Fault Prediction of Fan Gearbox Based on K-Means Clustering and LSTM

Hanfeng Zheng¹, Yiru Dai², Yiqing Zhou³

¹ CIMS Research Center, Tongji University, Shanghai, 201800, China
² CIMS Research Center, Tongji University, Shanghai, 201800, China
³ CIMS Research Center, Tongji University, Shanghai, 201800, China
*1399766691@qq.com

Abstract. The traditional fault prediction model of wind turbine equipment mainly aims to establishing the degradation characteristic curve of the wind turbine, and draws conclusions from the analysis of the curve. However, in the actual industry, the wind turbine operating data has a high degree of nonlinear complexity and the traditional enterprises pay less attention to this aspect, which leads to partial loss in the fault characterization and serious loss in the label data. This makes the previous data processing more difficult, and it also causes the accuracy of the later prediction results to be inferior. Therefore, this paper uses k-means clustering algorithm to process the original running data into K clusters, and then comprehensively analyzes the cluster results and the provided incomplete fault list to extract useful information. The data is built to fit the Long Short-Term Memory (LSTM) model predictions. Comparing the prediction results with the SVRM method reveals that LSTM has certain advantages in fault prediction for the data in this paper.

1. Introduction

As China's industrialization process continues to accelerate, more and more enterprises are adopting intelligent and mechanized production equipment, and the complexity of equipment is also increasing. This requires us to further improve the reliability and safety of equipment. Once the equipment fails, it will not only affect the production efficiency of the enterprise, but also cause huge economic losses, and may also lead to casualties. At present, equipment failure prediction and early warning is a hot spot in the industrial field. To this end, many scholars have used different methods to establish fault prediction models for equipment to improve prediction accuracy and have achieved rich research results.

At present, the methods for predicting equipment failure can be roughly divided into three categories: methods based on equipment mechanism model, methods based on statistics, and methods of machine learning. Among them, [1-2] is based on the device mechanism model to analyze the changes of strain, stress, cumulative fatigue damage, energy, etc. during equipment failure, and find out the change relationship. The statistical-based approach is based on a statistical analysis of historical data that reflects the state of the device in a normal state. Wang et al. [3] used the Weibull distribution and the proportional hazard model to reflect the operating state of the equipment and finally got a good prediction result. The machine learning-based method uses machine learning and other methods to establish a fault prediction model. Common methods include neural networks and support vector machines. Shin et al. [4] used the support vector machine to predict the fault of the
wind turbine. Wenbo Zhang et al. [5] used the dynamic cuckoo search algorithm to improve the BP neural network to predict the failure of industrial equipment. The improved prediction model has faster convergence and higher precision.

The prediction method based on the mechanism model needs to have a clearer understanding of the equipment mechanism, increasing the complexity of the model, and the prediction accuracy is not high. In addition, most statistical-based methods only consider the relationship between input and output, the application range is limited, and the prediction accuracy of the model is not high. Compared with the above two methods, the machine learning based method is more suitable, and the recurrent neural network can simulate the Markov decision process in time series. Due to the excessive external influence factors in the actual application, and the lack of attention to the data information and the lack of understanding of the equipment failure, the data marking of the machine is not clear enough or not marked in time, which causes troubles in post-processing. For example, Hui Han et al. [6] used the dynamic hierarchical clustering of rough set theory to process the original data. The improved algorithm based on the identifiable function attribute reduction reduces the redundant attributes of the data decision table and improves the validity of the input model data. Therefore, this paper adopts LSTM network as the method of fault prediction for wind turbine gearbox. It is divided into two steps. The first step is data processing. The k-means algorithm is used to cluster the original fault marker unclear data, and then compare and analyze with the actual fault label to extract more useful data. Then in the second step the LSTM structure and simulation are set. In recent years, support vector regression (SVRM) has been applied more in time series prediction, and many exciting conclusions have been obtained [7]. Therefore, the simulation results of LSTM network are compared with the results obtained by SVRM.

2. Cluster analysis

The cluster analysis in this paper analyzes the regularity in the sample data [8], and preliminarily divides some sample data in the original data with less clear fault labels, which is helpful for later model input.

