Short-term prediction model of photovoltaic power generation based on rough set-BP neural network

Xiaoling Ma¹, Qingle Pang²* and Qingsong Xie²*

¹ School of Information and Electronic Engineering, Shandong Institute of Business and Technology, Yantai, Shandong Province, 264003, China
² School of Information and Control Engineering, Qingdao Technological University, Qingdao, Shandong Province, 266400, China
*Email of the corresponding author: 2018420082@sdtbu.edu.cn

Abstract. It is very important to predict the photovoltaic power generation because of the great challenges to the safe and stable operation of the power grid. Based on the analysis of traditional forecasting models, a power generation forecasting model based on rough set and neural network is proposed. Firstly, the Hampel filtering algorithm and rough set theory are used to process outliers and redundant data for the collected information. The above algorithms solve the problem of big data processing before forecasting. Finally, the fuzzy C-means clustering was used to divide the data set and combined with the neural network to predict. The key weather feature variables combined with the forecast time were taken as the input and the power was taken as the output variable. The results show that the prediction method proposed in this paper is more accurate and fast in predicting the power generation at each time.

1. Introduction
In recent years, more and more researchers use the method of preprocessing sample data before forecasting. In terms of the treatment of outliers, the current methods include the combination of quant method [1], the optimal in-group variance cleaning algorithm [2], and Isolation Forest[3]. Literature [4] can effectively detection and carry data anomaly out by using Hampel filtering algorithm, and has achieved good outliers removal effect for different data sets. In terms of feature extraction, this paper uses rough set theory to calculate the influence degree of each influencing factor, and finds out the key features that have influence on the actual prediction, which has higher objectivity and credibility.

Among the prediction models, there are mainly the generalized weather type method [5], that is, the classification of weather types. At present, most of the researches mainly rely on the collection of weather characteristic information of photovoltaic power stations [6-8], and almost all ignore the time factor. When the prediction result is used as a reference for grid dispatching, we are more inclined to grasp the photovoltaic power generation at each moment. Therefore, this paper uses fuzzy C-means clustering algorithm to classify data according to time period. Finally, BP neural network is combined to build models respectively according different time period, and then the prediction results of each time period are fitted into the prediction results of one day. In the experimental part, the data of the No. 5 photovoltaic power station of the Solar Energy Intellectual Property Center in Alice Spring, Australia is used to perform an experiment, the results are compared with the weather type prediction method shows that the prediction model proposed in this paper is more accurate and faster, thus verifying the effectiveness of the model method.
2. Data processing and dimensionality reduction

2.1. Hample filter based outlier processing
The Hampel filter like a median filter\cite{9}, but it only replaces values equivalent to a few standard deviations away from the local median. This filter helps remove outliers from the data without over-smoothing the data.

2.2. Attribute reduction based on rough set theory
Rough set theory defines the whole composed of sample set, attribute set and attribute range as knowledge representation system. In system S, if \( C \) is a conditional attribute set and \( D \) is a decision attribute set, the system is called a decision table\cite{10-11}. Rough set knowledge reduction method can be used to determine which attributes are necessary, and then select several attributes of greater importance.

3. Model based on rough set - fuzzy C-means clustering- BP neural network algorithm
The detailed steps of rough set-fuzzy C-means clustering BP neural network prediction model are as follows:

1. Hample algorithm is used to process the outliers of the collected data.
2. Use rough set theory to extract key attributes from sample data.
3. Perform fuzzy C-means clustering on the data set according to the key attributes.
4. Data normalization, each type of weather data and time data are adjusted to \((0~1)\), this step can ensure the smooth operation of BP neural network.
5. BP neural network prediction. The normalized data, the corresponding time, and the power generation at the time to be predicted are used as input. Set the minimum error of the training target, the number of training times, the actual frequency and the learning rate. After outputting the result, the number of neurons can be adjusted to achieve the optimal result.
6. Reverse normalization of prediction results and output.
7. Compare the predicted output value with the real value, and find out the data with large error for analysis.

