A Feature Extraction Method for Fault Classification of Rolling Bearing based on PCA

Fengtao Wang¹, Jian Sun¹, Dawen Yan², Shenghua Zhang¹, Liming Cui¹ and Yong Xu³.

¹Institute of Vibration Engineering, Dalian University of Technology, Dalian, China
²School of Mathematical Sciences, Dalian University of Technology, Dalian, China
³West Pacific Petrochemical Company Ltd, Dalian, China

E-mail: wangft@dlut.edu.cn
E-mail: sunjian0804220@163.com

Abstract. This paper discusses the fault feature selection using principal component analysis (PCA) for bearing faults classification. Multiple features selected from the time-frequency domain parameters of vibration signals are analyzed. First, calculate the time domain statistical features, such as root mean square and kurtosis; meanwhile, by Fourier transformation and Hilbert transformation, the frequency statistical features are extracted from the frequency spectrum. Then the PCA is used to reduce the dimension of feature vectors drawn from raw vibration signals, which can improve real time performance and accuracy of the fault diagnosis. Finally, a fuzzy C-means (FCM) model is established to implement the diagnosis of rolling bearing faults. Practical rolling bearing experiment data is used to verify the effectiveness of the proposed method.

1. Introduction
Rolling bearing is widely used in modern rotating machine, but it is also one of the components which are easily damaged. Statistics show that 30% rotating machinery faults are caused by rolling bearings. Therefore, successful fault diagnosis of rolling bearings is of profound significance, which can reduce the incidents and economic loss. Feature extraction of rolling bearing vibration signals is the key of fault diagnosis [1, 2]. When rolling bearing faults occur, the time-frequency statistical features of vibration signal will change. Attempts at analyzing large numbers of original features are motivated by how to effectively extract the potential features responsible for fault characterization.

PCA has been proved to be a useful tool for feature reduction from the original features on fault diagnosis [3]. When PCA-based feature extraction is used for machine bearing fault classification, it is believed that reduced dimensional data which is arranged along the principal eigenvectors could represent features effectively. For example, a PCA-based feature selection scheme was presented to choose the most representative features from a multi-domain feature set for bearing defect
classification [4]. Studies in [5] used the low-dimensional principal component representations from the statistical features of the measured signals to characterize and monitor gearbox conditions. In another study, by adopting the first seven principal components as inputs, a Euclidian distance discriminating approach is utilized to distinguish the bearing fault data [6]. While the frequency statistical features based on Hilbert transformation is not considered in the above studies, this paper summarizes the multiple statistical time-frequency features to construct the high dimensional feature set, then the low-dimensional feature set are obtained by PCA. Experimental data sets which include different fault types and different fault sizes are considered in this study.

The rest of the paper is organized as follows. The theoretical background is briefly introduced in Section 2. The proposed method for fault classification of rolling bearing is described in Section 3. An application case is presented in Section 4. Finally, conclusions are drawn in Section 5.

2. A brief description of PCA

The principal component analysis technique is a statistical analysis approach to mapping multiple characteristic parameters to a few comprehensive features. These PCA-based comprehensive features are not related to each other and can represent original fault features effectively.

For a given feature vector set \( X = \{x_1, \cdots, x_m\} \), \( x_i \in \mathbb{R}^n \) which consists of \( m \) feature vectors, each with \( n \)-dimensional, the algorithm to extract sensitive features from the defect conditions is expressed as follows:

1) Calculate the average value
\[
\mu = \frac{1}{m} \sum_{i=1}^{m} x_i
\] (1)

2) Compute the covariance matrix \( C \) of eigenvectors
\[
C = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T
\] (2)

3) Compute the eigenvalues \( \lambda_i \) and eigenvectors \( v_i \) \((i = 1, 2, \ldots, n)\) of \( C \)
\[
Cv_i = \lambda_i v_i
\] (3)

4) Arrange the eigenvalues in descending order \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \), composite the first \( k \) eigenvalues \( \Delta = (\lambda_1, \lambda_2, \cdots, \lambda_k) \) and corresponding eigenvectors \( V = [v_1, v_2, \ldots, v_k] \). Thus the cumulative contribution rate is defined as
\[
R_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}
\] (4)

\( R_k \) indicates the percentage of the total variance by the first \( k \) principal components (PCs), and the threshold value should be carefully selected to choose the most representative PCs.

