The Journal of Engineering

Special Issue: Energy Internet and its Related Technology Research

IET Journals
The Institution of Engineering and Technology

Application of cluster analysis in short-term wind power forecasting model

eISSN 2051-3305
Received on 13th December 2018
Revised 8th February 2019
Accepted on 18th February 2019
E-First on 29th April 2019
doi: 10.1049/joe.2018.5488
www.ietdl.org

Aoran Xu[1] ✉, Tao Yang[1], Jianwei Ji[1], Yang Gao[2], Cailian Gu[2]

[1]College of Information and Electrical Engineering, Shenyang Agricultural University, 120 Dongling Road, Shenhe District, Shenyang, People's Republic of China

[2]Institute of Electric Power, Shenyang Institute of Engineering, 18 Puhe Road, Shenbei District, Shenyang, People's Republic of China

✉ *E-mail: 656085652@qq.com*

Abstract: At present, the method of predicting wind power generation is mainly based on data integration calculation. Although this method is simple, it has shortcomings in short-term and ultra-short-term predictions owing to low accuracy. In this study, the clustering analysis data processing method is used to pre-process the meteorological wind power generation data, thus improving the data quality. This method builds model samples based on historical data with similar numerical weather prediction (NWP) characteristic parameters of the original sample data and forecast date, takes the NWP information of the forecast date as the basis of similarity measurement, and extracts effective data for the neural network prediction model after the improved clustering processing. Therefore, short-term wind power prediction analysis can be performed. Herein, the proposed data processing method is combined with the neural network model to create a software product that is applied to a wind farm in northeast China. The combined clustering data processing method of the wind power prediction model improved power prediction by ∼12% compared with that of the traditional continuous model. This demonstrates an obvious improvement in the prediction accuracy, thereby further proving the validity of the proposed method.

1 Introduction

Current technologies on predicting wind power generation focus on physical methods combined with artificial intelligence methods, such as the time calculation combined with chaotic methods [1]. Due to the strict time-limit requirement, the time calculation based on meteorological data or the simple neural network method is the main application technology used in short- and ultra-short-term wind power prediction. However, the prediction accuracy of this technology is not high and is highly dependent on the original data. Herein, a data clustering analysis method is proposed to predict wind point power using the neural network model [2, 3]. We performed in-depth research and combined the characteristics of large-scale wind farms in northeast China and the daily similarity of wind speed and wind power for applications in short-term wind power prediction. The generation power curve in Fig. 1 shows the daily similarity of the generation power of a wind farm. This method considers the numerical weather prediction (NWP) information of the prediction day as its characteristic parameter. We propose an improved method to calculate the Euclidean distance of the characteristic parameter to determine all kinds of similar-day data and then establish the power prediction model through the similar samples after clustering. Therefore, the model input parameter is NWP data, and the model target value is actual wind power data. After model training, this method can create a short-term power prediction model of multiple wind farms. The proposed method aims to solve the data quality problem at the original data end, improve the data accuracy with less time-consuming data processing technology, and uses the neural network method for power prediction. The innovative application of the improved clustering data algorithm to pre-process the data and its comparison with the actual wind farm power generation data effectively improve the accuracy of power prediction.

2 Application of the cluster analysis method

This paper uses a clustering method based on metrics. This method is unsupervised and is analysed without pre-set values. In the field of automation, it is known as a method of unsupervised learning clustering [4]. The most classical approach dynamic clustering method is the *K*-means clustering method, which divides each sample into the categories closest to the mean, in order to perform clustering based on the distance. The process can be broken down into the following five processing steps:

(i) Divide all the data into K initial classes. One sample point is selected in each initial class as the initial cluster centre, denoted as $z_1(l), z_2(l), \ldots, z_k(l)$, where the initial value $l = 1$.

(ii) On the basis of the nearest neighbour rule, assign all samples to the K class $\omega_j(K)$ represented by each cluster centre, and $N_j(l)$ represents the number of samples included in each category.

(iii) Calculate the various mean vectors, representing the new cluster centres with the following mean vectors:

$$z_j(l+1) = \frac{1}{N_j(l)} \sum_{x(i) \in \omega_j(l)} x(i), \tag{1}$$

where

$$j = 1, 2, \ldots, k$$

$$i = 1, 2, \ldots, N_j(l).$$

(iv) If $z_j(l+1) \neq z_j(l)$ in the operation process, indicating that the clustering result is not optimal, return to step (2) and continue the iterative calculation.

