processes

MDPI

Article

Short-Term Wind Power Prediction Using GA-BP Neural Network Based on DBSCAN Algorithm Outlier Identification

Pei Zhang [1], Yanling Wang [1,*], Likai Liang [1,*], Xing Li [2] and Qingtian Duan [2]

[1] School of Mechanical, Electrical and Information Engineering, Shandong University (Weihai), Weihai 264209, China; zzpp950419@126.com

[2] Shandong Hanlin Technology Co., Ltd., Jinan 250000, China; wangdi612@163.com (X.L.); m18766141096@163.com (Q.D.)

* Correspondence: wangyanling@sdu.edu.cn (Y.W.); lianglikai@sdu.edu.cn (L.L.)

Received: 5 December 2019; Accepted: 21 January 2020; Published: 27 January 2020

check for updates

Abstract: Accurately predicting wind power plays a vital part in site selection, large-scale grid connection, and the safe and efficient operation of wind power generation equipment. In the stage of data pre-processing, density-based spatial clustering of applications with noise (DBSCAN) algorithm is used to identify the outliers in the wind power data and the collected wind speed data of a wind power plant in Shandong Province, and the linear regression method is used to correct the outliers to improve the prediction accuracy. Considering the important impact of wind speed on power, the average value, the maximum difference and the average change rate of daily wind speed of each historical day are used as the selection criteria to select similar days by using DBSCAN algorithm and Euclidean distance. The short-term wind power prediction is carried out by using the similar day data pre-processed and unprocessed, respectively, as the input of back propagation neural network optimized by genetic algorithm (GA-BP neural network). Analysis of the results proves the practicability and efficiency of the prediction model and the important role of outlier identification and correction in improving the accuracy of wind power prediction.

Keywords: short-term wind power prediction; outlier identification; DBSCAN algorithm; linear regression method; GA-BP neural network

1. Introduction

Wind is pollution-free, abundant, and widely distributed—which is one of the most important energy sources for generating electricity. However, the wind is intermittent and uncontrollable; the large-scale grid connection of wind power will bring great risks to the power system. Accurate prediction of wind power can not only provide support for making generation plan, but also effectively make the utilization rate of wind energy higher.

With the widespread use of wind energy, wind power prediction has become a hot topic [1]. Combining time autocorrelation with a neural network, a wind speed prediction model is established, which is a period ahead of time [2]. In Reference [3], a prediction model using reverse back propagation-artificial neural network (BP-ANN) is established, which is 10 min and 1 h ahead of time, and an improved model is established by system error revision and wake coefficient improvement. In Reference [4], a prediction model using back propagation (BP) neural network, which is combined with wavelet, is established based on weather forecast data. In Reference [5], based on the measured data from numerical weather report and supervisory control and data acquisition (SCADA) system, considering the spatial correlation, static and dynamic neural networks are used to establish the prediction model. In Reference [6], a combination model of long-term and short-term memory (LSTM) networks, wavelet decomposition (WT), and principal component analysis (PCA) to predict the ultra short-term probability of wind power is proposed.

At present, research on wind power prediction mainly focuses on algorithms, and there is less research on the identification and correction of outliers. For most wind power prediction, only simple pre-processing is adopted for the outliers in power data, such as using deviation rate or pauta criterion to identify the outliers, and using mean value, mode or hotdecking method to correct the outliers. Using these methods will cause the temporality of wind speed and power being ignored, which will lose the characteristic of the data and decrease the accuracy of the prediction model. In Reference [7], the output power curve is fitted by adjusting the parameters, and the upper and lower thresholds are set by moving the output power curve. If the value of historical data is between the upper and lower thresholds, the historical data is normal data, otherwise, it is an outlier. This method needs more artificial experience, which may misjudge the normal data under some extreme conditions as an outlier, or ignore some outliers included in the upper and lower thresholds. In Reference [8], the neural network is trained by using the known normal data and outliers, respectively, so that the neural network can get the characteristic difference between normal data and outliers before identifying the outliers. Because a large number of historical data which have been classified correctly are usually difficult to get, and too much data will prolong the training time of the neural network, so this method also has great defects. In Reference [9], a combined screening model of outlier based on the quartile method and cluster analysis is proposed. Firstly, quartiles method is used twice to remove the conventional scattered outliers, then cluster method is used to remove the stacked outliers, and the method of secondary clustering is used to solve the problem of k-means clustering. The drawback of this method is that the k-means clustering algorithm can only recognize spherical data clustering, which is not suitable for irregular and multi-density data sets.

To get rid of the limitations of the above methods, DBSCAN algorithm is used to identify the outliers, and linear regression method is used to correct the outliers in the wind speed and wind power data. The average value, the maximum difference and the average change rate of daily wind speed of each historical day are calculated by using the pre-processed wind speed. These three standards and Euclidean distance are used to select similar days. Finally, GA-BP neural network is used to make the short-term prediction. The error analysis proves the feasibility and effectiveness of the prediction model.

2. Outliers Identification and Correction

2.1. Outliers Identification Based on DBSCAN Algorithm

Outliers will reduce the potential value of data and the accuracy of the prediction model. It is necessary to identify it [10]. The distribution density of outliers is usually very low, and DBSCAN, as the most common density clustering algorithm, can divide the data with high distribution density into different clusters, and effectively filter out the data with low distribution density [11].

