Fault diagnosis of rolling bearing based on empirical mode decomposition and convolutional recurrent neural network

Mulin Huang, Tingting Huang, Yuepu Zhao and Wei Dai
School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
E-mail: DW@buaa.edu.cn

Abstract. Bearing is more important in mechanical parts. Many failures of rotating machinery are caused by bearing failure. It is very important to diagnose the rolling bearing fault and help the mechanical products to find out the failure of parts in operation. It can avoid danger and improve efficiency. To research the problem of rolling bearing fault diagnosis under different loads, a method using vibration signals based on empirical mode decomposition (EMD) and convolutional recurrent neural network (CRNN) is proposed. First, the EMD is used to deal with the vibration signal for noise reduction. Then, CRNN is built as the rolling bearing fault diagnosis classifier using the envelope of EMD processing. The Case Western Reserve University data sets are used to validate the method. The result shows that the method fits well.

Keywords: Empirical mode decomposition, convolutional neural network, long short time memory neural network, bearing fault diagnosis.

1. Introduction

Rolling bearing is an important component used in rotating machinery. A lot of faults in motors are because of bearings [1]. Rolling bearing is also one of the 11 categories of specific revitalization machinery products identified by the state, as a basic part it requires high precision and good fit. As an accessory, it is widely used in military, aerospace, satellite, vehicle and other fields. Many faults of rotating instruments are associated with rolling bearings, which are damaged in various forms, such as fatigue, surface wear of inner and outer rings of bearings, bearing breakage, bearing cage or rolling body fragmentation, etc. Failure of the rolling bearing can lead to collapse, so that fault diagnosis is vital for a rotating machinery. Researches on fault diagnosis are fruitful [2], [3]. Considering the environment of the equipment, vibration signal, which indicate the equipment status directly, is commonly used in rolling bearing fault diagnosis [4]. Machine learning methods are widely used in fault diagnosis. Azeddine et al. [5] use complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) to reduce the noise of the vibration signals and use an optimized thresholding operation to select intrinsic mode functions as result signals. Amar et al. [6] use vibration spectrum imaging as features, then enhance the signal noise rate and train an artificial neural network to classify the fault. Lei et al. [7] use sparse filter to learn hidden features in the vibration signals, and diagnose the faults by softmax classifier.

Empirical mode decomposition (EMD) is an adaptive signal decomposition method, which can decompose a signal into multiple intrinsic mode function (IMF) groups. Duan and Wang [8] use EMD to process vibration signals to do on-load tap-changer fault diagnosis. Sherr et al. [9] propose a
satellite fault diagnosis method based on predictive filter and EMD. Li et al. [10] use bandwidth EMD to process the vibration signals in incipient bearing fault diagnosis. Convolutional recurrent neural network (CRNN) is a combination of convolutional neural networks and recurrent neural networks. In [11], a CRNN model is proposed to deal with hand pose estimation. CRNN model can be used in indefinite length discrete series processing to deal with vibration signals and find out the hidden fault features. Ince et al. [12] use 1-D convolutional neural networks to detect early motor faults. Fu et al. [13] use the CRNN model for actuators fault diagnosis, which has a higher accuracy than CNN or DNN.

In the bearing fault diagnosis, the frequency spectrum of bearing vibration signal is decomposed, and the frequency spectrum after decomposition is corresponding to the characteristic frequency of bearing fault, so as to determine the fault state and fault location of bearing. Therefore, it is a method with specific physical significance to judge the fault type by EMD decomposition. CRNN is a neural network structure model, which is a data-driven method. When CRNN is trained, the neural network model will be affected by vibration signal noise, which will reduce the fitting accuracy of data. Therefore, the combination of EMD method and CRNN method can improve the fitting accuracy of neural network to a certain extent. In this paper, a method based on empirical mode decomposition and convolutional neural network for rolling bearing fault diagnosis is proposed. The EMD is for noise reducing and signal feature enhancement. Then, the CRNN model is built as a classifier for bearing fault using the processed signals. At last, the Case Western Reserve University data set is used for validation. The proposed method can achieve a high accuracy.

