Research on short-term wind power Prediction of GRU based on similar days

Yong Lin\textsuperscript{1,a*}, Haiing Zhang\textsuperscript{2,b}, JiYan Liu\textsuperscript{3,c}, WenJie Ju\textsuperscript{4,d}, JinYou Wang\textsuperscript{5,e}, XiaoDong Chen\textsuperscript{6,f}

\textsuperscript{1}State Grid Shandong Electric Power Company Weifang Power Supply Company, Weifang, China
\textsuperscript{2}State Grid Shandong Electric Power Company, Jinan, China
\textsuperscript{3}State Grid Shandong Electric Power Company, Jinan, China
\textsuperscript{4}State Grid Shandong Electric Power Company, Jinan, China
\textsuperscript{5}State Grid Shandong Electric Power Company Weifang Power Supply Company, Weifang, China
\textsuperscript{6}State Grid Shandong Electric Power Company Weifang Power Supply Company, Weifang, China

\textsuperscript{a}email:202014753@mail.sdu.edu.cn
\textsuperscript{b}email: 916344797@qq.com, \textsuperscript{c}email: 1251867972@qq.com,
\textsuperscript{d}email: 1634351866@qq.com, \textsuperscript{e}email: 987026307@qq.com,
\textsuperscript{f}email: 852870783@qq.com

Abstract. As the proportion of wind power generation continues to increase, accurate forecasting of wind power output is of great significance to the smooth operation of the entire power grid. However, due to the greater impact of environmental factors, wind power generation has strong randomness, and it becomes difficult to accurately predict the power generation. Thus, a new hybrid model for wind power generation prediction combining GRU neural networks and similar days’ characters analysis is proposed to address solve this problem. The prediction method employs grey relation analysis to screen similar days, which not only reduces the amount of data required to train the model, reduces the computational complexity, and improves the training speed, but also improves the prediction accuracy based on the selected datasets. In addition, this method also filters and processes the data through box-plot analysis and linear smoothing, which further improves the prediction accuracy of the model. The results show that compared with a single GRU network, the MAE of this method has dropped by 1.89, RMSE has dropped by 1.9, and MAPE has dropped by 11.07%. Obviously, the prediction model based on similar days extraction has obvious advantages.

1. Introduction

Different from the traditional thermal power generation, there is great uncertainty about wind power influenced by complicated meteorological factors. It is imperative to accurately predict the power
generated by wind farms. Wind power prediction methods can be generally divided into physical methods, statistical methods and machine learning methods [1].

The physical prediction method mainly uses meteorological information and performs power prediction based on the output power curve of the wind turbine [2]. In order to ensure the accuracy of the prediction, this method needs to consider the power generation principle of the wind turbine and many details of environmental factors. In [3], the wake effect between multiple wind turbine arrays in the convective atmospheric boundary layer is considered, and the wake loss of the computer group is used to correct the high fan output prediction. However, this method of considering physical factors requires the use of a large amount of detailed geographic and meteorological information. In actual application, it is also easily affected by factors such as the complex environment around the wind farm and the accuracy of measuring instruments.

The statistical forecasting method is to establish the functional relationship between meteorological information and wind farm output power from historical statistical data. Common statistical forecasting methods include Autoregressive Moving Average Model (ARMA) [4] and Autoregressive Integrated Moving Average Model (ARIMA) [5]. Rajesh G. and Kavasseri proposed the use of fractional ARIMA or f-ARIMA models to forecast wind speed in the day before (24 hours) and the previous two days (48 hours) [6]. For stationary time series, the accuracy of statistical prediction model is very high. However, when dealing with time series with large fluctuations like wind power output, there are still shortcomings such as weak fitting ability, difficult modeling, and low prediction accuracy.

Machine learning and deep learning methods have good prediction accuracy and operation speed in processing nonlinear and non-stationary data, so they have been widely used in the field of wind power output prediction. At present, the main models used for time series prediction are Artificial Neural Network (ANN) [7], Recurrent Neural Network (RNN) [8], Adaptive Wavelet Neural Network (AWNN) [9], Support Vector Machine (SVM) [10] etc. A hybrid prediction model based on SVR is proposed [11], and genetic algorithm is used to adjust the parameters of SVR. Experimental results show that the prediction accuracy of this method in wind speed prediction (WSF) and wind power prediction (WPF) is higher than statistical prediction models (AR, ARMA and ARIMA). However, and it is difficult to improve the accuracy only considering the mapping relationship between wind power and wind speed. Therefore, strengthening the feature extraction of data, fully considering other features hidden in multivariate data, and using hybrid forecasting models have become a research hotspot in the field of wind power forecasting in recent years.