4. Experimental simulation and result analysis

4.1. Data collection and outlier processing
The data selects the data from 2019 to 2020 of the No. 5 photovoltaic power station of the Alice Springs Solar Intellectual Property Center in Australia. The data resolution is 5 minutes, that is, every 5 minutes, it records the power generation and the weather characteristics of the corresponding time. There are 57,953 pieces of data, and some outliers are removed using the hample algorithm, leaving 57,938 pieces of data.

4.2. Analysis of influencing factors of photovoltaic power generation based on rough set
As an effective data analysis method, rough set can process uncertain evaluation data and dig out potentially valuable evaluation information from a large amount of uncertain information. In this paper, rough set attribute reduction and importance evaluation methods are mainly used to explore and analyze the relationship between photovoltaic power generation and influencing factors, and at the same time remove some redundant data.

4.2.1. Rough set analysis process
The process of extracting key features from collected data using rough set method can be divided into the following three steps:

Step1: Unify data types. Weather characteristic variables and power generation variables are continuous data of different dimensions, and time information is text-type, so firstly, the text-type time variables must be converted into data-type variables, and then the weather characteristic variables and
time variables must be discretized. And monotonic processing, and finally transform the collected data into the form of a rough centralized single decision information table. Before using rough set to reduce the attributes of samples, the repeated samples can be merged by equal frequency discretization to remove redundant sample information.

Step 2: Construct the information decision table. The processed sample data is transformed into the form of rough set decision table, that is, time variable and weather characteristic variable correspond to conditional attribute set \(C\), and power generation variable correspond to decision attribute \(D\). Let the time, relative humidity, temperature, global radiation level, diffuse radiation level, wind direction, precipitation, global radiation tilt and radiation diffusion tilt correspond in turn in the condition attribute \(C = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}\); The power generation decision attribute is \(D\). The conditional attributes are divided into 10 levels \([1, 2, 3, 4, 5, 6, 7, 8, 9, 10]\), and the decision attributes are divided into six levels \([1, 2, 3, 4, 5, 6]\). Table 1 is a small part of the decision table.

| sample | conditional attribute set C | decision attribute D |
|--------|----------------------------|---------------------|
| 17     | 2  6  7  4  3  5  10  3  3  1 |
| 18     | 3  6  7  4  3  5  10  3  3  2 |
| 20     | 3  6  7  4  4  5  10  3  3  1 |
| 21     | 3  6  7  4  3  4  10  3  3  1 |
| 22     | 3  6  7  4  4  4  10  4  4  2 |
| 24     | 3  7  7  4  4  4  10  4  4  2 |
| 25     | 3  7  7  5  4  3  10  4  4  1 |
| 26     | 3  7  6  5  4  3  10  4  4  2 |
| 28     | 3  7  6  5  5  3  10  4  4  1 |
| 29     | 4  7  6  5  5  3  10  5  5  3 |
| 31     | 3  7  6  5  5  3  10  5  5  1 |
| 33     | 4  7  6  6  5  3  10  5  5  3 |
| 36     | 4  8  6  6  5  3  10  5  5  3 |
| 37     | 4  7  6  6  6  2  10  6  6  1 |
| 38     | 4  8  6  6  6  2  10  6  6  1 |
| 39     | 4  8  5  7  6  2  10  6  6  3 |
| 42     | 5  8  5  7  6  2  10  6  6  3 |
| 45     | 5  8  5  8  6  2  10  6  6  1 |
| 46     | 5  8  5  8  7  2  10  7  7  2 |
| 48     | 6  8  5  8  7  2  10  7  7  3 |
| 51     | 6  9  5  8  7  2  10  7  7  3 |
| 53     | 6  9  4  8  7  2  10  7  7  6 |
| 56     | 6  9  5  9  7  2  10  8  8  2 |
| 57     | 6  9  4  9  7  2  10  8  8  6 |
| 58     | 6  9  4  9  8  2  10  8  8  2 |

Step 3: Use rough set to analyze the importance of attributes. Firstly, the positive domain of each conditional attribute \(C_i\) relative to the decision attribute \(D\) and the positive domain when the conditional attribute is missing are calculated. If \(\text{POS}_{(C_i | C_i)} (D) = \text{POS}_{C} (D)\), it indicates that the conditional attribute \(C_i\) is not necessary in the conditional attribute set \(C\) relative to the decision
attribute, and the unnecessary attribute can be deleted. The relative importance of each attribute is then calculated to find the key features for classification.

