Mechanical Systems and Signal Processing 72-73 (2016) 303–315

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data

Feng Jia, Yaguo Lei *, Jing Lin, Xin Zhou, Na Lu

State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form
5 September 2015
Accepted 26 October 2015
Available online 18 November 2015

Keywords:
Deep learning
Deep neural networks
Intelligent fault diagnosis
Rotating machinery
Massive data

ABSTRACT

Aiming to promptly process the massive fault data and automatically provide accurate diagnosis results, numerous studies have been conducted on intelligent fault diagnosis of rotating machinery. Among these studies, the methods based on artificial neural networks (ANNs) are commonly used, which employ signal processing techniques for extracting features and further input the features to ANNs for classifying faults. Though these methods did work in intelligent fault diagnosis of rotating machinery, they still have two deficiencies. (1) The features are manually extracted depending on much prior knowledge about signal processing techniques and diagnostic expertise. In addition, these manual features are extracted according to a specific diagnosis issue and probably unsuitable for other issues. (2) The ANNs adopted in these methods have shallow architectures, which limits the capacity of ANNs to learn the complex non-linear relationships in fault diagnosis issues. As a breakthrough in artificial intelligence, deep learning holds the potential to overcome the aforementioned deficiencies. Through deep learning, deep neural networks (DNNs) with deep architectures, instead of shallow ones, could be established to mine the useful information from raw data and approximate complex non-linear functions. Based on DNNs, a novel intelligent method is proposed in this paper to overcome the deficiencies of the aforementioned intelligent diagnosis methods. The effectiveness of the proposed method is validated using datasets from rolling element bearings and planetary gearboxes. These datasets contain massive measured signals involving different health conditions under various operating conditions. The diagnosis results show that the proposed method is able to not only adaptively mine available fault characteristics from the measured signals, but also obtain superior diagnosis accuracy compared with the existing methods.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In order to fully inspect the health conditions of rotating machinery, condition monitoring systems are used to collect real-time data from machines and therefore massive data are acquired after long time operation of the machines [1]. As the data is generally collected faster than diagnosticians can analyze it [2], there is an urgent need for diagnosis methods that

* Corresponding author.
E-mail address: yaguolei@mail.xjtu.edu.cn (Y. Lei).

http://dx.doi.org/10.1016/j.ymssp.2015.10.025
0888-3270/© 2015 Elsevier Ltd. All rights reserved.

can effectively analyze massive data and automatically provide accurate diagnosis results. This kind of methods is called intelligent fault diagnosis methods, in which artificial intelligence techniques, such as artificial neural networks (ANNs), support vector machine (SVM), fuzzy inference, etc., are used for distinguishing machinery health conditions [3–5]. Based on the results produced by the intelligent diagnosis methods, it is possible to take appropriate maintenance actions and ensure healthy operation of the machines [6]. Correspondingly, intelligent fault diagnosis methods have been widely investigated and applied in the field of fault diagnosis of rotating machinery [7]. Samanta [8] extracted time-domain features and employed three optimized neural networks to detect pump faults. In addition, Samanta et al. [9] utilized time-domain features to characterize the bearing health conditions and employed ANNs and SVM to diagnose faults of bearings. Statistical features were extracted by Tran et al. [10] for representing the health conditions of induction motor and then decision tree and adaptive neuro-fuzzy inference system (ANFIS) were utilized for distinguishing the faults. Moreover, Tran et al. [11] calculated features from thermal imaging based on bi-dimensional empirical mode decomposition, and then input selected features into relevance vector machine (RVM) for fault classification. Two features were proposed by Lei et al. [12] to characterize health conditions of planetary gearboxes and ANFIS was applied to recognize these health conditions. Widodo et al. [13] calculated statistical features from the measured signals and carried out RVM and SVM to diagnose the bearing faults. Lai et al. [14] introduced cumulants as input features and used radial basis function network as the fault classifier. A method was presented by Bin et al. [15], utilizing wavelet packets-empirical mode decomposition for feature extraction and multi-layer perceptron network for fault classification.

Through the literature review, we notice that ANNs are one of the most commonly used classifiers in the intelligent fault diagnosis methods, which generally include two main steps, i.e. fault feature extraction using signal processing techniques and fault classification using ANN classifiers. Feature extraction involves mapping of measured signals onto representative features characterizing the health conditions of machinery. And fault classification is to distinguish the health conditions based on the extracted features. Thanks to the representative features from the measured signals and adaptive learning capability of ANNs, the ANN-based methods are supposed to displace diagnosticians for making decisions and work well in intelligent fault diagnosis [7]. The ANN-based methods reported in literature, however, have two obvious deficiencies: (1) The features input into classifiers are extracted and selected by diagnosticians from the measured signals, largely depending on prior knowledge about signal processing techniques and diagnostic expertise. In addition, the features are selected according to a specific diagnosis issue and probably unsuitable for other issues. Thus it is necessary to adaptively mine the characteristics hidden in the measured signals to reflect the different health conditions of machinery, instead of extracting and selecting features manually. (2) The ANNs commonly adopted in intelligent fault diagnosis of rotating machinery have shallow architectures, which means that only one hidden layer is included in an ANN architecture, like the ANNs in Refs. [8,9,14,15]. Such simple architectures limit the capacity of ANNs to learn the complex non-linear relationships in fault diagnosis issues. Thus it is necessary to establish a deep architecture network for distinguishing the health conditions of machinery.

Deep learning [16] holds the potential to overcome the aforementioned deficiencies in current intelligent diagnosis methods. It refers to a class of machine learning techniques, where many layers of information processing stages in deep architectures are exploited for pattern classification and other tasks [17]. Using deep learning, deep neural networks (DNNs) with deep architectures can be established. Due to the deep architectures, DNNs are able to adaptively capture the representative information from raw data through multiple non-linear transformations and approximate complex non-linear functions with a small error. Since the idea of deep learning appeared in *Science*, it has attracted lot of attention from researchers in different fields [18]. Dahl et al. [19] proposed a pre-trained deep neural network hidden Markov model for large-vocabulary speech recognition and obtained an accuracy improvement compared with traditional models. Krizhevsky et al. [20] developed a DNN-based method in large scale visual recognition challenge involving millions of labeled images, and got the best result. Deep learning methods were utilized by Baldi et al. [21] to search for exotic particles in high-energy physics and the results demonstrated that the methods can improve the searching ability of collider. The aforementioned applications prove that deep learning is a promising tool in dealing with massive data. But it attracts few attentions in the field of fault diagnosis. Based on Teager–Kaiser energy operator and deep belief network trained by deep learning, Tran et al. [22] proposed a new method for diagnosing faults of reciprocating compressor valves. In this method, they treated deep belief network as a classifier and still manually extracted features to input the classifier, which ignored the ability of the network in mining fault characteristics.

Based on DNNs trained through deep learning, this paper proposes a novel intelligent diagnosis method to overcome the two deficiencies of the ANN-based methods in fault diagnosis of rotating machinery. In this method, DNNs are utilized to implement both fault feature extraction and intelligent diagnosis. The DNNs are first pre-trained by an unsupervised layer-by-layer learning and then fine-tuned with a supervised algorithm, where the unsupervised process helps the fault characteristic mining and the supervised process contributes to construct the discriminative fault characteristics for classification [23]. The merits of the proposed method are summarized as follows. (1) It is able to adaptively mine fault characteristics from the measured signals for various diagnosis issues. (2) The method is good at establishing the non-linear mapping relationship between the different health conditions of machinery and the corresponding measured signals. Therefore, the proposed method is expected to obtain higher diagnosis accuracy compared with the methods based on shallow ANNs. The rest of this paper is organized as follows. Section 2 briefly introduces the theoretical background of DNNs. Section 3 is dedicated to a description of the proposed intelligent diagnosis method. In Section 4, the effectiveness of the proposed method is validated using four rolling element bearing datasets and a planetary gearbox dataset. The bearing datasets contain thousands of signals with different fault categories and severities under various operating loads. And the gearbox dataset includes tens of thousands of signals with different fault modes and locations

under various operating conditions, like different rotating speeds and loads. In addition, the proposed method is compared with several intelligent methods using the same bearing datasets in this section. Conclusions are drawn in Section 5.