2.1. Process of k-means algorithm

The core idea of k-means clustering is to first find the distance between each sample point, and then divide the sample into K clusters according to the distance of all sample points. The division is based on minimizing the distance within each cluster, and the distance between the clusters is as large as possible, which can be expressed by mathematical expression (1), that is, the minimum squared error E, where \( \mu_i \) is the mean vector.

\[
E = \sum_{i=1}^{K} \sum_{x \in C_i} \| x - \mu_i \|^2
\]  

The purpose of the algorithm is to further divide the sample data into K clusters, combine the results with the provided incomplete fault list, and comprehensively consider the division of normal data and fault data for the sample data for subsequent data processing and model input. The algorithm input is the data set \( D=\{x_1,x_2,...,x_m\} \) of the sample set m-dimension, the cluster number K of the cluster, and the output is \( C=\{C_1,C_2,...,C_k\} \), which satisfy the minimum squared error E division.

Key steps for k-means clustering:

1. Create k points as the starting centroid (random selection);
2. Calculate the distance between the centroid and the data point for each data point in the dataset;
3. Assign data points to the nearest cluster, calculate the mean and square error;
4. Recalculate the centroid of each cluster;
5. Repeat 234 until the center of mass no longer changes;
2.2. Determination of the number of cluster centers

How to choose the value of the cluster number K is a key problem in the k-means algorithm, which has a relatively large impact on the late clustering results. In the past, it was judged according to experience, the efficiency was poor, and the reliability was not high. In this paper, the hierarchical clustering method is used to select the appropriate distance from the tree to determine the K value. As shown in the clustering result in Figure 1, the ordinate is the distance of each cluster center, and the K value is 3 is a good choice. The clustering result is tagged and output, combined with the incomplete fault list provided for comprehensive analysis, and the last required data is taken out.

![Dendrogram of hierarchy clustering](image)

Figure 1. Dendrogram of hierarchy clustering

3. Based on LSTM prediction model

According to the contents of the fault list provided by the analysis, it is found that the probability of failure of the fan caused by the gearbox failure is large, and in the gearbox fault data provided, the oil pressure of the oil port of the gearbox oil pump is basically different from the normal value. This failure prediction for the fan gearbox can be reflected by the oil pressure at the suction port of the gearbox oil pump.

Therefore, the fan data of the gearbox oil pump suction port is used as the output target, and the parameters including the L1 phase power factor of the grid side, the L1 phase current of the grid side, and the L2 phase current of the grid side, grid side L3 phase current, grid side frequency, grid side L1 phase voltage, grid side L2 phase voltage, grid side L3 phase voltage, generator active power, generator reactive power, generator speed, generator drive end bearing temperature, generator stator U-phase coil temperature, generator stator V-phase coil temperature, generator stator W-phase coil temperature, actual torque, set torque, instrument panel wind speed, cabin temperature, control cabinet Internal temperature, cabin temperature, wind direction, wind speed, cabin air direction angle, 1# paddle blade angle, 1# paddle setting angle, 2# paddle blade angle, 2# paddle setting angle, 3# paddle blade angle, 3# Paddle setting angle, 1# paddle motor...
temperature, 2# paddle motor temperature, 3# paddle motor temperature, hub inner temperature, gearbox oil pump suction port oil pressure, gearbox distributor position oil pressure, spindle speed, gearbox oil pump oil absorption Port oil pressure, gearbox distributor position oil pressure, spindle speed, gearbox intermediate shaft non-drive end Bearing temperature, gearbox intermediate shaft drive end Bearing temperature, gearbox sump temperature, main bearing outer ring temperature, availability, cabin side vibration (unfiltered), engine room axial vibration (filtered), cabin side vibration (unfiltered), cabin axial vibration (filtered), cabin position, total twist cable angle and other parameters. The above parameters are used as input variables for the model. Due to the large order of magnitude change between the daily operating data parameters of the fan and the partial distortion of the data, the direct input prediction effect is very poor. Therefore, the original data is normalized to scale the data and effectively eliminate it. The difference between the magnitudes of the data.

This study focuses on long- and short-term memory networks (LSTM), which evolved from recurrent neural networks (RNNs). At first, the creation of such a network to ensure that the model can maintain long-term dependence in the learning process, eliminating the problem of gradient disappearance in deep network learning [9]. In this paper, the long-term and short-term memory network (LSTM) is considered to be a long-term continuous change process of the fan gearbox failure. In the model of Fig. 2, the formula (2) is used to decide whether to discard the previously added information, formula (3) and (4) to decide whether to update the information, formula (5) is to determine the output of the cell state, formula (6) (7) to calculate the output information [10].

