Typical status recognition of gearbox based on big data

Lixia Zhang*, Zhaohong Yang and Hongyu Wang

Academy of armored forces of the PLA army, Beijing, China

*Corresponding author e-mail: zlx_work@163.com

Abstract. In this paper, the experimental data of five different working states of the gearbox during the normal operation and the failure of the gearbox are taken as the research object, the python-based data analysis tools Numpy and Pandas are used to establish a fault detection model, by collecting time-domain and frequency-domain data of several faults such as fracture, crack and wear of transmission gear, and analyzing the characteristic parameters of the data, the classification of transmission faults is realized, thus laying a foundation for the application of big data tools in fault diagnosis and analysis.

1. Introduction

In this paper, the experimental data of five different working states of the gearbox during the normal operation of the gearbox bench experiment and the gearbox failure are taken as the research objects. The typical fault diagnosis analysis of the gearbox is combined with the data analysis tools to explore the mechanical fault identification and diagnosis analysis based on big data. Data analysis tools in python were applied to the data set for data cleaning, data conversion, feature extraction, and other processing, and the logistic regression and support vector machine algorithms were combined with training data for model training. Through comparative experiments, the algorithm model with the highest accuracy was selected to improve the accuracy of fault detection. Finally, the trained model is used to diagnose the fault data, and the input parameters of the algorithm model are adjusted, which lays a foundation for using the massive data of the mechanical system for state recognition.

2. Description of experimental data sources

2.1. Test structure of transmission bench

![Diagram of test structure of transmission bench](image)

Figure 1. Schematic diagram of test structure of transmission bench

The schematic diagram of the experimental structure of the transmission bench is shown in figure 1. The experimental system is mainly composed of input shaft, two transmission shafts, output shafts...
and gears with different number of teeth. The shape of the gear includes spur gear and helical gear. The number of teeth of the first pinion is 16 and is installed at the input shaft. The number of teeth of the first transmission gear on the transmission shaft is 48, the number of teeth of the second transmission gear is 24; The number of pinion teeth on the output shaft is 40. Therefore, the gear reduction ratio of the gearing is 5:1.

2.2 Number and position of acceleration sensors

The signal acquisition system of this experiment includes two acceleration sensors and a speed sensor, two of which are installed at the position of the input axis and the output axis respectively, and the speed sensor is installed at the position of the input axis.

The main fault forms of the gearbox bench test are gear teeth missing, gear fracture, bearing eccentricity, input and output shaft unbalance, shaft bending and other different combination states. For different states, data are collected under high load and low load respectively, which increases the reliability of data. It lays a foundation for machine model training. There are six main states as follows: normal state, second transmission gear (24T) missing teeth, second transmission gear fracture, internal failure of the bearing assembly, eccentricity of the bearing ball, bending of the input shaft, fracture of the second transmission gear, internal failure of the bearing and bending of the input shaft. The main three characteristic items of the data are input acceleration, output acceleration and input speed. More than 30 million data in 6 states were collected in the experiment. During the processing, the data were divided into training data set and test data set.

3. Selection of transmission characteristic parameters

This paper is based on the analysis of the fault data of the transmission, and then the paper selects the data processing tools and algorithms to analyze the fault data so as to realize the classification diagnosis of the variable box fault. Therefore, in the time-domain and frequency-domain characteristic parameters of vibration signals collected and calculated, how to select appropriate characteristic parameters to make them suitable for the study of the problem in this paper as the data input of the model is one of the important tasks of fault diagnosis and analysis in this paper. The characteristic parameters are generally divided into time domain characteristic parameters and frequency domain characteristic parameters.

3.1 Time-domain characteristic parameters

5 dimensionless indexes are taken as the input parameters for the initial processing of big data.

\[
k_r = \frac{1}{n} \sum_{i=1}^{n} X_i^4 / nX_{rms}^4
\]

The above formula is the calculation formula of kurtosis, and the realization method in python is data.kurt(), which is the effective value of signal and signal. Kurtosis is often used as the characteristic parameter vector in signal time domain analysis.