2.1.1. Concept of DBSCAN

DBSCAN algorithm includes two main parameters: *eps* and *Minpts*. *eps* is the radius parameter, which is the radius of the neighborhood of a point. *Minpts* is the neighborhood density threshold, which indicates the number of sample points within the neighborhood with a radius of *eps*. These two parameters will directly affect the accuracy of clustering. The main concepts of DBSCAN are:

1. Neighborhood *eps*: If there is a region with radius *eps* and p is the center of this region, this region is called the neighborhood *eps* of p;
2. Core object: assuming that a p contains at least *Minpts* sample points in its neighborhood *eps*, p is a core object;

3. Direct density reachable: assuming that p is a core object and there is a q in the neighborhood *eps* of p, which means the distance from q to p is not greater than *eps*, q is direct density reachable to p, which is shown in Figure 1a;
4. Density reachable: assuming that there is a series of points $p_1, p_2, \ldots, p_n$, if p_{i+1} is direct density reachable to $p_i (i = 1, 2, \ldots, n-1)$, p_n is density reachable to p_1, which is shown in Figure 1b;
5. Density connected: assuming that o is a core object, both p to q are density reachable to o, p to q are density connected, which is shown in Figure 1c;
6. Border point: If a point is not a core object, but is located in the neighborhood *eps* of a core object, this point is a border point.
7. Cluster: assuming that U is a data set. For known *eps* and *Minpts*, a cluster C is a subset of U. C meets the following conditions:

   (1) $\forall p, q$: if $p \in C$, and p is density connected to q, then $q \in C$.

   (2) $\forall p, q \in C$: p is density connected to q.

8. Outlier: assuming that U is a data set. For known *eps* and *Minpts*, there are $C_1, C_2, \ldots, C_k$, a total of *k* clusters. Outliers are the data that does not belong to any cluster but belongs to U.

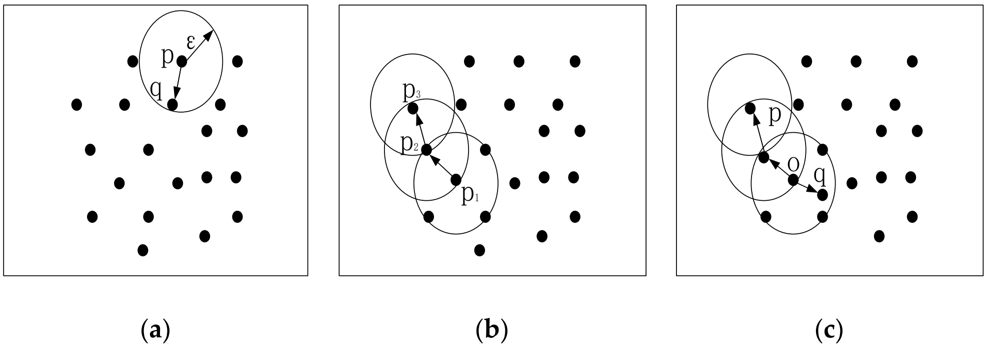

Figure 1. (**a**) description of direct density reachable; (**b**) description of density reachable; (**c**) description of density connected.

The flow of the DBSCAN algorithm is shown as follows:

Step 1: casually choose an unprocessed data p and determine whether it is a core object;

Step 2: find all the density reachable points in the data set and form a new cluster;

Step 3: get the final cluster result according to density connection;

Step 4: repeat step 2 and step 3 until all points in the data set are processed, and finally, clusters and outliers can be obtained.

2.1.2. Setting the Parameters of DBSCAN Algorithm

DBSCAN algorithm is extremely sensitive to input parameters *eps* and *Minpts*. The smaller *eps* or the larger *Minpts* is, the fewer core objects are, and some normal points will be misjudged as outliers, and the number of clusters will increase. On the contrary, the number of core objects will increase, some outliers will be missed, and the number of clusters will decrease [12].

The historical data used in this paper are the wind power data and wind speed data of 20 wind turbines in a wind power plant in Shandong Province. The period is 89 days, and the data's interval is 5 min. There are 25,632 groups of data totally. Figure 2 is a scatter diagram with wind speed as abscissa and wind power as ordinate. In this paper, the distribution of normal points and outliers is generally determined by observing the scatter diagram of the wind speed and power. Figure 2 shows that the wind speed and power have an obvious linear relationship, and the density of normal points is high and uniform, while the outliers are mainly distributed away from the clusters of normal points, with low and uneven density.

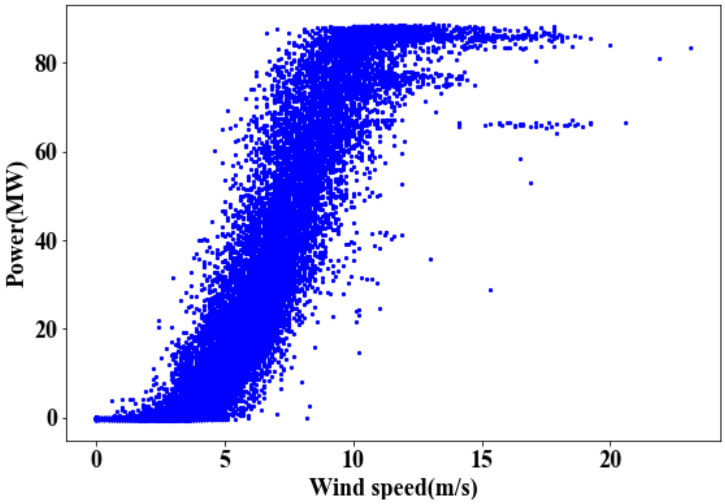

Figure 2. The scatter diagram describes the relationship between wind speed and wind power.

Firstly, standard deviation standardization is used to eliminate the influence of the difference in dimension and range between different types of data. Standard deviation standardization is shown in Equation (1):

$$x^* = \frac{x - \overline{x}}{\sigma}, \tag{1}$$

where x^* represents the standard deviation score which indicates how many standard deviations the original data has from the mean value of all data; $\overline{x}$ is the average value of the original data; σ is the standard deviation of the original data.

Secondly, the parameters *eps* and *Minpts* are iterated in equal steps. Since the distribution density of normal points is relatively uniform, this paper selects parameters that can include all normal points in the same cluster. After iterating the parameters, the partial iteration results are shown in Table 1. The number of clusters and outliers under different parameters are described in Table 1.

