Two Strategies to Reduce the Negative Effects of Abnormal and Missing Values for Wind Power Forecasting

Ruizhi Zhang
Shanghai University
Shanghai, China
1686431212@qq.com

Zeyu Long
Shanghai University
Shanghai, China
484106361@qq.com

Yusu Mao
Shanghai University
Shanghai, China
1908596260@qq.com

Nengjun Zhu*
Shanghai University
Shanghai, China
zhu_nj@shu.edu.cn

ABSTRACT

The spatial dynamic wind power prediction is vital for energy conservation and emission reduction. Baidu has organized a challenging competition for this problem in KDD CUP 2022. The released wind power generation dataset, namely SDWPF, was collected from 134 wind turbines in a wind farm for 245 days. Besides, the official also delivers a baseline which is a Gated Recurrent Unit (GRU)-based model.

SDWPF dataset consists of external and internal features. The former mainly includes the wind speed and the temperature of the surrounding environment. The latter includes the temperature inside the turbine nacelle and the nacelle direction, etc. Both are important and can help to predict the spatial dynamic wind power according to our experiments. Thus, the missing and abnormal values on the dataset can decrease the model's performance. However, these challenges always exist on SDWPF. Specifically, the proportion of the missing values is 0.155%. For the abnormal values, taking the Patv feature as an example, the proportion of abnormal values is 0.311%. To this end, we attempt to improve the performance of official GRU model from multiple aspects, including the interpolation of missing values and the process of abnormal values. The experiments conducted on PaddlePaddle verify the effectiveness of our strategies.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; • **Information systems** → *Data cleaning*.

KEYWORDS

Data cleaning, Time series analysis, Gated Recurrent Unit

*The 123123 team finished the work, and Nengjun Zhu was the team adviser.

Conference acronym 'XX, June 03–05, 2022, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Ruizhi Zhang, Zeyu Long, Yusu Mao, and Nengjun Zhu. 2022. Two Strategies to Reduce the Negative Effects of Abnormal and Missing Values for Wind Power Forecasting. In *Proceedings of KDD (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The variability of wind power supply can present significant challenges to integrating wind energy into the grid system. Therefore, wind power forecasting (WPF) has been widely regarded as one of the most critical problems in wind power integration and operation. However, dealing with the WPF problem is still challenging because high prediction accuracy is always required to ensure the grid's stability and supply security.

The KDD CUP 2022 organized by Baidu offers a unique spatial dynamic wind power forecast dataset: SDWPF, which includes wind power generation data from 134 wind turbines in a single wind farm for 245 days, as well as their relative position and internal state. We attempt to predict wind power generation through the GRU model and improve the quality of the dataset by some data processing strategies.

GRU is a simplified version of the LSTM (Long Short-Term Memory) recurrent neural network model which reduces vanishing gradient and exploding gradient [1]. GRU uses only one state vector and two gate vectors, reset gate and update gate. It has less training parameters and therefore occupies less memory, making it execute faster and trains faster than LSTM [2].

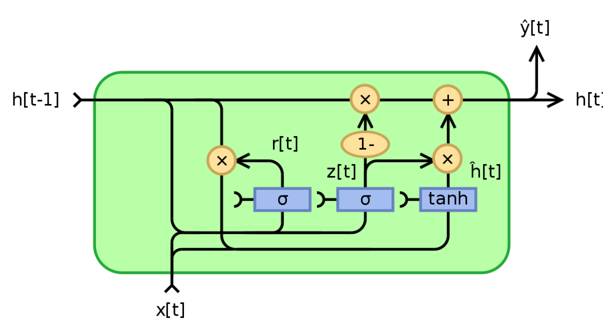

Figure 1: The struction of GRU [Wikimedia: https://en.wikipedia.org/wiki/Gated_recurrent_unit]

GRU is widely used in time series prediction. It can deal with more complex problems by combining with a high-dimensional feature extraction unit, such as a Convolutional neural network (CNN). For example, Wu et al. [3] combined GRU with CNN to propose a GRU-CNN hybrid neural network model. GRU part is responsible for extracting feature vectors of time series data and CNN extracting feature vectors of high-dimensional data. This model is applied to improve the power system's short-term load prediction (STLF). Compared with individual GRU or CNN, the hybrid model can better process time series data and extract data set features simultaneously.

In data processing, to deal with the missing values, all the missing values can be set to 0. After that, all data are normalized to scale the numeric values to a typical range, such as $[0, 1]$, without misshaping contrasts between instances. We use the following formula to perform a z-score normalization on each value in the dataset:

$$\widehat{x} = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original value, $\widehat{x}$ represents the normalized form of x, μ and σ are the mean and standard deviation of data, respectively.

