A Vibration Signal Anomaly Detection Method Based on Frequency Component Clustering and Isolated Forest Algorithm

Haoyan Xu*
College of Control Science and Engineering
Zhejiang University
Hangzhou, China
3160104027@zju.edu.cn

Pengyu Song*
College of Electrical Engineering
Zhejiang University
Hangzhou, China
pysong2019@163.com

Bocheng Liu
College of Control Science and Engineering
Zhejiang University
Hangzhou, China
3160105488@zju.edu.cn

***Abstract*—Vibration signal represents the health state of equipment operation. Abnormal detection of vibration signal can discover equipment faults in time, facilitate equipment maintenance and avoid safety accident. In view of the obvious difference in frequency components of vibration signal under different operation conditions, a clustering method based on frequency components and Cosine Distance is presented to establish more accurate models for samples under different operation conditions. For time sequence vibration signal with high dimension, time domain and frequency domain features are extracted to represent signal characteristics, and isolated forest model is used to detect abnormal data. The results show that, compared with the supervised fault diagnosis method, this method does not need fault sample data during modeling, and thus ensures detection accuracy while avoiding destructive experiments.**

Keywords—vibration signal, anomaly detection, frequency component, clustering method, isolated forest

I. INTRODUCTION

Vibration signal is one of the most common signals in industrial control systems. Motor, bearing and other equipment will generate high frequency vibration signals during operation. Vibration signal can reflect the operation state of the equipment and can be used to judge whether the equipment malfunctions. Reasonable use of vibration signals can replace manual detection, which is more convenient and quicker, saves a lot of manpower and material resources, and effectively improves the accuracy of fault detection.

At present, fault diagnosis methods utilizing vibration signal are mainly divided into two categories: signal processing method and machine learning method. Signal processing method mainly transforms the time domain and frequency domain of original signal to observe the difference with the signal generated when the equipment works normally. At present, such signal processing and transformation methods as Wavelet Transform [1-3] and Hilbert Transform [3] have been widely used in vibration signal fault detection. However, machine learning method is better at handling this situation because the working state of the equipment often changes dynamically with factors such as rotational speed and load, while it is difficult to detect and diagnose the fault of complex and changeable working modes accurately with signal processing method only. If a certain number of vibration signals generated by equipment operating under normal and fault conditions can be obtained at the same time, the anomaly detection problem will be converted into a classification problem in machine learning. Nandi et al. summarized several time and frequency domain factors used to characterize signal features, and used them as features to carry out fault detection with classification methods such as decision tree, artificial neural network, SVM, etc. [4]. Ahmed et al. used sparse autoencoder to extract signal features automatically and classified with deep neural network (DNN) [5]. Lei et al. used multiple fully connected layers to extract multiple local eigenvectors of vibration signal, and used the mean value of these vectors as the feature of the whole signal for classification with softmax regression [6]. All the above methods can achieve certain fault detection effect, but no separate model for samples under different working conditions are built, which lacks pertinence.

In some cases, such as when the equipment is just put into operation, it is difficult to obtain sufficient fault signal samples, and the sample categories are extremely unbalanced or have no fault data at all. As a result, the supervised classification method will be out of effect. Since only normal samples can be used for modeling, the problem turns into an unsupervised anomaly detection method. Anomaly detection is designed to make quantitative characterization for the mode of normal data by a certain method. Then, calculate the difference with the normal data used in training when a new sample is generated. If the difference exceeds a certain range, the sample will be determined abnormal. Fault detection can be completed without fault samples with anomaly detection method, so that data acquisition becomes simpler, and artificial destructive experiment is not necessary to obtain fault samples.

Most fault detection methods of vibration signal are supervised classification methods, while studies on unsupervised anomaly detection methods are limited. Isolated forest [7] is an effective anomaly detection method that calculates the degree of sample anomaly by establishing multiple isolated trees and conducts ensemble decision making. This paper presents a vibration signal anomaly detection method based on isolated forest. Firstly, cluster the training samples and extract the time and frequency domain features based on the frequency components of the original

978-1-7281-5029-1/19/$31.00 ©2019 IEEE

Shenyang, China
November 22-24, 2019

signal. Then, build models for the samples on each cluster respectively according to the clustering results with the isolated forest method. Clustering establishes a more targeted model for samples under different working conditions, and the introduction of isolated forests also enables the method to achieve a similar effect to the supervised method without the necessity of fault history data.