5) Finally, a \( k \)-dimensional feature vector \((k < n)\) is obtained
\[
P = V^T X
\] (5)
P is the first \( k \) PCs. Then the goal to reduce feature dimension achieves.

3. The proposed method for fault classification of rolling bearing based on PCA

3.1. Feature extraction
A general feature extraction scheme to classification is divided into two steps. Firstly, select relative statistical features that cover both time and frequency domain, and transform the original data space into feature space. Then, convert global feature space matrix by linear or nonlinear transformation.

In this study, PCA-based feature extraction scheme is introduced for fault classification of rolling bearing. The time-domain statistical features including mean, variance, peak, root mean square, skewness, kurtosis, crest factor and shape factor etc., totally 16 parameters (p1 ~ p16) as defined in table 1, have been considered for bearing fault diagnosis. Since it is difficult to recognize bearing conditions only with time-domain features, the characteristics of the frequency-domain also have been proposed. Frequency domain feature contains 11 parameters (p17 ~ p27) as defined in table 1, including the mean frequency, root mean square frequency and the centre frequency and the variance frequency, in which p17 reflects the size of the vibration energy in the frequency domain; p21 and p23 ~ p25 reflect changes in the frequency band positions; p18 ~ p20, p22 and p26 ~ p27 represent the degree of concentration or dispersion of the spectrum. Extract the 11 parameters (p28 ~ p38) in Hilbert envelope spectrum in the same way as p17 ~ p27. Finally, 38 feature parameters compose the initial feature set [7].

3.2. Feature selection
Apply PCA to the obtained high-dimensional feature set and select the most representative features from the original time-frequency statistics feature set for bearing defect classification based on the defined cumulative contribution rate in equation 4. Finally, adopting the selected representative principal components as inputs, a FCM approach is utilized to classify the bearing fault data. The flowchart of the proposed method is as follows:

**Figure 1.** Flowchart of the proposed method.
| No. | Feature Definition | No. | Feature Definition | No. | Feature Definition |
|-----|-------------------|-----|-------------------|-----|-------------------|
| 1   | \( p_1 = \frac{\sum_{n=1}^{N} x(n)^a}{N} \) | 11  | \( p_{11} = \frac{p_2}{p_4} \) | 21  | \( p_{21} = \frac{\sum_{k=1}^{K} (f_k y(k))}{\sum_{k=1}^{K} y(k)} \) |
| 2   | \( p_2 = \sqrt{\frac{\sum_{n=1}^{N} (x(n) - p_1)^2}{N-1}} \) | 12  | \( p_{12} = \frac{p_8}{p_2} \) | 22  | \( p_{22} = \frac{\sum_{k=1}^{K} (f_k - p_{21})^2 y(k)}{K} \) |
| 3   | \( p_3 = \frac{\sqrt{\sum_{n=1}^{N} |x(n)|}}{N} \) | 13  | \( p_{13} = \frac{p_8}{p_4} \) | 23  | \( p_{23} = \frac{\sum_{k=1}^{K} (f_k^2 y(k))}{\sum_{k=1}^{K} y(k)} \) |
| 4   | \( p_4 = \frac{\sum_{n=1}^{N} |x(n)|}{N} \) | 14  | \( p_{14} = \frac{p_8}{p_3} \) | 24  | \( p_{24} = \sqrt{\frac{\sum_{k=1}^{K} (f_k^4 y(k))}{\sum_{k=1}^{K} (f_k^2 y(k))^2}} \) |
| 5   | \( p_5 = \frac{\sum_{n=1}^{N} (x(n))^3}{N} \) | 15  | \( p_{15} = \frac{p_5}{(\sqrt{p_7})^3} \) | 25  | \( p_{25} = \frac{\sum_{k=1}^{K} (f_k^2 y(k))}{\sqrt{\sum_{k=1}^{K} (f_k^2 y(k))^2}} \) |
| 6   | \( p_6 = \frac{\sum_{n=1}^{N} (x(n))^4}{N} \) | 16  | \( p_{16} = \frac{p_8}{p_7^2} \) | 26  | \( p_{26} = \frac{p_{22}}{p_{21}} \) |
| 7   | \( p_7 = \frac{\sum_{n=1}^{N} (x(n))^2}{N} \) | 17  | \( p_{17} = \frac{\sum_{k=1}^{K} y(k)^b}{K} \) | 27  | \( p_{27} = \frac{\sum_{k=1}^{K} (\sqrt{f_k - p_{21}}) y(k))}{K\sqrt{p_{22}}} \) |
| 8   | \( p_8 = \max |x(n)| \) | 18  | \( p_{18} = \frac{\sum_{k=1}^{K} (y(k) - p_{17})^2}{K - 1} \) | 28  | \( p_{28} = \frac{\sum_{k=1}^{K} (y(k) - p_{17})^4}{K^2 p_{18}^2} \) |
| 9   | \( p_9 = \min |x(n)| \) | 19  | \( p_{19} = \frac{\sum_{k=1}^{K} (y(k) - p_{17})^3}{K(\sqrt{p_{18}})} \) |
| 10  | \( p_{10} = p_8 - p_9 \) | 20  | \( p_{20} = \frac{\sum_{k=1}^{K} (y(k) - p_{17})^4}{K^2 p_{18}^2} \) |