2. Brief introduction of EMD, CNN and RNN

2.1 Empirical mode decomposition

EMD is a time-frequency analysis method proposed by N.E. Huang. It can decompose the signal into several intrinsic mode functions (IMF). The frequency component of each intrinsic mode function is related to the signal. It is an adaptive time-frequency decomposition method. By EMD decomposition of fault vibration signal, the components of different frequencies can be obtained, which can reduce signal noise and strengthen signal fault characteristics.

The EMD operation is as figure 1:

![Figure 1. The EMD operation](image1)

Hilbert transform can demodulate amplitude modulation signal and form envelope, as the Figure 2. The Hilbert spectrum can be obtained by Hilbert transformation (HT) of each IMF. The envelope spectrum obtained by Hilbert transform is the instantaneous frequency energy of vibration wave in physical sense. The EMD decomposition and envelope processing of vibration signal can adaptively decompose the signal according to frequency, which reduces the interference of parameters and uncertain factors such as sampling frequency and bearing rotation frequency. The frequency spectrum analysis of processed results is as Figure 4. Compared with the frequency spectrum of the original data, as Figure 3, low frequency part has a more significant feature on the at 1, 2, 3 times of the inner ring fault characteristic frequency. The inner ring fault feature character is clearer to see in the Figure.
4. That is to say, EMD and Hilbert transformation enhance the features of fault, which can be used for noise reduction, to enhance the accuracy of the method.

Figure 3. Frequency spectrum of original data. Figure 4. Frequency spectrum of processed data

The adaptiveness of decomposition effectively extracts the high frequency vibration excitation response during fault impact, compared with Fourier transform and wavelet decomposition, the EMD method reduces the loss of fault feature information due to manual frequency and parameters selection. Using EMD to process the vibration signals can reduce the noise and increase the CRNN accuracy.

2.2 Convolutional neural network

In CRNN model, convolution neural network is set before recurrent neural network. 1D-convolution operation and maxpooling operation are used in CRNN. After these two operations, the feature length of vibration signals decrease, so that the long short time memory (LSTM) cell which is commonly used in recurrent neural network could enhance the model performance. The 1-D convolution operation [14] is as follow:

\[ Q_i(\tau) = f\left(\sum_{i=0}^{F} \omega_i(\tau) \ast D(\tau) + b_i\right) \]  

(1)

\( Q \) is the feature map, \( f \) is the convolution operation, \( \omega \) is a convolution kernel, \( D \) is the original data, \( b_i \) is the bias. The function of this layer is to segment the convolution result of the upper layer and take the part with the highest similarity with the convolution kernel as the value of this segment. The purpose of this method is to reduce the influence of waveform phase on the diagnosis result. When the impact fault event occurs, this part of the signal will be affected by the occurrence time. The model does not have the ability of frequency analysis, so the occurrence of fault impact events at different time points will affect the diagnosis ability of the model. So, more functions needed to be used to solve these problem.

After the convolution operation, a maxpooling operation follows to reduce the dimension of the feature map. The max pooling operation is as follow:

\[ P = \max(x_i) \]

(2)

\( P \) is the max feature value in \( x_i \). \( x_i \) is part of the feature map. The size of \( x_i \) is set according to experience. By adding the pooling layer, the convolution result of a signal is taken as the maximum value, which represents whether the signal has had a fault impact event, and the time sequence of the fault event is aligned, so as to reduce the influence of signal phase.

2.3 Recurrent neural network

In a recurrent neural network, the output at the last time will be passed along with the input at the next time, simultaneous interpreting the output of the next time. This makes the neural network have the ability of processing time series data. For recurrent neural network part, LSTM [15] cell is used. Long short-term memory is a kind of recurrent neural network. Through its unique design structure,
it can process and predict the important events with very long interval and delay in the time series. That is to say, the network has a certain "memory" ability for the time series and can process the sequence information with a long span.