II. Clustering of Vibration Signals Based on Frequency Components

To a certain extent, frequency components of vibration signal represent the working state of the equipment, and the frequency spectrum of vibration signal changes with different rotating speeds and loads. FFT method is often used at present as the basis to extract frequency domain features in vibration signal anomaly detection methods. However, the frequency components of vibration signals are different for equipment with different working modes, so it is almost impossible to extract time domain and frequency domain features accurately for samples under different working modes without differentiation when analyzing and modeling multiple signal samples. Therefore, it is necessary to use cluster analysis method to cluster samples with similar frequency components before establishing the anomaly detection model.

For original vibration signal X, the frequency component vector V shall be obtained by FFT[8] method at first. FFT is a fast implementation method of Discrete Fourier Transform (DFT) with the calculation process as follows:

$$X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n) e^{-j\pi nt/N} \tag{1}$$

The frequency component vector is:

$$V(k) = |X(k)| \tag{2}$$

For long vibration signal sequence, the high dimension of the frequency component vector is not conducive to clustering samples with similar frequency component, where the method of summation and reconstruction is adopted for processing. For the original frequency component vector V with length L, if the reconstructed dimension is designated as m, the calculation process of obtaining the reconstructed frequency component vector V' is as follows:

$$V'(i) = \sum_{j=Li/m}^{L(i+1)/m} V(j) \tag{3}$$

Then, this paper selects K-means method [9] to cluster for V'. K-means usually measures the distance between samples by Euclidean Distance, which measures by summing the squares of various dimensions and ignores the difference between different dimensions. In this paper, Cosine Distance is used to calculate the similarity of frequency components between samples. Cosine Distance refers to the cosine value of the angle between different vectors. This measure focuses more on the distribution of the sample frequency components and ignores the difference in vector amplitude, and the calculation method is as follows:

$$d_{ij} = 1 - \cos\langle V_i, V_j \rangle = 1 - \frac{V_i^T V_j}{|V_i||V_j|} \tag{4}$$

Moreover, the K-Means method needs to manually specify the number of clusters K. In order to select the best super parameter K, the clustering effect is measured by the silhouette coefficient [10]. The silhouette coefficient comprehensively considers the degree of cohesion and separation of clustering results. For the i-th sample in cluster C, the average Cosine Distance between the i-th sample and other samples in cluster C is a_i, and the average Cosine Distance between the i-th sample and samples in clusters other than C is b_i, then the silhouette coefficient of the sample s_i is:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{5}$$

The value of s_i is between -1 and 1, and the closer to 1, the more reasonable the clustering result is. When evaluating the overall clustering result, only the mean value s of all samples' silhouette coefficients is required.

During online monitoring, the cosine similarity between each new sample's frequency component vector and the center of each cluster is calculated, and then the new sample is classified into the cluster with the largest cosine similarity.

III. Extraction of Time-Frequency Domain Feature

TABLE I. Expressions of Time Domain Features

Features	Expressions
PK	$x_{\max} - x_{\min}$
var	$\frac{\sum_{i=0}^{N-1}(x_i - \bar{x})^2}{N-1}$
rms	$\sqrt{\frac{1}{N}x_i^2}$
K	$\frac{1}{N}\frac{\sum_{i=0}^{N-1} x_i^4}{rms^4}$
I	$\frac{\max(\lvert x_i \rvert)}{\frac{1}{N}\sum_{i=0}^{N-1}\lvert x_i \rvert}$
L	$\frac{\max(\lvert x_i \rvert)}{(\frac{1}{N}\sum_{i=0}^{N-1}\sqrt{\lvert x_i \rvert})^2}$
S	$\frac{rms}{\lvert \bar{x} \rvert}$

In order to describe the state and characteristics of vibration signal, corresponding time domain and frequency domain features shall be extracted. In this paper, seven time domain features including peak-to-peak value (PK), variance (var), root mean square (rms), kurtosis factor (K), pulse index (I), margin factor (L), waveform factor (S) and five frequency domain features including frequency center (FC), mean square frequency (MSF), root mean square frequency (RMSF), variance frequency (VF) and root variance

frequency (RVF) are extracted[4]. The calculation methods are shown in Table I and Table 2 respectively, where $X(k)$ is the amplitude at frequency k after Fourier transformation of the original signal x.