Table 1. The number of clusters and outliers under different parameters.

eps	*Minpts*	Cluster Number	Outliner Number
0.001	2	4120	10,183
0.051	8	17	693
0.101	3	12	55
0.101	8	2	187
0.101	9	1	203
0.151	2	7	19
0.201	9	2	40
0.851	8	1	2
0.901	5	1	1
0.951	2	1	0

2.1.3. Setting the Parameters of DBSCAN Algorithm

As shown in Table 1, when the parameters *eps* = 0.101 and *Minpts* = 9, the number of clusters is 1. The distribution and quantity of outliers are also roughly consistent with the estimation. Figure 3 shows the scatter diagram of wind speed and power after the identification under these parameters. The blue points are normal data, and the red '+' points are the outliers identified by the model. It shows that the identification effect of outliers is comparatively ideal.

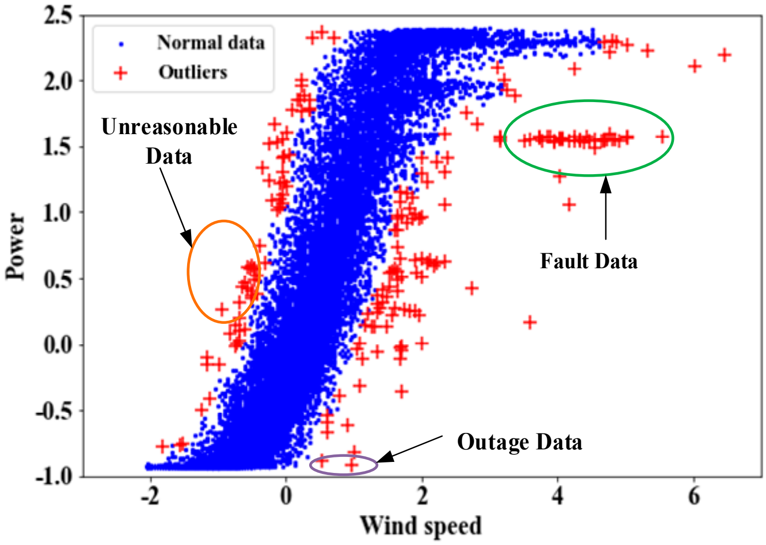

Figure 3. The scatter diagram of the wind speed and power after being identified.

According to Figure 3, outliers can be divided into four categories: lost data, unreasonable data, fault data and outage data. Lost data shows that the data is empty, which may be due to the data transmission interruption, storage system exception, etc. Unreasonable data shows that wind power is high when the wind speed is low, which may be due to the mistakes in data collection or transmission. Fault data shows that wind power is low when the wind speed is high, which may be caused by the loss of connection and data transmission error of some wind turbines. Outage data shows that the output power is 0 when the actual wind speed is greater than the starting up wind speed of the wind turbine, which is the wind speed that can make the wind turbine start to rotate, and the reason may be that wind power plant is abandoning the wind or overhauling [13].

2.2. Outliers Correction Based on Linear Regression Method

Discarding the outliers or using the mean value or mode instead of the outliers will reduce the available information of the data and affect the accuracy of the data prediction model. In order to maintain the integrity and temporality of the data, the linear regression method is used to correct the outliers.

2.2.1. Concept of Linear Regression Method

Linear regression is a method that uses regression analysis to determine the quantitative relationship between variables [14]. Independent variable is represented by x, and the dependent variable is represented by y. The relationship between x and y is shown in Equation (2):

$$y = a\ + bx + \phi, \tag{2}$$

where ϕ is random error, which follows Gaussian distribution $\mathrm{N}(0, \sigma^2)$. a, b and σ^2 are unknown parameters, which are not affected by the value of x. In practical problems, values of y are observed independently according to the different n values of x. These n pairs of observation values are called sample. If the unknown parameters a and b can be estimated by sample, the linear regression equation of y with respect to x can be obtained from Equation (3):

$$\hat{y} = \hat{a} + \hat{b}x, \tag{3}$$

where $\hat{a}$ is intercept and $\hat{b}$ is regression coefficient. $\hat{y}$ is the estimation of dependent variable. The estimation of the unknown parameters of the regression model is described below. The parameters

$\hat{a}$ and $\hat{b}$ in the above samples can be estimated by the least square method, which are shown in Equations (4) and (5):

$$\hat{b} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}, \tag{4}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}, \tag{5}$$

where $\overline{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$.

The unbiased estimation $\hat{\sigma^2}$ based on probability statistics can be obtained by Equation (6):

$$\hat{\sigma^2} = \frac{Q_\phi}{N-2}, \tag{6}$$

where $Q_\phi = \sum\limits_{i=1}^{n} \left(y_i - \hat{a} - \hat{b}x_i\right)^2$, Q_ϕ is the residual sum of squares, $\hat{\sigma}$ is the error of regression equation.

2.2.2. Correction Effect of Outliers

By repeatedly adjusting the sample number and observing the correction effect, 6 points before each outlier and 6 points behind each outlier are used as the samples. The correction effect is shown in Figure 4. The red "+" points represent the outliers which have been corrected. It shows that with the exception of a few outliers, the linear regression method can effectively correct most of the outliers.

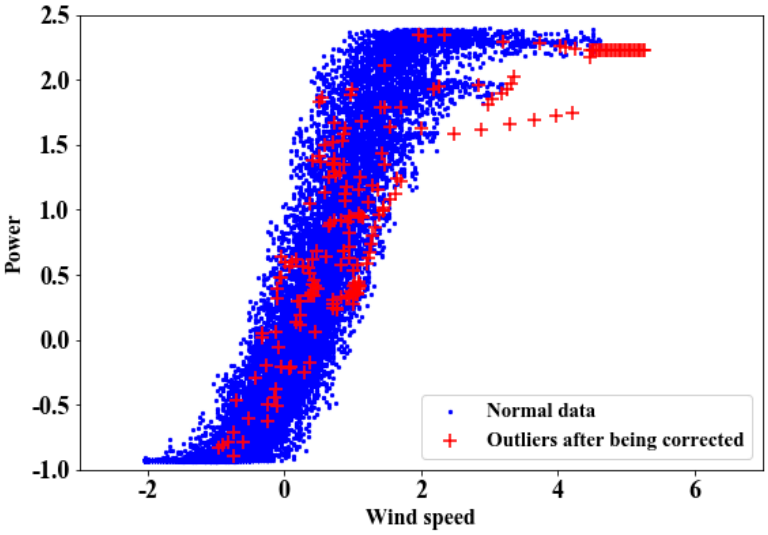

Figure 4. The correction effect of outliers.

