2964. An intelligent fault diagnosis method of rotating machinery using L1-regularized sparse filtering

Weiwei Qian[1], **Shunming Li**[2], **Jinrui Wang**[3], **Zenghui An**[4], **Xingxing Jiang**[5]
[1, 2, 3, 4]College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, P. R. China
[5]School of Urban Rail Transportation, Soochow University, Suzhou, 215137, P. R. China
[1]Corresponding author
E-mail: [1]*qianweiwei33@163.com*, [2]*smli@nuaa.edu.cn*, [3]*wangjr33@163.com*, [4]*me_anzh@163.com*, [5]*jiangxx@suda.edu.cn*

Received 1 November 2017; received in revised form 14 May 2018; accepted 24 May 2018
DOI https://doi.org/10.21595/jve.2018.19339

Abstract. Traditional intelligent fault diagnosis methods take advantage of diagnostic expertise but are labor-intensive and time-consuming. Among various unsupervised feature extraction methods, sparse filtering computes fast and has less hyperparameters. However, the standard sparse filtering has poor generalization ability and the extracted features are not so discriminative by only constraining the sparsity of the feature matrix. Therefore, an improved sparse filtering with L1 regularization (L1SF) is proposed to improve the generalization ability by improving the sparsity of the weight matrix, which can extract more discriminative features. Based on Fourier transformation (FFT), L1SF, softmax regression, a new three-stage intelligent fault diagnosis method of rotating machinery is developed. It first transforms time-domain samples into frequency-domain samples by FFT, then extracts features in L1-regularized sparse filtering and finally identifies the health condition in softmax regression. Meanwhile, we propose employing different activation functions in the optimization of L1SF and feedforward for considering their different requirements of the non-saturating and anti-noise properties. Furthermore, the effectiveness of the proposed method is verified by a bearing dataset and a gearbox dataset respectively. Through comparisons with the standard sparse filtering and L2-regularized sparse filtering, the superiority of the proposed method is verified. Finally, an interpretation of the weight matrix is given and two useful sparse properties of weight matrix are defined, which explain the effectiveness of L1SF.

Keywords: fault diagnosis, automatic feature extraction, sparse filtering, L1 regularization, activation function.

1. Introduction

With the rapid development of technology, industrial Internet of Things (IoT) and data driven techniques [1] have been revolutionizing manufacturing. As machines are becoming more efficient and automatic than ever before, it is more significant to have a safe and sound operation [2]. Meanwhile, rotating speed oscillation usually occurs during machine operation which can accelerate the damage on machines [3]. Therefore, condition monitoring systems are adopted to collect the massive amount of data from monitored machines [4, 5]. At the same time, with substantial development of computing systems and sensors, machinery fault diagnosis has fully embraced the extensive revolution in modern manufacturing.

Traditionally, there are mainly three sequential steps in the framework of fault diagnosis systems: 1) signal acquisition; 2) feature extraction and selection and 3) fault classification [2, 6, 7]. In the first step, vibration signals are widely used as the data source since essential information about machine health condition is embedded in them. However, feature extraction is always needed because the raw signals contain various kinds of noise. In the second step, representative features are extracted for fault classification. The more discriminative features are extracted from vibration signals, the more accurate fault diagnosis result would be. Signal processing techniques have been proposed and widely used in feature extraction and selection such as time-domain analysis [8], frequency transform [9], high resolution time-frequency analysis [10], wavelet transform [11], and envelope demodulation algorithms [12]. Besides, as entropy is effective in detection the dynamic characteristics of time series, some methods use entropy in feature extraction and selection. Modified multi-scale symbolic dynamic entropy (MMSDE) [13] and modified hierarchical permutation entropy (MHPE) [14] were proposed to extract features for gearbox fault diagnosis. Zhao et al. [15] combined the EEMD and multi-scale fuzzy entropy to extract more discriminative features. However, the further analysis of features extracted by these algorithms requires manpower and prior knowledge. At the same time, although these techniques can effectively extract primary features from signals, these features may prevail among all signals of various health conditions. Therefore, dimension-reduction strategies such as nonlinear PCA [16], independent component analysis (ICA) [17] are required to select sensitive features. It means that these extracted features should be sensitive to different health conditions but be insensitive to different operating conditions such as large speed oscillation. Intelligent feature learning methods can also be utilized in feature extraction along with feature selection. Among various intelligent feature learning methods, unsupervised ones have been successfully applied to extract useful features in many image, video and audio tasks.

Well developed as unsupervised feature learning algorithms are, there are still two main drawbacks [18]. Firstly, unsupervised learning algorithms generally have plenty of parameters, which are difficult to tune and crucial to the performance of the network. The comparison of tunable parameters in various unsupervised learning algorithms is shown in Table 1 [18, 19]. It can be seen that there are massive parameters in sparse restricted Boltzmann machines (RBMs) [20], sparse autoencoder [21], sparse coding [22]. Secondly, computational complexity is also a major drawback. Independent component analysis (ICA) has just one hyperparameter needing tuning [23] and provides commendable results in object recognition experiments, but it scales poorly to large datasets and high-dimension inputs. Sparse coding has an expensive inference process, which requires a costly optimizing iteration. Stacked autoencoder (SAE) [24] has to do a greedy layer-wise training process which includes decoding and encoding networks, and the parameters of decoding network are useless in fine tuning which will increase the computational complexity. In the final step, the learned features are utilized to train the classifiers such as support vector machine (SVM) and softmax regression [25], which are always used to identify the health condition. In Ref. [26], the improved Particle Swarm Optimization (PSO) algorithm was proposed to optimize the parameters of least squares support vector machines (LS-SVM) in order to construct an optimal classifier.