TABLE II. EXPRESSIONS OF FREQUENCY DOMAIN FEATURES

Features	Expressions
FC	$\frac{\sum_{i=0}^{N-1} kX(k)}{\sum_{i=0}^{N-1} X(k)}$
MSF	$\frac{\sum_{i=0}^{N-1} k^2 X(k)}{\sum_{i=0}^{N-1} X(k)}$
RMSF	$\sqrt{MSF}$
VF	$\frac{\sum_{i=0}^{N-1} (k - FC)^2 X(k)}{\sum_{i=0}^{N-1} X(k)}$
RVF	$\sqrt{VF}$

For each original signal, a total of 12 time domain and frequency domain feature is extracted in the above manner, and the longer signal is converted into a feature vector with dimension 12 as a detection eigenvector.

IV. ANOMALY DETECTION METHOD BASED ON ISOLATED FOREST

Anomaly detection usually refers to modeling with normal samples only when there is a lack of fault samples to fit the data distribution of normal samples, and after obtaining new samples, it will be judged as abnormal if there is a big difference from the fitted normal sample distribution. Isolated forest (iForest) is an effective anomaly detection method for continuous variables by constructing multiple isolated trees (iTrees) to judge the anomaly of samples in an ensemble way.

Isolated Tree (iTree) is a random binary tree. For a given data set D, the specific construction method is as follows:

Step1:Input data set D and tree height limit L;

Step2:Select an attribute Q of D and a value V of the attribute (V is between the maximum value and the minimum value of Q) randomly. Then take samples with the attribute Q less than V in D as left sub-nodes, and take samples with the attribute Q greater than or equal to V as right sub-nodes;

Step3:For the sample set on the left and right sub-nodes, if the number of samples in the set is not 1, perform recursive operation of *Step2* until the tree height reaches the limit L. End when reaches the tree height limit L or the number of samples on all leaf nodes is 1.

For isolated forest model, specify the total number t of iTree and the downsampled number Ψ to construct t iTrees in total. For each iTree, sample Ψ samples randomly in the original data set D as input samples, and set the tree height limit L of each iTree as:

$$L = ceiling(\log_2 \psi) \quad (6)$$

Where *ceiling* means roundup.

For a certain sample x, the isolated forest model uses the expected value of the number $h(x)$ of edges passed from the root node to the leaf node on each iTree to measure the abnormality degree. The more abnormal the sample point is, the shorter the path length it passes by. The average path length of the tree for a data set containing n samples is:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (7)$$

Where $H(i)$ is the harmonic number, and the anomaly score of sample x is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (8)$$

This value is between 0 and 1, and the closer to 1, the higher the sample abnormality degree is. The detection of abnormal samples can be realized by setting a reasonable anomaly discrimination threshold according to the anomaly distribution of normal samples.

V. VIBRATION SIGNAL ANOMALY DETECTION PROCEDURE AND CASE STUDY

In practical application, firstly collect a certain number of vibration signal sequences generated by normal operation of equipments under different working conditions as training samples, then cluster according to frequency components to extract time and frequency domain features, and train the isolated forest models separately for the samples on each cluster. During online monitoring, collect a vibration signal sequence every other period of time to obtain its frequency component vector and calculate its Cosine Distance from the center of each cluster, classify it into the cluster with the smallest distance. Then extract the time and frequency domain features of the signal sequence, and use the isolated forest model trained by the cluster samples to judge the abnormality.

In order to verify the model's effect, an experiment is carried out by using the vibration signal data set generated during actual production. The data set covers 792 vibration signal sequences with 6000 sampling points in length collected from a certain type of rolling bearing equipment, including 177 vibration signal sequences during normal operation and 615 signal sequences during equipment failure. The failure types cover outer ring failure, inner ring failure and ball failure occurring at three different diameters. Examples of normal signal and abnormal signal are shown in Fig.1. Since each signal sequence is long and may contain different working conditions, sliding interception is carried out on each signal with 500 sampling periods as the window size and 300 sampling periods as the step size. Finally, 15048 signal fragments with a length of 500 are obtained, of which 3363 are normal samples. Among the normal samples, 2685 samples are selected as training samples, the rest of the normal samples are taken as test data together with some fault samples, and the final test set totally contains 857 normal samples and 2191 abnormal samples.