This paper proposes a new SD-GRU model for wind power prediction, which combines GRU (Gated Recurrent Unit) neural networks and similar days' characters analysis. The contributions are as follows:

1. Comprehensively analyzing various factors affecting wind power output and select highly correlated NWP data as the modeling basis.
2. By selecting similar days based on the gray correlation degree to improve the calculation speed and prediction accuracy of the prediction model.
3. Using wind power and NWP (numerical weather prediction) data sets from Laizhou Huadan Wind Farm to compare the prediction results with traditional models for index analysis. And the results show that the predictive accuracy of the SD-GRU model is higher than that of the single model.

2. Feature analysis of wind power generation

2.1. Analysis of wind power generation model

The output power of the wind turbine obtained in [12] can be obtained from equation (1):

\[
P(t) = \frac{P}{8R_e} e^{-\left(\frac{gh}{RT}\right)} C_p \eta C_D \pi D^3 \sin^3 \theta
\]

Where \( P \) is local atmospheric pressure; \( T \) is the ambient temperature; \( H \) is altitude; \( R_e \) is the molar gas constant; \( G \) is the acceleration of gravity; \( D \) is the wheel spoke diameter of the generator; \( V \) is the
actual wind speed, $\theta$ is the Angle between wind speed and wheel surface; $C_p$ is wind energy utilization coefficient; $\eta_g$ is the power conversion efficiency of the generator, $\eta_m$ is the mechanical transmission efficiency.

According to equation (1), the output power of wind turbine is affected by meteorological factors such as local atmospheric pressure $P$, local temperature $T$, local altitude $H$, wind speed $V$, wind direction $\theta$, among which wind speed $V$ has the greatest influence. The other atmospheric pressure $P$ and temperature $T$ are minor influencing factors, which need further data analysis.

2.2. Meteorological factors affecting wind power generation

In the part of 2.1, various factors affecting the output power of wind turbines are simply analysed from the aspect of physical model. In terms of probability and statistics, the output power of wind turbines can be obtained from the following equation [13]:

$$W_0 = 8760 \int_0^\infty E(v) \sum_{i=1}^N p(v_i)$$  \hspace{1cm} (2)

Where $E(v)$ is the probability density function of average wind speed; $P(v)$ is the wind turbine power curve; $v$ is the working wind speed of the fan.

From the wind speed forecast data actually obtained from the Laizhou Huadian Wind Farm, there are two sets of surface wind speed forecast data and high altitude wind speed forecast data. Calculate the frequency distribution of these two sets of predicted data and the actual working wind speed of the wind turbine hub respectively, and draw the histograms as shown in Fig.1.

![Fig.1 Wind speed frequency distribution on sample machine, ground surface and high altitude](image)

It can be found that the upper altitude wind speed forecast is very close to the actual working wind speed frequency distribution. In the actual situation, the fan’s center height of the wind turbines is between 70-85m. At this height, the air flow is less affected by surface obstacles, so there is a large gap between the environmental wind speed and the surface wind speed, so the high-altitude wind speed should be closer than the surface wind speed.

In order to further determine the correlation between the wind speed forecast data and the actual data, take 500 days of measured wind speed and corresponding surface wind speed prediction and high-altitude wind speed prediction data, and the results of regression fitting in MATLAB are shown in Fig. 2.
From the comparison of the fitting results, it can be seen that the fitting correlation coefficient $R$ of the surface prediction data is only 0.3992, while the fitting correlation coefficient $R$ of the high-altitude prediction data is as high as 0.9016, and the scattered points are relatively concentrated. Therefore, in the subsequent construction of the forecast model, the high-altitude forecast wind speed should be used as the main NWP data, and a higher weight should be given to it.

In addition to wind speed $v$, temperature $T$, relative humidity $RH$, altitude $h$ and atmospheric pressure $p$ are all influencing factors on wind power generation. This paper takes the temperature, humidity and atmospheric pressure forecast data of the wind farm within 500 days for correlation analysis. Perform regression fitting analysis, and the fitting results are shown in Fig.3.