Waveform factor:

\[
Q_0 = X_{\text{peak}} / |x_m|
\]

The above equation is the calculation formula of waveform factor. The realization code in python is Max (data[p1:p2]) / (abs(data[p1:p2])). Mean (), which is the ratio of peak value \(X_{\text{peak}}\) to average amplitude \(x_m\). When the fault occurs, the peak will increase while the average amplitude will change little, and the waveform factor will increase to about 10. However, when the early fault is not effectively handled, the peak reaches the limit, and the average amplitude increases with the severity of the fault, the value of waveform factor will slowly fall back to the level when there is no fault.

Peak factor:

\[
C = X_{\text{peak}} / X_{\text{rms}}
\]

\(\text{SQRT (pow(df_mean,2) + pow(df_std,2))}\) is a relative ratio between signal peak value \(X_{\text{peak}}\) and signal effective value \(X_{\text{rms}}\), which has nothing to do with mechanical structure and working conditions. It reflects the waveform peakness of vibration signal, is a dimensionless parameter, and has
good anti-interference, so it is convenient to be applied in mechanical fault diagnosis.

Pulse factor: \[ s_h = \frac{x_{rms}}{|x|} \] (4)

The above formula is the calculation formula of the impulse factor, and the python code is implemented as (Max (data[p1:p2])/(abs(data[p1:p2])). Mean ()), which is the ratio of the effective value of the signal \( x_{rms} \) to the absolute value of the signal .

Margin factor: \[ L_t = \frac{X_{peak}/X_d}{} \] (5)

The above formula is the calculation formula of margin factor, and the python implementation code is (Max (data[p1:p2]))/pow((sum/(p2-p1)),2), where \( X_d \) is the square root amplitude and the formula is \[ X_d = \left( \frac{\sum_{i=1}^{n} |X_i|}{n} \right)^2 \]

3.2 Frequency domain characteristic parameters

The time-domain characteristic parameters of the vibration signal are often used to judge whether the machine has a fault and the time domain characteristic parameter calculation process is simple, the reaction speed is fast and the reaction often used for the real-time monitoring of mechanical running state. However, once a mechanical fault is detected, it is necessary to determine the type of fault and the specific location of the fault, as well as the severity of the fault, etc. It is necessary to analyze the mechanical fault in the frequency domain of vibration signal. The frequency-domain characteristic parameters frequently used in vibration signals are briefly introduced below:

- The mean square frequency: \[ \text{MSF} = \frac{\sum_{i=1}^{N} \phi^2(i) / 4 \pi^2 \sum_{i=1}^{N} x^2(i)}{N} \]
- Frequency variance: \[ \text{VF} = \text{MSF} - FC^2 \]
- Frequency standard deviation: \[ \text{RVF} = \sqrt{\text{VF}} \]
- Root mean square frequency: \[ \text{RMSF} = \sqrt{\text{MSF}} \]

Where, \( \phi_i = \frac{x(i) - x(i-1)}{\Delta}, \Delta \) is the sampling period, and \( N \) is the number of samples collected. In this paper, frequency standard deviation and root mean square frequency characteristic parameters are taken as input parameters.

4. Recognition and realization of typical state of gearbox based on big data

Big data research and development is very rapid, the new big data processing tools are constantly emerging, the more popular is the machine learning python library, there are a lot of libraries in the python, each library can realize many different functions, including sklearn library calls and data processing for the machine learning algorithm provides machine convenient interface (API), and the training of the support vector machine (SVM) algorithm, classification algorithm, the problem such as kernel function is discussed.

The main algorithms used in this paper are logistic regression and linear support vector machine, both of which use linear rows as input to achieve the purpose of prediction.

Model training and testing are key steps in gearbox status recognition and diagnosis, and all the previous work is in preparation for this process. Only through learning and training the model of sample data can the model be able to recognize the state after generalization.