Though conducting the z-score normalization, the negative influence of outlier is suppressed when training the model. Then, we use the data of first 214 days as training data and the remain 31 days as validation data. Finally, the overall score of GRU model is -46.1125 in phase 2 of KDD CUP 2022.

Besides, to handle abnormal values, we utilize the boundary value of each feature to replace the abnormal ones. This strategy can also improve the performance of GRU model.

2 RELATED WORK

With the increasingly prominent problems of the energy and environment, wind energy, as a pollution-free and sustainable clean energy, has become the focus of energy development. Domestic wind power generation has also maintained steady growth. With the steady growth of the wind power scale, its problems have been gradually amplified. Influenced by a variety of environmental factors, the volatility and randomness of wind power generation have become important factors restricting its development. Currently, wind power prediction is conducted to evaluate the requirements in advance.

The physical model forecasting method and data model forecasting method are two main methods for wind power forecasting at present. The physical model prediction method mainly considers the influence of numerical weather prediction (NWP) and other factors. Wind power prediction based on this method requires many parameters, such as historical and meteorological data, which leads to its relatively general performance in short-term forecasting ability. Its results are more suitable as a reference standard for long-term forecasting.

The mathematical model prediction method is currently the most widely studied wind power prediction method. In the mathematical prediction model, wind power prediction can be divided into two forms. One prediction method is to consider the influence of various factors on wind power generation, get the corresponding nonlinear regression curve by establishing a mathematical model, and then predict the wind power generation. Traditional mathematical model prediction methods, such as wavelet algorithm [4] and support vector machine [5], are based on the short-term prediction of wind power-related features. They forecast wind power by analyzing the features that affect wind speed and wind power. The other prediction method is to ignore the influence of other environmental conditions and only consider the trend of wind power itself over time, and then carry out the time series prediction analysis of wind power. Autoregressive Moving Average [6] is a major application algorithm for sequential forecasting. It quantifies the relationship between current data and previous data and solves the problem of solving random variation terms so as to obtain the data at the next moment of demand forecasting.

The artificial neural network is often used in the short-term forecasting of wind power. Due to the diversity of the neural network, its corresponding prediction methods are also different. For example, Li Bin et al. [7] proposed a prediction method based on an artificial neural network. It takes a variety of features as neural network input to predict the probability distribution curve of wind power. Jiang Yuechun et al. [8] combine the traditional BP neural network with a variety of algorithms to predict the time series of wind power in a short time sequence and achieve good prediction results. Due to the excellent nonlinear data fitting ability of the neural network, in many short-term and ultra-short-term prediction methods for wind power, a neural network algorithm is used to analyze historical data. Then other prediction methods are used to predict wind power and its distribution probability [9].

3 SOLUTION OVERVIEW

We have tried combining CNN, GRU, and the bidirectional GRU model to predict wind power. We also tried adjusting the baseline model's parameters, such as the number of hidden layers and the number of neurons. Unfortunately, the score of the result has not been improved.

Therefore, not only is the model's structure important, but the pre-processing of the dataset is also necessary, especially when there are many missing and abnormal values. In the official GRU model, the missing value is simply set to 0, and the abnormal values are not processed. Therefore, we start from the missing and abnormal values in the dataset to study how to process these data to improve the model training.

In dealing with the missing values, we compare the method of backward fill and interpolation method. The experimental results show that the backward fill method can achieve better performance.

Meanwhile, to handle abnormal values, we are going to start by deleting abnormal values directly. However, deleting abnormal values is a way to reduce data anomalies by reducing historical data, which will cause a large waste of data. There are many abnormal values in the dataset. For example, the number of abnormal values of the feature Patv accounts for 0.311% of the total number. If directly deleted, the training data will be greatly reduced, which may affect the results' accuracy. Therefore, we replace abnormal values with boundary values. By modifying the feature number of abnormal values, we find that only dealing with the abnormal values of Pab1, Pab2, Pab3, and Patv works better.

4 DETAILED METHOD

This section covers two ways to fill in the missing values and modify abnormal values in the dataset.

Missing data is defined as the values or data that is not stored in the given dataset. It is important to handle the missing values appropriately because many algorithms will fail if the dataset contains missing values or the model will lead to incorrect results if the missing values are not handled properly.

The first step in handling missing values is to look at the data carefully and find out all the missing values. We take a detailed look at the percentage of missing values in the dataset. The missing value exists in the feature of Wspd, Wdir, Etmp, Itmp, Pab1, Pab2, Pab3 and Patv and the number of missing values for each feature is 49518, accounting for 0.155% of the total, as shown in Figure 2. In general, the distribution of missing values is evenly distributed across all features.

