Research on the Method of Rotary Machinery Fault Diagnosis based on PCA and DBN

H P Li, Zh L Qi*, J P Hu and X Y Zhang
Institute of Systems Engineering, Academy of Military Sciences, Beijing, 100141, China
E-mail: zhuoli_qi2020@163.com

Abstract. Aiming at the difficulty of complex vibration signal transmission path, great influence of different sensor positions on diagnosis results and difficulty in feature extraction of rotary machinery fault diagnosis, a new fault diagnosis method based on Principal Component Analysis (PCA) and Deep Belief Network (DBN) is proposed. The framework of the method is constructed. The theory of PCA and DBN are introduced. And the validity and superiority of the proposed method are verified by the experimental data of typical rotary machinery.

1. Introduction
Rotary mechanical equipment fault signal is a typical non-linear and non-stationary signal. And with the increase of equipment complexity and the development of functional diversity, the vibration signal becomes more and more complex. The traditional shallow classification method lacks the ability to process the complex signal.

Lei et al. [1, 2] conducted multiple simulation tests to study common fault modes of planetary gearboxes in oil fields. Wu et al. [3] combined the Hilbert-Huang Transform (HHT) and Support Vector Machine (SVM) to diagnose the different fault modes of gears under the condition of variable rotation speed. Yan et al. [4] improved HHT, reduced iteration error, and made fault diagnosis better. Yang [5] uses EMD to decompose fault signals processed by band pass filter, so as to suppress many shortcomings of traditional HHT methods. Yang et al. [6] applied EMD to decompose vibration signals of large rotating machines, and analyzed the decomposed results to realize fault diagnosis. Lei et al. [7] applied the improved EMD method to fault diagnosis of rotating machinery. Feng et al. [8] proposed a demodulation analysis method based on EEMD and energy separation for fault diagnosis of planetary gearbox. Local Mean Decomposition (LMD) is an adaptive non-stationary signal processing method similar to EMD, which is firstly proposed by Jonathan S. Smith. Liu et al. [9] applied LMD to fault diagnosis of wind power gearbox. Yang et al. [10] proposed an improved LMD method, which performs well in fault diagnosis of the counter rotating subsystem. Feng et al. [11] proposed a new method for frequency and amplitude demodulation of planetary transmission, which has a good diagnosis effect but is highly dependent on expert experience.

Through the analysis of the above research status, it can be seen that the research on fault diagnosis of rotary machinery has achieved good results. However, most of the fault diagnosis methods need signal processing, feature extraction and other preprocessing, which have high technical requirements, great difficulty in application. Therefore, simple and easy fault diagnosis methods still need to be
studied. In order to tackle this dilemma, a new fault diagnosis method based on PCA-DBN is proposed in this paper. PCA is used to comprehensively analyze the signals collected by multiple sensors to reduce the influence of sensor position on diagnosis results. At the same time, the original data is directly processed to obtain the principal components to reduce the dimension of the data and the feature extraction is not needed. And the model training time can be greatly reduced. DBN is a classification method with self-learning ability developed rapidly in recent years. It can extract fault features effectively from quantitative complex signals for learning and classification, and this plays an important role in solving the problem of mechanical equipment fault diagnosis. Although DBN has successfully applied in the fields of image recognition and character recognition, the research and application of mechanical equipment fault diagnosis is not enough, and some related technical problems need to be studied and solved. For this reason, this paper refers DBN to the field of rotary machinery fault diagnosis.

The remaining sections of this paper are organized as follows. Section 2 introduces the framework of the method proposed in this paper. In section 3, the experiment is described and the proposed method is applied to planetary gearbox fault diagnosis. Finally, section 4 gives the conclusions.