(v) If $z_j(l+1) = z_j(l)$ in the operation process, the iterative process ends, and the clustering result at this time is considered the optimal clustering result.

The similarity measure in this paper must be used before grouping the samples [5, 6]. This metric means that each sample must be compared to other samples. Hence, the samples in the same class are very 'similar', and the different classes have highly dissimilar samples. The similarity between two elements can be measured by various methods, such as distance metrics, correlation metrics, and

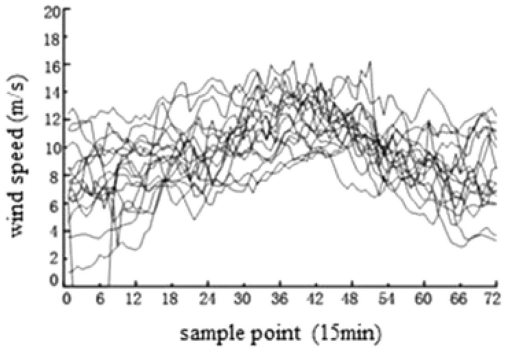

Fig. 1 *Part of the daily wind speed curve family*

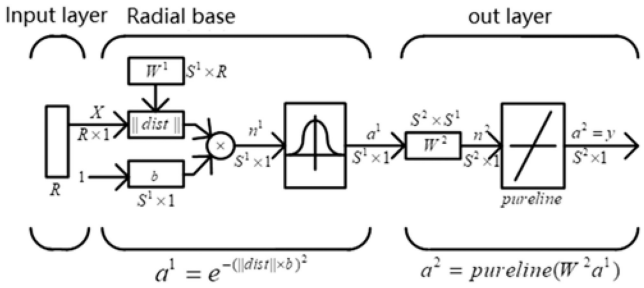

Fig. 2 *GRNN architecture*

information metrics. In the similarity measure, Minkowski, Mahalanobis, and Canberra are the more common distances [7].

The Minkowski distance is the minimum number of substitutions needed to change one into the other between two equal-length strings, s1 and s2. However, this method is suitable for data processing with fewer vector parameters and is not suitable for the requirements of multi-parameter conditions in wind farms. The Mahalanobis distance is an independent and identically distributed Euclidean distance extension between each sample vector. Although it has high data processing accuracy, it suffers from poor timeliness and is not suitable for the time requirement of short-term power prediction [8, 9]. Moreover, the Orchid-type distance is a dimensionless quantity to overcome the Minkowski distance, and is associated with the dimension of each index of faults [10]. The Canberra distance is not sensitive to large singular values, making it especially suitable for high migration and data, but it does not take into account the Orchid-type distance parameter correlation, but the wind between each meteorological data (i.e. data correlation). Hence, the Canberra distance is not suitable for wind power prediction data processing [11, 12]. Given the difference in measurement method, the trend of wind speed of the similar degree between Euclidean distance measurement, the measurement method for simple timeliness and strong similarity measure method, as well as the proposal based on the number of cumulative variance contribution rate of wind data selection principal component, when the contribution rate is >85%, you can give up the rest of the parameters directly into the neural network operation link, that is, the effective prediction in time.

The metric used in this paper is the European distance measure of the degree of similarity between wind speed trends on different days. The metric is a daily NWP vector defined as a seven-dimensional vector given by

$$\boldsymbol{X} = [Pav, T\min, T\max, V\min, V\max, D\sin, D\cos].$$

The above variables represent the daily mean value, the daily minimum, the daily maximum, the daily wind speed, the daily wind sine average, and the daily wind cosine average. In this formula, the dimensions of the components in the NWP vector are different; thus, the air pressure, wind speed, and temperature need to be normalised and divided by their respective historical maximum values. Moreover, the wind direction sine and cosine values need not be processed anymore and are used as normalised values.

The distance is defined as

$$d_i = \left(\sum_{k=1}^{7} (x_m(k) - x_i(k))^2 \right)^{1/2}, \tag{2}$$

where d_i represents the distance, i.e. the Euclidean distance between the predicted date and the historical sample i; x_m denotes the day NWP vector of the prediction day; x_i denotes the day NWP vector of the history data, $i = 1, 2, \ldots, n$; and n refers to the number of samples [13].