The values of $R$ for wind power generation and NWP data are shown in Table 1, which describes the correlation between them.

| NWP data         | Temperature | Relative humidity | Atmospheric pressure |
|------------------|-------------|-------------------|----------------------|
| Correlation coefficient | 0.0634      | 0.1317            | 0.0360               |

The scale coefficient of the power-temperature fitting curve in Fig 6 is negative, and the power-temperature scatter distribution also shows a negative correlation. The correlation coefficient $R$ between the two in the figure is only 0.0634, indicating that the temperature has a small effect on the power output of the fan in actual conditions. Similarly, from equation (2), it can be seen that the output power
is proportional to the atmospheric pressure, which is consistent with the positive value of the proportional coefficient of the power-atmospheric pressure fitting curve in Figure 8, which proves the validity of the atmospheric pressure data. But in the same way, the correlation coefficient R between the two is only 0.0360, which shows that the actual atmospheric pressure has little effect on the power output. This conclusion is also related to the small change in atmospheric pressure throughout the year.

In summary, considering the relevance of actual data, temperature and atmospheric pressure prediction data are not used in the subsequent prediction model construction.

3. Wind power prediction methods based on GRU model and similar days search

3.1. Prediction data preprocessing

In the actual data collection process, because the collection equipment itself is easily affected by environmental changes, extreme bad values and invalid data appear. Therefore, data pre-processing is the only way in the process of building a model.

Data preprocessing has four processing methods: data cleaning, data integration, data transformation, and data reduction [14]. The data cleaning process mainly deals with missing data and bad data. Among them, the method of box-plot analysis can be used to identify the bad value of the data.

Box plot is a statistical chart used to display a set of data dispersion information, it can display a set of data maximum, minimum, median, and upper and lower quartiles. According to the upper and lower quartiles in the box plot, we can define the upper and lower bounds of the error judgment as:

\[
U = U_q + 1.5IQR \\
L = L_q - 1.5IQR
\] (3)

In the equation, \(U_q\) is the upper quartile of the data; \(L_q\) is the lower quartile of the data; \(IQR\) is the difference between the upper quartile and the lower quartile.

Smoothing refers to the weighted average of surrounding data, which is a processing method that replaces the data close to the surrounding data. In the wind power forecasting problem, since the actual weather data is physical information, it has the characteristics of not being able to change suddenly, so the use of smoothing method can better restore the missing data.

If the data at the \(i\)-th point is missing (the value is zero), replace it with:

\[
\hat{x} = ax_{i-1} + bx_{i+1}
\] (4)

In the equation, \(x_{i-1}\) is the data of the \(i-1\)-th point; \(x_{i+1}\) is the data of the \(i+1\)-th point; \(a\) is the weight of \(x_{i-1}\); \(b\) is the weight of \(x_{i+1}\). Here \(a\) and \(b\) are both taken as 0.5.

3.2. GRU model and similarity day search based on grey relation analysis

GRU (Gated Recurrent Unit) is a variant of LSTM. Each neuron contains two gate structures: update gate and reset gate. The expressions are as follows:

Update gate:

\[
z_j = \sigma(\begin{bmatrix} W_z x' \end{bmatrix}_j + \begin{bmatrix} U_z h^{t-1} \end{bmatrix}_j)
\] (5)

Reset gate:

\[
r_j = \sigma(\begin{bmatrix} W_r x' \end{bmatrix}_j + \begin{bmatrix} U_r h^{t-1} \end{bmatrix}_j)
\] (6)

In the equation, \(\sigma(\cdot)\) represents the \(j\)-th element of the vector; \(x'\) is the input at node \(t\); \(h^{t-1}\) is the hidden value at the previous moment; \(W_z, U_z, W_r, U_r\), respectively is the calculated weight of each vector; \(\sigma(\cdot)\) is the sigmoid function.

Since GRU has fewer gates than LSTM (three for LSTM), and has only one transmission parameter \(h'\), its calculation amount is much smaller than that of LSTM, which can greatly save computing resources and improve model training speed.
Grey Relation Analysis [15] (GRA) is a multi-factor statistical analysis method. In a grey system (Grey System), if you want to find the correlation between a certain factor and other influencing factors, the grey correlation analysis is proposed. The basic idea of this method is to find the factor that has the greatest impact on the target factor by analysing the consistency of the target factor and the curve change trend of each influencing factor. The basic calculation steps are:

1. Each item of the original data shall be pre-processed step by step as described in Chapter 3 and normalized.

2. The sequence of target factors and influencing factors is arranged as follows:

\[
(X_0, X_1, ..., X_m) = \begin{bmatrix}
x_0(1) & x_1(1) & ... & x_m(1) \\
x_0(2) & x_1(2) & ... & x_m(2) \\
... & ... & ... & ...
\end{bmatrix}
\]

Where \(X_0\) is the target sequence and \(X_i\) is the comparison sequence.