2. Framework of the proposed method

2.1. Flowchart of the proposed method

The PCA-DBN based rotary machinery fault diagnosis method presented in this paper is shown in Figure 1, and the specific steps are as follows:

![Flowchart of the method proposed in this paper](image-url)
(1) Vibration data collection: In this step, multiple sensors should be set up to collect signals simultaneously to improve the diagnostic accuracy. Set the data samples collected as \( x_{mn} \), where \( m \) is the number of sensors and \( n \) is the number of data points collected by each sensor.

(2) Data dimension reduction: PCA is used to analyze the data of each sensor, and the first \( p \) principal components (usually accounting for 95% of the total components) are taken as required to obtain the PCA result matrix \( y_{mp} \). Using the advantages of PCA that is simple and easy to implement to extract features to improve the computational accuracy and reduce the computational time.

(3) One-dimensional vector generation: Expand the matrix \( y_{mp} \) in row order into a one-dimensional vector \( z_{1mp} \).

(4) Data preparation: Firstly, the results obtained from PCA processing need to be normalized, and then the data is divided into training data and test data and corresponding classification labels are set.

(5) DBN model training: DBN classification model is trained by training data. First, the initial DBN parameters are set, and then the weights and biases are continuously corrected by using error back propagation. After reaching the pre-set maximum number of iterations, the iteration stops and the model training is completed, as Figure 1 shows.

(6) Diagnosis results output: The effect of DBN is tested using test data and the diagnosis result is given.

2.2. Theory of PCA
PCA is an important multivariate statistical analysis method, which can project the high-dimensional data with correlation to another low-dimensional space, and use the discrete K-L transform algorithm to remove the linear correlation between the data, so as to realize the dimensionality reduction of high-dimensional data. In this process, the most important information in the high-dimensional data is retained, that is the principal components finally obtained. The geometric meaning of this process can be expressed as follows:

\[
\begin{align*}
    u &= x \cos \theta + y \sin \theta \\
    v &= -x \sin \theta + y \cos \theta
\end{align*}
\]

As Figure 2 shows, take sample points of two variables as an example. In the coordinate system \((x, y)\), data points have some obvious linear correlation. After rotating the coordinate system with Angle \( \theta \) counter clockwise, it becomes \((u, v)\). The rotation formula is as follows:

\[
\begin{align*}
    u &= x \cos \theta + y \sin \theta \\
    v &= -x \sin \theta + y \cos \theta
\end{align*}
\]

Figure 2. Schematic map of the geometrical meaning of PCA.

In the coordinate system \((u, v)\), the sample points are basically concentrated on the \( u \) axis, and there is no correlation. At this point, the \( v \) axis contains very little sample information, which can be ignored to some extent. That is, \( u \) is the first principal component and \( v \) is the second principal component. At this time, the first principal component can represent the original sample information. This process eliminates the linear correlation between sample points, and uses one variable to represent the information expressed by the original two variables, so as to achieve the purpose of dimension reduction.

For its specific algorithm process, it can be expressed as follows:

Suppose the original data matrix is \( X_o \),
\[
X_n = \left( x_{ij} \right)_{n \times m} = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1m} \\
  x_{21} & x_{22} & \cdots & x_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\]  

(2)

Where \( m \) is the number of variables, \( n \) is the sample length collected, and \( x_{ij} \) is the value of the \( i \)th sampling point of the \( j \)th variable.

When using PCA, the original data should be normalized to eliminate the difference between variables of different dimensions, as follows:

\[
\overline{x}_{ij} = \frac{x_{ij} - m_j}{\sigma_j} \quad i = 1,2,\ldots;n; \ j = 1,2,\ldots,m
\]  

(3)

\[
m_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}
\]  

(4)

\[
\sigma_j = \left( \frac{1}{n} \sum_{i=1}^{n} \left( x_{ij} - m_j \right)^2 \right)^{1/2}
\]  

(5)