3. Selection of Similar Days

Using all historical data as training samples will greatly extend the training time and reduce the accuracy of the neural network. Preserving the most effective historical data according to the similar day principle can reduce the workload of neural network training and improve the efficiency and accuracy of the prediction. In this paper, DBSCAN clustering and Euclidean distance are used to select similar days.

Among all meteorological factors, wind speed has the greatest impact on wind power. Therefore, the average value, the maximum difference and the average change rate of daily wind speed are taken as the selection criteria of a similar day. The number of daily wind speed samples is 288, and these three values can be calculated from the following equations:

$$\overline{v} = \frac{\sum\limits_{i=1}^{n} v_i}{n}, \tag{7}$$

$$v_{\mathrm{d}} = v_{\max} - v_{\min}, \tag{8}$$

$$\overline{p_{\mathrm{c}}} = \frac{\sum\limits_{i=1}^{n} \left| \frac{v_i - v_{i-1}}{v_{i-1}} \right|}{n}, \tag{9}$$

where v_i is the wind speed collected at the ith sampling point of the historical day; $\overline{v}$ is the average value of daily wind speed; n is the total number of daily sampling points; $v_{\max}$ is the maximum wind speed collected on the historical day; $v_{\min}$ is the minimum wind speed collected on the historical day; v_{d} is the maximum difference of daily wind speed; $\overline{p_{\mathrm{c}}}$ is the average change rate of daily wind speed.

Firstly, using the wind speed data identified previously, which include wind speed data of 88 historical days and 1 forecasted day, and calculate the daily average value, the daily maximum difference value and the average change rate value of the daily wind speed. Normalizing these three types of data to eliminate the difference of magnitude between different types of data. The normalization method is shown in Equation (10):

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{10}$$

where x_i^* is the normalized value of the ith sample data; x_i is the ith sample data; $x_{\min}$ is the minimum value of the data set; $x_{\max}$ is the maximum value of the data set. Figure 5 shows the distribution of the historical days and the forecasted day in three dimensions.

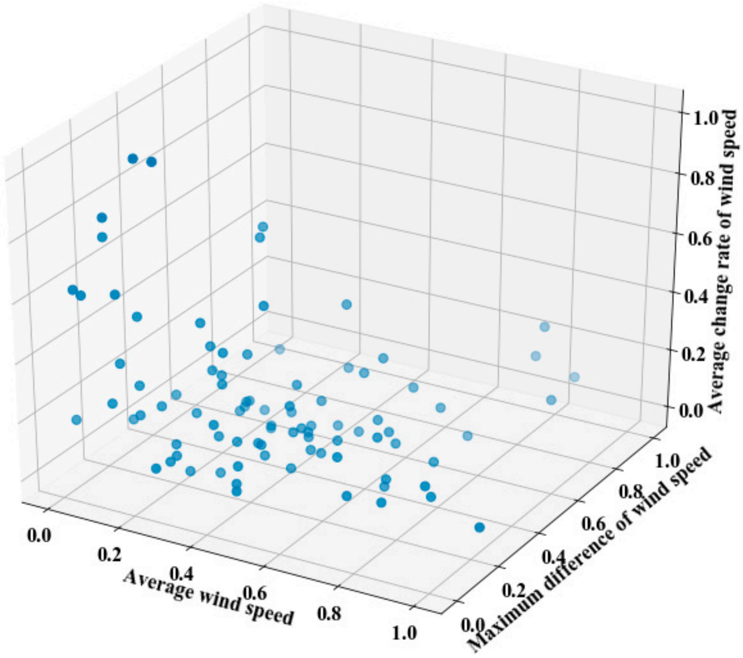

Figure 5. The scatter diagram of the historical days and the forecasted day in three dimensions.

Secondly, the parameters *eps* and *Minpts* of DBSCAN are iterated in equal steps. In this paper, parameters that can divide all sample points evenly into two clusters are used, and historical days in the cluster which the forecasted day is in is selected as similar days. Figure 6 is the scatter diagram of the historical days and the forecasted day when parameter *eps* = 0.147 and *Minpts* = 9. The blue star point represents the forecasted day, the blue circle points represent the similar days, and the red '+' points represent the unsimilar days.

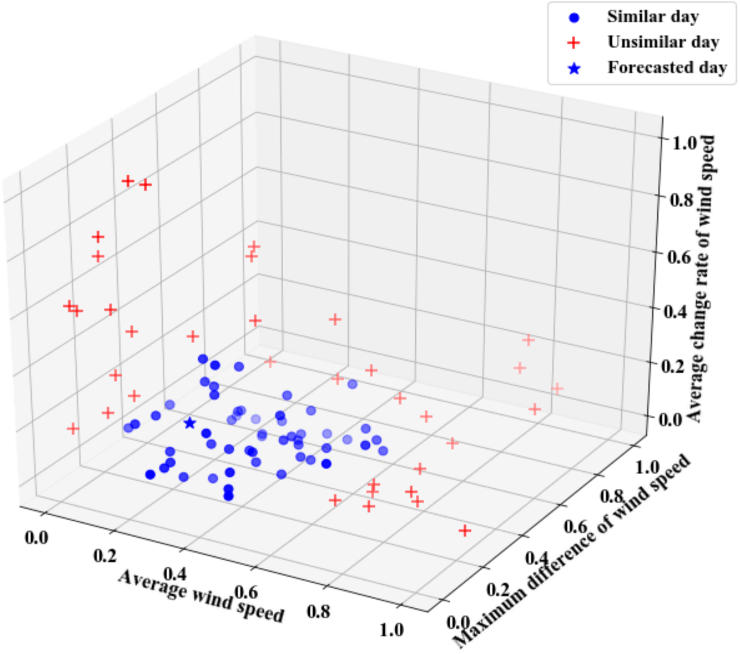

Figure 6. The scatter diagram of the historical days and the forecasted day in three dimensions.

