Ultra Short Term Wind Power Prediction Model Based on WRF Wind Speed Prediction and CatBoost

Xingdou Liu¹, Li Zhang¹,*, Zhirui Zhang¹, Tong Zhao¹ and Liang Zou¹

¹School of Electrical Engineering, Shandong University, Jinan250061, Shandong, China

*zhleee@sdu.edu.cn

Abstract. Due to the urgent requirements of clean energy development, the installed capacity of wind turbine is gradually increasing. In order to achieve effective wind turbine grid control, accurate wind power prediction is very important. In this paper, based on the US Atmospheric Research Center (NCAR) Weather Research and Forecast (WRF) model, the predicted wind field is obtained. Then, a super short-term fast wind power prediction model is proposed, which is composed of Spearman correlation coefficient method for feature extraction and catboost algorithm. With historical power, historical wind speed and predicted wind speed as inputs, the predicted power in the next 4 hours (15 minutes interval) is output. In order to evaluate the prediction ability of the proposed model, four algorithms, LSTM, CatBoost, XGBoost and LightBGM, are adopted. Taking root mean square error (RMSE), mean absolute error (MAE) and training time as evaluation indexes, the results show that CatBoost has the best comprehensive performance.

1. Introduction

The rapid development of wind power generation can gradually replace thermal power generation, which can not only reduce greenhouse gas emissions, but also reduce the cost of power generation with the progress of technology, which is a sustainable green energy. By the end of 2019, China's total installed capacity of wind power is 236 GW [1], reaching 651 GW [2] in the world. However, due to the randomness and intermittence of wind power, large-scale integrated grid connection makes power system operation regulation face challenges. Accurate prediction of wind farm output power is the basis of power system regulation decision, which is of great significance to ensure the safe and economic operation of power system and energy conservation and emission reduction [3].

Algorithm models that currently perform well in predicting accuracy, such as Deep Neural Network (DNN) [4], Recurrent Neural Network (RNN) [5], Long and Short-term Memory Network (LSTM) [6], Convolutional Neural Network (CNN) [7], Hybrid model CNN-LSTM [8-9] and other methods are widely used. However, the deep neural network has many parameters, and each training takes a long time, and the maintenance workload is relatively large when the parameters are updated in actual deployment.

Predicting power requires a more accurate and customized forecast of wind speed. Using statistics or machine learning methods to predict wind speed, the prediction time is short and lacks stability and interpretability [10]. Using numerical weather prediction (NMP) [11] from nearby weather stations may cause time and space position deviations.
This paper uses the forecast data released by the Global Forecast System (GFS) to drive the WRF model to obtain the predicted wind speed of the horizontal component. Use the Spearman correlation coefficient method to select the input feature quantity, take the historical power, historical wind speed and predicted wind speed as the input of the model, adopt LSTM, XGBoost, LightBGM, CatBoost four algorithms, output the predicted power for the next 4 hours (interval 15 minutes). Results in root mean square error (RMSE) and mean absolute error (MAE), except LSTM, CatBoost performs best among the three ensemble algorithms. For the same dataset, the training efficiency of the three algorithms is much higher than LSTM, which is conducive to real-time updating of model parameters in daily prediction.

2. Symbol Definition and Algorithm Framework
In order to better demonstrate the actual work of this article, firstly, the symbol definition of the parameters involved in the article is given. The defined symbols include predicted power, predicted wind speed obtained by WRF, and wind turbine operation and meteorological data recorded by the data acquisition and monitoring control system (SCADA), as shown in Table 1.

| Characteristic quantity       | Symbol   |
|------------------------------|----------|
| Predicted power              | $P_{\text{predict}}$ |
| Predicted wind speed         | $W_{\text{predict}}$ |
| Historical wind speed        | $W_{\text{history}}$ |
| Historical power             | $P_{\text{history}}$ |
| Future power                 | $P_{\text{future}}$ |
| Wind direction                | $W_{\text{direction}}$ |
| Temperature                  | $K$      |
| Humidity                     | $H$      |
| pressure                     | $p$      |

After analysis by Spearman correlation coefficient method, select $W_{\text{history}}$, $P_{\text{history}}$, and $W_{\text{predict}}$ as the model input features, and output $P_{\text{predict}}$. Apply LSTM, XGBoost, LightBGM and CatBoost algorithms for model training, calculate RMSE and MAE based on $P_{\text{predict}}$ and $P_{\text{future}}$, and record the training time at the same time. XGBoost, LightBGM, CatBoost, as improved algorithms of GBDT, have similar principles and structures. The frame structure of the entire model is shown in Figure 1. The time series corresponding to all data is given, where $t$ is the current time and the interval $T$ is 15 minutes.