4.2.2. Rough set analysis results
The calculation results are shown in Table 2. From the final calculation results, it can be seen that among the factors affecting the power generation, temperature, time and relative humidity have the greatest influence on the power generation.

| Table 2. Table of influence degree of various factors on power generation |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| C1   | C2   | C3   | C4   | C5   | C6   | C7   | C8   | C9   |
| 0.212 | 0.124 | 0.62  | 0.015 | 0.0119 | 0.0134 | 0  | 0.0015 | 0.003 |

4.3. Time-segment prediction model based on fuzzy C-means clustering and BP neural network
According to the temperature and time related data extracted from the rough set, the clustering feature vector scale was constructed, and the fuzzy C-means clustering was used to divide it into three categories. The first category is in the noon period, the peak period of power generation. The second category is in the morning and afternoon, when the power generation increases or decreases monotonously. The third category is in the evening, when the power generation is low and the fluctuation is small. Model the classified data and set the parameters separately. For each type of sample, 90% of the data is selected as the training data, and 10% of the data is used as the test data.

4.4. Results analysis
The model was trained in different time periods, and the fitted prediction results were compared with those of the generalized weather model. The comparison results are shown in Table 3 and Table 4:

| Table 3. Comparison of forecast results in different periods of fine weather |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| period | MSE/kW | MAE/kW | MAPE/% | MSE/kW | MAE/kW | MAPE/% |
| 6: 00-7:00 | 0.0033 | 0.0120 | 26.89 | 0.016 | 0.0504 | 20.15 |
| 7: 01-11:00 | 0.0055 | 0.0278 | 1.16 | 0.0122 | 0.0695 | 3.53 |
| 11:01-13:00 | 0.0206 | 0.0800 | 1.80 | 0.0269 | 0.1165 | 2.65 |
| 13:01-17:00 | 0.0130 | 0.0569 | 2.09 | 0.0155 | 0.0788 | 2.72 |
| 17:01-19:00 | 0.0070 | 0.0190 | 0.49 | 0.0107 | 0.0462 | 66.94 |

| Table 4. Comparison of forecast results in different periods of precipitation weather |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| period | MSE/kW | MAE/kW | MAPE/% | MSE/kW | MAE/kW | MAPE/% |
| 6: 00-7:00 | 0.0036 | 0.0235 | 39.87 | 0.0067 | 0.0213 | 41.72 |
| 7: 01-11:00 | 0.0052 | 0.0127 | 2.11 | 0.0077 | 0.0421 | 4.48 |
| 11:01-13:00 | 0.0769 | 0.2999 | 7.74 | 0.0539 | 0.1866 | 4.59 |
| 13:01-17:00 | 0.0193 | 0.1401 | 7.59 | 0.0713 | 0.3449 | 18.77 |
| 17:01-19:00 | 0.0491 | 0.1689 | 16.38 | 0.0776 | 0.1322 | 34.9 |

According to the analysis of the figure, there are obvious differences in the law of power generation in different time periods. The model was trained in different time periods, and then compared with the predicted results without classification. The results are shown in Table 3 and Table 4.

Table 3 and Table 4 are the power prediction results on September 8 and 17 respectively. Model A is the Rough Set-FCM-BP prediction model proposed in this paper, which is classified by time period; Model B represents the prediction model of FCM-BP, which is traditionally classified by weather. In fine weather, the average mape of Model A was 6.486%, and that of Model B was 19.20%. The mean mse of Model A is 0.010kW, and that of Model B is 0.020kW. The average mae of Model A is 0.04kW, and that of Model B is 0.072kW. There are large errors between 6:00 and 7:00, which may be due to the
greater influence of more moisture in the air the day before the forecast date. Compared with traditional methods, the prediction results of Model A presented in this paper are more accurate.

