mean absolute percentage error) value is calculated for the predicted power generation and the actual generation and the performance is evaluated.

2.2. Clustering analysis model

Clustering analysis is a process of dividing an object of observation into several groups or classes based on certain quantitative features. The division is done in such a way that the same class within the data object has a high degree of similarity and low degree of similarity between various types of data object. The main purpose of clustering is to bring similar things together.

2.2.1. Clustering method. According to the different clustering methods, clustering analysis can be divided into system clustering method, adding method, decomposition method and dynamic classification method, etc. [14]. In this paper, the purpose of clustering analysis is to deal with the outlier data (abnormal samples) from the original data to improve the quality of the original data. Therefore, this paper uses the system clustering method. The main algorithm steps are as follow:

- To convert the objects in the sample set;
- The processed objects are divided into n classifications and every class contains an object;
- Calculating the distance between each of the two classes;
- Merging the nearest two classes as a new class;
- Calculating the distance between the new class and the current class. If the distance is close enough, then merge into a class. This process continues until all classes (or samples) are combined. If this step works well then we will move further to step 6, otherwise, to step 4.
- Drawing the clustering pedigree of the sample set, as shown in figure 2;
- According to the number of classes to be classified, the corresponding classification results are obtained.

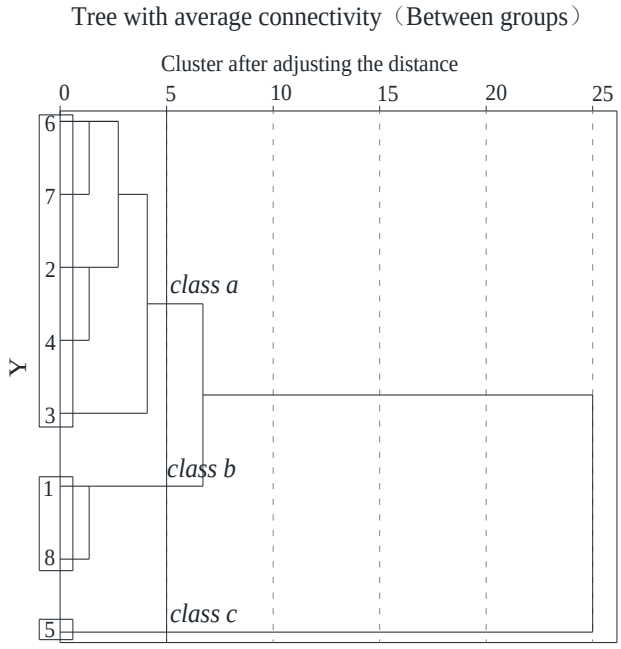

Figure 2. Clustering pedigree chart.

Figure 2 shows the data clustering results for randomly selected eight groups on rainy day. The sample can be clustered into two, three, four or five categories. As shown in the figure, when the class

distance is 5, the samples can be clustered into three classes. The samples 6, 7, 2, 3 and 4 are class a, sample 1 and sample 8 are b, and sample 5 is c.

2.2.2. Select the metric. The metric directly affects the quality of the cluster, so the statistical analysis of the historical data of the database is carried out to determine the metric. In the figures 3-5 of cloudy weather, some samples are found away from the main data sets which are outliers. While, the data distribution exhibits a distinct oblique strip.

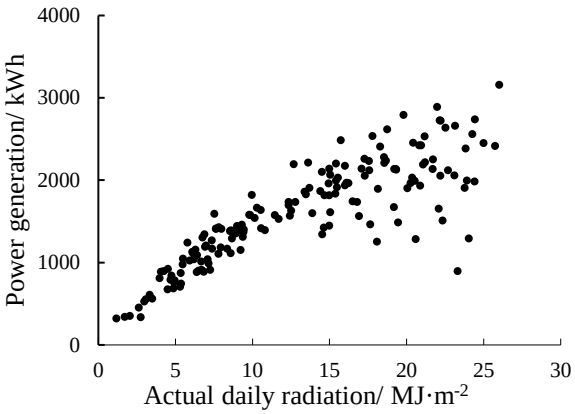

Figure 3. Scatter plots of actual radiation and power generation.