![Figure 1. Forecast framework and process](image)
3. Method introduction

3.1. Introduction to WRF mode

The WRF model is a new generation of mesoscale forecasting model and assimilation system jointly researched by the US Atmospheric Research Center. The forecast data issued by the Global Forecast System (GFS) can be downscaled to obtain the forecast wind speed of a specific wind field. The basic parameters are as follows:

(1) The selected wind farm is located in Jining City, Shandong Province, China. The simulation area selected in this paper is shown in Figure 2. With (35.5°N, 117°E) as the center, a three-layer nesting scheme is adopted, and the number of horizontal grids (resolution) are 100×100 (18 km), 100×100 (6 km), 121×121 (2 km). The selection of important physical parameters is shown in Table 2.

(2) The recorded data from the wind farm is converted to UTC from January 1, 2019 to September 28, 2020. Download the historical GFS data at the corresponding time with a resolution of 0.25°×0.25° (approximately 28 km).

Figure 2. Nesting range of target area in WRF

Figure 3. Predicted and actual wind speed

Table 2. Symbol definition

| Physical scheme       | Mode scheme selection                      |
|-----------------------|--------------------------------------------|
| Microphysics          | WRF Single-Moment 5-class scheme           |
| Longwave radiation    | RRTM scheme                                |
| Shortwave radiation   | Dudhia                                     |
| Surface layer         | MM5 scheme                                 |
| Land surface          | Noah land surface model                    |
| Planetary boundary layer | Yonsei University scheme                  |
| Cumulus parameterization | Kain-Fritsch scheme                      |

The comparison between the obtained predicted wind speed and the real wind speed of 200 continuous wind speed samples is shown in Figure 3. The curve shows that the recorded real wind speed fluctuates greatly, but the trend is basically consistent with the predicted wind speed.

3.2. Spearman correlation coefficient analysis

The Spearman correlation coefficient is a non-parametric rank statistical parameter used to measure the strength of the relationship between two variables. Assuming that there are two variables $X$ and $Y$
in the sample, the Spearman correlation coefficient of $X$ and $Y$ is shown in formula (1).

$$P_s = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2)^{1/2}}$$  \quad (1)

According to Table 3, based on the Spearman correlation coefficient value, $P_{\text{future}}$ has a strong correlation with History, $W_{\text{predict}}$, $W_{\text{history}}$, and $P_{\text{history}}$. At the same time, it has a certain correlation with temperature and pressure, while it has a weak correlation with weather elements such as wind direction and humidity. This paper selects History, Phistory and Wpredict as the input features of the prediction model.

**Table 3. Spearman correlation coefficient value**

|               | $P_{\text{future}}$ |
|---------------|---------------------|
| $W_{\text{predict}}$ | 0.77                |
| $P_{\text{history}}$  | 0.62                |
| $W_{\text{history}}$  | 0.52                |
| $p$             | 0.019               |
| $K$             | 0.028               |
| $W_{\text{direction}}$ | -0.01             |
| $H$             | -0.12               |

3.3. *Introduction to CatBoost algorithm*

CatBoost algorithm is an improved gradient lifting tree (GBDT) algorithm, which is an algorithm based on boosting method. Gradient descent operation is added on the basis of boosting [12], that is, every time the regression tree is established, the information of the former regression tree is used to calculate the gradient descent direction of the previous decision tree. The enhancement of catboost function is mainly reflected in three aspects:

1. CatBoost adopts the "order principle" to avoid the inherent conditional displacement problem in the iterative process of gbdt algorithm, and makes it possible to use the whole data set for training and learning.

2. CatBoost transforms the traditional gradient enhancement algorithm into ordered boosting algorithm, which solves the inevitable problem of gradient migration in the iterative process, improves the generalization ability, reduces the possibility of over fitting of the model, and enhances the robustness of the model.

3. CatBoost constructs the combination of classification features through greedy strategy, and takes these combinations as additional features, which helps the model to capture higher-order dependencies more easily and further improve the prediction accuracy.

4. Research case

The case in this article comes from a wind farm in Jining City, Shandong Province, China, to verify the superiority of the proposed method in prediction accuracy and training prediction time. The installed capacity of the target wind farm is 93MW, and the time span for collecting data from SCADA is from January 1, 2019 to September 28, 2020. It is divided into training set and test set for model training and testing, using four algorithms model predicts wind power for the next four hours (15 minutes apart).