\[ a \quad x(n), n = 1, \cdots, N \] is the time-domain signal sequence, \( N \) is the number of samples.

\[ b \quad y(k), k = 1, 2, \cdots, K \] is the spectrum of the signal \( x(n) \), \( K \) is the number of spectral lines.

\[ c \quad f_k \] is the value of \( k\text{-th} \) frequency spectrum.
4. Case study

To verify the effectiveness of the proposed method, apply it to the Case Western Reserve University Experimental data [8]. The test bearings are 6205-2RS JEM SKF deep groove ball bearings. The driver end bearing is tested under 2 hp loads and speed of 1772 r/min. The vibration data set including four different fault types (normal, inner race fault, outer race fault, and ball fault) is collected at 12K sampling frequency. Single point faults were made on the test bearings using electro-discharge machining with fault diameters of 0.007, 0.014 and 0.021 inches.

The three fault types each with three fault diameters and the normal compose 10 cases. In each case, there are 15 samples with each having 2048 data points. Choose 5 samples randomly from the 15 ones for each case to construct the training data set. The rest ones are used as the testing data set. Data sets in this paper are described in table 2. And we use the D₀7, D₁₄ and D₂₁ data sets to identify the defective components of bearing and use the D₀, D₁ and D₀ data sets to differentiate the defective bearings on the basis of defect size. The time domain acceleration signals for defect classification (four types) and severity classification (four degrees) are depicted in figure 2.

| Dataset | Diameter(mil) | Type  | Size |
|---------|--------------|-------|------|
| D₀₇    | 0 7 7 7      | N I B O | 15 15 15 15 |
| D₁₄    | 0 14 14 14   | N I B O | 15 15 15 15 |
| D₂₁    | 0 21 21 21   | N I B O | 15 15 15 15 |
| D₀     | 0 7 14 21    | N B B B | 15 15 15 15 |
| D₁     | 0 7 14 21    | N I I I | 15 15 15 15 |
| D₀ O   | 0 7 14 21    | N O O O | 15 15 15 15 |

The dimensions of original feature set are 38, including 16 time-domain statistic parameters, 11 FFT-based statistic parameters and 11 Hilbert-based parameters. Totally we get 60*38 feature set for each dataset. Then the PCA is performed on the obtained high dimensional feature set to acquire low dimensional feature set. The results for all datasets show that the accumulate contribution $R_1$ of the first two principal component reaches 99%. So the first 2 PCs are selected to represent the fault. The other PCs are not considered as key contribution of the fault.