Where \( m_j \) and \( \sigma_j \) are the mean and standard deviation of the \( j \)th variable respectively. The mean matrix and standard deviation matrix of variables can be obtained, \( \overline{M} = [m_1 \ m_2 \ \cdots \ m_m] \) and \( \sigma = [\sigma_1 \ \sigma_2 \ \cdots \ \sigma_m] \), respectively. Then the normalized matrix is:

\[
X = \left[ X_n - \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \overline{M} \right] \text{diag} \left( \frac{1}{\sigma_1} \ \frac{1}{\sigma_2} \ \cdots \ \frac{1}{\sigma_m} \right)
\]  

(6)

At this point, the normalized data matrix is

\[
X = \left( \overline{x}_{ij} \right)_{n \times m} = \begin{bmatrix}
  \overline{x}_{11} & \overline{x}_{12} & \cdots & \overline{x}_{1m} \\
  \vdots & \vdots & \ddots & \vdots \\
  \overline{x}_{n1} & \overline{x}_{n2} & \cdots & \overline{x}_{nm}
\end{bmatrix}
\]  

(7)

The matrix \( X \) is decomposed, and according to the linear algebra analysis, it can be obtained:

\[
X = t_1 p_1^T + t_2 p_2^T + \cdots + t_m p_m^T = TP^T
\]  

(8)

Where \( T = [t_1 \ t_2 \ \cdots \ t_m] \) and \( P = [p_1 \ p_2 \ \cdots \ p_m] \) are score matrix and load matrix, respectively. \( t_i \in \mathbb{R}^{n \times 1} \) and \( p_i \in \mathbb{R}^{m \times 1} \) are score vector and load vector, respectively. And the relationship between them is as follows:

\[
t_i^T t_j = 0 \quad (i \neq j)
\]  

(9)

\[
p_i^T p_j = 0 \quad (i \neq j)
\]  

\[
p_i^T p_i = 1 \quad (i = j)
\]  

The length of \( t_i \) represents the projection of the matrix \( X \) in the \( p_i \) direction, and the larger the value is, the larger the components of the matrix \( X \) in the \( p_i \) direction. If \( t_i \) is sorted in terms of vector length, then there is

\[
\|t_1\| > \|t_2\| > \cdots > \|t_m\|
\]  

(10)
The load vector $p_1$ corresponding to the first score vector $t_1$ is the direction with the largest projection of matrix $X$, while $p_m$ is the opposite direction. Generally, the matrix $X$ projection is mainly concentrated in the direction of the first few load vectors, and the changes in some vectors after that are very small and negligible. So the matrix $X$ can be decomposed into

$$X = t_1p_1^T + t_2p_2^T + \cdots + t_mp_m^T + E = TP^T + E = \hat{X} + E$$

(11)

where, $r$ is the number of principal components. In general, $r \leq m$.

$E$ is the error matrix, and the matrices $[12]_r^T t t t = [12]_r^T P p p p$ are the score matrix of principal components and the corresponding load matrix, respectively. At this point, the covariance matrix is:

$$R_i \approx \frac{1}{n-1}X^TX$$

(13)

The process of analysing matrix $X$ with PCA can be transformed into the feature value analysis of its covariance matrix $R_i$. A set of feature values $\lambda_i$ can be obtained by feature value decomposition of $R_i$, which will be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ sorted by size. The corresponding feature vectors are $p_1, p_2, \cdots, p_m$, and this set of vectors is all the load vectors of $X$. Assuming that the first $r$ principal components are taken after the analysis, then the first $r$ feature values after the principal component analysis are $\lambda_1, \lambda_2, \cdots, \lambda_r$, and the corresponding load vector matrix is $P = [p_1 \ p_2 \ \cdots \ p_r]$. Determining the specific value of $r$ is a very important process, and there are three most commonly used methods: Scree test, cumulative variance contribution rate and parallel analysis. In this paper, the cumulative variance contribution rate method is used to determine $r$ value, so only this method is introduced.

The cumulative variance contribution rate is to determine $r$ by the feature value contribution rate of $R_i$, where the ratio of one feature value to the sum of all feature values is the feature value contribution rate. The cumulative variance contribution rate of the first $r$ principal components is:

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \times 100\%$$

(14)

In application, an acceptance value (such as 95%) is usually set, and it is considered that when the cumulative variance contribution rate of $r$ principal components reaches this value, it is the number of principal components to be retained.