2. A brief introduction to DNNs

DNNs have deep architectures containing multiple hidden layers and each hidden layer conducts a non-linear transformation from the previous layer to next one [18,24]. Through deep learning addressed by Hinton et al. [16], DNNs are trained according to the following two main procedures: (1) Pre-train the DNNs layer by layer with unsupervised techniques, like autoencoders. (2) Further fine-tune the DNNs with back propagation (BP) algorithm for classification.

2.1. Autoencoders

An autoencoder is one type of unsupervised neural networks with three layers [24,25] and the output target of the autoencoder is the input data. As depicted in Fig. 1, the autoencoder comprises two parts, i.e., encoder network and decoder network. The encoder network transforms the input data from a high-dimensional space into codes in a low-dimensional space and the decoder network reconstructs the inputs from the corresponding codes.

The encoder network is explicitly defined as an encoding function denoted by f_θ [24]. This function is called the encoder. For each measured signal $\boldsymbol{x}^m$ from a dataset $\{\boldsymbol{x}^m\}_{m=1}^M$ of rotating machinery, we define

$$\boldsymbol{h}^m = f_{\boldsymbol{\theta}}(\boldsymbol{x}^m) \tag{1}$$

where $\boldsymbol{h}^m$ is the encode vector obtained from $\boldsymbol{x}^m$.

The decoder network is defined as a reconstruction function denoted by $g_{\boldsymbol{\theta}'}$, namely the decoder. It maps $\boldsymbol{h}^m$ from the low-dimensional space back into the high-dimensional space, producing a reconstruction

$$\hat{\boldsymbol{x}}^m = g_{\boldsymbol{\theta}'}(\boldsymbol{h}^m) \tag{2}$$

The parameter sets of the encoder and decoder are learned simultaneously on the task of reconstructing as well as possible the original input, attempting to incur the lowest possible reconstruction error $L(\boldsymbol{x}, \hat{\mathbf{x}})$ over the M training examples, where $L(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is a loss function that measures the discrepancy between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ [24].

In summary, the autoencoder training aims to find the parameter sets Q and $\boldsymbol{\theta}'$ minimizing reconstruction error:

$$\phi_{\mathrm{AE}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{M}\sum_{m=1}^{M} L(\boldsymbol{x}^m, g_{\boldsymbol{\theta}'}(f_{\boldsymbol{\theta}}(\boldsymbol{x}^m))) \tag{3}$$

The commonly used forms for the encoder and decoder are affine mappings [26], optionally followed by a non-linearity:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = s_f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{4}$$

$$g_{\boldsymbol{\theta}'}(\boldsymbol{x}) = s_g(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}) \tag{5}$$

$$L(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 \tag{6}$$

where s_f and s_g are the encoder and decoder activation functions, respectively. Thus, the parameter sets of the autoencoder are $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}\}$ and $\boldsymbol{\theta}' = \{\boldsymbol{W}^{\mathrm{T}}, \boldsymbol{d}\}$, where $\boldsymbol{b}$ and $\boldsymbol{d}$ are bias vectors, and $\boldsymbol{W}$ and $\boldsymbol{W}^{\mathrm{T}}$ are the weight matrices.

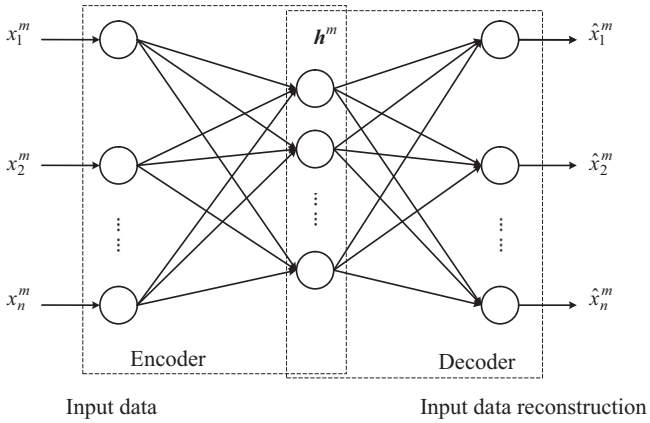

Fig. 1. Architectural graph of an autoencoder.

2.2. Pre-training and fine-tuning

N autoencoders could be stacked to pre-train an N-hidden-layer DNN. When given input signal $\boldsymbol{x}^m$, the input layer and the first hidden layer of the DNN are regarded as the encoder network of the first autoencoder. After the first autoencoder is trained through minimizing the reconstruction error in Eq. (3), the trained parameter set $\boldsymbol{\theta}_1$ of the encoder network is used to initialize the first hidden layer of the DNN. And the first encode vector $\boldsymbol{h}_1^m$ of the $\boldsymbol{x}^m$ is calculated as follows:

$$\mathbf{h}_1^m = f_{\boldsymbol{\theta}_1}(\boldsymbol{x}^m) \tag{7}$$

Then the encode vector $\boldsymbol{h}_1^m$ is the input data, the first hidden layer and the second hidden layer of the DNN are regarded as the encoder network of the second autoencoder. Correspondingly, the second hidden layer of the DNN is initialized by the second trained autoencoder. The process is conducted in the sequence until the Nth autoencoder is trained for initializing the final hidden layer of the DNN. And the Nth encode vector $\boldsymbol{h}_N^m$ of the $\boldsymbol{x}^m$ is calculated as

$$\boldsymbol{h}_N^m = f_{\boldsymbol{\theta}_N}(\boldsymbol{h}_{N-1}^m) \tag{8}$$

where $\boldsymbol{\theta}_N$ is the parameter set of the Nth autoencoder.

In this way, through training N stacked autoencoders, all the hidden layers of the DNN are pre-trained. This pre-training process is proven to yield significantly better local minima than random initialization of the DNN and helps achieve better generalization in classification tasks [26,27], as well as in fault diagnosis of rotating machinery.

After the DNN is pre-trained, fine-tuning process is utilized in next step of the DNN training. The output layer of the DNN is employed to contain the output targets for classification tasks. The output of the DNN calculated from the input signal $\boldsymbol{x}^m$ is

$$\boldsymbol{y}^m = f_{\boldsymbol{\theta}_{N+1}}(\boldsymbol{h}_N^m) \tag{9}$$

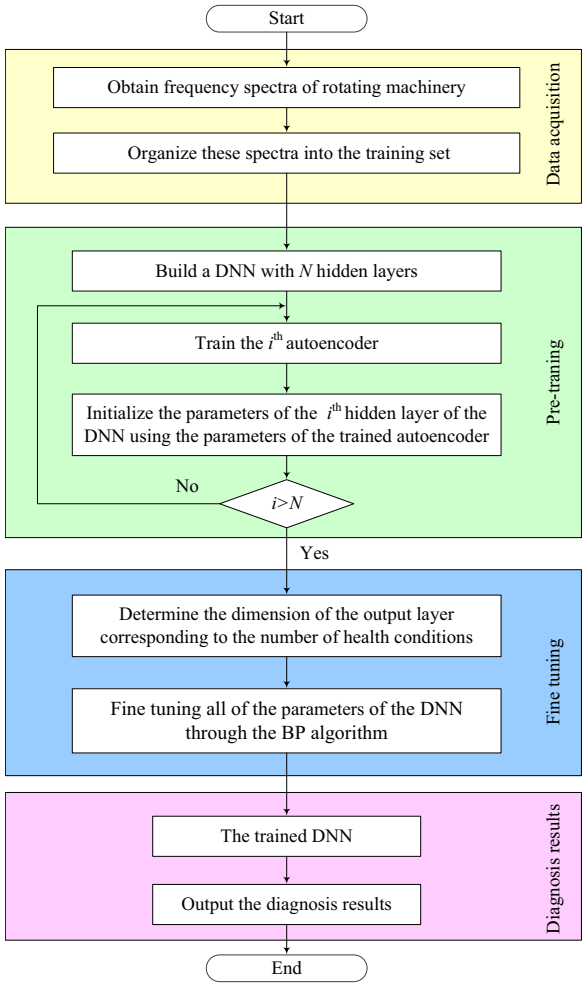

Fig. 2. Flowchart of the proposed method.