Table 1. Tunable parameters in various algorithms

Algorithm	Tunable parameters
Sparse filtering	Feature dimension
ICA	Feature dimension
Sparse coding	Feature dimension, mini-batch size, sparse penalty
Sparse autoencoders	Feature dimension, target activation, sparse penalty, weight decay
Sparse RBMs	Features, target activation, sparse penalty, weight decay, momentum learning rate

As sparsity is one of the desirable properties of a good feature representation [27], Ngiam et al. [19] proposed an unsupervised algorithm called sparse filtering to solve the problems pointed out above, which is a two-layer neural network. Only the dimension of features learned from sparse filtering needs tuning and its main idea is the sparsity of the feature matrix [18]. Meanwhile, it was validated on image recognition and phone classification tasks, which obtained the state-of-the-art performance. For its simplicity and good performance, Lei et al. [2] introduced sparse filtering into rotating machine fault diagnosis and constructed a two-stage intelligent fault diagnosis method, which got better performance than other networks. Sparse filtering works by imposing sparse constraints on the distribution sparsity of the feature matrix but the overefitting risk of it is quite high, namely having poor generalization ability. Yang et al. [28] utilized L2 norm to regularize the weight matrix of sparse filtering. However, its effectiveness in preventing overfitting is limited and it takes no consideration of the sparsity of the weight matrix. Meanwhile, the activation functions used in weight matrix optimization of sparse filtering [2] are soft-absolute function, which computes fast and can partly avoid the gradient vanishing problem. Nevertheless, when it is used in feedforward, its robustness to noise is weak. L1 norm can also be utilized to reduce the overfitting risk. However, when introducing L1 regularization into the original network, the optimization of the new network is difficult sometimes. Recently, Deng et al. [29, 30] introduced new algorithms into solving complex optimization problems, which can be beneficial to the network with many terms to be optimized. With the help of them, L1 norm can be utilized better.

Inspired by Ref. [28-30], aiming at reducing the overfitting risk, an improved sparse filtering network with L1 regularization is proposed, which can improve the performance of the network by exploring the sparsity of the weight matrix. Meanwhile, for considering the different requirements of non-saturating and anti-noise properties, we propose using different activation functions in feature extraction network training and feedforward, which is inspired by the training process of deep learning that different activation functions can be adopted in greedy layer-wise pre-training and further fine tuning.

The main contributions of this paper are described as follows:

(1) A three-stage intelligent fault diagnosis method of rotating machinery is developed. In the first stage, time-domain samples are automatically pre-processed into frequency-domain samples by FFT. In the second stage, an improved feature extraction network called L1-regularized sparse filtering is proposed to extract and select discriminative features. In the last stage, features are fed into softmax regression to identify the health condition. Experiments conducted on a bearing dataset and a gearbox dataset verify the effectiveness of the proposed method.

(2) L1-regularized sparse filtering is proposed to constrain the weight matrix sparsity and improve the generalization ability of standard sparse filtering. We also construct the L2-regularized sparse filtering and do experiments on it. The comparison with standard sparse filtering demonstrates its effectiveness. The comparison with L1-regularized sparse filtering shows it is inferior when using frequency-domain samples.

(3) An interpretation of the weight matrix learned by sparse filtering is given, which is the supplement to the interpretation given in Ref. [2]. Based on it, two sparse properties of the weight matrix are defined, which are validated significant in obtaining discriminative features. They also explain why L1-regularized sparse filtering performs better when frequency-domain samples are adopted.

This paper is organized as follows. Section 2 describes sparse filtering and L1 regularization briefly. Section 3 details the proposed L1-regularized sparse filtering and the using of the new activation function. In Section 4, diagnosis cases of a bearing dataset and a gearbox dataset are studied separately using the proposed method and other models. Section 5 discusses the properties of the weight matrix learned by the improved sparse filtering. Finally, the conclusion is drawn in Section 6.

2. Theoretical background

2.1. Sparse filtering

Generally, unsupervised feature learning algorithms can be divided into two main categories [18]: explicitly modeling the input distribution or not. ICA, sparse coding, SAE and sparse RBMs explicitly model the input distribution and reconstruct mechanical dynamic signal structure by minimizing the reconstruction error. Although it’s desirable to learn a good approximation of the input distribution, algorithms such as sparse filtering show that this is not so important if the aim is to get discriminative features. The fundamental idea of sparse filtering focuses on constraining the sparsity of output features instead of explicitly modeling the input distribution. Sparse filtering tries to learn sparse feature distribution that satisfies three principles:

(1) Population sparsity, which refers that all the features of one sample should be activated sparsely at a time;

(2) Lifetime sparsity, which refers that the same kind of features belonging to all samples should be activated sparsely at a time;

(3) High dispersal, which refers that all kinds of features should have similar statistics.