3. Find the absolute value of the difference between each item in the comparison sequence and the corresponding item in the target sequence, and find the maximum and minimum of all the differences:

\[
\Delta_{ui}(k) = |x_0(k) - x_i(k)|
\]

In the equation, \(k \in \{1, 2, ..., n\}\), \(i \in \{1, 2, ..., m\}\)

4. Calculate the correlation coefficient of each item in the comparison sequence according to the following equation:

\[
\rho_{ui}(k) = \frac{\Delta_{\text{min}} + \rho \Delta_{\text{max}}}{\Delta_{ui}(k) + \rho \Delta_{\text{max}}}
\]

In the equation, \(\rho\) is the resolution coefficient, generally between 0 and 1, and the empirical value is 0.5.

5. Take the mean value of each item in the comparison sequence as the final degree of association of the item:

\[
r_i = \frac{1}{n} \sum_{k=1}^{n} \rho_{ui}(k)
\]

\(r_i\) is considered to be the correlation between the relevant factor \(i\) and the target factor. The value of \(r_i\) is between 0 and 1. The larger the value, the stronger the correlation between the two factors.

### 3.3. Procedure of forecast model

The GRU neural network for short-term prediction is shown in Fig.4.
The basic procedure of short-term forecast based on similar days is:

1. Select the feature vector of the prediction day as the target sequence, and select the feature vector 120 days before the prediction day as the comparison sequence. The feature vector is:

   \( x_i (1) \) — previous 12h average low altitude forecast wind speed; 
   \( x_i (2) \) — previous 12h average low altitude forecast wind direction; 
   \( x_i (3) \) — previous 12h average high altitude forecast wind speed; 
   \( x_i (4) \) — previous 12h average forecast humidity; 
   \( x_i (5) \) — 12h average output power before the previous day; 
   \( x_i (6) \) — after 12h average low-altitude forecast wind speed; 
   \( x_i (7) \) — late 12h average low-altitude forecast wind direction; 
   \( x_i (8) \) — last 12h average high-altitude forecast wind speed; 
   \( x_i (9) \) — the average forecast humidity in the next 12h; 
   \( x_i (10) \) — the average power output in the previous day and after the 12h.

   In this paper, averaging the data of the first and second half of the day is to consider the large changes in the weather data within a day, and calculating the first and second half of the day can increase the recognition of the data.

2. The selected sequence is processed as described in 3.1 to obtain the similarity between each day and the forecast day. Sort the 120 comparison days in ascending order of similarity, and select the largest 30 as the similarity days of the prediction days.

3. Use the selected low-altitude wind speed forecast, low-altitude wind direction forecast, high-altitude wind speed forecast, relative humidity forecast and the previous day's output power of the selected similar days as training data, and use them for the training of the GRU model. After the training is completed, the above-mentioned data on the prediction day is input into the model to obtain the final short-term prediction result.

4. Experiments and results

   The datasets from Laizhou Huadan Wind Farm is applied to verify the proposed prediction model. The NWP data and historical datasets of wind power generation from January 1, 2017 to May 31, 2018 are employed to train and test samples for the forecast model.

   In order to ensure the accuracy of the prediction model, it is necessary to preprocess the forecast data based on the method in 3.1. Box plot analysis of various raw data in MATLAB, some of the original data box plots are as follows:

   ![Fig.5 Part of the original data box plot](image)

   Based on the above analysis, for the bad values in each dataset above the upper bound \( U \) or below the lower bound \( L \), it can be replaced by the corresponding upper bound and lower bound \( L \) values; for missing data values like zero, they can be filled with the linear smoothing method introduced in 3.1.

   After processing the data, the next step is to determine the feature vector according to the method in 3.3, select the feature vector of the prediction day as the target sequence, and the feature vector of the
120 days before the prediction as the comparison sequence for similarity calculation. In this example, May 2, 2017 is the forecast day, and January 2, 2017 to May 1, 2017 is selected as the comparison day. The 120 comparison days are sorted in descending order of similarity, and the largest 30 of them are selected as the similarity days of the prediction day. January 9, January 14, January 27, January 28, February 15 and other 30 days are regarded as similar days, of which March 3 has the highest degree of relevance.

Select NWP data of similar days for model training. Input the NWP data on the forecast day into the model to get the forecast result on the forecast day.