where $\boldsymbol{\theta}_{N+1}$ is the parameter set of output layer. In order to approximate the output target properly, BP algorithm is utilized to minimize the error of the output by adjusting the parameters in the DNN backwards [28]. Supposing that the output target of the $\boldsymbol{x}^m$ is $\boldsymbol{d}^m$, the error criterion is described as

$$\phi_{\mathrm{DNN}}(\Theta)=\frac{1}{M}\sum_{m}L(\boldsymbol{y}^m,\boldsymbol{d}^m) \tag{10}$$

where $\Theta=\{\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2,\ \cdots,\ \boldsymbol{\theta}_{N+1}\}$. The parameter set Θ can be updated as follows.

$$\Theta=\Theta-\eta\frac{\partial\phi_{\mathrm{DNN}}(\Theta)}{\partial\Theta} \tag{11}$$

where η is the learning rate of the fine-tuning process, which is introduced to guarantee a convergence in the update procedure [29].

3. DNN-based intelligent diagnosis method

Based on DNNs, this study proposes a novel intelligent fault diagnosis method that adaptively mines the fault characteristics from raw signals of rotating machinery and automatically classifies machinery health conditions with these fault characteristics. The raw signals refer to the measured signals in the frequency domain, i.e., frequency spectra. And the main reason of using frequency spectra is that the frequency spectra of rotating machinery show how their constitutive components are distributed with discrete frequencies and may provide clear information about the health conditions of rotating machinery [30].

As shown in Fig. 2, the proposed method includes the following four procedures: (1) Obtain the frequency spectra of rotating machinery under different health conditions. These spectra comprise the training set $\left\{\boldsymbol{x}^i,\boldsymbol{d}^i\right\}_{i=1}^{M}$, where $\boldsymbol{x}^i$ is the ith frequency spectrum for training, $\boldsymbol{d}^i$ is the health condition label of the $\boldsymbol{x}^i$ and M is the number of the frequency spectra. (2) Build a DNN with multiple hidden layers, in which the number of the input units is the dimension of the frequency spectrum $\boldsymbol{x}^i$. Then utilize the unlabeled training set $\boldsymbol{x}=\{\boldsymbol{x}^i\}_{i=1}^{M}$ to pre-train the DNN layer by layer with a stack of autoencoders, where the number of autoencoders refers to the number of hidden layers inside the DNN. The process is specifically displayed in Fig. 3. Firstly, regard first hidden layer of the DNN as hidden layer of the first autoencoder and utilize the unlabeled training set $\boldsymbol{x}$ as input data and output target to train the first autoencoder, as shown in Fig. 3(a). The trained parameters $\{\boldsymbol{W}_1,\boldsymbol{b}_1\}$ of the autoencoder are used to initialize the parameters of the first hidden layer of the DNN, and $\boldsymbol{h}_1$ is the encode vector computed from frequency spectra of rotating machinery by the first autoencoder. Then, use $\boldsymbol{h}_1$ as the inputs and outputs to train the second autoencoder for initializing parameters of the second hidden layer of the DNN, and obtain $\boldsymbol{h}_2$ in Fig. 3(b). Finally, continue the training steps in the sequence until the Nth autoencoder is trained and the frequency spectra are coded into $\boldsymbol{h}_N$ in Fig. 3(c). In this way, all of the hidden layers of the DNN are pre-trained. (3) Determine the dimension of the output layer according to the number of the machinery health conditions. And implement the BP algorithm to fine-tune the parameters of the DNN through minimizing the error between the

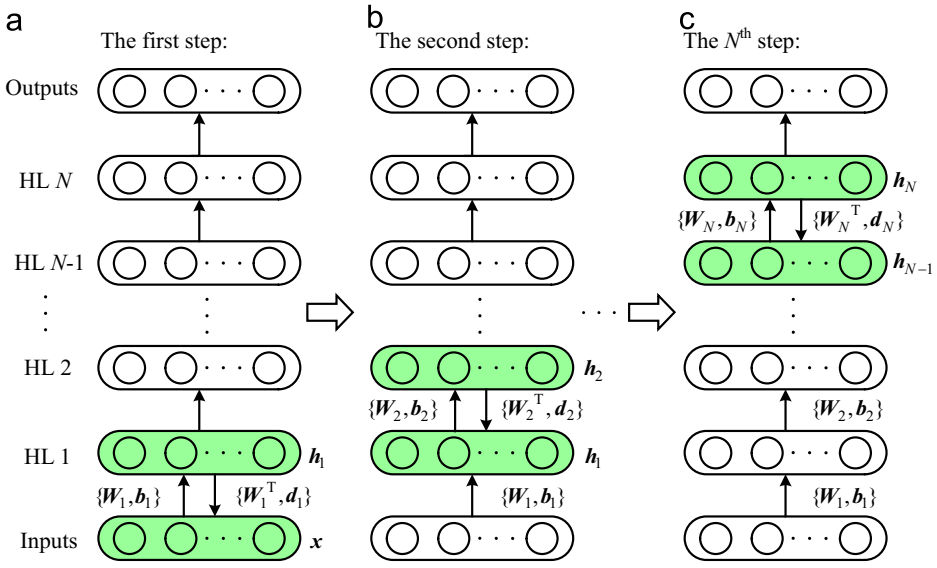

Fig. 3. Diagram of illustrating the pre-training process (HL is short for hidden layer): (a) train the first autoencoder of the DNN, (b) train the second autoencoder and (c) train the Nth autoencoder.

output calculated from the frequency spectra and health condition labels. (4) Employ the trained DNN to diagnose faults of the machinery.

In the proposed method, the pre-training process helps DNNs learn multiple nonlinear transformations, capturing the main variations of the frequency spectra [26]. Then the fine-tuning process helps the DNNs discover the discriminative information from these spectra. These processes enable the trained DNNs to mine the essential characteristics from the frequency spectra and establish complex non-linear mapping between the frequency spectra and health condition labels. Therefore, the proposed method could implement fault characteristics mining and intelligent diagnosis of rotating machinery.

4. Fault diagnosis using the proposed method

Rolling element bearings and gears are the key components in rotating machinery [31,32]. The health conditions of these components often affect the performance, reliability and service life of the machinery. Whereas due to tough working environment, these components easily suffer from different kinds of damage, leading to breakdowns and heavy economic losses [33,34]. In this section, two diagnosis cases of rolling element bearings and planetary gearboxes are used to validate the proposed method, respectively.

4.1. Case 1: Fault diagnosis of rolling element bearings

4.1.1. Data description

The bearing data used here are provided by the Case Western Reserve University (CWRU) [35]. The data were collected from a motor driving mechanical system under four different loads with the sampling frequency of 48 kHz. The bearing data set was obtained from the experimental system: (1) under normal condition (N), (2) with outer race fault (OF), (3) with inner race fault (IF) and (4) with roller fault (RF). The faults were introduced into the drive-end bearing of the motor with fault diameters of 0.18 mm, 0.36 mm and 0.54 mm, respectively.