The architecture of the sparse filtering network is shown in Fig. 1. Firstly, the collected raw vibration data is alternately separated into K samples. Then they are transformed from original time-domain samples into Fourier spectra $\{\mathbf{x}^i\}_{i=1}^{K}$ by fast Fourier transformation (FFT), where $\mathbf{x}^i \in R^{N\times 1}$ is ith sample containing N Fourier coefficients, which are the amplitudes of the spectral lines. Next, the training dataset is used to train sparse filtering to get an optimized weight matrix $\mathrm{W} \in R^{N\times N^{out}}$, where L is the dimension of output feature vectors. Finally, the frequency-domain sample $\mathbf{x}^i \in R^{N\times 1}$ is mapped into a feature vector $\mathbf{f}^i \in R^{N^{out}\times 1}$ in feedforward. The mapping can be realized through $\mathbf{W}$ and activation function 2.

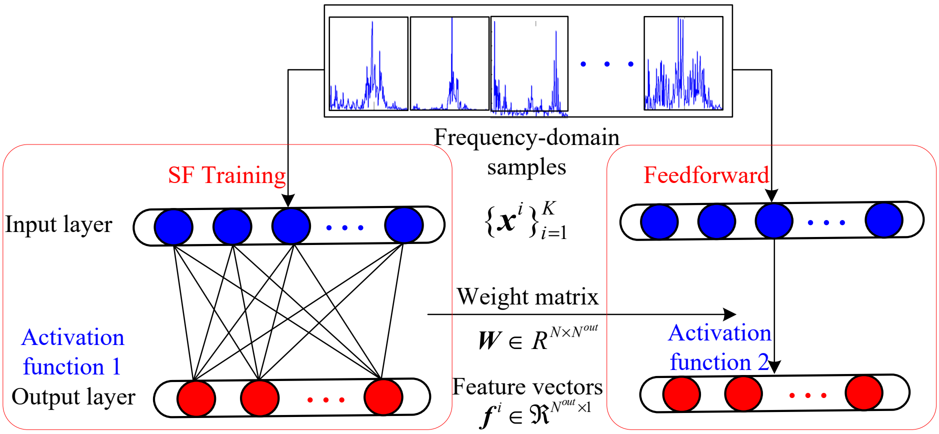


Fig. 1. Architecture of the sparse filtering network

In the optimization, the standard sparse filtering uses a soft-absolute function as activation function 1 [19], so the features can be computed as follows:

$$f_j^i = \sqrt{\varepsilon + \left(\mathbf{W}_j^T\mathbf{x}^i\right)^2} \approx \left|\mathbf{W}_j^T\mathbf{x}^i\right|, \tag{1}$$

where f_j^i refers to the jth feature of the ith sample and ε is set to 10^{-8}. The feature matrix learned from training dataset is $\mathbf{f} \in R^{N^{out}\times K}$, where K is the number of training samples and N^{out} is the dimension of output feature vectors. Three steps should be applied to realize the sparse principles mentioned above, which are shown as follows.

Firstly, by means of dividing each kind of feature using its L2 norm [2] across all samples, each kind of feature $\mathbf{f}_j$ is normalized to be equally active:

$$\tilde{\mathbf{f}}_j = \frac{\mathbf{f}_j}{\left\|\mathbf{f}_j\right\|_2}, \quad (j = 1,2,\dots,N^{out}). \tag{2}$$

Then, the new feature matrix is normalized by columns, so that features of each sample $\tilde{\mathbf{f}}^i$ lie on the unit L2 ball:

$$\hat{\mathbf{f}}^i = \frac{\tilde{\mathbf{f}}^i}{\left\|\tilde{\mathbf{f}}^i\right\|_2}, \quad (i = 1,2,\dots,K). \tag{3}$$

At last, the normalized features are optimized for sparsity by L1 norm. The objective function of sparse filtering can be described as follows:

$$\underset{\mathbf{W}}{\text{minimize}} \sum_{i=1}^{K} \left\|\hat{\mathbf{f}}^i\right\|_1 = \sum_{i=1}^{K} \left\| \frac{\tilde{\mathbf{f}}^i}{\left\|\tilde{\mathbf{f}}^i\right\|_2} \right\|_1 . \tag{4}$$

In this paper, we take a novel view on sparse filtering: considering it as a greedy layer-wise training process [31], which means that the activation function used in fine-tuning process could be different from the one used in weight matrix initialization and that is analogous to the training process in SAE. Therefore, to get a more noise-robust feature representation, we adopt Eq. (5) as activation function 2:

$$\mathbf{f}_j^i = \log\left(1 + (\mathbf{W}_j^T \mathbf{x}^i)^2\right), \tag{5}$$

instead of the widely used soft-absolute function, as shown in Fig. 1.

2.2. L1 and L2 regularization

Overfitting can be simply described as an undesirable phenomenon that the training accuracy is higher than testing accuracy. The network with higher overfitting risk means it has lower generalization ability. It always occurs when the network model is too complex or when the training dataset is too small. However, it is desirable to use a smaller dataset in training and get a higher diagnosis accuracy in testing. To reduce the overfitting risk of the sparse filtering, we use strategies like L2 and L1 regularization. L2 norm penalty, which can constrain the magnitudes of parameters is usually used in overfitting problem as a regularization approach [32] and is generally called L2 regularization. It works by penalizing the bigger parameters in the weight matrix to make them smaller. In addition, it should be noticed that L2 regularization cannot make weight matrix sparse theoretically [33].