The structure and operations are as follows:

![Figure 5. LSTM structure](image)

The $\sigma(x)$ means sigmoid function. The sigmoid function is as the following function:

\[
\sigma(x) = \frac{1}{1 + e^{-x}}
\]

In the LSTM structure, the forget gate operation is based on sigmoid function. $X$ is the input sequence, and $h$ is the output sequence. In the forgetting gate structure, the cell state $C_{t-1}$ at last moment is calculate by dot product with the current input $\sigma(Xt+b)$ to determined how much the cell state value to keep. Then in the input gate, the value plus the current input $\tanh(Xt)*\sigma(x+b)$ as current cell state $C_t$. At last, in the output gate structure, the current cell state $C_t$ processed with tanh function and sigmoid function, to get the output $ht$. The formulas are as following functions:

\[
\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}
\]

\[
a_i = \sigma(W_iX_t + b_i)
\]

\[
a_o = \sigma(W_oX_t + b_o)
\]

\[
c_t = C_{t-1} \cdot a_i + \tanh(C_{t}) \cdot a_o
\]

\[
h_t = \tanh(C_t) \cdot \sigma(C_t)
\]

With the LSTM structure, the recurrent neural network can perform better on accuracy and training time. In the data-driven fault diagnosis strategy, for the use of historical data, it is often used as model training data and validation data; in the fault diagnosis, only a section of sensor signal under the existing state is collected for fault diagnosis. By using LSTM model, some historical data can be "memorized" to form a priori fault state, so that the judgment of current equipment state may be more accurate, and the service life of equipment or the time point of fault occurrence can be predicted. Long short time memory model is a kind of cyclic neural network model, which is used to solve the problem of gradient vanishing and gradient exploding in the process of long sequence training.
2.4 Convolutional recurrent neural network structure

Use tensorflow2.0 to build the model. The model is a shallow neural network, which is composed of sequence data input layer, batch normalization layer, 1-D convolution layers, 1-D pooling layers and short-term memory neural unit layer, two full connection layers, dropout layer and softmax classifier. The model structure is as figure 6.

![Figure 6. The structure of the CRNN model](image)

In this model structure, the function of the normalization layer is to normalize the input vibration signal data so that its variance is 1 and its mean value is zero. Through certain transformation, the distribution of input and output is consistent. The functions are as following formulas.

\[
\mu = \frac{1}{m} \sum_{i=1}^{m} x_i
\]

\[
\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2
\]

\[
\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}
\]

\[
y_i = \hat{x}_i + \beta = BN_{\gamma, \beta}(x_i)
\]

Through the above formulas, first calculate the mean value of the input data \(\mu\); then calculate the variance of the data \(\sigma^2\) and normalize the input data to get the distribution data with the mean value of 0 and the variance of 1. \(\varepsilon\) is the bias number. Then multiply the processed data by the coefficient and add the offset term to get the distribution data \(\hat{x}, y\) with the mean value \(\beta\) and the variance \(\gamma^2\). Through this operation, the input data of each layer can be transformed into data of the same distribution, so as to reduce the impact of data distribution offset and improve the overall fitting speed and accuracy of the model.

The 1-D convolution layer is used to extract the characteristics of vibration signals through the training of convolution kernel. The initialization parameters of the convolution kernel are random, and the details of the waveform are close to each other gradually by gradient descent. Then, by fitting a large number of data, a more general convolution kernel is obtained, which reflects the general characteristics of the vibration signal. In this layer, convolution operation is carried out by convolution kernel and processed vibration signal. By setting parameters such as convolution kernel size and convolution step length, a one-dimensional numerical vector is obtained. The physical meaning of the vector is the similarity between vibration signal and convolution kernel. Maxpooling layer and LSTM layer are set for a better fit and a higher accuracy. These two layers' function has described in 2.3. Then, the two full connection layers are set. Through the setting of two dense layers, the depth of the network is deepened, so as to improve the fitting accuracy.
The purpose of the dropout layer is to prevent the model from over fitting. Because the parameters of neural network are too many, in the process of training, there are often over fitting states. Dropout can prevent over fitting by randomly removing some neuron parameters of a layer. Softmax layer is a kind of classifier, which is mostly used to output classification results by neural network. There it is used to output the bearing fault state.