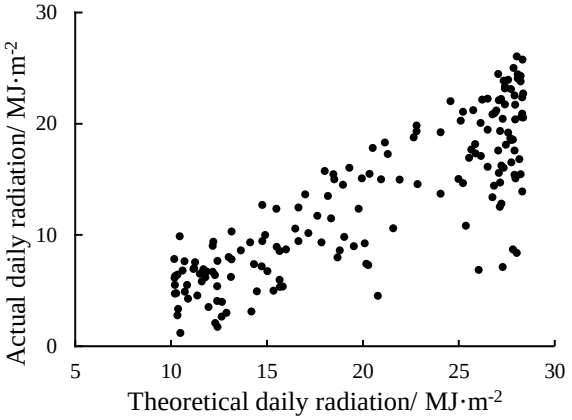

Figure 4. Scatter plots of theoretical and actual radiation.

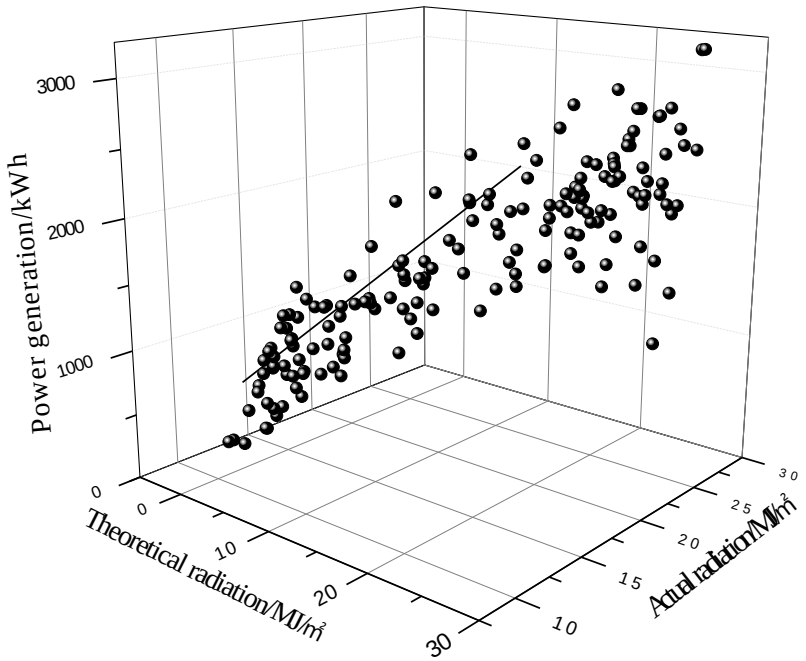

Figure 5. Spatial scatter plots of actual radiation, theoretical radiation and power generation.

According to the previous study results [15], the data covered by the oblique range belong to a class. Selecting the cosine coefficient as the criteria, the data on the oblique bar is classified as valid data and other data is classified as the outlier which is far from the oblique bar. After the outlier deletion, the discrete degree of the sample is optimized. The cosine distance is measured according to

formula (1):

$$d_{ij} = \cos(\theta_{ij}) = \frac{\sum_{k=1}^{0} X_{ik} X_{jk}}{\sqrt{\sum_{k=1}^{n} X_{ik}^2 \sum_{k=1}^{n} X_{jk}^2}} \quad (1)$$

In the above formula, d_{ij} -the distance between simples X_{ik} and X_{jk}; X_{ik} - the k-the variable of the *i* sample; X_{jk} - the k-the variable of the *j* sample; *n* - number of samples.

2.3. Neural network model

Fundamentals. The neural network is a network system that performs the parallel processing and non-linear transformation of information same as human brain. The BP neural network chosen in this paper is a forward process using error back propagation algorithm of the network. Its structure is shown in figure 6.