PCA is applied to the fault diagnosis method proposed in this paper. The specific process is as follows: PCA method is used to analyse the data collected by each sensor, and the first $p$ principal components are taken according to the requirements (the cumulative variance contribution rate is usually 95%) to obtain the PCA result matrix, which is expanded into a one-dimensional vector according to the row order.

2.3. Theory of DBN

The concept of deep learning was first proposed in 2006 in an article by Geoffrey Hinton published in Science [12]. After more than ten years of development, many deep learning models have been proposed. It is mainly based on three kinds of deep networks: Convolutional Neural Network (CNN), Deep Belief Network (DBN) and Deep Neural Network (DNN). CNN can be input in the form of matrix, so it is mostly used in the field of image recognition. Moreover, it is usually carried out on the computing platform with GPU, a microprocessor specially used for image computing work, and the training time is relatively long. In the field of mechanical equipment fault diagnosis, it mainly
Deals with the monitoring data, and usually does not involve image processing. Both DNN and DBN can be used to process vibration signals, but DNN belongs to a fully connected structure, that is, layers are connected to each other and within layers. Compared with DBN, DNN has many parameters, long training time and low diagnostic analysis efficiency, so it is difficult to realize real-time monitoring and diagnosis. Therefore, DBN is referred to the field of mechanical equipment fault diagnosis in this paper.

DBN consists of several Restricted Boltzmann Machine (RBM) and classifiers. RBM is the basic unit of DBN, it is a neural network with two layers, and it is a model of probability generating with visual layer ($v$) and hidden layer ($h$). Visual layer and hidden layer are connected by weight $w$ and the bias vector $b$, but there is no connection between the units of visual layer or hide between, each unit in the network (including the visual layer and hidden layer units) has a bias. Its structure is shown in Figure 3.

![Figure 3. The structure of RBM.](image)

As described above, DBN is a multi-layer random generation model stacked by multiple RBM, with the output of low-level RBM as the input of high-level RBM.

Figure 4 is a DBN structure stacked by three RBMs. The input data is the input layer, namely the visual layer of RBM 1. Then the hidden layer of RBM 1 is the visual layer of RBM 2. And it can recurrence to the last RBM, the last output layer gives the sample classification result.

![Figure 4. The fault diagnosis process of DBN.](image)

As shown in Figure 4, the specific workflow of using DBN for fault diagnosis can be divided into the following steps:
Data preprocessing: Preprocess the collected equipment operation data, including data normalization and setting class labels. For example, the class labels of normal, minor and severe fault are set to 1, 2 and 3 respectively.

RBM stacking: Multiple RBMs are stacked layer by layer, and then pertaining and reverse fine-tuning are carried out to adjust the weight and bias of each RBM.

Fault diagnosis: Access the classifier at the top level of RBM, classify the samples and give corresponding classification result to achieve the goal of fault diagnosis.

The training process of DBN classification model is shown in Figure 5. There are two main steps: (1) pre-training of unsupervised learning between each RBM layer; (2) fine-tuning is performed by using the error back-propagation algorithm.

![Figure 5. The flowchart of DBN training process.](image)

3. A case study

3.1. Experimental setup and data acquisition

The experimental system is shown in Figure 6, which is mainly composed of planetary gearbox for experiment, drive motor, speed and torque sensor, and air-cooled magnetic powder brake.

![Figure 6. The experimental system.](image)

The gearbox is a single-stage planetary gear box, consisting of one sun gear (teeth number: 13), three planetary gears (teeth number: 64) and one ring gear (teeth number: 146). And its transmission ratio is 12.5.