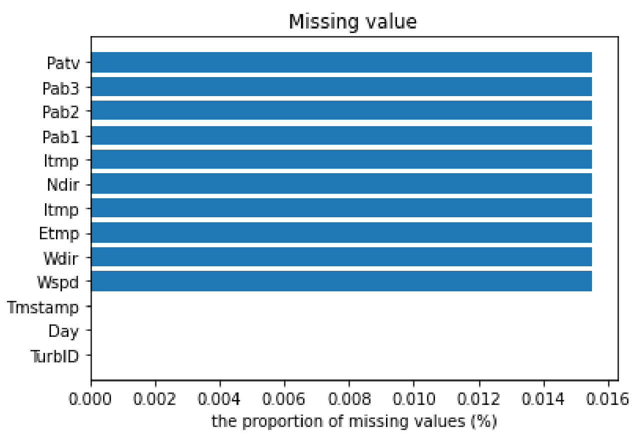

Figure 2: The proportion of the missing values in the dataset

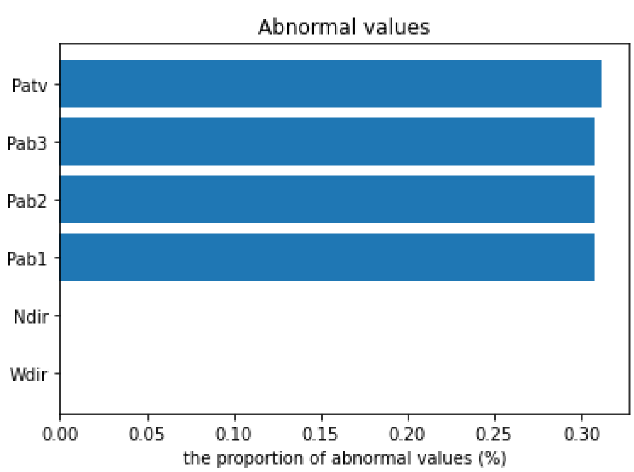

Figure 3: The proportion of the abnormal values in the dataset

Abnormal data is a value that does not belong to the normal data distribution, which occurs when there is a measurement error. The result of the model will be affected due to the presence of abnormal values. So, before training the model, we should eliminate them.

According to the official document [10], we find the abnormal values in the dataset and study the number and proportion of the abnormal values in the feature Patv, Pab1, Pab2, Pab3, Ndir and Wdir. As shown in Figure 3, each feature has a different number of abnormal values. There are 995,377 abnormal values of feature Patv, 980,991 abnormal values of feature Pab1, 981,831 abnormal values of feature PAB2, and 980,665 abnormal values of feature Pab3. Meanwhile, there are only 78 abnormal values of feature Ndir and only 49 abnormal values of feature Wdir. Because the distribution of abnormal values is extremely uneven, the focus of abnormal values modification is mainly on Patv, Pab1, Pab2 and Pab3, followed by Ndir and Wdir.

4.1 Missing values

There are many different ways of handling missing values, such as replacing With arbitrary value, replacing with previous value – Forward fill and replacing with next value – Backward fill.

All the missing values are set to 0, which is replacing With arbitrary value. And we think this method can't reflect the changes in the time series. According to the official CSV file, there are missing values in the first line of data, but there is no missing value in the last line. Therefore, we replace the missing value with the next value rather than the previous value.

However, in the process of training, we find that different epoch numbers have different results for train loss and validation loss. As shown in Figure 4, the validation loss of turbine 0 is constantly increasing. So the over-fitting appeared when epoch is set as 3. Therefore, the epoch is set to 1 during the training.

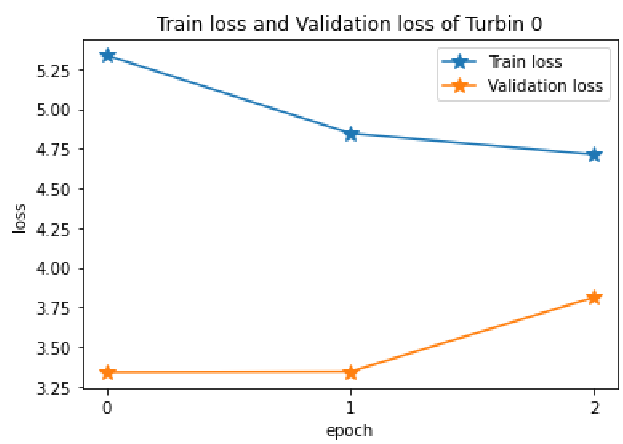

Figure 4: The train loss and validation loss of the turbine 0 when the epoch is set as 3

In the mathematical field of numerical analysis, interpolation is a type of estimation, a method of constructing finding new data points based on the range of a discrete set of known data points [11, 12].