Different from L2 norm, L1 norm can impose a sparse constraint on the parameters. Therefore, L1 norm is used to realize the sparsity of the feature matrix in sparse filtering, as shown in the subsection above. However, it is noticed that the sparsity of the weight matrix is not constrained. As weight matrix is closely linked with the output features, heuristically we propose utilizing L1 norm to directly constrain the sparsity of the weight matrix and then further constrain the sparsity of the feature matrix. It is called L1 regularization and can make the sparse filtering perform better. In this paper, we will also compare the performance of L1 regularization and L2 regularization. Specifically, L1 regularization and L2 regularization are added to the objective function of the standard sparse filtering in Eq. (4) respectively, and the new objective functions are defined separately in Eq. (6) and Eq. (7) as follows:

$$\underset{\mathbf{W}}{\text{minimize}} \sum_{i=1}^{K} \left\|\hat{\mathbf{f}}^i\right\|_1 + \lambda_1 \sum_{i=1}^{N} \sum_{j=1}^{N^{out}} \left|w_{ij}\right| = \sum_{i=1}^{K} \left\| \frac{\tilde{\mathbf{f}}^i}{\left\|\tilde{\mathbf{f}}^i\right\|_2} \right\|_1 + \lambda_1 \sum_{i=1}^{N} \sum_{j=1}^{N^{out}} \left|w_{ij}\right|, \tag{6}$$

$$\underset{\mathbf{W}}{\text{minimize}} \sum_{i=1}^{K} \left\|\hat{\mathbf{f}}^i\right\|_1 + \lambda_2 \sum_{i=1}^{N} \sum_{j=1}^{N^{out}} w_{ij}^2 = \sum_{i=1}^{K} \left\| \frac{\tilde{\mathbf{f}}^i}{\left\|\tilde{\mathbf{f}}^i\right\|_2} \right\|_1 + \lambda_2 \sum_{i=1}^{N} \sum_{j=1}^{N^{out}} w_{ij}^2, \tag{7}$$

where λ_1 and λ_2 are the regularization parameters, λ_1 controls the relative importance of feature sparse distribution and weight matrix sparse distribution; λ_2 controls the relative importance of feature sparse distribution and weight decay; N is the input dimension of samples and N^{out} is dimension of learned feature vector; w_{ij} is the ith row and the jth column element of the weight matrix $\mathbf{W}$. The gradient of the regularization term can be derived correspondingly and combined into the original gradient.

2.3. Softmax regression

The unsupervised feature learning process of sparse filtering is described above and it is always followed by a classifier. Softmax regression [31] is often applied at the last stage for multiclass classification and performs well. Therefore, we implement softmax regression as the classifier in this paper.

Suppose that the output feature dataset $\{\mathbf{f}^i\}_{i=1}^{K}$ has label set $\{y^i\}_{i=1}^{K}$, where $y^i \in \{1,2,\dots,R\}$. For each $\mathbf{f}^i$, the output layer attempts to estimate the probability $p(y^i = r|\mathbf{f}^i)$ for each label of $r = 1, 2, ..., R$. Thus, the output layer is optimized by minimizing the objective function:

$$J(\mathbf{W}_2) = -\frac{1}{M}\left[\sum_{m=1}^{K}\sum_{r=1}^{R} 1\{y_m = r\}\log\frac{e^{\mathbf{W}_2^r\mathbf{f}^m}}{\sum_{l=1}^{R} e^{\mathbf{W}_2^r\mathbf{f}^m}}\right] + \lambda_3 \sum_{k=1}^{R}\sum_{j=1}^{N^{out}} \left(\mathbf{W}_2^{k,j}\right)^2, \tag{8}$$

where $1\{\cdot\}$ represents the indicator function, which outputs 1 if the condition is true, and 0 otherwise; $\mathbf{W}_2$ is the weight matrix of the output layer and $\mathbf{W}_2^r$ is the rth row of $\mathbf{W}_2$. The second term is the weight decay term, and $\mathbf{W}_2^{k,j}$ is the kth row and the jth column element of $\mathbf{W}_2$; λ_3 is the regular parameter.

3. Proposed framework

This section details how the proposed method automatically extracts features from raw data and then briefly describes how the diagnosis results are obtained. The framework of our method is illustrated in Fig. 2 and the process can be described as follows.

(1) Data sampling and preprocessing. The time-domain raw signals are sampled into dataset $\{\mathbf{z}^i\}_{i=1}^{M}$, where $\mathbf{z}^i \in R^{2N\times 1}$ is the ith sample containing $2N$ data points. Then each sample $\mathbf{z}^i$ is transformed into frequency-domain sample $\mathbf{t}^i$ by FFT, where $\mathbf{t}^i \in R^{N\times 1}$ is the ith sample containing N Fourier coefficients. These samples are combined into dataset $\{\mathbf{t}^i\}_{i=1}^{M}$.

(2) Training dataset selection and whitening. K samples are randomly selected from $\{\mathbf{t}^i\}_{i=1}^{M}$ composing the training dataset $\{\mathbf{x}^i\}_{i=1}^{K}$ and the label set of the training samples is $\{l^i\}_{i=1}^{K}$, where l^i is the label of sample $\mathbf{x}^i$. The rest samples compose the testing dataset $\{\mathbf{s}^i\}_{i=1}^{M-K}$. A whitening approach [2] is utilized to preprocess the training samples and obtain the whitened dataset $\mathbf{S}_w \in R^{N\times K}$, where K is the number of samples and N is the dimension of samples.