In this study, four datasets, i.e. A–D, of the bearings are used to test the performance of the proposed method, and the detailed description for the datasets is shown in Table 1. Datasets A, B and C contain 10 bearing health conditions under loads of 1, 2 and 3 hp, respectively. There are 200 signals for each health condition and each signal contains 2400 data points. We implement fast Fourier transformation on each signal to get the 2400 Fourier coefficients. Since the coefficients are symmetric, the first 1200 coefficients are used in each sample. So each of dataset A–C totally contains 2000 samples and each sample has 1200 Fourier coefficients. Dataset D contains 10 bearing health conditions under loads of 1–3 hp and there are 600 samples for each condition, where the same health condition under different loads is treated as one class. So dataset D includes 6000 samples.

4.1.2. Diagnosis results

The designed DNN has five layers, in which the unit number of the input layer is determined by the dimension of the samples, the unit number of the first hidden layer, the second hidden layer and the third hidden layer are 600, 200 and 100 respectively, and the unit number of the output layer is determined by the number of the health conditions. The active functions of the DNN are hyperbolic tangent functions. The weights of the DNN are initialized randomly and the biases are initialized to zero. The maximum training epoch is 100, the learning rate is 0.05 and the momentum is 0.05. Fifty percent of samples are randomly selected for pre-training the DNN, then these samples are used to fine-tune the parameters of the whole network, and the other 50% of samples are used to test the performance. Twenty trials are carried out for diagnosing each bearing dataset. The diagnosis results of datasets A–D using the proposed method are shown in Fig. 4. In these trials, all of the diagnosis accuracies are over 99% and some are even 100%, which means that the proposed method is able to distinguish the ten health conditions of bearings with a high accuracy.

For comparison, we also deal with the datasets utilizing the method based on back propagation neural network (BPNN), which is commonly used in intelligent fault diagnosis of rotating machinery. The BPNN has the same architecture as the DNN and is

Table 1
Description of bearing datasets

Datasets	Load (hp)	The number of samples	Fault type	Fault diameter (mm)	Classification label
A/B/C/D	1/2/3/1 − 3	200/200/200/600	N	0	1
		200/200/200/600	RF	0.18	2
		200/200/200/600	RF	0.36	3
		200/200/200/600	RF	0.54	4
		200/200/200/600	IF	0.18	5
		200/200/200/600	IF	0.36	6
		200/200/200/600	IF	0.54	7
		200/200/200/600	OF	0.18	8
		200/200/200/600	OF	0.36	9
		200/200/200/600	OF	0.54	10

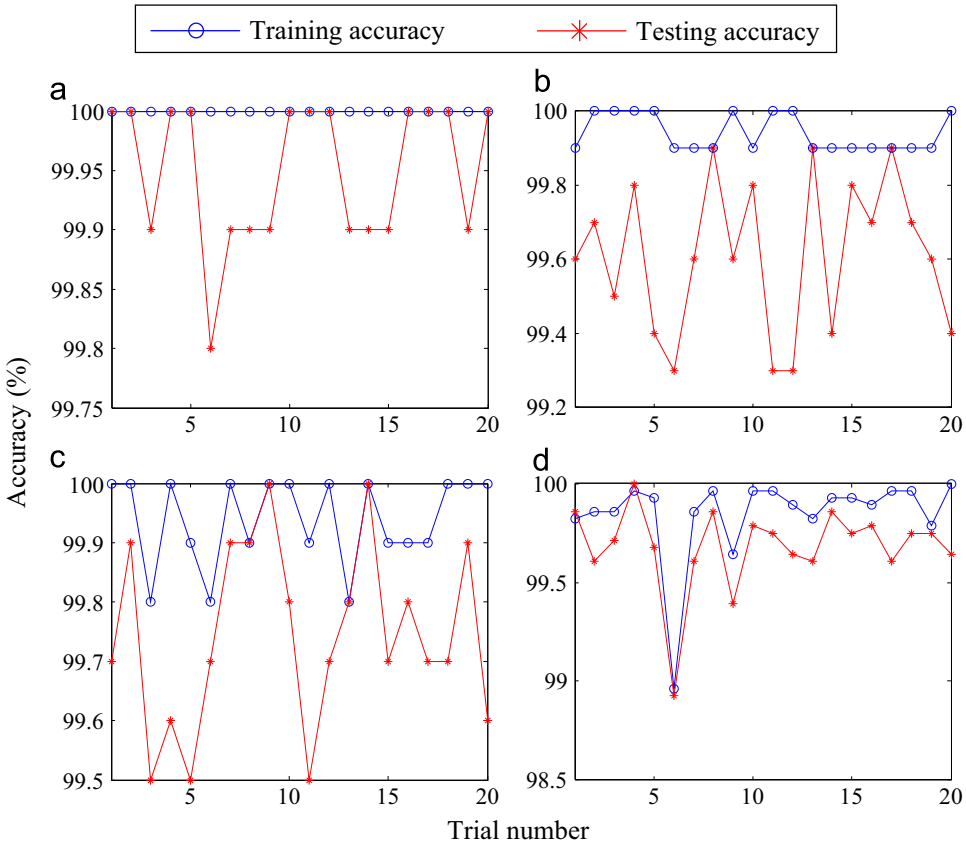

Fig. 4. Diagnosis results of 20 trials of bearing datasets using the proposed method: (a) dataset A, (b) dataset B, (c) dataset C and (d) dataset D.

trained by the same parameters. In each trial, it should be noted that the same training set is used to train the DNN and the BPNN, respectively. The diagnosis results using the BPNN-based method are depicted in Fig. 5. It shows that the diagnosis accuracies vary greatly. Most of the accuracies are around 70%, a few of accuracies are good, and the others are terrible. To quantitatively compare the results between the proposed method and the BPNN-based method, the average accuracies and the standard deviations of 20 trials are calculated for each dataset, as shown in Table 2. It presents that the average training and testing accuracies using the proposed method range from 99.61% to 100% and the standard deviations below 0.22%, which means the proposed method can effectively and stably distinguish not only bearing fault categories but also fault severities. But the average accuracies using the BPNN-based method are much lower and the standard deviations are larger, which illustrates the superiority of the proposed method to the BPNN-based methods. It is worth mentioning that dataset D includes massive samples for 10 bearing health conditions under different loads, and its diagnosis accuracies illustrate that the proposed method is able to diagnose the bearing faults regardless of the load fluctuation.

In Fig. 5, an interesting phenomenon is observed that the 18th trial of dataset A and the 8th trial of dataset D using the BPNN-based method perform as well as the same trials using the proposed method. The comparisons of training errors are shown in Fig. 6(a) and (b). It is noticed that the training errors are close to 0 in the two trials using both methods. This demonstrates that neural networks with deep architectures indeed have great capability of recognizing nonlinear characteristics of the bearing failure so as to facilitate the diagnosis. Unfortunately, the BPNN-based method happens to perform well because of random factors, but not always. Taking the 12th trial of dataset B and the 10th trial of dataset C for examples, the results of both trials using the BPNN-based method are extremely poor, and both of testing accuracy is 10.03%. It seems that in the two trials, the trained networks have learned nothing useful for fault diagnosis, and the training errors of these trials are shown in Fig. 6(c) and (d). They show that the trainings get stuck in a premature convergence, leading to the unsatisfied diagnosis accuracies. However, the training errors of the proposed method are updated smoothly into good solutions. It indicates that the proposed method is more robust than the BPNN-based method, which is helpful for achieving better diagnosis results of bearings.