Use the frequency component clustering method presented in this paper to cluster the original normal signals. When m=50 is selected, the variation of the silhouette

coefficient *s* of the clustering result with the number *k* of clustering clusters is shown in Fig. 2. This shows that the clustering has the best effect when *k*=2, and the silhouette coefficient is 0.536, which indicates a reasonable clustering result. The frequency component vectors of the two cluster centers, i.e. the amplitudes at each frequency stage after Fourier transform, are shown in Fig. 3 that the frequency components of the two cluster centers have certain differences. The main frequency's inconsistence reflects the differences in the frequency components of vibration signals under different working conditions.

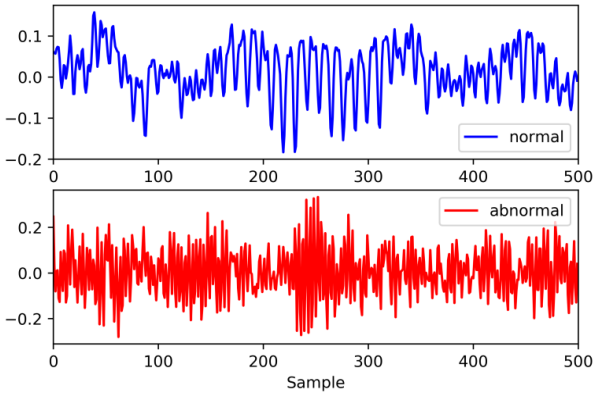

Fig. 1. Examples of normal signal and abnormal signal

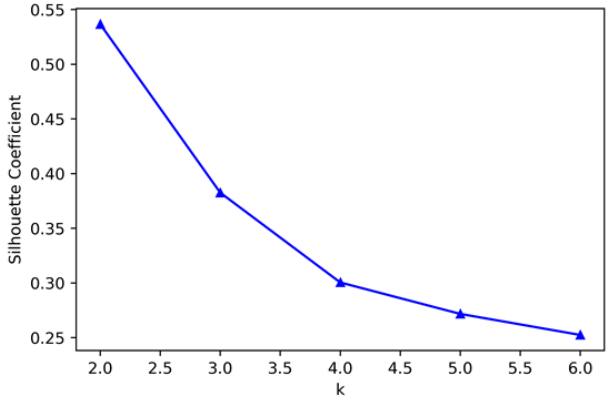

Fig. 2. Silhouette coefficient *s* as a function of clustering clusters numbers *k*