(3) Sparse filtering training. Firstly, the whitened dataset $\mathbf{S}_w$ is fed into the L1-regularized sparse filtering model. Then, the weight matrix $\mathbf{W_1} \in R^{N\times N^{out}}$ is obtained by minimizing the new objective function as Eq. (6), where N^{out} is the output dimension. It should be noted that Eq. (1) is used as activation function 1.

(4) Output features calculation. A feedforward process is carried out on the training dataset $\{\mathbf{x}^i\}_{i=1}^{K}$ with the optimized weight matrix $\mathbf{W_1}$ and activation function 2. The new function described in Eq. (5) is adopted as activation function 2. Each sample $\mathbf{x}^i$ is mapped into feature vector $\mathbf{f}^i \in R^{N^{out}\times 1}$, so the feature matrix of training samples is $\{\mathbf{f}^i\}_{i=1}^{K}$.

(5) Softmax regression training. The feature matrix $\{\mathbf{f}^i\}_{i=1}^{K}$ together with the labels of the training samples $\{l^i\}_{i=1}^{K}$ is fed into the softmax regression to train the network. The trained softmax regression will be used to identify the health conditions of feature vectors.

(6) Diagnose the testing samples. The trained network is employed to identify health conditions of testing samples$\{\boldsymbol{s}^i\}_{i=1}^{M-K}$, as shown in Fig. 2.

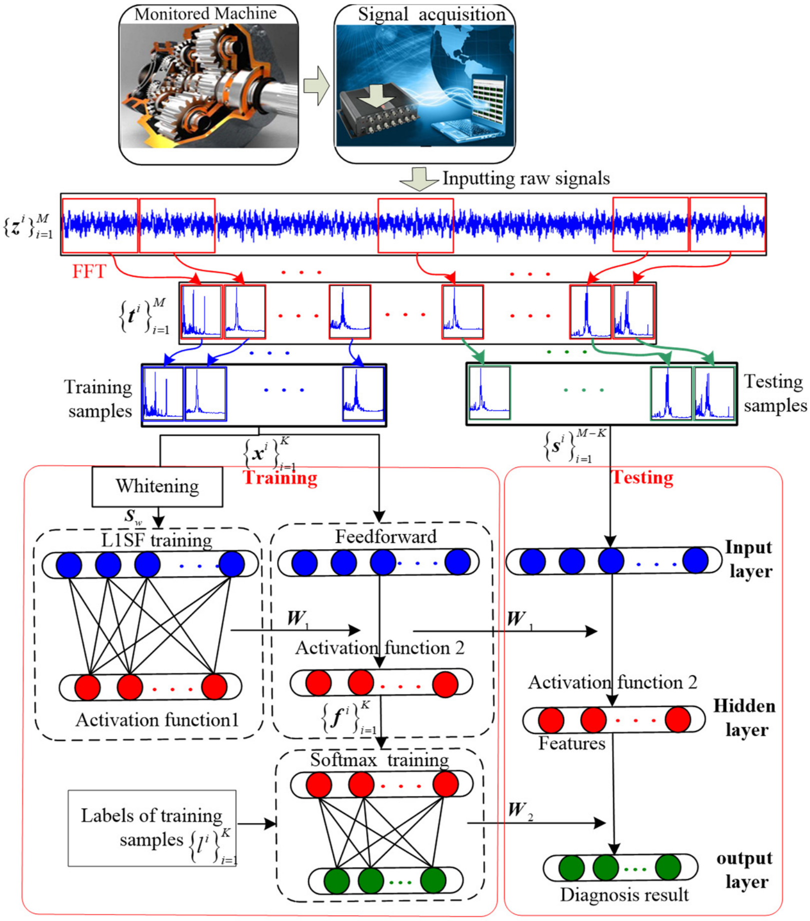


Fig. 2. Flowchart of the proposed method

4. Fault diagnosis using the proposed method

In this section, a bearing dataset and a gearbox dataset are employed to validate the effectiveness of our method. In order to further verify the superiority of our method, six different networks are constructed, as shown in Table 2, where SF denotes sparse filtering; L1SF denotes L-regularized sparse filtering; L2SF denotes L2-regularized sparse filtering. Meanwhile, 20 trials are carried out for each experiment to reduce the effect of randomness. The parameters are all tuned to the best, and the regular parameter of softmax regression is tuned to 1E-5 [2]. The computation platform is a PC with an Intel I5 CPU and 8G RAM.