In this experiment, a total of three local faults, namely, the wear on the tooth surface of a single gear tooth of the sun gear, ring gear and planet gear, are pre-set. And the fault degree is uniformly set as 1/2 of the tooth length and 1/2 of the width, as shown in Figure 7. Experiments are carried out in normal state and three failure states. The sampling frequency and sampling time are 20 kHz and 12s, respectively, and the rotational speeds are set at 400, 800 and 1200rpm. Each rotational speed load is set at 0, 0.4, 0.8 and 1.2nm.
3.2. Results analysis and discussion

3.2.1. Application effect analysis of fault diagnosis method based on original data-DBN.

The fault diagnosis method based on original data-DBN is to normalize the collected vibration signals and input them directly into the DBN model for fault diagnosis. It is worth noting that data normalization must be carried out before data is input into DBN model.

In order to compare with the previous research results of the author [13], four state signals with a rotation speed of 1200 rpm and a load of 1.2 nm are also selected for analysis in this section. Each state has 33 samples, each sample has 4 columns of vibration signals, and each column of signals has $12 \times 20,000 = 240,000$ data points. In order to improve the calculation speed and ensure the diagnostic accuracy, the number of sampling points (2000) which can contain one running cycle of each gear is selected as a sample, so that there are 3960 samples for each state. The method of $K$-fold cross-validation is adopted (i.e., all samples are divided into $K$ pieces, $K-1$ for training and 1 for testing). Set $K=10$, the number of training samples is 14,256 and the number of test samples is 1,584. The data are normalized and input into the DBN model. For DBN parameter setting, two calculation combinations are adopted to illustrate, one is the number of hidden layer $l=2$, the number of hidden layer units $u_1=1024$, $u_2=512$ and batch size $b_1=99$, $b_2=99$, the other is the number of hidden layer $l=3$, the number of hidden layer units $u_1=1024$, $u_2=512$, $u_3=256$, and the batch size $b_1=99$, $b_2=99$. The results are shown in Figure 8. As can be seen from the results in the figure, the accuracy of using the original vibration data for diagnosis is only about 25%. In addition, because the dimension of the input data is too large, the training time of the models is particularly long. The average training time of the two models is $1.32 \times 10^4$ s and $1.40 \times 10^4$ s respectively.

3.2.2. Application effect analysis of fault diagnosis method based on FFT-DBN.

Each fault of rotating machinery has its unique characteristic frequency. Using FFT to convert signals from time domain to frequency domain can better reflect the characteristics of each fault, which is conducive to fault diagnosis. The fault diagnosis method based on FFT-DBN is to convert data from
time domain to frequency domain based on the data in the previous section, thus reducing the input data dimension from 2000 to 1000. In the calculation process, except that the number of units in each hidden layer is reduced by half, other parameters are set as in the previous section. Taking the experimental data of planetary gearbox seeded fault for analysis, the number of hidden layer \( l = 2 \), number of hidden layer units \( u_1 = 512, u_2 = 256 \) and batch size \( b_1 = 99, b_2 = 99 \) are set, and the number of hidden layer \( l = 3 \), number of hidden layer nodes \( u_1 = 512, u_2 = 256, u_3 = 128 \) and batch size \( b_1 = 99, b_2 = 99 \) are set. The analysis results were shown in Figure 9. It can be seen from the results that, compared with the original time-domain signal as input, the diagnosis accuracy is significantly improved (about 95%) when the FFT transform is applied to the frequency domain signal as input, and the results are rapidly convergent and stable, indicating that the fault characteristic information represented by the frequency domain signal is more accurate and more suitable for the model. In addition, the training time is greatly reduced, about 10min, and the 50 iterations basically reach convergence. It can be said that the FFT-DBN method has a good effect.

![Figure 9. The diagnosis result of FFT-DBN.](image)

3.2.3. Application effect analysis of fault diagnosis method based on original PCA-DBN.

Data composition, training sample number and test sample number are the same as above. Through PCA analysis, it can be seen that the first three principal components of each column in the four columns of signals account for more than 95% of all components, as shown in Figure 10. Therefore, \( p = 3 \), namely, the size of PCA result matrix is \( y_{43}x3 \), and the one-dimensional vector is expanded to be \( z_{1x43} \). Thus, the sample dimension is reduced from 2000 to 12.