The number of the similar days obtained by DBSCAN algorithm is 53, which still needs to be further screened to improve the efficiency and accuracy of the prediction model. The Euclidean distances can be calculated as Equation (11) [15]:

$$d_j = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{11}$$

where d_j is the Euclidean distance between the two groups of data. x_i and y_i are the corresponding sample data in the two groups of data. Here, x_i and y_i represent the corresponding wind speed data of a similar day and the forecasted day, respectively. As shown in Figure 7, v_{fi} is the wind speed at the ith sampling point of the forecasted day, and v_{ij} is the wind speed at the ith sampling point of the jth similar day.

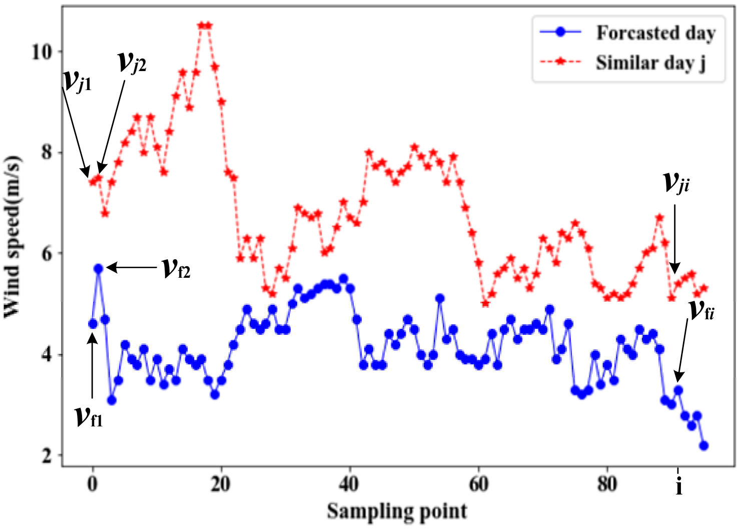

Figure 7. The wind speed of the forecasted day and similar days.

The Euclidean distances of each similar day and the forecasted day are sorted from smallest to largest. The first 30 similar days with the smallest Euclidean distance are selected as the samples for training.

4. Selection of Meteorological Factors Affecting Wind Power Generation

The principle of wind power is that the wind drives the blades of the wind generator to rotate, and then the speed of the blades is increased by the speed increaser to promote the generator to generate electricity. In essence, the kinetic energy of the air is converted into electrical energy. The output power of the wind generator can be calculated by Equation (12).

$$P = \frac{1}{2}\rho C_{\mathrm{P}} A V^3, \tag{12}$$

where P is the output power of the wind power equipment, ρ is the density of the air, C_{P} is the wind energy utilization factor of the wind generator, A is the area swept by the blades of the wind generator, and V is the wind speed. It can be known that C_{P} and A of the same type of wind turbines in the same wind power plant are constant, so wind speed and air density are the main meteorological factors affecting wind power.

Air density refers to the mass of air per unit volume at a certain temperature and pressure. When the wind speed is constant, the greater the air density, the greater the kinetic energy of the wind, and the greater the output power of the wind power equipment. The meteorological factors, such as temperature, pressure and humidity of the environment where the wind power equipment is located can affect both the air density and the operating performance of the power generation equipment, so these meteorological factors will also have a negligible impact on the output power of the wind power equipment. The wind direction mainly affects the output power of the wind power plant through the angle and wake effect of the blades of the wind turbine. In summary, in addition to considering historical power, wind speed, and wind outward, it is also necessary to use temperature, pressure, humidity, air density and other meteorological factors as inputs to the prediction model.

5. Short-term Wind Power Prediction Based on GA-BP Neural Network

5.1. GA-BP Neural Network Algorithm

The BP neural network has a strong mapping ability to nonlinear relation, simple structure and strong operability [16]. BP neural network consists of an input layer, hidden layer, and output layer. The number of input layers and output layers is 1. The number of hidden layers depends on the actual situation. Each layer contains one or more neurons. The structure of the BP neural network is shown in Figure 8.

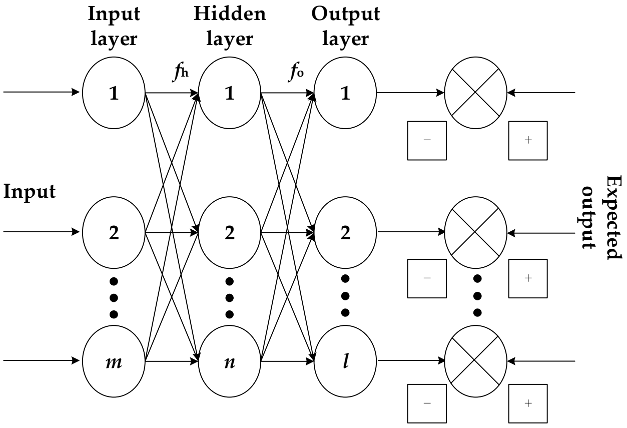

Figure 8. The structure of the BP neural network.

In Figure 8, m represents the number of hidden layer neurons; n represents the number of input layer neurons; l represents the number of output layer neurons; f_h is the activation function of the hidden layer and f_o is the activation function of the output layer. The number of hidden layer neurons can be calculated according to Equation (13):

$$m = \sqrt{l+n} + \alpha, \tag{13}$$

where α is the adjustment constant, which usually takes a range of $1-10$.

Genetic algorithm is a globally parallel random search method with good global optimization capabilities [17]. The genetic algorithm can effectively solve the defects of the traditional BP neural network. For example, it can speed up the convergence rate of the neural network, avoid it falling into local minimum value, and improve the learning accuracy of the network [18]. Genetic algorithm uses three operations, which are selection, crossover and mutation, to keep the individuals with high fitness, so that the new group not only has the information of the previous generations, but also better than them [19].

In this paper, using the group optimization strategy of genetic algorithm, the network's weights and thresholds are searched globally, so that the network's weights and thresholds are able to converge to the optimal global solution or infinitely be close to it [20]. The flow of the GA-BP neural network algorithm is shown in Figure 9.