Table 2. Description of networks

Method	Main model	Activation function 1	Activation function 2
SF-ABS	SF	Eq. (1)	Eq. (1)
SF-LOG	SF	Eq. (1)	Eq. (5)
L1SF-LOG	L1SF	Eq. (1)	Eq. (5)
L2SF-LOG	L2SF	Eq. (1)	Eq. (5)
L1SF-ABS	L1SF	Eq. (1)	Eq. (1)
L2SF-ABS	L2SF	Eq. (1)	Eq. (1)

4.1. Case 1. Fault diagnosis of rolling element bearings

4.1.1. Data description

The motor bearing experimental dataset provided by Case Western Reserve University [34] is employed in this section. The vibration signals were collected from the drive end of the motor through the tri-axial accelerator. The dataset contains four different health conditions: 1) normal condition (NC); 2) outer race fault (OF); 3) inner race fault (IF) and 4) roller fault (RF). Vibration signals of three different damage severity levels (0.07, 0.14 and 0.21 inch) were separately collected for health condition OF, IF and RF. In addition, the signals were all collected under three different loads (1, 2 and 3 hp) and the sampling frequency was 48 kHz. The total dataset is composed of ten health conditions under three loads, and we treat the same health condition under different loads as the same class. 300 samples are obtained from each health condition under one load, so the dataset is composed of 9000 samples and each sample includes 1600 time-domain data points.

4.1.2. Parameter selection and diagnosis result

In this subsection, the dataset will be processed by the proposed method. Firstly, FFT is implemented on each time-domain sample and 1600 Fourier coefficients are obtained. As the coefficients are symmetric, we utilize the first 800 coefficients in each sample, so each sample contains 800 Fourier coefficients. As there is no obvious periodic property in each frequency-domain sample, the coefficients of each sample are treated as a whole. Therefore, there is no need to segment it in the way mentioned in Ref. [2] and the dataset $\{\mathbf{x}^j\}_{j=1}^{9000}$ is used to train and test the network, where $\mathbf{x}^j \in R^{800\times 1}$.

First, we investigate the selection of the parameters of sparse filtering. We randomly select 10 % of samples to train the proposed method and the rest of the samples are used for testing. The averaged diagnosis results of L1SF-LOG with different output dimensions are shown in Fig. 3. It presents that the testing accuracy is higher than 99.2 % when N^{out} is bigger than 600. It is evident that the testing accuracy goes higher and the corresponding standard deviation drops continuously with the increase of N^{out}, but the averaged training time keeps increasing. It indicates that we should make a tradeoff between the diagnosis accuracy and the time spent while selecting N^{out}. It can be seen that the increase of accuracy is unobvious when N^{out} is bigger than 700, so we choose 700 as N^{out}. The testing accuracy reaches the highest and the standard deviation reaches the lowest when the regular parameter of L1SF-LOG is around 100, so it is set to 100, as displayed in Fig. 4. The output dimension and regular parameter of L2SF-LOG are tuned to 400 and 10 respectively. The output dimensions of SF-LOG and SF-ABS are tuned to 700 and 100 separately.

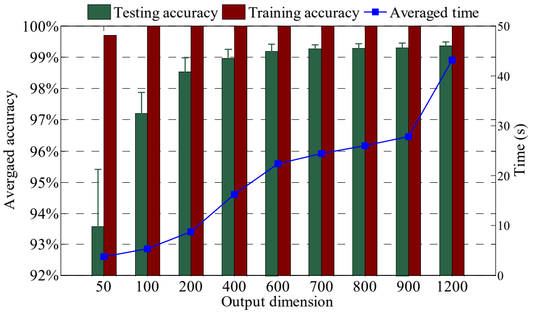


Fig. 3. Output dimension selection

The diagnosis results using various numbers of samples for training are presented in Fig. 5. It presents in Fig. 5(a) that the testing accuracy rises moderately with the increase of training sample number because the testing accuracy is higher than 96 % with just 1 % samples for training. Fig. 5(a) also shows that L1SF-LOG can diagnose ten health conditions with averaged testing accuracy up to 98 % when only 2 % samples are used for training, and the testing accuracy rises to 99 % when the percentage of training samples increases to 5 %. At the same time, it can be observed from Fig. 5 that the test accuracy is up to 96 % when the training samples for L2SF-LOG, SF-LOG, and SF-ABS increase to 10 %, 10 % and 20 % respectively. It indicates that L1SF-LOG can achieve excellent performance even while lacking training samples and L1SF-LOG performs better than L2SF-LOG.

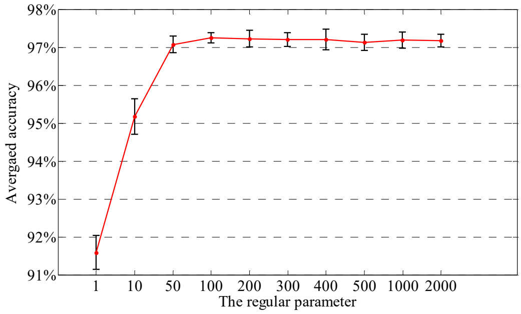
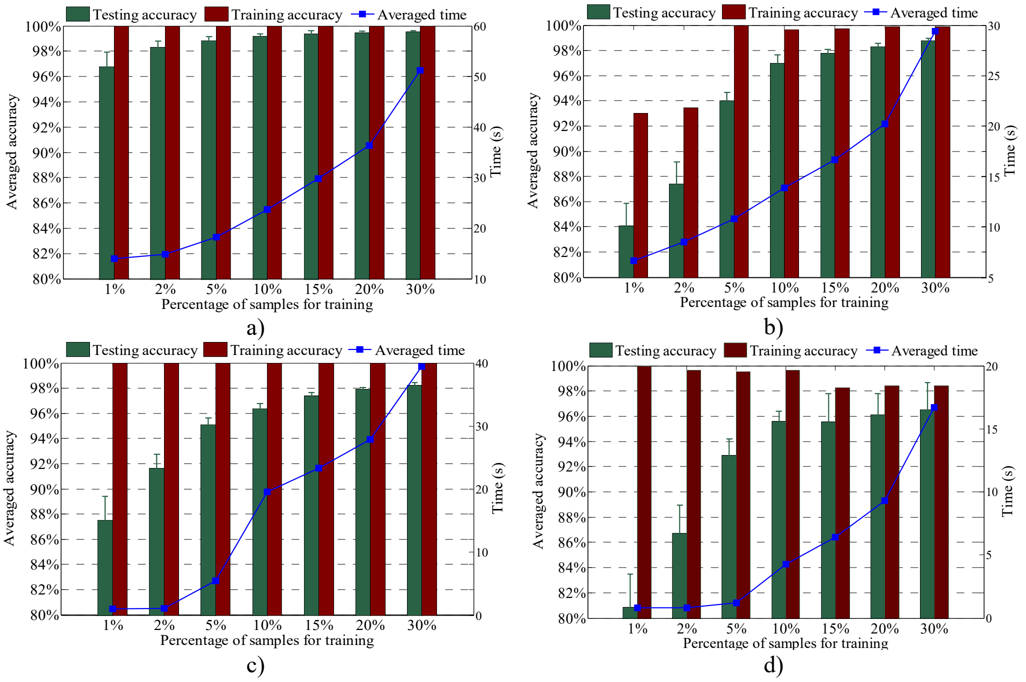