3 Neural network prediction model

The generalised regression neural network (GRNN) used in this paper is a typical representative of the radial basis function neural network. Its advantage is that it has great potential in nonlinear fitting.

The GRNN is a three-layer feedforward network with a single hidden layer. The input layer node only transmits the input signal to the hidden layer. The hidden layer node is composed of the radial function (generally taking Gauss function), while the output node is a simple linear function. The structure of the GRNN is shown in Fig. 2.

The parameters of the GRNN input layer mainly refer to wind speed, which can be determined by the formula, $P = C_P A \rho v^3/2$, where P is the output power of the wind wheel, C_P is the power coefficient of the wind wheel, ρ is the air density, A is the sweep area of the wind wheel, and v is the wind speed. From the formula, it can be seen that the output power P of wind turbines is proportional to the cubic power of the wind speed v. Moreover, the output power of wind turbines can vary greatly in response to the slight change of wind speed; thus, the main factor affecting the output power is wind speed.

In order to obtain better prediction accuracy, in addition to wind speed, the influence of wind direction, temperature and air pressure on output power should also be considered.

The network output is given by

$$a^2 = \text{purelin}(W^2 a^1), \tag{3}$$

$$a^1 = \text{radbas}(n^1), \tag{4}$$

$$n^1 = \| IW^1 - X \| \cdot b^1. \quad (5)$$

In (4), radbas is a radial basis function and is generally a Gaussian function expressed as

$$a^1 = \text{radbas}(n^1) = e^{-n^2}. \quad (6)$$

In addition, b^1 is the threshold of the network, which can be obtained from the distribution density spread

$$b^1 = [-\lg(0.5)]^{1/2}/\text{spread}. \quad (7)$$

The GRNN network has a single adjustable parameter spread, and its learning depends on the data samples. This characteristic determines that the GRNN network can minimise the impact of subjective assumptions on prediction results. The smaller the selected spreads, the higher the approximation accuracy of the function. However, this does not guarantee that the approximation process would be smooth. If the selected spreads are larger, the approximation process is relatively smooth, but the approximation accuracy is poor and the error is large.

In the design process, selecting appropriate spreads can lead to better prediction accuracy. The GRNN network is trained with the training samples, and the spreads are gradually increased. The errors between the actual values and the predicted values are calculated. By repeating this process for each training sample, a set of error sequences can be obtained, and the 2-norm of the error sequence can be calculated as follows:

$$\| e \|_2 = \sqrt{|e_1|^2 + |e_2|^2 + \cdots + |e_n|^2} \quad (8)$$

The future wind speed can be predicted by choosing the spreads that minimise the error sequence 2-norm.

In this paper, the GRNN method is selected to overcome some problems, such as the slow convergence speed of the back propagation neural network method and the tendency to easily fall into the local minimum. Compared with the Elman neural network method, the GRNN method avoids the problem wherein the feedback gain value is not easy to be determined. When applied to short-term wind speed prediction, the GRNN method shows high engineering value in nonlinear fitting.

The NWP information provided by the European mesoscale meteorological prediction centre is used as the meteorological information in this paper. The time point is selected based on the follow-up experimental data. Currently, the European mesoscale meteorological prediction centre is the world's source of accurate information, and its satellite remote sensing data recognition and accuracy are recognised by the authorities, thus ensuring the quality of data.

In this paper, the NWP neural network prediction model is used for sample training. As shown in Fig. 3, the input data are NWP air pressure, NWP wind speed, NWP air temperature, NWP wind direction sine, and NWP wind direction cosine. The output data are the predicted values of wind power.

As shown in Fig. 4, the principle of predicting the short-term model of wind power is as follows. First, the historical samples are classified, that is, the *K*-means clustering method is used to classify the samples into *K* categories (Category 1, Category 2, …, Category *K*), which are, respectively, classified by the *K*-means clustering method. We use MATLAB programming to automatically search the classification of the forecast day, and use the sample data in the *K* class as the training sample to establish the daily NWP vector neural network model. The predicted value of wind power can be obtained after multiple training runs.

When the model is trained, the input quantity comprises the NWP wind direction cosine, NWP wind direction sine, NWP air pressure, NWP air temperature, and NWP wind speed of the classification sample to be predicted. The output of the model is the output power of the wind farm. When predicting the model, the input is the NWP information of the forecast day, and the predicted value of the wind power is obtained.