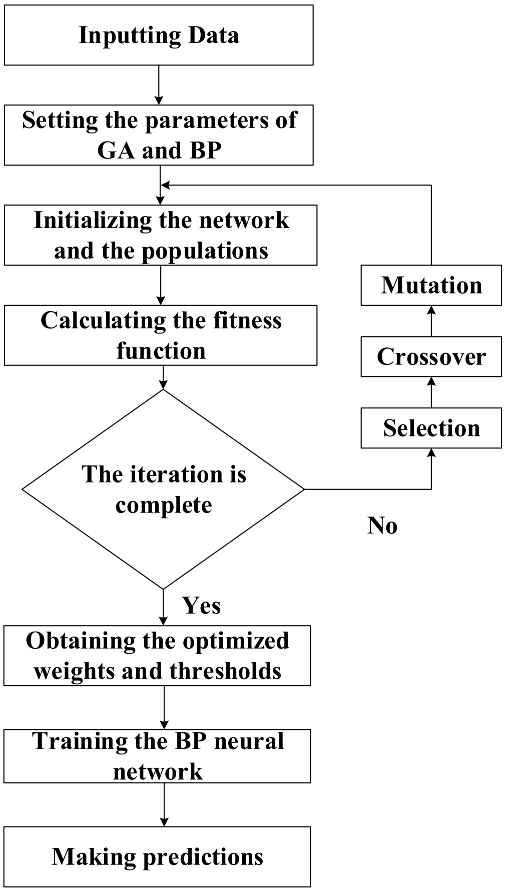
Inputting Data
Setting the parameters of GA and BP
Initializing the network and the populations
Calculating the fitness function
The iteration is complete
Yes
No
Selection
Crossover
Mutation
Obtaining the optimized weights and thresholds
Training the BP neural network
Making predictions

Figure 9. The flow of the GA-BP neural network algorithm.

5.2. Setting the Parameters of GA-BP Neural Network

5.2.1. Setting the Parameters of BP Neural Network

The neural network of the prediction model adopts three layers of neurons, which are an input layer, a hidden layer and an output layer, respectively. Since there is no air density data in the data set used in this paper, the input are the wind speed, temperature, humidity, cosine value of wind direction, air pressure and power of the first three days of the forecasted day and the wind speed, temperature, humidity, air pressure and wind direction of the forecasted day, the number of input layer neurons is $3.6 + 5 = 23$. The output is the wind power of the forecasted day, so the number of output layer neurons is 1. In this paper, the number of hidden layer neurons is set to 10. The activation function of the hidden layer is tan sig, and the activation function of output layer is purelin. The learning rate is set to 0.06, and the training target is 0.001. tan sig is shown as Equation (14) and purelin is shown as Equation (15):

$$y = \frac{2}{1 + \mathrm{e}^{-2x}} - 1, \tag{14}$$

$$y = x, \tag{15}$$

where x and y represent the input and output of hidden layer and output layer, respectively.

5.2.2. Selecting the Parameters of Genetic Algorithm

The weights and thresholds of the genetic algorithm are set as follow [21]:

$$N_{\mathrm{w}} = n_{\mathrm{input}}.n_{\mathrm{hide}} + n_{\mathrm{hide}}.n_{\mathrm{output}}, \tag{16}$$

$$N_{\mathrm{t}} = n_{\mathrm{hide}} + n_{\mathrm{output}}, \tag{17}$$

$$l = N_{\mathrm{w}} + N_{\mathrm{t}}, \tag{18}$$

where N_{w} represents the number of weights; N_{t} represents the number of thresholds; l represents the length of the individual coding of the genetic algorithm; n_{input} represents the number of input layer neurons; n_{output} represents the number of output layer neurons; n_{hide} represents the number of hidden layer neurons. The number of weights used in this paper is $23.10 + 10.1 = 240$; the number of thresholds is $10 + 1 = 11$; the length of the genetic algorithm is $240 + 11 = 251$; the initial population size is 50; the sampling method is roulette sampling; the probability of arithmetic crossover is 0.6; the probability of variation is 0.1; the largest evolutionary algebra of fitness is 50 generations.

5.3. Power Prediction and Result Analysis

The training data of the neural network include the power data of 20 wind turbines in a wind power plant and numerical weather forecast data in Shandong Province, with a period of 31 days, including the forecasted day and the 30 similar days which are screened previously. There are 25,632 groups of data totally. The previous 7776 groups data are used to train the neural network, and the last 288 groups are used as samples to test the accuracy of the short-term wind power prediction model.

The GA-BP neural network is trained by the data that have been identified and corrected and the unprocessed data, respectively. The data that have been identified and corrected are also called the pre-processed data. The comparison between the predictive and the actual values in these two cases is shown in Figure 10. The degree of fit in these two cases is shown in Figure 11. The comparison of the error between the predictive and actual values in these two cases is shown in Figure 12. In order to make observation more convenient, all the error values are taken as absolute values.

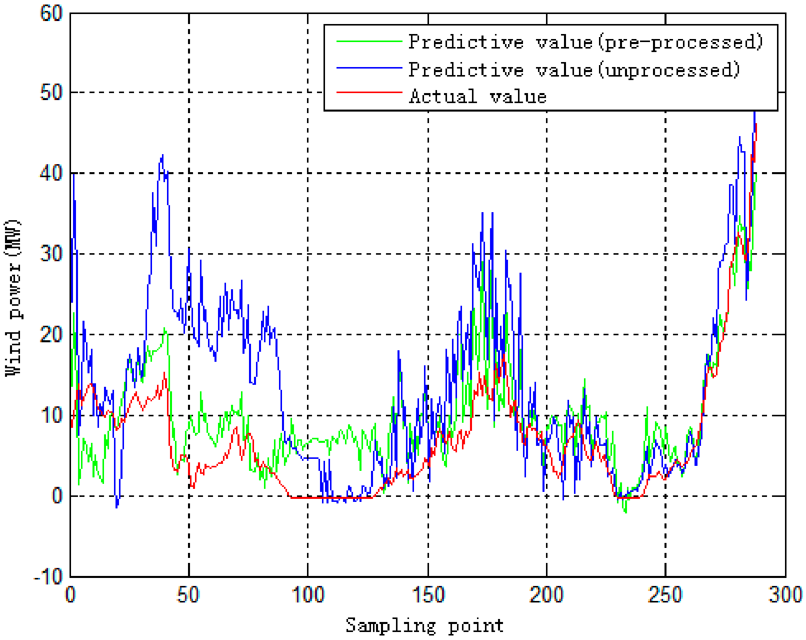

Figure 10. The comparison between the predictive values and the actual values.