4.1 Data classification

Before model training, data need to be divided into training set and test set. This step can be manually classified. Of course, sklearn machine learning tool provides a very convenient data classification API, which can randomly divide data into training set and test set according to specific parameter proportion. The specific implementation method is as follows:

```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Where \( X \) and \( y \) are the feature data and label data of the data set, test_size=0.2 is 20% of the data as the test set, the rest as the training set, random_state=0, which represents the random number seed is
0. Data is randomly selected according to some random rule. The data after splitting is: X_train, y_train (to train the model), X_test, y_test (to test and verify the model).

4.2 Model training

Model training is in using selection algorithm model and has good data of split for data fitting or learning. First choose the appropriate algorithm, in this case logistic regression algorithm and linear support vector machine algorithm. The actual results of the two algorithms are compared to determine the final application algorithm.

(1) logistic regression:

from sklearn.linear_model import LogisticRegression
arithmetic = LogisticRegression(C=10000)
arithmetic.fit(X_train, y_train)

Specifically, it calls the logistic regression algorithm in the sklearn library, instantiates the algorithm, and then uses the fit method to fit the training data. In this way, the identification ability of sample data has been obtained in arithmetic instance. After algorithm generalization, it can perform well in test data.

(2) Linear support vector machine:

from sklearn.svm import SVC
arithmetic = SVC(kernel='rbf', C=10, gamma=0.1)
arithmetic.fit(X_train, y_train)

In the same way, call, instantiate, and fit the data. Here, the parameter gamma is inversely proportional to sigma. The smaller the gamma, the farther the affected training sample is, which can be regarded as the reciprocal of the influence radius of the support vector.

Parameter C is used to weigh the accuracy and complexity of the model. The smaller the value of C is, the smaller the number of samples in the support vector is, making the decision surface smooth, the model simple and the accuracy decreased. With a large C value, the model can choose more samples as support vectors, and the accuracy increases and becomes more complicated.

The model is very sensitive to the gamma parameters. If the gamma value is too large, the influence range of the support vectors only includes the support vectors themselves, without proper regularization C to prevent overfitting.

When gamma is very small, the model will not be able to fit the shape and complexity of the data, so that the influence region of any selected support vector will include the entire training set. The resulting model is similar to a linear model composed of a hyperplane set, separating any two high-density regions.

4.3 Model test

Model test is a way to check the accuracy of the model by using the test data. There is a convenient interface in sklearn for testing. The result finally printed out three data, namely TruePositive, recall and F1 Score.

Forecast data/actual data      The actual value is positive  The actual value is negative
The prediction is positive    TruePositive              FalsePositive
The predicted value is        FalsePositive             TruePositive
negative

precision:

\[
\text{precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}
\]

The apis in sklearn are:
from sklearn.metrics import precision_score

The recall rate:

\[\text{recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}\]

The apis in sklearn are:

from sklearn.metrics import precision_recall_fscore_support

F1 Score:

\[\text{F1 Score} = 2 \frac{PR}{P + R}\]

The apis in sklearn are:

from sklearn.metrics import f1_score

Where, P is the accuracy rate and R is the recall rate. It is used to judge the algorithm model directly.

Specific test implementation steps:

from sklearn.metrics import classification_report

classification_report(y_test, y_pred)

It can be seen that under the processing of big data tools, the accuracy of model recognition is good, reaching around 98%. The accuracy of both algorithms is similar, and the results of linear support vector machine are slightly better than those of logistic regression algorithm.

5. Conclusion

In this paper, based on the structure of the gearbox bench test and the possible fault forms, the fault state identification and diagnosis are studied, this paper makes a comprehensive use of different algorithms in machine learning to analyze the fault state data of the gearbox. Finally, the data is imported for training and test verification, and the classification of transmission faults is realized, thus laying a foundation for the application of big data tools in fault diagnosis and analysis.

References

[1] SHI J, LIANG M, NECSELESCU D-S, et al. Generalized stepwise demodulation transform and synchrosqueezing for time–frequency analysis and bearing fault diagnosis[J]. Journal of Sound and Vibration, 2016, 368: 202-222.

[2] Yaguo Lei. mechanical fault diagnosis based on improved hilbert-huang transformation [J]. Journal of mechanical engineering. 2011. 47(5) : 71-77.

[3] Lin Zhang. Anquan Sun. Tianyi Wang, etc. Fault intelligent diagnosis of certain missile equipment [J]. Journal of central south university. 2013(S1) : 216-220.

[4] Jingde Chen, Gehao Sheng, Jijian Wu, Yougang Xu, Fujie Wang. big data technology in the application status and prospects of the smart grid [J]. High voltage electric equipment, 2018 (01)

[5] LEI Yaguo, LIU Zongyang, WU Xionghui, et al. Health condition identification of multi-stage planetary gearboxes using a mRVM-based method[J]. Mechanical Systems and Signal Processing, 2015, 60: 289-300.