Fig. 4. The regular parameter of L1SF-LOG



a)

b)

c)

d)

Fig. 5. Diagnosis results of: a) L1SF-LOG, b) L2SF-LOG, c) SF-LOG, d) SF-ABS with different percentage of training samples

It also indicates in Fig. 5 that L1SF-LOG can prevent the overfitting dramatically since the gaps between training accuracy and testing accuracy are always smaller than 3.7 %. By comparison, it shows that L1SF-LOG and L2SF-LOG both have higher testing accuracies and smaller gaps between training accuracy and testing accuracy than SF-LOG under the same training sample number. Meanwhile, it also presents that SF-LOG outperforms SF-ABS, which benefits directly from the using of the new activation function. Furthermore, when using same numbers of samples in training, L1SF-LOG always holds the lowest standard deviation, which indicates L1SF-LOG is more stable than the other three networks. In the following experiments, we randomly select 15 % samples from the whole dataset for training.

To further present the efficiency of the proposed method, 20 trials averaged accuracies and their standard deviations of all the six networks are shown in Table 3. It shows that L1SF-LOG holds the highest accuracy and lowest standard deviation, which means L1SF-LOG is more stable than the others. Meanwhile, it is worth mentioning that SF-LOG performs better than SF-ABS. Comparing the results of SF-LOG and SF-ABS, L2SF-LOG and L2SF-ABS, L1SF-LOG and L1SF-ABS, it presents that the networks which adopt the new activation in feedforward can achieve higher testing accuracies and are much more stable.

The 20 trials averaged confusion matrix of L1SF-LOG shows that the health conditions IR7 and OR7 will be misclassified as each other most probably, as displayed in Table 4. It can be assumed that the health conditions misclassified as each other, mainly because signals of them are similar to each other. The normal health condition is classified perfectly mainly because it has no fault and is most distinct from other health conditions.

Table 3. The setting and performance of 6 networks

Method	Output dimension	Training accuracy (%)	Testing accuracy (%)	Training time (s)
SF-ABS	100	98.26±2.94	95.57±2.21	6.41±0.34
SF-LOG	700	100	97.37±0.27	23.30±0.58
L1SF-LOG	700	100	99.39±0.23	29.84±0.39
L2SF-LOG	400	99.73±0.27	97.61±0.27	16.68±0.38
L1SF-ABS	700	99.50±0.68	98.08±0.89	25.30±1.06
L2SF-ABS	700	99.49±0.24	97.23±0.42	23.30±0.74
Note: The format of the result is: mean value ± standard deviation				

Table 4. Confusion matrix of 20 trials averaged diagnosis results

Predict true	NOR	BALL7	BALL14	BALL21	IR7	IR14	IR21	OR7	OR14	OR21
NOR	765	0	0	0	0	0	0	0	0	0
BALL7	0	765	0	0	0	0	0	0	0	0
BALL14	0	0	761.65	0	0	0.15	3.2	0	0	0
BALL21	0	0	0	764.85	0	0	0	0	0	0.15
IR7	0.25	1.85	0.15	0.7	748.9	3.45	0	1.05	6.15	2.5
IR14	0	0	0	0	1.1	760.3	0.3	0	3.25	0.05
IR21	0	0	4.25	0.25	0.1	1.05	759.3	0	0.05	0
OR7	0	0	0	0	1.4	0	0	763.2	0.15	0.25
OR14	0	0	0.7	0.1	13.9	1.3	0.1	1.05	747.8	0.05
OR21	0	0.85	0.5	1.15	0.05	0.4	0	0	0	762.05

4.2. Case2. Fault diagnosis of gearbox

4.2.1. Data description

The vibration signals of a four-speed motorcycle gearbox [35] used in this section were collected from the test rig as shown in Fig. 6. The test rig was also equipped with a tri-axial accelerometer, a data acquisition system, a load mechanism, an electrical motor, a tachometer and four shock absorbers. The vibration signals of four conditions were collected: 1) normal condition (NC); 2) medium-worn (MW); 3) broken-tooth (BT) and 4) slight-worn (SW), as shown in Fig. 6 and each health condition has one load. The sampling frequency was 16384 Hz. 80 samples are sampled for each health condition from raw data, where each time-domain sample contains 800 data points.