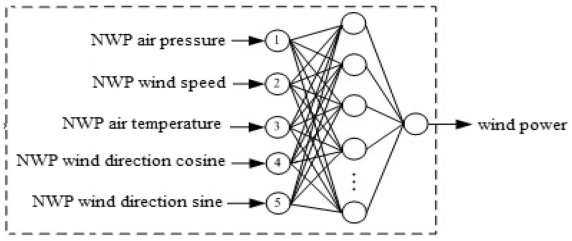

Fig. 3 *Daily NMP vector neural network mode*

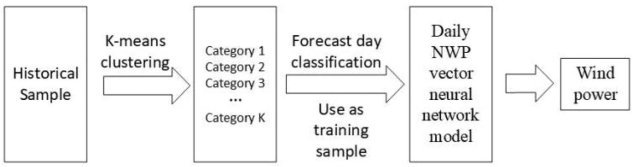

Fig. 4 *Schematic of the prediction model*

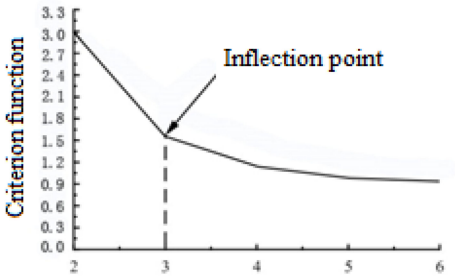

Fig. 5 *Relationship between criterion function and classification number K*

4 Theoretical application

We performed historical data contrast experiment to choose the actual wind farm. The selected wind farm is located in northeast China (N42°03′, E121°36′) and has a typical inland monsoon climate in northeast China. As the time selection for the region in the early winter and the moment wind meteorological condition are relatively complex, the weak meteorological data characteristic value can help us verify the effectiveness of the proposed method. The NWP data of the wind farm from November to December 2013 and the actual measured wind power data were analysed, modelled, and predicted. The data resolution was recorded as 15 min, the predicted step size was recorded as 96 points, and the forecasted day was selected as 9 December 2013.

According to the *K*-means clustering method, that is, the classification based on the distance, it is only necessary to select the historical data ∼20 days before the forecast for cluster analysis, i.e. from 19 November 2013 to 18 December 2013. We marked them as sample days 1, 2, …, 19, 20. The relation between the criterion function and the classification number *K* is obtained, as shown in Fig. 5. According to the method of the optimal classification number, the optimal classification number *K* is obtained, i.e. the inflection point of the criterion function curve, *K* = 3 [14, 15].

When $K = 3$, the analysis of the category of the historical sample day of the 20-day forecast is known. Among them, the second category has one day (the sample day is 18). Meanwhile, the third category has four days, which are the sample days 2, 4, 5, and 7, respectively. Other sample days belong to Category 1. Various mean vectors are calculated according to (1), and the cluster centres of the first class, second class, and third class are normalised as [0.987 0.185 0.436 − 1.132 − 0.807 0.044 0.054], [0.987 0.552 0.862 − 1.153 − 0.851 0.118 0.186], and [0.994 0.048 0.267 − 0.857 − 0.552 − 0.024 − 0.127], respectively.

The NWP vector of the normalised day on the forecast day (December 9) and the forecast date is calculated as [0.982 0.341 0.802 − 0.933 − 0.578 0.114 − 0.054], and the relative three types of cluster centres are calculated according to (2). The Euclidean distances are 0.52, 0.47, and 0.64, respectively. The results show

that the data is the second type of clustering centre. Therefore, the forecasting day belongs to the second category, which belongs to the 18th sample day, i.e. 6 December 2013.

According to the NWP data of 6 December 2013, we used the input quantity, namely NWP wind direction cosine, NWP wind direction sine, NWP air pressure, NWP air temperature, and NWP wind speed, and the output is the measured power, which was modelled by the GRNN structure. We adjusted the data resolution to 15 min, the training sample data to 96, and the window width parameter to 0.15.

When the model training was completed, the input prediction date (the NWP wind direction cosine, NWP wind direction sine, NWP air pressure, NWP air temperature, and NWP wind speed on 9 December 2013), we obtained the predicted power value and the prediction errors normalised mean absolute error (NMAE) and normalised root mean square error (NRMSE) as 10.65 and 14.04%, respectively.