![Figure 2. The time-domain vibration signal (a) D₀₇ and (b) D₁.](image)
Figure 3 shows that 2D PCA-based representation of the D₀₇, D₁₄ and D₂₁. X, Y coordinate axes of each figure are represented by first and second principal eigenvector. It can be seen that corresponding to the four types: inner race fault, ball fault, outer race fault and normal condition, the selected first 2 PCs perform well in classifying the three dataset.

Figure 3. The defect classification with data set (a) D₀₇, (b) D₁₄ and (c) D₂₁.

Figure 4 shows that 2D PCA-based representation of the Dᵣᵢ, Dᵢ and D₀. X, Y coordinate axes of each figure are also represented by first and second dominant eigenvector. As depicted in figure 4, corresponding to the four degrees: 0 inches fault, 0.007 inches fault, 0.014 inches fault, 0.021 inches fault, the difference is clearer in the Dᵢ and D₀ than in Dᵣᵢ.

Figure 4. The severity classification with data set (a) Dᵣᵢ, (b) D₀ and (c) D₀.

To give quantitative estimation of the classification results, the Fuzzy C-means clustering method, is applied to the first 2 PCs data set. The number of clusters is 4 and other clustering parameters all select the default value. Finally we classify the clusters according to the membership function matrix. The classifier accuracy with FCM for multiclass data sets is shown in Table 3.
Table 3. Accuracy of FCM for each of the first 2 PCs data set.

| Dataset | D_{07} | D_{14} | D_{21} |
|---------|--------|--------|--------|
| PCA     | 100%   | 100%   | 95%    |
| Dataset | D_B    | D_I    | D_O    |
| PCA     | 100%   | 100%   | 90%    |

The result shows the proposed method has a rather good performance for defect classification (D_{07} and D_{14}) and severity classification (D_B and D_I). And the low accuracy of dataset D_{21} and D_O due to the wrong classification of the ball fault. Overall, the proposed method works well in defect classification and distinguishing inner race fault and outer race fault better than ball fault in severity classification.

5. Conclusion

The primary objective of this paper was to realize defect-classification and severity-classification of rolling element bearings. A PCA-based approach to selecting the most representative features for the classification is developed. Based on the defined time-frequency statistical features, the efficacy of the low-dimensional representations obtained by the proposed method was evaluated via a FCM classifier. The results of the practical rolling bearing experiment data validated the suitability of the PCA-based feature selection scheme.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.51375067), the Aviation Science Foundation of China (No. 20132163010) and the Fundamental Research Funds for the Central Universities of China (No. DUT13JS08).

Reference

[1] Zio E, Baraldi P and Gola G 2008 Feature-based classifier ensembles for diagnosing multiple faults in rotating machinery *Applied Soft Computing.* 8 1365-80
[2] Zhang YZ, Xu GH, Liang L and Wang J 2009 Feature extraction methods for fault classification of rolling element bearing based on nonlinear dimensionality reduction and SVMs *Artificial Intelligence and Computational Intelligence AICI International Conference on.* 3 228-234
[3] Widodo A, Yang B S and Han T 2007 Combination of independent component analysis and support vector machines for intelligent faults diagnosis of induction motors *Expert Systems with Applications.* 32 299-312
[4] Malhi A and Gao R X 2004 PCA-based feature selection scheme for machine defect classification *IEEE Transaction on Instrumentation and Measurement.* 53 1517-25
[5] He QB, Yan RQ, Kong FR and Du RX 2009 Machine condition monitoring using principal component representations *Mechanical Systems and Signal Processing.* 23 446-66
[6] Su Y, Yuan ZH and Qi XX 2013 A feature extraction method of rolling bearings faults based on PCA and Power Spectrum Analysis *International Journal of Advancements in computing Technology.* 5 681-83
[7] Ou L and Yu DJ 2014 Rolling bearing fault diagnosis based on supervised laplaian score and principal component analysis *Journal of Mechanical Engineering.* 50 (Preprint gr-qc/0303038)
[8] Loparo KA 2013 Bearing vibration data set Case Western Reserve University (http://csegrou ps.case.edu/bearingdatacenter/home)