3 Experiment and Validation of the proposed method

3.1 Data sets introduction

To validate the method, the vibration signals data set from Case Western Reserve University is used\cite{16}. The test bearing is a SKF6205 motor bearing. A single-point fault with multiple pitting diameters is introduced by the electric discharge machining technology, and an acceleration sensor is applied to collect the vibration signal while the bearing is used. The data set includes 4 different load levels on three failure levels, and 4 different kinds of fault, as the Table 1.

| Label | Failure level              | Fault Diameter (mm) | Load A (HP) | Load B (HP) | Load C (HP) | Load D (HP) |
|-------|----------------------------|---------------------|-------------|-------------|-------------|--------------|
| 1     | Normal                     | 0                   | 0           | 1           | 2           | 3            |
| 2     | Minor Inner Race Fault     | 0.18                | 0           | 1           | 2           | 3            |
| 3     | Medium Inner Race Fault    | 0.36                | 0           | 1           | 2           | 3            |
| 4     | Serious Inner Race Fault   | 0.54                | 0           | 1           | 2           | 3            |
| 5     | Minor Outer Race Fault     | 0.18                | 0           | 1           | 2           | 3            |
| 6     | Medium Outer Race Fault    | 0.36                | 0           | 1           | 2           | 3            |
| 7     | Serious Outer Race Fault   | 0.54                | 0           | 1           | 2           | 3            |
| 8     | Minor Ball Fault           | 0.18                | 0           | 1           | 2           | 3            |
| 9     | Medium Ball Fault          | 0.36                | 0           | 1           | 2           | 3            |
| 10    | Serious Ball Fault         | 0.54                | 0           | 1           | 2           | 3            |

The frequency of sample is 12000Hz. Considering the different failure levels under different loads have different features, the data set is divided into two parts by means of isometric sampling, so that the training set has the same fault modes as the validation set. This will increase the recognition accuracy of the method. The ratio of the training set number and the validation set number samples is 0.7:0.3.

3.2 Input sequence processing

The sequence input is the vibration signals processed by EMD, Hilbert transformation and data processing. Considering the signals sequence processed by EMD and HT are still informative, the sequence needs to be compressed by the data processing function. Set a window and calculate the root mean square (RMS) value of the processed vibration signals. The window size is set considering the rotation frequency, sample frequency and the output length. The length of data is reduced by windowing and calculating RMS. The processing operation is as Figure 7.
Figure 7. Set windows and calculate the RMS

\[ L = f_{\text{input}} \times f_{\text{sample}} \times f_{\text{rotation}} \]

\[ \text{RMS} = \left( \frac{1}{n} \sum X_i^2 \right)^{1/2} \]

(13) \hspace{0.5cm} (14)

\( L \) is the length of the window, \( f_{\text{input}} \) is the input length, \( f_{\text{sample}} \) is the sampling frequency, \( f_{\text{rotation}} \) is the rotation frequency. Considering the Gibbs sampling frequency, \( N \) is taken as 2.5. The RMS formula is as (10).

3.3 Training and validation of the proposed method

Use Shannon entropy as loss function. The training procession is as following figure 8.

Figure 8. The loss value of the training process

After about 500 epochs of training, the loss value decrease to a low level. The accuracy of the validation set is 98%. Comparing with some other machine learning method, the proposed method has a higher accuracy rate. The comparison is as the following Table 2:

Table 2. Comparison of different method

| Method                  | Logistic Regression | SVM  | LSTM | CNN  | KNN  | EMD and CRNN model |
|-------------------------|--------------------|------|------|------|------|-------------------|
| Accuracy                | 94%                | 96%  | 42%  | 95%  | 94%  | 98%               |

The LSTM neural network performs the worst both on training time and accuracy rate. This is because a too long sequence input causes gradient exploding and gradient vanishing, though the LSTM structure is designed to solve the problems.
4 Conclusion
This paper proposes a method for rolling bearing fault diagnosis based on EMD and CRNN model. The proposed method uses EMD to enhance the fault feature in frequency domain. The convolutional neural network part could extract hidden data features from the vibration signals and reduce data dimensions and input sequence length. Through the LSTM gate structure, the temporal features of the vibration signals can be fully utilized. And the compared with some other machine learning algorithm, the proposed method based on EMD and CRNN has a higher accuracy but also needs a longer training time, because the proposed method has more parameters to be trained and the recurrent neural network part needs to train every single input of the sequence. An optimal method still needs research, and the training time needs a quantitative research in further time.