Fig.6 illustrates the prediction result with different methods, including the statistical forecasting methods and machine learning methods. The blue line is the real output of the wind farm, and the red line achieved the best accuracy on the entire window.

The corresponding evaluation index comparisons are shown in Table 2:

| Evaluation index | SD-GRU  | GRU    | RNN    | ARMA   | ARIMA   |
|------------------|---------|--------|--------|--------|---------|
| MAE              | 10.08   | 11.97  | 16.60  | 17.09  | 17.43   |
| RMSE             | 12.81   | 14.71  | 18.95  | 20.12  | 20.55   |
| MAPE             | 24.97%  | 36.04% | 51.35% | 47.08% | 48.07%  |

The prediction results show that when the similar day optimization analysis is not used, the predicted MAE is 11.97. When adding the similar days model, the predicted MAE drops to 10.08, and the other error indicators also decrease to varying degrees. This result fully demonstrates the superiority of the similar days model in terms of prediction accuracy.

5. Conclusion

In this paper, a new hybrid model for wind power generation prediction combining GRU neural networks and similar days’ characters analysis is proposed. This prediction method uses grey relation analysis to screen similar days, which not only reduces the amount of data required to train the model, reduces the computational complexity, and improves the training speed, but also improves the prediction accuracy based on the selected datasets. In addition, this method also filters and processes the data through box-plot analysis and linear smoothing, which further improves the prediction accuracy of the model. The results show that compared with a single GRU network, the MAE of this method has dropped by 1.89, RMSE has dropped by 1.9, and MAPE has dropped by 11.07%. Obviously, the prediction model based on similar days extraction has obvious advantages.
Acknowledgements
This study was supported by science and technology project of Shandong Electric Power Co., Ltd. of State Grid (no.2020A-138)

References
[1] You Jing, Research on wind power prediction model based on multilayer perceptron and support vector machine. Shandong University, 2018.
[2] Lin Ye, Yongning Zhao, Cheng Zeng, Cihang Zhang, Short-term wind power prediction based on spatial model, Renewable Energy, 2017,101(1067-1074).
[3] Nielson J, Bhaganagar K. Using field data-based large eddy simulation to understand role of atmospheric stability on energy production of wind turbines. Wind Eng 2019,43(6):625–38.
[4] Liu Shuai, Zhu Yongli, Zhang Ke, Gao Jiacheng. Short-term wind power prediction based on error-corrected ARMA-GARCH model. Acta Energiae Solaris Sinica, 2020, 41(10): 268-275.
[5] Xue Xiang, Zhao Dongmei, Wang Jinyi, Zhang Jiajun. Wind power forecasting method based on the combination of ARIMA and two-layer BP neural network. Electric Power Science and Engineering, 2012, 28(12):50-55.
[6] Kavasseri RG, Seetharaman K. Day-ahead wind speed forecasting using f-ARIMA models. Renewable Energy, 2009, 34(5):1388-1393.
[7] Noorollahi Y, Jokar MA, Kalhor A. Using artificial neural networks for temporal and spatial wind speed forecasting in Iran. Energy Conversion and Management, 2016,115:17–25.
[8] Park C, Lee C, Hong L, et al. $S^2$-Net: Machine reading comprehension with SRU-based self-matching networks. Etri Journal, 2019, 41(3).
[9] Chandra DR, Kumari MS, Sydulu M, et al. Adaptive wavelet neural network based wind speed forecasting studies. Journal of Electrical Engineering and Technology, 2014, 9(6):1812–1821.
[10] Zhongda T, Xianwen G, Tong S. Combination kernel function least squares support vector machine for chaotic time series prediction. Acta Physica Sinica, 2014,63 (16): 160508.
[11] G.Santamaria-Bonfil, A.Reyes-Ballesteros, C.Gershenson. Wind speed forecasting for wind farms: a method based on support vector regression. Renewable Energy, 2016, 85:790–809.
[12] Li Chong, Zheng Yuan, Zhu Dasheng, Liu Youying, Li Guo. Research on the output power model of wind turbines. Hydropower, 2014,40(8):123-125+128.
[13] Wang Dan, Sun Changhui. The physical model of wind farm power generation calculation. China Electric Power, 2011, 44(1):94-97.
[14] X.J. Shen, X.J. Fu, C.C. Zhou. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm. IEEE Trans Sustainable Energy, 2019,10(1):46-54.
[15] Zhang Han. Research on PV power prediction based on LSTM neural network based on air quality. Xi’an University of Technology, 2020.