It can be concluded from Figure 10 that the trend of both curves of prediction is basically consistent with the curve of the actual values, while the curve of wind power prediction with data pre-processing is more consistent with the actual values's curve compared with the curve of wind power prediction without data processing, and the fluctuation amplitude is also significantly reduced. Especially between the 25th sampling point and the 95th sampling point, the maximum error of wind power prediction with data pre-processing is 9.97 MW, and the average error is 4.07 MW, while the maximum error of wind power prediction without data processing is 27.83 MW, and the average error is 13.24 MW. The reason is that during this period, the number of outliers is 9, and there are six consecutive outliers in this period, which proves the importance of identifying and correcting outliers.

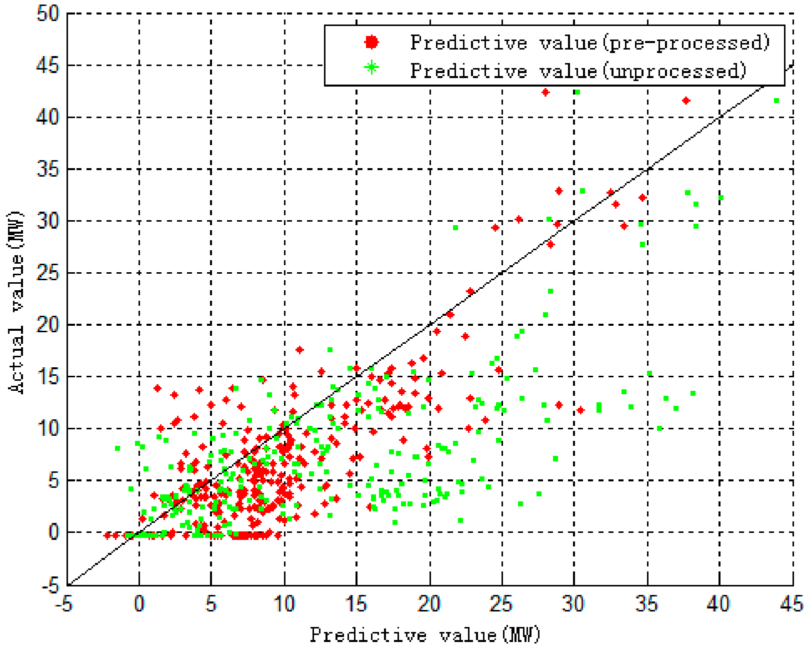

Figure 11. The degree of fit between the predictive values and the actual values.

In Figure 11, each scatter corresponds to a sampling point. The abscissa is the predictive value of the sampling point, and the ordinate is the actual value of the sampling point. The predictive values of the points on the straight line are equal to the actual values. It can be seen that compared with the wind power prediction without data processing, the scatters distribution of the prediction with data pre-processing are closer to the straight line, that is to say, the predictive value is closer to the actual value.

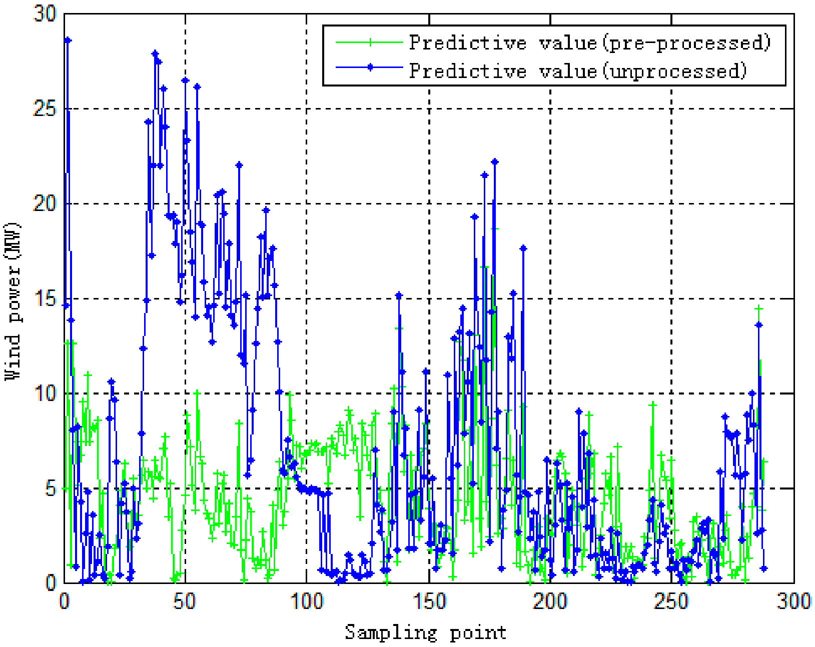

Figure 12. The comparison of the error of the two cases.

Figure 12 shows that the error of the unprocessed wind power prediction is significantly larger. Nearly half of the error values reach over 10 MW, and the maximum error value reaches 28.54 MW. However, most of the error values of the prediction with data pre-processing are below 10 MW, and the maximum error value is 18.64 MW, which shows that wind power predictive values with data pre-processing are more consistent with the actual wind power values than those without data processing, and the error is smaller.

It is necessary to analyze and evaluate the error of wind power prediction. Normalized root mean square error (NRSME) is the evaluation of the average value of error and normalized mean absolute error (NMAE) can measure the dispersion of error, which can evaluate the overall accuracy of the prediction model essentially. The NRSME and the NMAE of predictions with data pre-processing and wind power prediction without data processing are calculated, respectively. Table 2 shows the result.