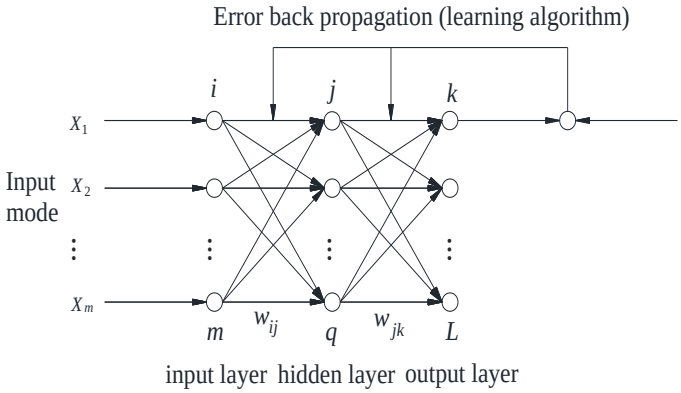

Figure 6. Structure of BP neural network.

Where w_{ij} is the connection weight between the input layer node and the hidden layer node, and w_{jk} is the connection weight between the hidden layer node and the output layer node. The input of the hidden layer and the output layer node is the weighted sum of the output of the previous node [16]. In this paper, a three-layer neural network is used which contains only one hidden layer.

Evaluation of photovoltaic power generation forecasting model. In this paper, a two-layer forward neural network based on Levenberg-Marquardt (LM) algorithm is adopted. The neural network toolbox of Matlab 2010b is used to build and train the network and obtain the simulation results. The training data and test data are randomly selected by neural network as 90% and 10% respectively of the total data. In order to ensure the credibility of the results, test data is not used in the training part.

The simulation results use the average absolute percentage error (MAPE) to evaluate the entire system, which is calculated as (2):

$$MAPE = \left(\frac{100}{N}\right) \sum_{i=1}^{N} \left| \frac{P_f{}^i - P_a{}^i}{P_a{}^i} \right| \% \quad (2)$$

In the above formula, *N* - the total number of data; P_f - the actual power generation; P_a - predicted power generation; *i* - the data serial number. The lower MAPE value ensures the better predictive effects.

3. Clustering process and results

The original data of three kinds of weather types such as sunny, cloudy and rainy days are extracted by SPSS (statistical analysis software). The cosine coefficient is used as the criteria to cluster and analyze the different weather types respectively. When the number of clusters is 4, for sunny days, according to the clustering pedigree, a_s class has 119, b_s class has 23, c_s class has 5, and d_s class has 3; for rainy days, according to the cluster spectrum, a_r class has 4, b_r class has 2, c_r class has 19, and d_r class has 36; for cloudy days, a_c class has 10, b_c class has 30, c_c class has 61, and d_c class has 67, comparing the characteristics of the number of samples contained in each cluster and screening a few samples with the least number of clusters. Figure 7 shows the filtered cloudy sample scatter plots.

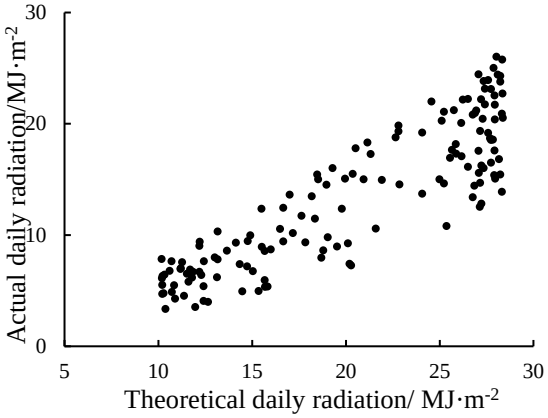

Figure 7. The theoretical and actual daily radiation scatter plots after screening in cloudy.

The clustering results of the various data samples are shown in table 1. In the sunny days, the atmospheric state is relatively stable and the reverse radiation of the atmosphere is weak. Therefore, the theoretical daily radiation is close to the actual amount of radiation in the whole day, and the degree of data discrimination is small and the screening ratio is 10.7%. In the cloudy days will have large random variation in terms of clouds and the similarity of the data is low, the screening ratio is 13.1%. In the rainy days, the atmospheric state fluctuates at a higher degree and the screening ratio is also large which is 21.3%.

Table 1. The clustering result statistics for all kinds of sample data.

Weather type	Sunny	Cloudy	Rainy	Total
Before screening	150	168	61	379
After screening	134	146	48	328
Screening ratio	10.7%	13.1%	21.3%	13.5%

4. Predictive results and analysis

4.1. Forecast results

In order to verify the effectiveness of clustering analysis for the prediction of photovoltaic power generation, the prediction model of PV generation is established with raw data as a comparison. Firstly, cluster analysis is used in data cleaning for different kinds of weather. Then, corresponding type of model is selected from forecasting models and simulation data is entered. Finally, the power generation is predicted.