Although the method exhibits a higher accuracy compared with other methods, hybrid fault modes should be further concerned. Besides, fault prediction based on life cycle and residual service life prediction based on different loads will be studied further.

Reference
[1] IEEE Committee Report 1985 Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part I. IEEE Transactions on Industry Applications. vol.1A-21(4) 853-64
[2] Gao, Zhiwei, C. Cecati and S. X. Ding 2015 A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. IEEE Transactions on Industrial Electronics. 62(6) 3757-67
[3] Gao, Zhiwei, C. Cecati and S. X. Ding 2015 A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part II: Fault Diagnosis With Knowledge-Based and Hybrid/Active Approaches. IEEE Transactions on Industrial Electronics. 62(6) 3768-74
[4] A. Bellini, F. Immovilli, R. Rubini and C. Tassoni 2008 Diagnosis of Bearing Faults of Induction Machines by Vibration or Current Signals: A Critical Comparison. IEEE Industry Applications Society Annual Meeting, Edmonton 1-8
[5] Azeddine Bendiabdellah, Ziane Derouiche, Rabah Abdelkader and et al 2018 Rolling Bearing Fault Diagnosis Based on an Improved Denoising Method Using the Complete Ensemble Empirical Mode Decomposition and the Optimized Thresholding Operation. IEEE sensors journal 18(17) 7166-72
[6] M. Amar, I. Gondal and C. Wilson 2015 Vibration Spectrum Imaging: A Novel Bearing Fault Classification Approach. IEEE Transactions on Industrial Electronics. 62(1) 494-502 doi: 10.1109/TIE.2014.2327555
[7] Lei Yaguo, Jia Feng, Lin Jing and et al 2016 An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. IEEE Transactions on Industrial Electronics. 63(5) 3137-47
[8] R. Duan and F. Wang 2016 Fault Diagnosis of On-Load Tap-Changer in Converter Transformer Based on Time–Frequency Vibration Analysis IEEE Transactions on Industrial Electronics 63(6) 3815-23 doi: 10.1109/TIE.2016.2524399
[9] Y. Sherr, Y. Zhang and Z. Wang 2011 Satellite fault diagnosis method based on predictive filter and empirical mode decomposition. Journal of Systems Engineering and Electronics. 22(1) 83-7 doi: 10.3969/j.issn.1004-4132.2011.01.010
[10] Y. Li, M. Xu, X. Liang and W. Huang 2017 Application of Bandwidth EMD and Adaptive Multiscale Morphology Analysis for Incipient Fault Diagnosis of Rolling Bearings IEEE Transactions on Industrial Electronics vol.64(8) 6506-17 doi: 10.1109/TIE.2017.2650873
[11] Hu Zhongxu, Hu, Youmin, Liu Jie and et al 2019 A CRNN module for hand pose estimation Neurocomputing 333 157-68
[12] Ince Turker, Kiranyaz Serkan, Eren Levent and et al 2016 Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks. *IEEE Transactions on Industrial Electronics* 63(11) 7067-75

[13] J. Fu, C. Sun, Z. Yu and L. Liu 2019 A hybrid CNN-LSTM model based actuator fault diagnosis for six-rotor UAVs. *2019 Chinese Control And Decision Conference (CCDC), Nanchang, China*. 410-4 doi: 10.1109/CCDC.2019.8832706

[14] M. Qiao, S. Yan, X. Tang and C. Xu 2020 2019 Deep Convolutional and LSTM Recurrent Neural Networks for Rolling Bearing Fault Diagnosis Under Strong Noises and Variable Loads *IEEE Access*. vol.8 66257-69 doi: 10.1109/ACCESS.2020.2985617

[15] F. A. Gers and E. Schmidhuber 2001 LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*. vol.12(6) 1333-40 doi: 10.1109/72.963769

[16] Case Western Reserve University Bearing Data Center. Bearing data file.218-01-11. http://csegroups.case.edu/bearingdatacenter/home