To verify the ability of the proposed method in adaptively mining fault characteristics, principal component analysis (PCA) is utilized to visualize the mined features. These features are considered as the inputs of the output layer of the DNN in this study since the output layer of the DNN can be treated as a classifier. Therefore, 100 features are obtained from a sample through the DNN. The PCA is implemented on the 100 features of each sample and the first three principal components (PCs) of the features of datasets A–D are shown in Fig. 7(a)–(d), respectively. It is seen that most features of the same health condition are gathered in the

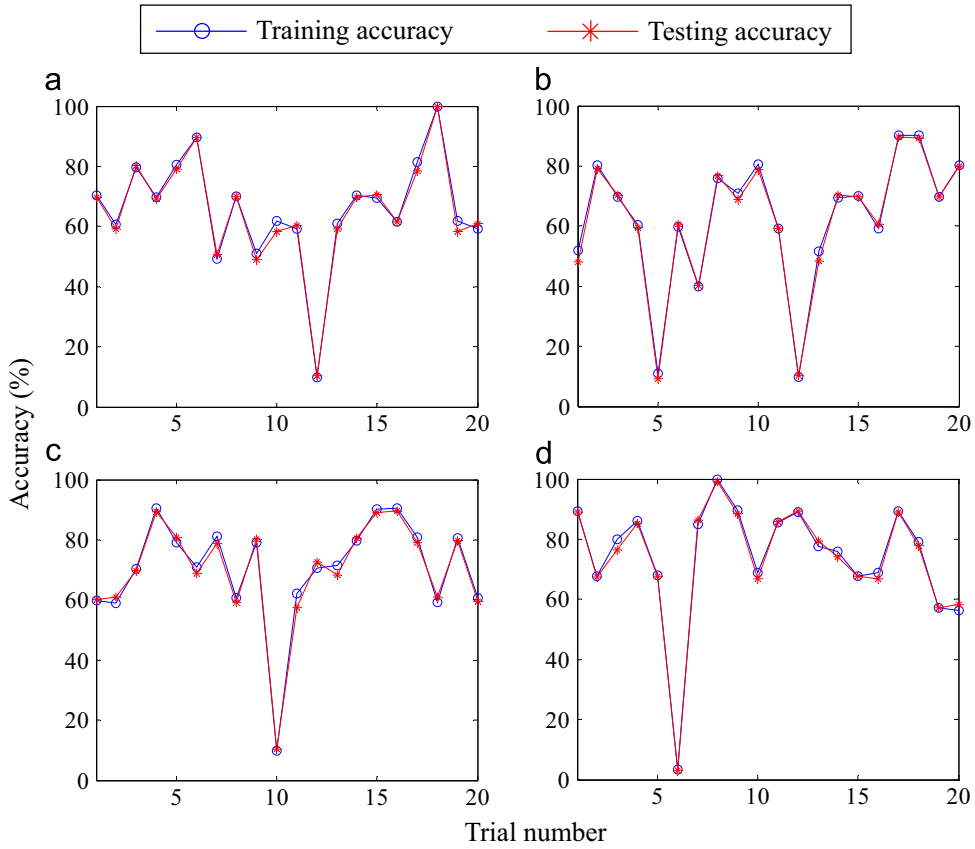

Fig. 5. Diagnosis results of 20 trials of bearing datasets using the BPNN-based method: (a) dataset A, (b) dataset B, (c) dataset C and (d) dataset D.

Table 2
Diagnosis results of bearing datasets.

Datasets	The proposed method		The BPNN-based method	
	Training accuracy	Testing accuracy	Training accuracy	Testing accuracy
Dataset A	100	99.95 ± 0.06	65.81 ± 18.2	65.2 ± 18.09
Dataset B	99.94 ± 0.05	99.61 ± 0.21	62.51 ± 21.97	61.95 ± 22.09
Dataset C	99.94 ± 0.08	99.74 ± 0.16	70.31 ± 17.86	69.82 ± 17.67
Dataset D	99.85 ± 0.22	99.68 ± 0.22	74.24 ± 20.31	73.74 ± 20.23
Dataset E	100	100	81.34 ± 17.92	81.35 ± 17.86

The format of the results is: average accuracy ± standard deviation.

corresponding cluster and most features of the different health conditions are separated. In Fig. 7(c) and (d), RF with 0.36 mm defect and RF with 0.54 mm defect are not separated well by the first three PCs, but the other principal components contain the essential information for clustering the samples of the two fault severities. Therefore, datasets C and D still achieve the high diagnosis accuracies. The results reveal that the proposed method could adaptively mine the fault characteristics of rolling element bearings under varying operation conditions.

4.2. Case 2: Fault diagnosis of planetary gearboxes

4.2.1. Experiments and data description

The planetary gearbox data used in this section is collected from the test rig shown in Fig. 8. It contains a two-stage planetary gearbox, a two-stage fixed-axis gearbox, a 3-hp motor for driving the gearboxes and a magnetic brake for loading. The accelerometer mounted on the first stage bearing end cover is used to acquire vibration signals with a sampling frequency of 5120 Hz.

Seven experiments are carried out under different health conditions. These conditions involve normal, a pitted tooth on the sun gear of the first stage, a cracked tooth on the sun gear of the first stage, a chipped tooth on the planetary gear of the first stage, a chipped tooth on the sun gear of the second stage, a missing tooth on the sun gear of the second stage and a

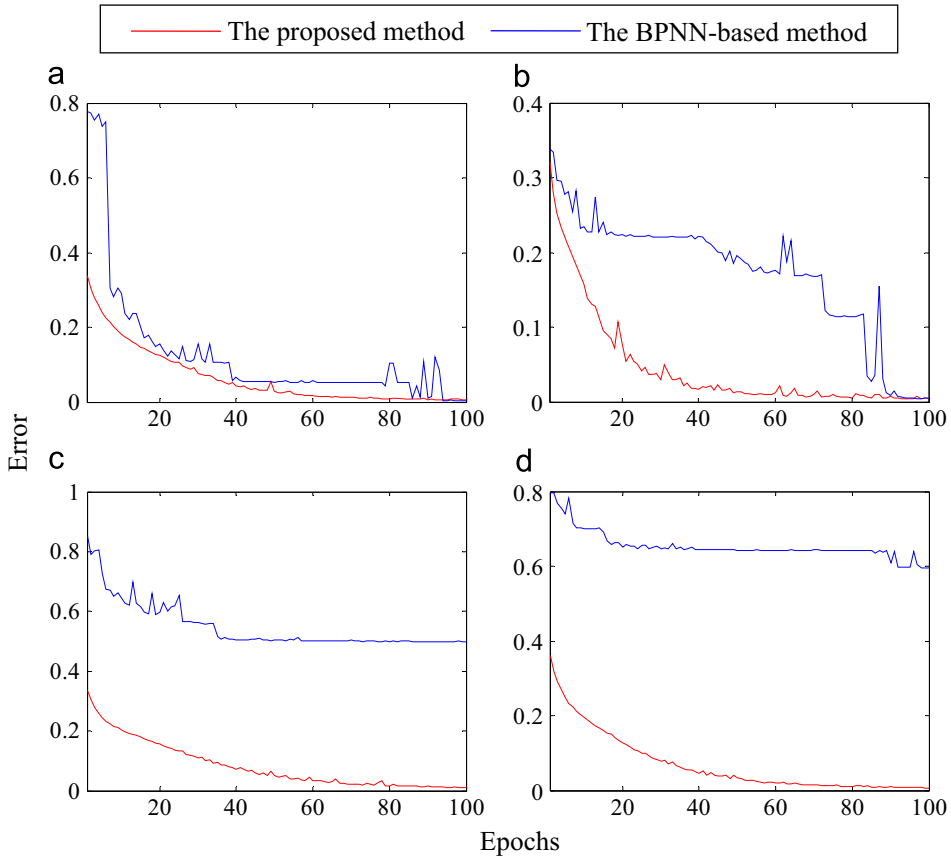

Fig. 6. Curves of the training error of the proposed method and the BPNN-based method: (a) the 18th trial of dataset A, (b) the 8th trial of dataset D, (c) the 12th trial of dataset B and (d) the 10th trial of dataset C.