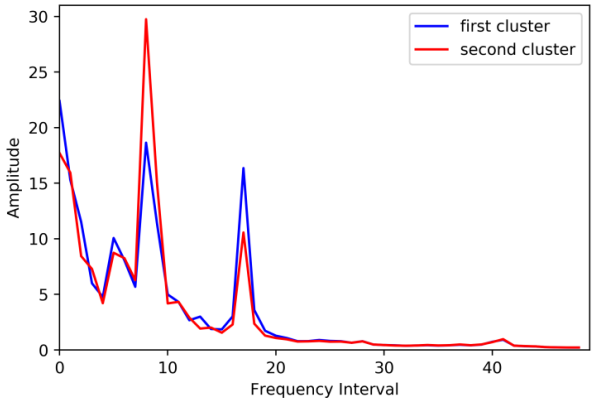

Fig. 3. Frequency component amplitude of the two cluster centers

Extract the 12 features above from the original signal, and establish isolated forest models respectively for the samples on each cluster. The number of iTree is selected as 100 here. In order to contrast with the common fault diagnosis methods of vibration signals, the same tree-based classification algorithm is used. Random forest(RF)[11] classifies the samples after extracting time-frequency domain features, and one-dimensional convolution neural network (1D-CNN)[12] is used to directly classify the original signals. Since RF and CNN are supervised classification algorithms, a certain number of fault samples are required, but the number of fault samples that can be obtained in actual production process is much smaller than that of normal samples. 300 fault samples are randomly selected in this paper from unused fault samples for RF and CNN training. With the same test set, the accuracy and recall of the three methods are shown in Table 3 that under the condition of not using fault samples, the detection accuracy rate of iForest is less than 2% lower than that of supervised learning RF model, and the recall rate is all up to 100%, so the detection result is quite accurate. Fault in industrial processes often brings serious loss, but false alarm only increases a bit manpower demand, so iForest can replace the supervised method in the case of lacking fault samples. For equipment without fault sample records, no manual destructive experiment is needed to obtain fault signal sequences. While for CNN model, serious over-fitting occurs on training samples under the condition of small training samples and serious imbalance of categories due to the large amount of model parameters and data, which causes all test samples are determined as normal on the test set and loses the diagnostic ability. Therefore, although deep learning can extract time sequence features automatically for large data volume without the necessity of artificial feature engineering and with high accuracy, the effect will be greatly reduced for small samples and imbalanced categories.

TABLE III. COMPARISON OF ACCURACY AND RECALL RATE OF THREE METHODS

Model	Accuracy	Recall
iForest	92.35%	100%
RF	94.13%	100%
CNN	28.12%	0%

VI. CONCLUSION

In order to realize high precision anomaly detection of vibration signal in the absence of fault samples, this paper designs an anomaly detection method based on frequency component clustering and isolated forest algorithm. The frequency component vector is used to distinguish samples under different working conditions, and cluster analysis is carried out to establish models for samples under different working condition. Furthermore, the introduction of the isolated forest method also enables the model to better capture the data characteristics of normal samples and accurately identify the faults. Compared with the supervised fault sample method, this model is not easy to over-fit, thus ensuring higher accuracy and recall rate.

ACKNOWLEDGEMENT

Supported by the Science and Technology Major Project of Ningbo City (No. 2018B10047)

*Haoyan Xu and Pengyu Song contributed equally to this work

REFERENCES

[1] W. J. Wang and P. D. Mcfadden, “Application of wavelets to gearbox vibration signals for fault detection,” Journal of Sound & Vibration 192.5(1996): pp.927–939.

[2] Baydar, N. , and A. Ball, “Detection of gear failures via vibration and acoustic signals using wavelet transrofm,” Mechanical Systems and Signal Processing17.4(2003):787-804.

[3] Xianfeng Fan , and M. J. Zuo , “Gearbox fault detection using Hilbert and wavelet packet transform,” Mechanical Systems & Signal Processing 20.4(2006):966-982.

[4] A.K. Nandi, C. Liu, M.L.D. Wong, Intelligent vibration signal processing for condition monitoring, Surveill. 7 Int. Conf. – Oct. 29-30, 2013 Inst. Technol. Chartres, Fr. Plenary Session, 2013

[5] H.O.A. Ahmed, M.L.D. Wong, A.K. Nandi, Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete features, Mech. Syst. Signal Process. 99 (2018) 459–477.

[6] Yaguo Lei, Feng Jia, Jing Lin, Saibo Xing, Steven X. Ding, “An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data,” IEEE Transactions on Industrial Electronics (2016):1-1.

[7] Liu, Fei Tony , K. M. Ting , and Z. H. Zhou . "Isolation Forest."Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on IEEE, 2009.

[8] W.T. Cochran, J.W. Cooley, D.L. Favin, H.D. Helms, R.A. Kaenel, W.W. Lang, et al, “What is the fast Fourier transform? ,” Proceedings of the IEEE 55.10(2005):1664-1674.

[9] J. A. Wong, Hartiganm. A., “Algorithm AS 136: A K-Means Clustering Algorithm,” Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1(1979):100-108.

[10] Aranganayagi, S., and K. Thangavel, “[IEEE International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) - Sivakasi, Tamil Nadu, India (2007.12.13-2007.12.15)] International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) - Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure,”(2007):13-17.

[11] L. Breiman, “Random Forests. Machine Learning,” 2001, 45(1):5- 3.

[12] C. Z. Wu, P. C. Jiang, F. Z. Feng, T. Chen, X. L. Chen, “Faults diagnosis method for gearboxes based on a 1-D convolutional neural network,” Journal of vibration and shock, 2018, 37(22): 51-56.