Missing values can be imputed using interpolation. Pandas interpolate method can be used to replace the missing values with different interpolation methods like 'polynomial', 'linear', 'quadratic'. Default method is 'linear'. In this experiment, we adopted the simplest initial parameter, the method 'linear', which means that the interpolation function is linear, and the filling value is the average of the upper and lower values.

As it is shown in Figure 5, given the two red points, the red line is the linear interpolation between the points, and the value y at x, such as the blue points, may be found by linear interpolation.

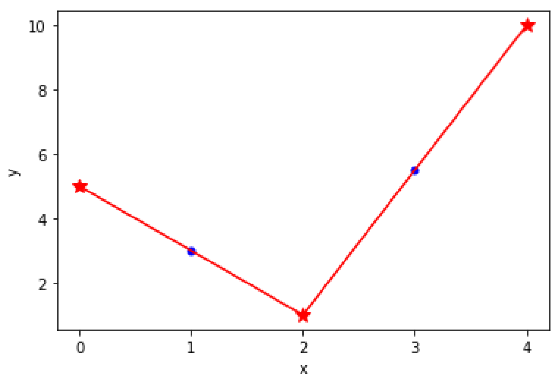

Figure 5: The linear interpolation

If the two known points are given by the coordinates (x_0, y_0) and (x_1, y_1), the linear interpolation is the straight line between these points. To find the unknown value y at x, gives:

$$y = \frac{(x_1 - x)y_0 + (x - x_0)y_1}{x_1 - x_0}, \tag{2}$$

which x is between x_0 and x_1.

4.2 Abnormal values

According to the official document [10], Pab1, Pab2, Pab3 should be less than 89 degrees. Patv should be greater than 0. The normal range of Ndir should be in [-720°, 720°]. The normal range of Wdir should be in [-180°, 180°]. But considering that training data will be reduced a lot if directly deleting abnormal values, so we replace the abnormal values with boundary values. For example, if the Ndir value is -884.86, then change it to -720.

5 EXPERIMENT

5.1 Filling missing values

When the abnormal values of Patv, Pab1, Pab2 and Pab3 are replaced by boundary values, we compare the method of backward fill and the method of interpolation. The evaluation results are shown in Figure 6.

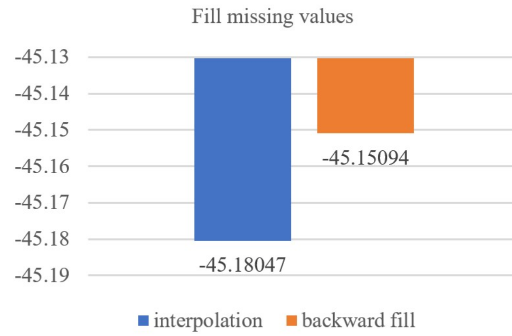

Figure 6: The result of filling the missing values in Phase 2

As the Figure 6 shows, filling the missing values with with next values is better than interpolation. The interpolation method is linear interpolation and the effect is not as good as the other method probably because the interpolation method is not suitable for weather prediction. In terms of correlation, as shown in Figure 7, the correlation between wind speed and wind power is the highest, which is about 0.814801. However, wind speed does not change linearly, so linear interpolation is not suitable to fill the missing value. Therefore, the linear interpolation model is a little less effective.

	TurbID	Day	Wspd	Wdir
TurbID	1.000000	0.000000	-0.041406	0.001285
Day	0.000000	1.000000	-0.163576	0.013096
Wspd	-0.041406	-0.163576	1.000000	-0.025661
Wdir	0.001285	0.013096	-0.025661	1.000000
Etmp	-0.049863	-0.121619	0.042256	-0.004776
Itmp	-0.055087	-0.423885	0.170341	-0.011901
Ndir	0.004805	0.076267	-0.058581	0.006929
Pab1	0.022065	0.049154	-0.414923	0.037285
Pab2	0.022758	0.049783	-0.415604	0.037333
Pab3	0.021608	0.049542	-0.414949	0.037359
Prtv	0.039763	-0.010860	0.185128	0.004143
Patv	-0.062815	-0.063410	0.814801	-0.025926

Figure 7: Correlation coefficients of TurbID, Day, Wspd, Wdir and Patv

5.2 Modifying abnormal values

The abnormal values are replaced by boundary values, and the missing values are replaced by the data in the following line. The different number of features of abnormal values is compared, and the evaluation results is shown in Figure 8.