In the case of rainfall weather, the average mape of Model A was 14.74%, and that of Model B was 18.87%. The mean mse of Model A is 0.031kW, and that of Model B is 0.035kW. The average mae of Model A is 0.13kW, and that of Model B is 0.11kW. In rainy days, there are large errors in the two periods of 6:00-7:00 and 17:01-19:00, which may be because the instrument is affected by water vapor in the air, leading to large error in data. However, the model proposed in this paper eliminated outliers from the data before the prediction, making the overall data more accurate, and the overall prediction result is more accurate than the traditional weather classification prediction.

5. conclusion
In this paper, a Rough set-FCM-BP short-term photovoltaic power prediction model is proposed based on the traditional weather classification prediction and considering the variation of photovoltaic power generation with time, which firstly sorts out the shutting characteristics of the samples and then predicts them with Rough Set theory. The following conclusions are drawn from the analysis of the simulation results:

The modified model has good applicability. The prediction model looks for the key features again, which makes the classification more accurate, reduces the classification difficulty caused by the absence of weather-related data, and effectively improves the accuracy of power generation prediction.

The model still has limitations for the prediction of morning and evening periods or the time with high relative humidity, but the prediction accuracy of sunny days is significantly improved, and generally the prediction accuracy is still higher than that of the traditional weather classification. Next, the research is aimed at improving the prediction method of rainy day or relatively high humidity.

Acknowledgments
When this paper is completed, I first express my heartfelt gratitude and sincere respect to Teacher Pang Qingle and Teacher Xie Qingsong. The two teachers’ persevering scientific research spirit, profound knowledge, truth-seeking work attitude, magnanimous and modest behavior, and selfless dedication have always inspired me to continuously learn knowledge and pursue the realm of life.

There are too many people who have helped me. I sincerely thank all those who care about and help me. Thank them for giving me the strength to overcome myself, surpass myself, and pursue continuous pursuit. Thank you.

References
[1] Wu, J.H., Shao, Z.G., Yang, S, H. (2020) Research review and prospects of data cleaning in new energy power prediction. J. Electrical Technology., 21: 1-6.
[2] Lou, J. L., Xu, J., Lu, H. (2016) Data cleaning algorithm for wind turbines based on power curve. J.Automation of Electric Power Systems., 40:116-121.
[3] Yang, J., Wang, L., Song, D. G. (2021) A wind turbine yaw angle zero drift diagnosis method based on isolated forest and sparse Gaussian process regression. J. Proceedings of the Chinese Society of Electrical Engineering., 1: 1-19.
[4] Wang, Y. Z., Duan, C. S., Wu, K. (2019) Design of high-precision through-heart ammeter. J. Acta Metrology., 40: 56-59.
[5] Zhang, J. J., Zhang, Q., Ma, Y. (2020) Fuzzy clustering-Elman neural network short-term photovoltaic power generation prediction model. J. Electrical Measurement and Instrumentation., 57: 46-51.
[6] Lai C. W., Li, J. H., Chen, B. (2019) Summary of research on photovoltaic power generation output prediction technology. J. Transactions of the Chinese Society of Electrical Engineering., 34: 1201-1217.
[7] Sun, D. L., Wang, Y., Yu, Y. (2020) Short-term photovoltaic cluster power interval prediction based on BP neural network. J. Journal of Shandong University., 50: 70-76.
[8] Zhao, K. N., Pu, T. J., Wang, X. Y. (2019) Forecast of photovoltaic output probability based on improved Bayesian neural network. J. Power System Technology., 43: 4377-4386.

[9] Wang, X. F. (2021) Median filtering detection using LBP encoding pattern. J. Multimedia Tools and Applications., 2: 1-24.

[10] Zhang, X. H., Pei, D. W., Dai, J. H. (2013) Fuzzy Mathematics and Rough Set Theory. M. Tsinghua University Publishing. Beijing.

[11] Li, L. Y., Wu, J. M. (2020) Elman neural network rural water demand prediction based on rough set-fuzzy C-means clustering. J. Science Technology and Engineering., 20:373-380.