bearing inner race fault of planetary gear of the first stage. In each experiment, a damaged component is installed inside the test rig and other components are normal. Each experiment is conducted under eight operating conditions: four different drive motor speeds (2100 rpm, 2400 rpm, 2700 rpm and 3000 rpm) and two loading conditions (no load and the maximum load). There are 203 signals for each experiment under an identical operating condition and each signal contains 2560 data points. The frequency spectrum of each signal is used as each sample, which contains 1280 Fourier coefficients. The samples of seven health conditions under eight operating conditions compose the dataset E, which totally has 12,992 samples. Fig. 9 gives an example of signals and their spectra for each health condition. Although there are some differences of the seven vibration signals and the corresponding spectra, it is difficult to manually distinguish the different health conditions by those differences. With thousands of samples needing to be distinguished, the manual diagnosis becomes more impractical. Therefore, the proposed method is applied to diagnose the faults of the planetary gearbox.

4.2.2. Diagnosis results

Both the proposed method and the BPNN-based method are used to process the dataset. Twenty trials are carried out, where 50% of samples are randomly selected for training and the other samples are used for testing.

The average accuracies and the corresponding standard deviations of 20 trials are calculated for dataset E, as shown in Table 2. In the trials using the proposed method, the average accuracies of both training and testing is 100%, which means all the samples are correctly classified. Using the BPNN-based method, however, the average training accuracy is only 81.34% with considerably large standard deviation of 17.92% and the average testing accuracy is in the similar situation. This implies that the proposed method obtains higher diagnosis accuracies and shows better robustness than the BPNN-based method does in distinguishing the various fault modes and different fault locations of the planetary gearbox.

The scatter plots of principal components of the mined features are displayed in Fig. 10. It is noticed that the features characterizing the same health conditions are clustered well and each cluster is separated. It deserves to be mentioned that dataset E contains massive samples under eight operating conditions for each health condition, which is ordinarily difficult for extracting and selecting features from such dataset. Nevertheless, the results of Fig. 10 indicate that the proposed method still extracts the essential characteristics for distinguishing the health conditions of the planetary gearbox. Thus the proposed method has a great ability in adaptively mining the fault characteristics of planetary gearboxes.

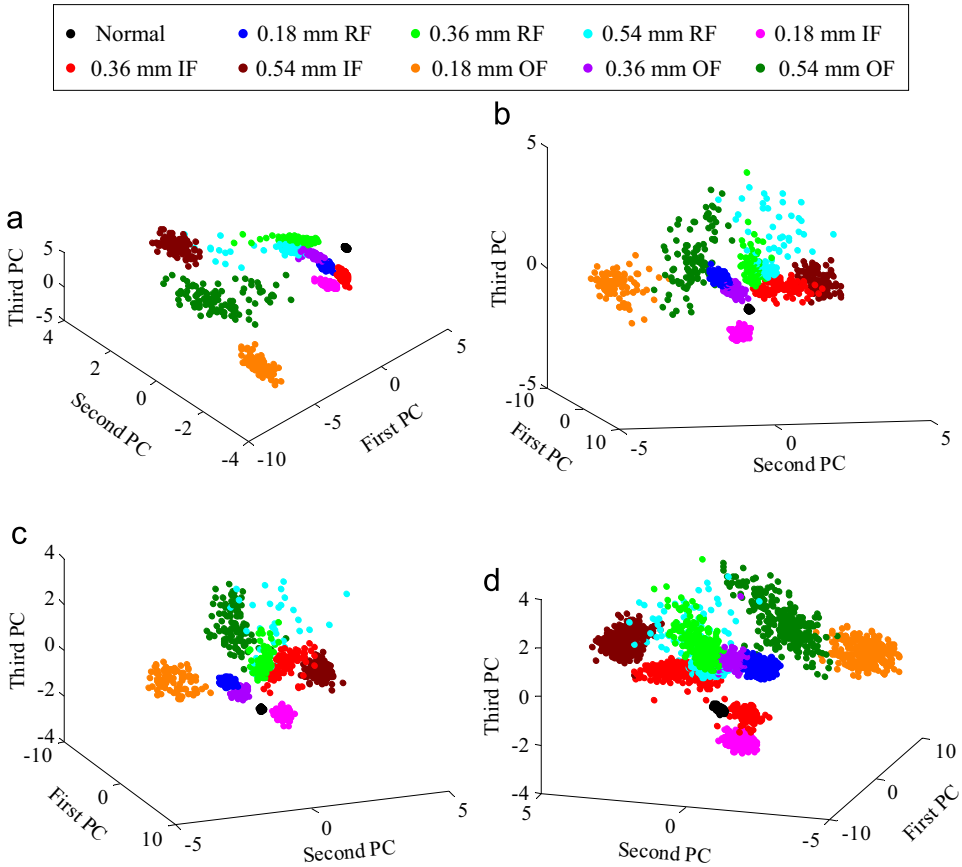

Fig. 7. Scatter plots of principal components for the features: (a) dataset A, (b) dataset B, (c) dataset C and (d) dataset D.

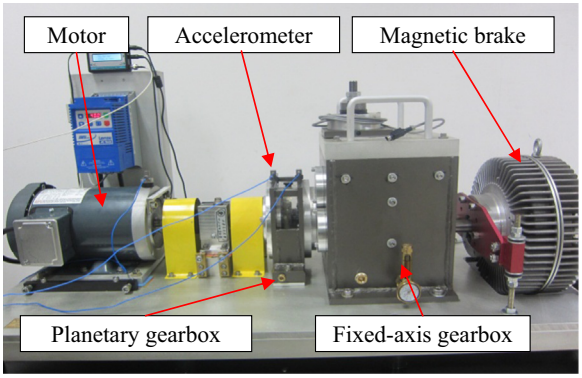

Fig. 8. The test rig of the planetary gearbox.

4.3. Discussion

(1) Different from the ANN-based methods using the signal processing techniques for feature extraction and shallow ANNs for fault classification, this work presents a new idea that utilizes DNNs for both fault characteristic mining and intelligent fault diagnosis. The proposed method is compared with the shallow ANN-based methods using the CWRU bearing dataset. In Ref. [36], a method based on 19 time-domain and frequency-domain features with semi-supervised distance-preserving self-organizing map was proposed. The method was used to differentiate three bearings with ball defects (0.18, 0.36 and 0.54 mm in diameter, respectively) under different loads and 95.8% classification accuracy was obtained. Nevertheless, Dataset D in this study contains 10-class faults under different loads and the classification is more complex, and our method still gets 99.68% accuracy. In Ref. [37], a method based on time-domain features and