In order to make the effectiveness of the method more significant, the neural network model using cluster analysis is compared with the traditional continuous model. According to statistics, the prediction errors NMAE and NRMSE of the continuous model are 22.12 and 26.81%, respectively. As shown in Fig. 6, the superiority of the method can be verified by comparing the prediction results with the prediction curves of the models.

5 Conclusion

In the case of mature wind power prediction model technology, the big data processing problem has become a key technological issue for wind power prediction in the wind power generation of power systems. Based on the similarity of wind speed and wind power, this paper proposes a cluster analysis method applied to the neural network to predict wind power model based on the short-term prediction technology of the wind farm output power. The method studied in this paper can accurately predict the wind power according to different weather conditions. The proposed method has been applied to a wind farm in northeast China and has shown significant improvement in terms of the prediction accuracy of short-term wind power. Moreover, it has brought considerable economic benefits to the wind farm.

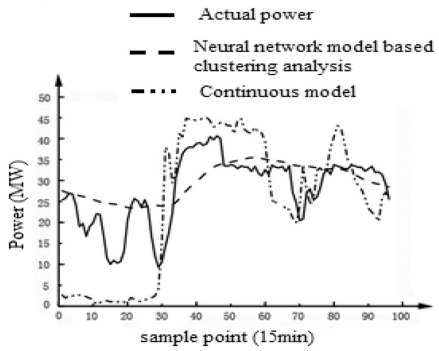

Fig. 6 *Comparison of the power prediction curves for each model*

6 References

[1] Zhong, H.Y., Gao, Y., Wu, L.: 'Analysis of power prediction technology in wind power', *Electr. Energy Efficiency Manag. Technol.*, 2015, **10**, pp. 51–55 (in Chinese)

[2] Gao, Y., Zhong, H.Y., Xu, A.R.: 'Study on the application of power prediction technology in photovoltaic power generation', *Electr. Energy Efficiency Manag. Technol.*, 2015, **17**, pp. 38–43o+58

[3] Gao, Y., Zhong, H.Y., Chen, X.Y.: 'Ultra-short term wind speed forecasting based on neural network and wavelet analysis', *Renew. Energy*, 2016, **35**, (5), pp. 705–711

[4] Gao, Y.: 'Study on wind power prediction method for wind farm', Shenyang Agricultural University, 2011

[5] Gao, Y., Ouyang, Q., GuanL, H.M.: 'Review of wind farm access power grid technology', *Northeast Electr. Power Technol.*, 2015, **31**, (2), pp. 14–17

[6] Gao, Y., Chen, H.Y., Ouyang, Q.: 'Summary of the research on the prediction technology of wind farm power generation', *Power Grid Clean Energy*, 2014, **26**, (4), pp. 60–63o+67

[7] Gao, Y., Piao, Z.L., Zhang, X.P.: 'Prediction of wind power generation based on ARMA model in noisy situation', *Power Syst. Prot. Control*, 2013, **38**, (20), pp. 164–167

[8] Gao, S., Dong, L., Gao, Y.: 'Medium and long term wind speed forecasting based on rough set theory', *Chin. J. Electr. Eng.*, 2012, **32**, (1), pp. 32–37o +21

[9] Liang, Z.C., Wei, H., Li, L.: 'An approximate dynamic programming method for the value function of medium and long term power generation', *Chin. J. Electr. Eng.*, 2015, **35**, (20), pp. 5199–5209

[10] Meng, X.X., Tian, C.W., Dong, L.: 'Rey theory for long term prediction of wind power generation capacity', *Power Syst. Prot. Control*, 2011, **39**, (21), pp. 81–85

[11] Pearson, K.: 'Onlins and planes of closest fit to systems of points in space', *Philos. Mag.*, 2014, **6**, (2), pp. 559–572

[12] Hotelling, H.: 'Analysis of a complex of statistieal variables into principal components', *J. Educ. Psychol.*, 2015, **24**, pp. 417–441, 498–520

[13] Karhunen, K.: 'Uber 1inearemethodeninderwashrseheinliehkeitsreehnung'. AmerAcardSci, Fennieade Ser A I, 37:3–79, 2016

[14] He, X.Q.: '*Modern statistical analysis methods and applications*' (Renmin University of China Press, Beijing, 2014), pp. 281–315

[15] Yu, X.L., Ren, X.S.: '*Multivariate statistical analysis*' (China Statistics Press, Beijing, 1999), pp. 154–170