![Diagram of Long Short-Term Memory Block](image)

**Figure 2. Long short-term memory block**

\[
\begin{align*}
 f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 C_{t} &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
 C_t &= f_t \ast C_{t-1} + i_t \ast C_c \\
 O_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 h_t &= O_t \ast \tanh(C_t)
\end{align*}
\]

What is used herein is the structure of the LSTM shown in Figure 3. The input contains 50 dimensional parameter data. After several experiments and comparisons, three LSTM layers are set, three Drop layers are interposed, and one output layer is added. The gradient is optimized by adam, and the final output is n rows. One-dimensional data.
4. Simulation result analysis

The LSTM prediction model is compared with the prediction results of the SVRM prediction model, and the RMSE and MAE values of the prediction results are calculated to reflect the pros and cons of the model. The SVRM and LSTM of Table 1 are compared in the training set and the test set. As a result, it is apparent that the prediction effect using LSTM is better.

Table 1: Training set and test set results for both models

| method | Train | Test |
|--------|-------|------|
|        | RMSE  | MAE  | RMSE  | MAE  |
| SVRM   | 0.078 | 0.03 | 0.82  | 0.81 |
| LSTM   | 0.0378| 0.014| 0.29  | 0.24 |

In addition, it can be obtained from the error results of the oil pressure prediction model of the oil pumping port of the gearbox shown in Figure 4. The red line indicates the error of using the LSTM to establish the prediction model, and the blue line indicates the use of the commonly used SVRM to establish the prediction model error.

5. Conclusion

This paper describes the process of establishing a predictive model in the case of faulty labels that are often faced in the real industry. By integrating the K-means clustering results and the missing fault label data, the effective data is extracted from the original data as much as possible, and the appropriate LSTM recurrent neural network model is constructed to predict the fan gearbox failure. In the simulation, comparing the results of the LSTM prediction model with the results of the commonly used SVRM prediction model, it is obvious that the prediction effect of LSTM is good and significantly better than SVRM. From the error analysis graph, it can be found that the error fluctuation in the later stage of the LSTM model prediction result is large. In later studies, further improvements can be made to the LSTM model structure to reduce this error.
Acknowledgments
This work was supported by “Shanghai Natural Science Foundation Project” (19ZR1461500).

References
[1] Peng Y, Dong M. (2011) A hybrid approach of HMM and grey model for age-dependent health prediction of engineering assets J. Expert Systems with Applications, 38(10):12946-12953.
[2] Kobayashi K, Kaito K, Lethanh N. (2012) A statistical deterioration forecasting method using hidden Markov model for infrastructure management. J. Transportation Research Part B, 46(4):544-561.
[3] Fengtao Wang, Shouhai Chen, Jian Sun, Dawen Yan, Lei Wang and Lihua Zhang. Time-Frequency Fault Feature Extraction for Rolling Bearing Based on the Tensor Manifold Method [J]. Mathematical Problems in Engineering, 2014: 1-15.
[4] Jun-Hyun Shin, Yun-Seong Lee, Jin-O Kim, (2014) Fault prediction of wind turbine by using the SVM method, In: Information Science Electronics and Electrical Engineering (ISEEE), Sapporo pp. 1923-1926.
[5] Wenbo Zhang, Guangjie Han, Jing Wang, Yue Liu, (2019) A BP Neural Network Prediction Model Based on Dynamic Cuckoo Search Optimization Algorithm for Industrial Equipment Fault Prediction. J IEEE Accessm 7:11736-11746.
[6] Han H, Gao X. (2017) Fault Prediction of Shield Machine Based on Rough Set and BP Neural Network. In: International Conference on Information Science & Control Engineering. IEEE Computer Society, Changsha, pp.994-998.
[7] Sak, Haşim, Senior A , Beaufays, Françoise. (2014) Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition J. Computer Science, pp. 338-342.
[8] Wang Jun, Wang Shitong, Deng Zhaohong, (2012) Some problems in cluster analysis,J, Control and decision-making, vol. 27, no. 03, pp. 321-328,
[9] K. Greff, R. Srivastava, J. Koutnik, B. R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, March 2015.
[10] Aydin, O., Guldamlasioglu, S. (2017) Using LSTM networks to predict engine condition on large scale data processing framework. In: 4th International Conference on Electrical and Electronic Engineering (ICEEE), Ankara ,pp. 281–285,