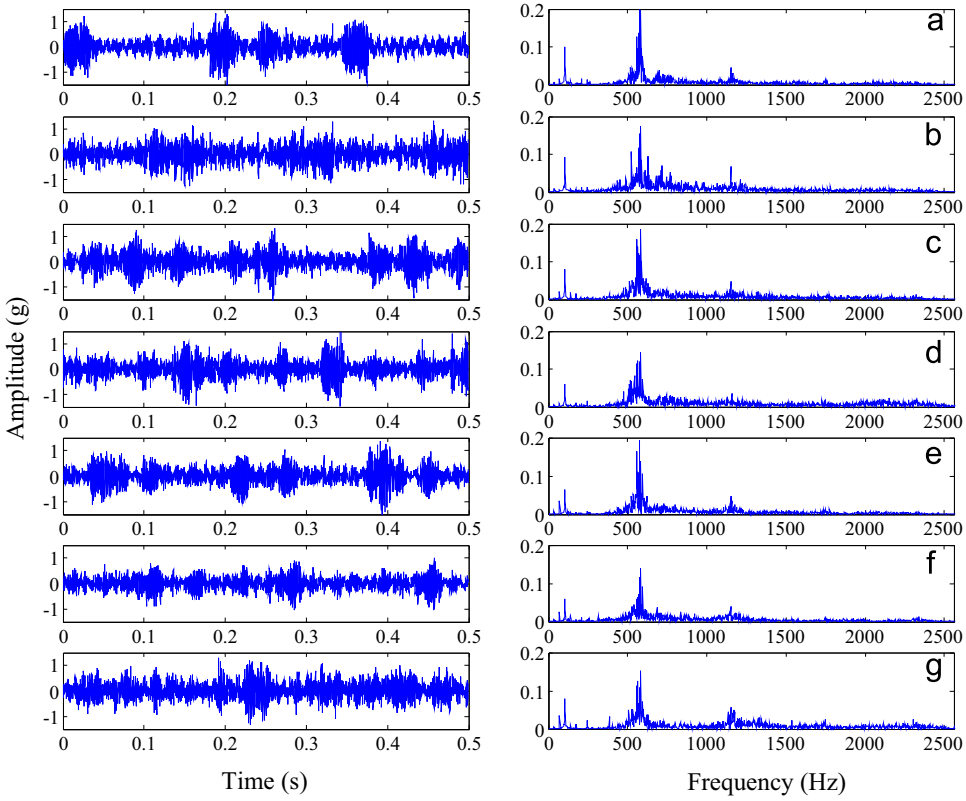

Fig. 9. Typical signals and their spectra for each health condition: (a) normal, (b) a pitted sun gear in the first stage, (c) a cracked sun gear in the first stage, (d) a chipped planetary gear in the first stage, (e) a chipped sun gear in the second stage, (f) a missing tooth sun gear in the second stage and (g) an inner race fault bearing of planetary gear in the first stage.

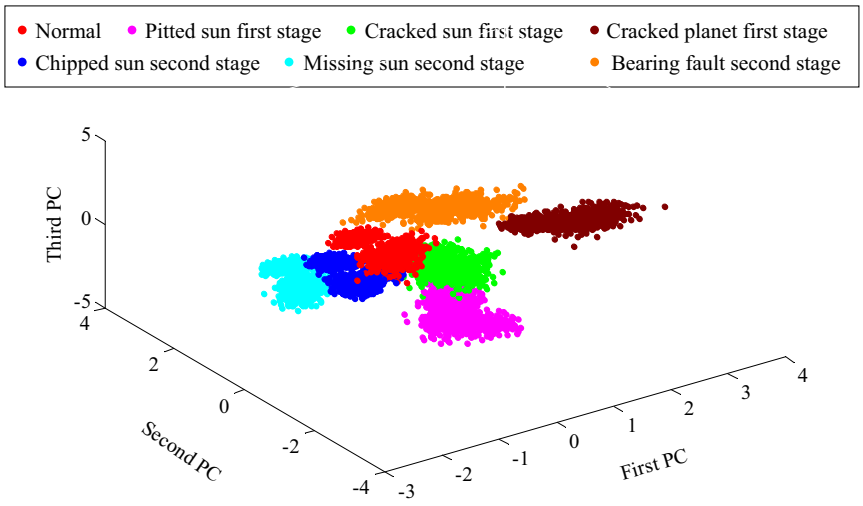

Fig. 10. Scatter plots of principal components for the features of dataset E.

neural networks with a single hidden layer was proposed to diagnose three datasets of bearings. These datasets contain six health conditions under the 1 hp, 2 hp and 3 hp, respectively, which is similar to datasets A, B and C in this work. The method in Ref. [37] obtained 95.7%, 99.6% and 99.4% classification accuracies for each dataset, whereas our classification accuracies are 99.95%, 99.61% and 99.74%. It shows that our results are better, even though more bearing conditions are classified in this study. Ref. [38] developed a method using singular spectrum analysis for feature extraction and a single-hidden-layer neural network for classification. The method was utilized to diagnose four bearing health conditions (normal condition, ball fault, inner race fault and outer race fault) under various loads and the diagnosis

accuracy of 100% was obtained. The diagnosis of dataset D is the similar problem and our diagnosis accuracy is 99.68%. It is a little lower than the diagnosis accuracy of the method in Ref. [38], but their method only classified four bearing health conditions, which is much simpler than our diagnosis issue. Thus the difference of 0.32% between the two results is acceptable. Through the results above, it show that the proposed method classifies obtains fairly high diagnosis accuracies compared with the shallow ANN-based method in Refs. [36–38].

(2) In the existing ANN-based methods, the features extracted by diagnosticians are used for characterizing the health conditions and the ANNs are used for classification. As we know, choosing discriminative features is a crucial step. So diagnosticians need to spend lots of time in analyzing collected signals and grasping their properties. Then according to the diagnosis issues, they design the feature extraction algorithms to calculate discriminative features from measured signals (like vibration signals and frequency spectra). These features manually concentrate the information important for classification in the measured signals and discard the variations irrelevant. Such processes take advantage of human ingenuity but largely depend on much prior knowledge about signal processing techniques and diagnostic expertise, which is time-consuming and labor-intensive. In contrast, the proposed method aims to use a deep neural network for both fault feature mining and intelligent fault diagnosis. Through the scatter plots of principal components for the mined features in Figs. 7 and 10, it is shown that the DNN is able to automatically mine the information important for classification from the frequency spectra according to the diagnosis issues. Since mining the features automatically, the proposed method reduces the need of human labor or the prior knowledge about signal processing techniques and diagnostic expertise. So based on the proposed method, new applications can be achieved easily. Therefore, the main advantage of the proposed method is that the features are mined from the frequency spectra using a general-purpose learning procedure instead of being extracted by diagnosticians. However, the proposed method suffers from a disadvantage: The DNN needs more training time than the shallow ANNs due to the deep architecture. But with the development of hardware technology, we can build DNNs more rapidly in future.

(3) The results of the two diagnosis cases indicate that the proposed method is able to effectively mine fault characteristics and classify mechanical health conditions. The main reason lies in utilizing the autoencoders for pre-training DNNs in the proposed method. Generally, an autoencoder could be regarded as a non-linear function that encodes the input data to a code vector. And with the code vector of the previous trained autoencoder as input for training the next autoencoder, the pre-trained DNNs actually presents certain characteristics of the input data. Therefore, the pre-training process facilitates the DNNs to recognize the characteristics of the rotating machinery signals and helps the DNNs to discover the discriminative information of these signals in the fine-tuning process, resulting in the high diagnosis accuracy.

(4) In this paper, we have not studied the architecture selection of the DNNs. When selecting the architecture, we follow a simple idea: The unit number of the next layer is smaller than that of the previous layer so that the encoding process of the DNN can be viewed as a data compression process or a feature extraction process. The architecture selection is still an open problem for neural networks and we will focus on this problem in future work.