Table 2. The NRSME and the NMAE of wind power prediction with data pre-processing and wind power prediction without data processing.

	NRSME	**NMAE**
Power prediction with data pre-processing	6.46%	5.24%
Power prediction without data processing	9.78%	7.98%

At present, in the “Frequency Specifications for Wind Power Forecasting System” issued by State Grid Corporation of China, the RSME of the short-term output power of wind turbines should be about 15%, generally not more than 20%. Table 2 shows that the predictive results obtained in both cases meet the requirements. The error of wind power prediction with data pre-processing is significantly smaller than that of wind power prediction without data processing, which means more accuracy.

6. Conclusions

According to the weather forecast data and the real data of a wind power plant in Shandong Province, DBSCAN algorithm is used to identify the outliers in wind speed and power data, and linear regression method is utilized to correct the outliers to improve the accuracy of the prediction. The iterative method is used to select the appropriate parameters of DBSCAN so that the outliers can be effectively identified. According to the temporality between the data, most of the outliers can be effectively corrected. Based on the data of similar days, which is the input of GA-BP neural network, the wind power prediction model is established. The simulation result based on actual data shows that the model can essentially ensure the accuracy of short-term wind power prediction.

Author Contributions: Conceptualization, Y.W.; Methodology, P.Z.; Writing—original draft, L.L.; Writing—review and editing, X.L. and Q.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41904169, and the Science and Technology Development Project of Weihai City, China, grant number 2018DXGJ05.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Costa, A.; Crespo, A.; Navarro, J.; Lizcano, G.; Madsen, H.; Feitosa, E. A review on the young history of the wind power short-term prediction. *Renew. Sustain. Energy Rev.* **2008**, *12*, 1725–1744. [CrossRef]
2. Yang, X.Y.; Xiao, Y.; Chen, S.Y. Wind speed and generated power forecasting in wind farm. *Proc. CSEE* **2005**, *25*, 1–5.
3. Chen, Y.; Zhou, H.; Wang, W.P.; Cao, X.; Ding, J. Improvement of ultra-short-term forecast for wind power. *Autom. Electr. Power Syst.* **2011**, *35*, 30–33.
4. Shi, H.T.; Yang, J.L.; Ding, M.S.; Wang, J.M. A short-term wind power prediction method based on wavelet decomposition and BP neural network. *Autom. Electr. Power Syst.* **2011**, *35*, 44–48.
5. Barbounis, T.G.; Theocharis, J.B.; Alexiadis, M.C.; Dokopoulos, P.S. Long-term wind speed and power forecasting using local recurrent neural network models. *IEEE Trans. Energy Convers.* **2006**, *21*, 273–284. [CrossRef]
6. Sun, Y.H.; Wang, P.; Zhai, S.W.; Hou, D.C. Ultra-short-term probability prediction of wind power based on LSTM network and condition normal distribution. *Wind Energy* **2019**, *23*, 63–76. [CrossRef]
7. Kusiak, A.; Zheng, H.Y.; Song, Z. Models for monioring wind farm power. *Renew. Energy* **2009**, *34*, 583–590. [CrossRef]
8. Marvuglia, A.; Messineo, A. Monitoring of wind farms' power curves using machine learning techniques. *Appl. Energy* **2012**, *98*, 574–583. [CrossRef]
9. Liu, Z.Q.; Gao, W.Z.; Wan, Y.H.; Muljadi, E. Characteristics and processing method of abnormal data clusters caused by wind curtailments in wind farms. *Autom. Electr. Power Syst.* **2014**, *21*, 39–46.
10. Wei, D.Q.; Wang, B.; Liu, D.C.; Luo, J.H.; Ji, X.P. A method for WAMS big data modeling and abnormal data detection with large random matrices. *Proc. CSEE* **2015**, *35*, 629–636.
11. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X.W. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **2017**, *42*, 1–21. [CrossRef]
12. Lenco, D.; Bordogna, G. Fuzzy extensions of the DBScan clustering algorithm. *Soft Comput.* **2018**, *22*, 1719–1730.
13. Ye, X.; Lu, Z.X.; Qiao, Y.; Min, Y.; Malley, M.O. Identification and correction of outliers in wind farm time series power data. *IEEE Trans. Power Syst.* **2016**, *31*, 4197–4205. [CrossRef]
14. Pinson, P.; Nielsen, H.A.; Madsen, H.; Nielsen, T.S. Local linear regression with adaptive orthogonal fitting for the wind power application. *Stat. Comput.* **2007**, *18*, 59–71. [CrossRef]
15. Gan, J.H.; Tao, Y.F. On the hardness and approximation of Euclidean DBSCAN. *ACM Trans. Database Syst.* **2017**, *42*, 1–45. [CrossRef]
16. Wang, Z.; Wang, B.; Liu, C.; Wang, W.S. Improved BP neural network algorithm to wind power forecast. *J. Eng.* **2017**, *2017*, 940–943. [CrossRef]

17. Faria, H.; Resende, M.G.C.; Ernst, D.I. A biased random key genetic algorithm applied to the electric distribution network reconfiguration problem. *J. Heuristics* **2017**, *23*, 533–550. [CrossRef]
18. Wang, S.X.; Zhang, N.; Wu, L.; Wang, Y. Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method. *Renew. Energy* **2016**, *94*, 629–636. [CrossRef]
19. Zhang, R.D.; Tao, J.L. A nonlinear fuzzy neural network modeling approach using improved genetic algorithm. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5882–5892. [CrossRef]
20. Contaldi, C.; Vafaee, F.; Nelson, P.C. Bayesian network hybrid learning using elite-guided genetic algorithm. *Artif. Intell. Rev.* **2019**, *52*, 245–272. [CrossRef]
21. Liang, H.B.; Zou, J.L.; Liang, W.L. An early intelligent diagnosis model for drilling overflow based on GA-BP algorithm. *Clust. Comput.* **2019**, *22*, 10649–10668. [CrossRef]