4.1. Data analysis

After being screened by Spearman's correlation coefficient method, the final generated data set includes 60,141 pieces of data, and the characteristic quantities include $W_{\text{predict}}$, $W_{\text{history}}$, $P_{\text{history}}$, and $P_{\text{future}}$. There are certain omissions or erroneous data in the data recorded in the wind field. Now it is stipulated that 16 consecutive points (four hours) are zero or unchanged data as abnormal data. After
analysis, the abnormal recorded power is 818 and the abnormal recorded wind speed is 1178. The overall effective rate of the data is 96.68%.

4.2. Evaluative index
In order to evaluate the accuracy of the proposed method, several error calculation standards need to be determined. This paper selects the absolute value error (MAE) and the root mean square error (RMSE) as the error evaluation indicators, which are $P_{\text{future}}$ and $P_{\text{predict}}$. The final error measurement is defined as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

$$\text{RMSE} = \left( \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right)^{1/2}$$

4.3. Model training and prediction effect
Four algorithms: LSTM, XGBoost, LightBGM, and CatBoost are used to predict the wind power in the next 4 hours at intervals of 15 minutes. The predicted power of the four algorithms is drawn into a curve and compared with the actual power in the corresponding time period. Taking 200 consecutive power points as samples, the drawn power prediction curve is shown in Figure 4.

![Predicted and actual wind power](image)

Figure 4. Predicted and actual wind power

According to the graph, the actual recorded power fluctuates greatly, so it is difficult to compare the actual prediction effect of the algorithm. MAE and RMSE of predicted power and actual power of the four algorithms are calculated and shown in Table 4 together with training time. And draw the error contrast curve as shown in Figure 5.

| Algorithm   | MAE (MW) | RMSE (MW) | Training time (s) |
|-------------|----------|-----------|-------------------|
| LSTM        | 7.299    | 9.603     | 894               |
| CatBoost    | 7.504    | 9.741     | 9.2               |
| LightBGM    | 8.003    | 9.920     | 1.8               |
| XGBoost     | 7.618    | 9.910     | 19.7              |

Table 4. MAE and RMSE
From the prediction error index of the four algorithm models, LSTM is the best, but its shortcomings are very prominent. The training time is about 15 minutes, which is close to a prediction interval. In the integration algorithm, the error values of Mae and RMSE of CatBoost are the lowest, and the three algorithms are far lower than LSTM in terms of training time, which is very conducive to online real-time training in practical application.

5. Conclusion
(1) Wind speed prediction is very important for wind power prediction. Using WRF model to forecast wind speed can get more accurate wind field forecast wind speed. Compared with the NWP issued by the weather station, the wind speed scale of this paper is reduced to 2 km, which is closer to the center of the wind field.
(2) For the continuous power prediction task, LSTM has the structure of memory unit and forgetting gate, and the prediction accuracy is better than the three integrated algorithms, but the training time is long, and the efficiency of updating model parameters is low.
(3) Compared with XGBoost and LightGBM, CatBoost embeds an innovative algorithm that automatically processes category features into numerical features. The training speed and memory consumption are between xgboost and lightgbm, and the prediction accuracy is higher than the former two algorithms. It can quickly update model parameters in actual deployment.

References
[1] China Wind Energy Association 2020 China Wind Power Industry Mapping 2019
[2] World Wind Energy Council 2020 GLOBAL WIND REPORT 2019
[3] Thor S E and Weis-Taylor P 2020 Wind Energy 5 73-75
[4] Marino D L, Amarasinghe K and Manic M 2016 Italy Proceedings of the IECON 42nd Annual Conference of the IEEE Industrial Electronics Society 7046-51
[5] Liu H, Mi X W and Li Y F 2018 Energy Convers. Manag 498-14
[6] Aydin O and Guldamlasiglu S 2017 Turkey Proceedings of the 2017 4th International Conference on Electrical and Electronic Engineering (ICEEE) 281-85
[7] Mkholdyar M and Wang J H 2019 IEEE Trans. Sustain. Energy 10 670-81
[8] Woo S, Park J, Park J and Manuel L 2020 IEEE Trans. Sustain. Energy 11 2294-04
[9] Sari A P, Suzuki H, Kitajima T and Yasuno T 2020 Thailand 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE) 958-63
[10] Zhang C Y, Chen C L P, Gan M and Chen L 2015 IEEE Trans. Sustain. Energy 6 1416-25
[11] Du P 2019 IEEE Trans. Sustain. Energy 10 2133-41
[12] Gou X and Xiao X Y 2020 Journal of Xi’an University of Technology 36 129-34