![Figure 10. The PCA result.](image)

After processing the data, enter it into the DBN model. First, set the model parameters. With reference to the results above, select the number of hidden layers \( l = 3 \), the number of hidden units \( u_1 = 8, u_2 = 8, u_3 = 8 \), and batch sizes \( b_1 = 99, b_2 = 99 \) for analysis. The settings of other parameters are the same as above. The results are shown in Figure 11. After 2000 iterations, the calculation results tend to be stable. At this time, the diagnosis accuracy of PCA-DBN is 98.42%, and the average time is 173.86s.
4. Conclusion
A rotary machinery fault diagnosis method based on PCA-DBN is proposed. Firstly, the method framework of fault diagnosis method based on PCA-DBN is constructed. Then, the influence of original time domain signal and FFT frequency domain signal as DBN input on the diagnosis effect is analysed. Finally, using the experimental data of planetary gearbox seeded failure verify the better comprehensive performance of the method proposed in this paper, that is, the method has high diagnosis accuracy, stable performance, and short training time.

References
[1] Lei Y, Hoseini M, Pandey M and et.al. 2009 Planetary Gearbox Test Rig: Report of the Run-to-failure Test. University of Alberta.
[2] Lei Y, Hoseini M, Pandey M and et.al. 2009 Planetary Gearbox Test Rig: Experiments on Gears with Pitting Damage. University of Alberta.
[3] Wu T Y, Chen J C and Wang C C. 2012 Characterization of gear faults in variable rotating speed using Hilbert-Huang Transform and instantaneous dimensionless. Mechanical Systems and Signal Processing, 30 103-22.
[4] Yan J H and Lu L. 2014 Improved Hilbert-Huang transform based weak signal detection methodology and its application on incipient fault diagnosis and ECG signal analysis. Signal Processing 98 74-87.
[5] Yang W X. 2008 Interpretation of mechanical signals using an improved Hilbert-Huang transform. Mechanical Systems and Signal Processing 22(5) 1061-71.
[6] Yang W X and Tavner P J. 2009 Empirical mode decomposition, an adaptive approach for interpreting shaft vibratory signals of large rotating machinery. Journal of Sound and Vibration, 321(3-5) 1144-70.
[7] Lei Y G, He Z J and Zi Y Y. 2009 Application of the EEMD method to rotor fault diagnosis of rotating machinery Mechanical Systems and Signal Processing, 23 1327-38.
[8] Feng Z P, Liang M, Zhang Y and et.al. 2012 Fault diagnosis for wind turbine planetary gearboxes via demodulation analysis based on ensemble empirical mode decomposition and energy separation. Renewable Energy, 47 112-26.
[9] Liu W Y, Zhang W H, Han J G and Wang G F. 2012 A new wind turbine fault diagnosis method based on the local mean decomposition. Renewable Energy, 48 411-5.
[10] Yang Y, Cheng J S and Zhang K. 2012 An ensemble local means decomposition method and its application to loca rub-impact fault diagnosis of the rotor systems. Measurement, 45 561-70.
[11] Feng Z P, Zuo M J, Qu J and et.al. 2013 Joint amplitude and frequency demodulation analysis based on local mean decomposition for fault diagnosis of planetary gearboxes. Mechanical Systems and Signal Processing, 40 56-75.
[12] Hinton G E and Salakhutdinov R R. 2006 Reducing the dimensionality of data with neural networks J. Science 313(5786) 504-7.

[13] Li H P, Zhao J M, Liu J and Ni X L. 2016 Application of Empirical Mode Decomposition and Euclidean Distance Technique for Feature Selection and Fault Diagnosis of Planetary Gearbox J. Journal of Vibroengineering 18 (8) 5096-113.