Figures 8-10 shows the predicted power generation, actual power generation, and the data for the

three kinds of weather before and after the clean-up process (the day number in the x axis is the data of the forecast in a year). From the results, it can be seen that the difference between the predicted value and the actual value of the model established by the weather type is relatively large before the data is cleaned up. After data clearing, the difference between the predicted value and the actual value become small. The predicted value of the model after the cluster analysis is closer to the actual value.

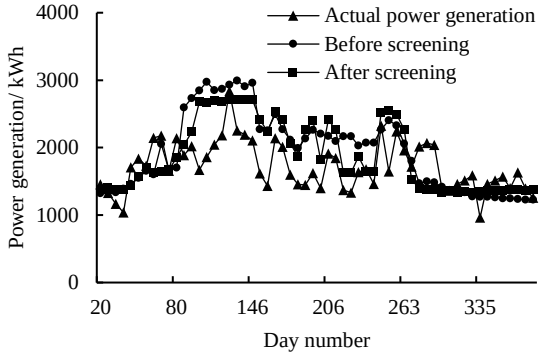

Figure 8. The relationship between the actual power generation and the prediction of power generation in sunny weather.

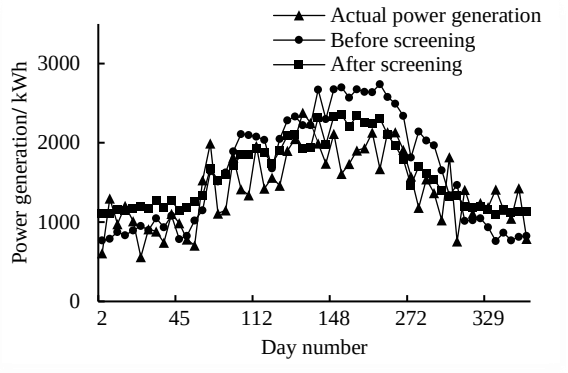

Figure 9. The relationship between the actual power generation and the prediction of power generation in cloudy weather.

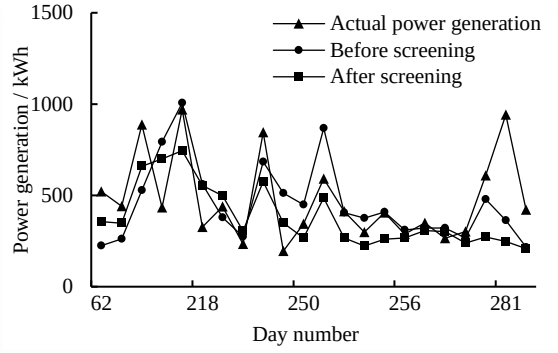

Figure 10. The relationship between the actual power generation and the prediction of power generation in rainy weather.

4.2. Evaluation of the results

Table 2. MAPE values of various weather neural network models.

	Weather type	sunny	cloudy	rainy	average value
Prediction error	Before screening	32.4%	45.6%	85.8%	46.8%
	After screening	20.7%	22.6%	34.5%	23.6%

Table 2 shows the difference between the predicted results of the six neural network models. Compared with the MAPE mean value, it can be seen that the predicted value of the model after the cluster analysis is closer to the actual value in each case. The average MAPE value of the model after clustering analysis is 23.2% which is much smaller than the conventional models. Indicating that the

clustering analysis is effective, and the prediction model of photovoltaic generation is more accurate base on the data after clustering analysis. Therefore, it is necessary to screen data.

In the case of sunny days, the prediction results of two models are closer, because the sunny weather atmospheric state is relatively stable and the degree of data discrete is small, the MAPE value of the prediction model that after screening is 11.7% lower than that before screening, the prediction effect is better. In the case of cloudy days, the thickness and location of the clouds is difficult to predict, the predicted results of the two models are relatively different, the MAPE value of the prediction model that after screening is 23% lower than that before screening. In the case of rainy days, the MAPE value of the prediction model that after screening is 51.3% lower than that before screening, which difference is the biggest. It can be seen from the table 2, the prediction of rainy days is far less accurate than sunny and cloudy days. The reason is the rain weather atmospheric changes are more complex, in different rainy days, clouds, rain and vapor will be very different, and the absorption of solar radiation will be very different, this will produce a large relative error, so it is difficult to achieve the desired prediction effect in rainy days.