5. Conclusions

This paper presents a DNN-based intelligent method for diagnosing the faults of rotating machinery. The effectiveness of the proposed method is verified using five datasets from rolling element bearings and planetary gearboxes. These datasets contain massive samples involving different health conditions under various operating conditions. Through the diagnosis results of these datasets, it is shown that the proposed method is able to mine fault characteristics from the frequency spectra adaptively for various diagnosis issues and effectively classify the health conditions of the machinery. Since mining fault characteristics automatically, the proposed method is less dependent on human labor or prior knowledge about signal processing techniques and diagnostic expertise than the current ANN-based methods. So based on the proposed method, new applications can be achieved easily. There may be some guidance for the utilization of the proposed method. First, hyperbolic tangent function is advised as the active function of DNNs. Then half coefficients of the frequency spectra should be used because the coefficients are symmetric in the spectra. Finally, a deeper network could be tried for your applications although the five-layer DNN has performed well in this paper.

In the proposed method, DNNs are trained by frequency spectra. Therefore, it only works for rotating machinery and reciprocating machinery whose measured vibration signals are periodic. It is interesting to directly train the DNNs using raw signals in the time domain so as to apply them to other machines. The authors would investigate this topic in future study.

Acknowledgments

This research is supported by National Natural Science Foundation of China (51222503 and 51475355), Provincial Natural Science Foundation Research Project of Shaanxi (2013JQ7011) and Fundamental Research Funds for the Central Universities (2012jdgz01 and CXTD2014001).

References

[1] C.Q. Shen, D. Wang, F.R. Kong, P.W. Tse, Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier, Measurement 46 (2013) 1551–1564.
[2] D. Wang, W.T. Peter, Prognostics of slurry pumps based on a moving-average wear degradation index and a general sequential Monte Carlo method, Mech. Syst. Signal Process. 56 (2015) 213–229.
[3] Y.G. Lei, Z.J. He, Y.Y. Zi, Q. Hu, Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs, Mech. Syst. Signal Process. 21 (2007) 2280–2294.
[4] D. Wang, W.T. Peter, W. Guo, Q. Miao, Support vector data description for fusion of multiple health indicators for enhancing gearbox fault diagnosis and prognosis, Measur. Sci. Technol. 22 (2011) 025102.
[5] A. Widodo, B.-S. Yang, Support vector machine in machine condition monitoring and fault diagnosis, Mech. Syst. Signal Process. 21 (2007) 2560–2574.
[6] J. Lee, F.J. Wu, W.Y. Zhao, M. Ghaffari, L.X. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems—reviews, methodology and applications, Mech. Syst. Signal Process. 42 (2014) 314–334.
[7] K. Worden, W.J. Staszewski, J.J. Hensman, Natural computing for mechanical systems research: a tutorial overview, Mech. Syst. Signal Process. 25 (2011) 4–111.
[8] B. Samanta, Artificial neural networks and genetic algorithms for gear fault detection, Mech. Syst. Signal Process. 18 (2004) 1273–1282.
[9] B. Samanta, C. Nataraj, Use of particle swarm optimization for machinery fault detection, Eng. Appl. Artif. Intell. 22 (2009) 308–316.
[10] V.T. Tran, B.-S. Yang, M.-S. Oh, A.C.C. Tan, Fault diagnosis of induction motor based on decision trees and adaptive neuro-fuzzy inference, Expert Syst. Appl. 36 (2009) 1840–1849.
[11] V.T. Tran, B.-S. Yang, F.S. Gu, A. Ball, Thermal image enhancement using bi-dimensional empirical mode decomposition in combination with relevance vector machine for rotating machinery fault diagnosis, Mech. Syst. Signal Process. 38 (2013) 601–614.
[12] Y.G. Lei, J. Lin, Z.J. He, D.T. Kong, A method based on multi-sensor data fusion for fault detection of planetary gearboxes, Sensors 12 (2012) 2005–2017.
[13] A. Widodo, E.Y. Kim, J.-D. Son, B.-S. Yang, A.C. Tan, D.-S. Gu, B.-K. Choi, J. Mathew, Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine, Expert Syst. Appl. 36 (2009) 7252–7261.
[14] W.X. Lai, P.W. Tse, G.C. Zhang, T.L. Shi, Classification of gear faults using cumulants and the radial basis function network, Mech. Syst. Signal Process. 18 (2004) 381–389.
[15] G.F. Bin, J.J. Gao, X.J. Li, B.S. Dhillon, Early fault diagnosis of rotating machinery based on wavelet packets-Empirical mode decomposition feature extraction and neural network, Mech. Syst. Signal Process. 27 (2012) 696–711.
[16] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[17] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inf. Process. 3 (2014) e2.
[18] J. Schmidhuber, Deep learning in neural networks: an overview, arXiv preprint arXiv 1404 (2014) 7828.
[19] G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Trans. Audio, Speech Lang. Process. 20 (2012) 30–42.
[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z.H. Huang, A. Karpathy, A. Khosla, M. Bernstein, ImageNet large scale visual recognition challenge, arXiv preprint arXiv 1409 (2014) 0575.
[21] P. Baldi, P. Sadowski, D. Whiteson, Searching for exotic particles in high-energy physics with deep learning, Nat. Commun. 5 (2014) 1–9.
[22] V.T. Tran, F. AlThobiani, A. Ball, An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks, Expert Syst. Appl. 41 (2014) 4113–4122.
[23] R.M. Yan, L. Shao, Image blur classification and parameter identification using two-stage deep belief networks, Br. Mach. Vis. Conf. (BMVC) (2013). Bristol, UK.
[24] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.
[25] G.E. Hinton, R.S. Zemel, Autoencoders, minimum description length, and Helmholtz free energy, Adv. Neural Inf. Process. Syst. (1994). 3-3.
[26] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11 (2010) 625–660.
[27] G. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[28] R.B. Palm, Prediction as a candidate for learning deep hierarchical models of data, Technical University of Denmark, Palm, 2012.
[29] Y.S. Wang, G.Q. Shen, Y.F. Xing, A sound quality model for objective synthesis evaluation of vehicle interior noise based on artificial neural network, Mech. Syst. Signal Process. 45 (2014) 255–266.
[30] S.S. Ng, P.W. Tse, K.L. Tsui, A. One-Versus-All Class, Binarization strategy for bearing diagnostics of concurrent defects, Sensors 14 (2014) 1295–1321.
[31] R.B. Randall, J. Antoni, Rolling element bearing diagnostics—a tutorial, Mech. Syst. Signal Process. 25 (2011) 485–520.
[32] Y.G. Lei, J. Lin, M.J. Zuo, Z.J. He, Condition monitoring and fault diagnosis of planetary gearboxes: a review, Measurement 48 (2014) 292–305.
[33] D. Wang, C.Q. Shen, P.W. Tse, A novel adaptive wavelet stripping algorithm for extracting the transients caused by bearing localized faults, J. Sound Vibr. 332 (2013) 6871–6890.
[34] D. Wang, Q. Miao, R. Kang, Robust health evaluation of gearbox subject to tooth failure with wavelet decomposition, J. Sound Vibr. 324 (2009) 1141–1157.
[35] X.S. Lou, K.A. Loparo, Bearing fault diagnosis based on wavelet transform and fuzzy inference, Mech. Syst. Signal Process. 18 (2004) 1077–1095.
[36] W.H. Li, S.H. Zhang, G.L. He, Semisupervised distance-preserving self-organizing map for machine-defect detection and classification, IEEE Trans. Instrum. Measur. 62 (2013) 869–879.
[37] L.F. de Almeida, J.W. Bizarria, F.C. Bizarria, M.H. Mathias, Condition-based monitoring system for rolling element bearing using a generic multi-layer perceptron, J. Vibr. Control (2014).
[38] B. Muruganatham, M. Sanjith, B. Krishnakumar, S. Satya Murty, Roller element bearing fault diagnosis using singular spectrum analysis, Mech. Syst. Signal Process. 35 (2013) 150–166.