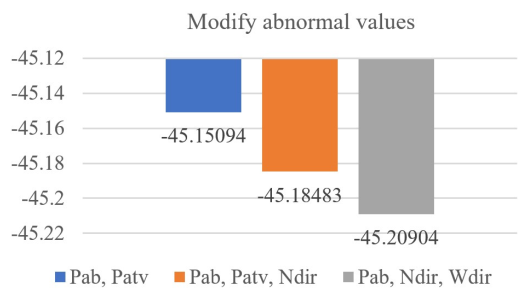

Figure 8: The result of modify the abnormal values

The method that only modifies the abnormal values of Pab and Patv is optimal. This probably because too fewer abnormal values (outliers) and less noise is bad for forecasting. Because a certain amount of noise can improve the generalization of the model. For example, during text classification, punctuation marks will be randomly inserted into the original text to argument data [13]. So when all the abnormal values are modified, the noise is greatly reduced, so the robustness of the model is decreased, and the final prediction effect is also decreased.

6 CONCLUSION

Baidu has organized a challenging competition in KDD CUP 2022. The task of the competition is to predict long-term wind power. In this paper, we attempted to improve the performance of the GRU model released by the official. Through carefully analyzing the SDWPF dataset, we found that there are many missing and abnormal values, which have a negative influence on the model's performance. Thus, we proposed two strategies to handle these two problems respectively. For missing values, they are first set to 0, followed by a z-score normalization. At the same time, abnormal values are replaced by boundary values in different features. Finally, the experiments implemented on PaddlePaddle have verified the effectiveness of these two strategies. There are some directions for future work, such as modeling the Spatio-temporal information by graph neural networks and modeling the tendency of long-term wind power instead of modeling multiple points.

REFERENCES

[1] Cho, Kyunghyun; van Merrienboer, Bart; Bahdanau, DZmitry; Bengio, Yoshua (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv:1409.1259
[2] X. Li and S. Wang. Energy management and operational control methods for grid battery energy storage systems. In CSEE Journal of Power and Energy Systems, vol. 7, no. 5, pp. 1026-1040, Sept. 2021, doi: 10.17775/CSEEJPES.2019.00160
[3] Lizhen Wu, Chun Kong, Xiaohong Hao, Wei Chen. A Short-Term Load Forecasting Method Based on GRU-CNN Hybrid Neural Network Model. Mathematical Problems in Engineering, vol. 2020, Article ID 1428104
[4] Mingli Yang, Sanming Liu, Zhijie Wang. Kalman Filter and Wavelet Neural Network Wind Speed Prediction[J]. In Proceedings of the CSU-EPSA. 2015(12): 42-46
[5] Li Guoqing, Zhang Yu, Zhang Mingjiang et al. The wind power real-time on the EEMD and SVM of the MRMR[J]. Journal of Northeast Electric Power University,2017,37(2):39-44
[6] Cao JunboZHOU, Renjun, Deng Xuehua, Fan Wenshuai, Liu Lili, Sun Jiagan. Wind Power Forecast Considering Differential Times of Optimal ARIMA Model. In Proceedings of the CSU-EPSA. 2019, 31(1): 105-111
[7] Li Bin, Peng Shurong, Peng Junzhe, Huang Shijun, Zheng Guodong. Wind power probability density forecasting based on deep learning quantile regression model[J]. Electric Power Automation Equipment. 2018, 38(9): 15-20
[8] Jiang Yue-chun, Zhang Bing-jiang, Xing Fang-fang, et al. Super-short-term multi-step prediction of wind power based on GA-VNN model of chaotic time series[J]. Power System Technology. 2015, 39(8): 2160–2166.
[9] Li D, Ren Z Y, Yan W, et al. Month-ahead wind power curve probabilistic prediction based on factor analysis and quantile regression neural network[J]. In Proceedings of the CSEE, 2017, 37(18): 5238-5248
[10] Zhou, et al. SDWPF: A Dataset for Spatial Dynamic Wind Power Forecasting Challenge at KDD Cup 2022.Baidu KDD Cup 2022. Mar. 16 – Jul. 17, 2022
[11] Sheppard, William Fleetwood (1911). "Interpolation" . In Chisholm, Hugh (ed.). Encyclopædia Britannica. Vol. 14 (11th ed.). Cambridge University Press. pp. 706–710.
[12] Steffensen, J. F. (2006). Interpolation (Second ed.). Mineola, N.Y. ISBN 978-0-486-15483-1. OCLC 867770894.
[13] Akbar Karimi, Leonardo Rossi, Andrea Prati. AEDA: An Easier Data Augmentation Technique for Text Classification. In Proceedings of EMNLP 2021