Based on overall assessment, the MAPE value of the predicted model is 23.6% which is less than 46.8% of the pre-screening prediction model. The comparison results show that photovoltaic power generation models are more accurate when using the clustering analysis of the original data. Therefore, it is necessary to carry out data cleanup.

5. Conclusion

In this paper, the prediction model is established by combining the clustering analysis and the BP neural network. The abnormal samples in the historical monitoring data of the PV power plant are screened out, and the prediction results are evaluated and analyzed. Research indicates:

- Using the cosine coefficient as the evaluation criteria, the abnormal samples are removed using clustering analysis and screening of historical data, which can effectively filter the discrete operating data.
- After the data is cleaned up as the training data, the BP neural network prediction models can be established which can improve the accuracy of the prediction results.
- The data cleaning is necessary for establishing the prediction model using the historical data. The cluster analysis can be effectively applied to the photovoltaic power generation forecasting system.

The above research results can provide an effective method for the forecasting technology of PV power plants, which is beneficial to the popularization and application of photovoltaic power generation technology.

References

[1] Li R and Li G M 2008 Photovoltaic power generation output forecasting based on support vector machine regression technique *Electr Pow* **41(2)** 74-8
[2] Shi J, Lee W J, Liu Y Q, *et al* 2012 Forecasting power output of photovoltaic systems based on weather classification and support vector machines *IEEE Trans Ind Appl* **48(3)** 1064-9
[3] Lan H, Liao Z M and Zhao Y 2011 ARMA model of the solar power station based on output prediction *Electr Meas Instrum* **48(54)** 31-4
[4] Wang S X and Zhang N 2012 Short-term output power forecast of photovoltaic based on a grey and neral network hybrid model *Autom Electr Pow Syst* **36(19)** 1-5
[5] Christophe P, Cyril V, Marc M *et al* 2010 Forecasting of preprocessed daily solar radiation time series using neural networks *Sol Energy* **84(12)** 2146-60
[6] Zhang Y X and Zhao J 2011 Application of recurrent neural networks to generated power forecasting for photovoltaic system *Pow Syst Prot Control* **39(15)** 96-101
[7] Şenkal O and Kuleli T 2009 Estimation of solar radiation over Turkey using artificial neural network and satellite data *Appl Energ* **86(7-8)** 1222-8
[8] Wang F, Mi Z Q, Zhen Z *et al* 2013 A classified forecasting approach of power generation for

photovoltaic plants based on weather condition pattern recognition *Proc CSEE* **33(34)** 75-82
[9] Lu J, Zhai H Q and Liu C 2010 Study on statistical method for predicting photovoltaic generation power *East China Electr Pow* **38(4)** 563-7
[10] Lang Y and Zhang W T 2016 Research on the application of MATLAB in the data processing of photovoltaic power generation *Electr Technol Softw Eng* **2016** 87
[11] Zhang H N, Zhang J T, Yang L B *et al* 2016 Date processing method for PV power station data acquisition system *Electr Energ ManageTechnol* **2016** 9-13
[12] Chen C S, Duan S X, Cai T *et al* 2011 Online 24-h solar power forecasting based on weather type classification using artificial neural network *Sol Energy* **85(11)** 2856-70
[13] Dai Q, Duan S X, Cai T *et al* 2011 Short-term PV generation system forecasting model without irradiation based on weather type clusting *Proc CSEE* **31(34)** 28-35
[14] Wu S, Pan F M *et al* 2014 *SPSS Statistical Analysis* (Tsinghua University Press) pp 315-30
[15] Mohamed A and Badrul C 2015 Solar power forecasting using artificial neural networks *Proc of The 47th Annual North American Power Symp* (USA: Charlotte, North Carolina)
[16] Chen C S, Duan S X and Yin J J 2009 Design of photovoltaic array power forecasting model based on neutral network *Trans China Electrotech Soc* **24(9)** 153-8