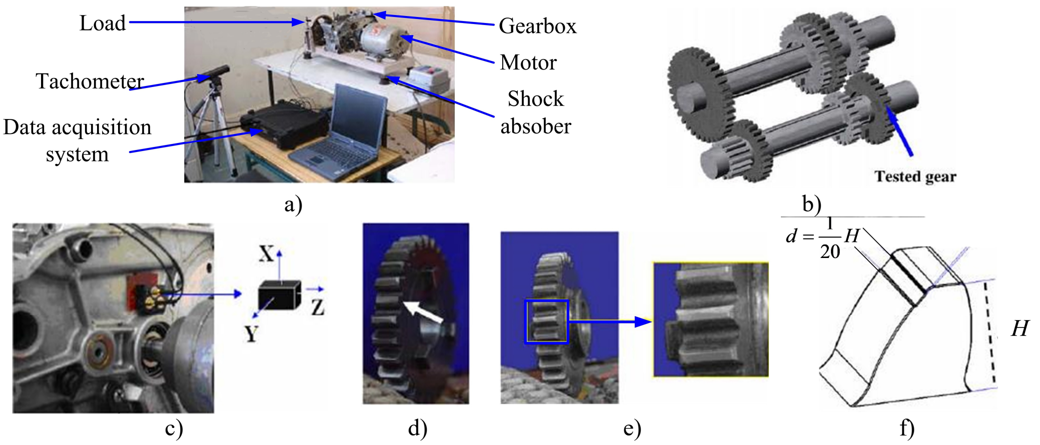


Fig. 6. a) test rig, b) schematic of the gearbox, c) accelerometer location, d) broken teeth defect, e) worn teeth defect, f) worn model defect

4.2.2. Diagnosis result

The preprocessing of each sample is the same as the bearing case, 400 Fourier coefficients are obtained from each sample. 10 % samples are randomly selected for training and the rest for testing. The output dimensions N^{out} for experiments of various networks are all tuned to 400 and the regular parameters λ_1, λ_2 are all tuned to 100.

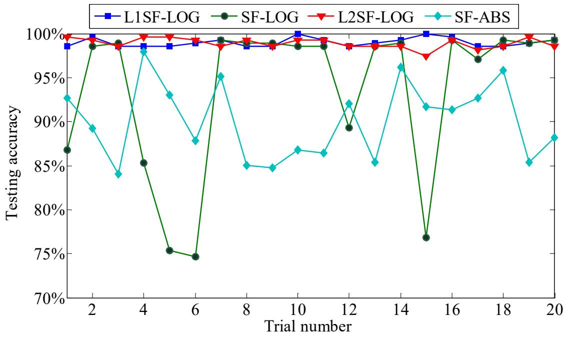


Fig. 7. Diagnosis results of 20 trials

20 trials are carried out for each experiment, and the comparison of diagnosis results among 4 networks is depicted in Fig. 7. It shows that the averaged accuracies are mostly over 99.5 % for L1SF-LOG and 99.0 % for L2SF-LOG, and both of them also hold quite low accuracy fluctuations. By comparison, it indicates that the new activation function can extract better features than the soft-absolute function, and L1 regularization is effective in improving the performance of standard sparse filtering.

Table 5. Training and testing accuracies of 6 networks

Method	Training accuracy (%)	Testing accuracy (%)
SF-ABS	100	88.35±7.85
SF-LOG	100	94.51±6.19
L1SF-LOG	100	99.13±0.50
L2SF-LOG	100	98.80±0.56
L1SF-ABS	100	94.82±4.24
L2SF-ABS	100	98.05±0.61
Note: The format of the result is: mean value ± standard deviation		

To quantitatively show the efficiency of L1SF-LOG and L2SF-LOG in preventing overfitting, the averaged accuracies and standard deviations of each network are shown in Table 5. It shows that the improved networks all perform better than SF-ABS, and L1SF-LOG holds the highest classification accuracy and the lowest standard deviation once again. Meanwhile, it also presents that the difference between averaged training accuracy and testing accuracy of L1SF-LOG is the smallest, namely 0.87 %. Furthermore, it is worth mentioning that the standard deviation is 0.5 % for L1SF-LOG and 0.56 % for L2SF-LOG, which means L1SF-LOG and L2SF-LOG are stable, and L1SF-LOG is superior to L2SF-LOG.

5. Discussion

Experiments on these networks show that proposed L1SF-LOG can obtain much higher testing accuracy, stability and generalization ability than the others, and networks with the new activation function perform better than the ones without. It also demonstrates that although inferior to L1-regularized sparse filtering, the L2-regularized sparse filtering is also effective in improving some aspects of performance.

We focus on the process of feature calculation and investigate why L1SF-LOG holds the best performance. As sparse filtering has no bias term, if we use $F(x) = x$ as the activation function, then we can describe the feature calculation process as follows:

$$f_l^{\,i} = F\left(\mathbf{W}_l^{\,T} x^i\right). \tag{9}$$

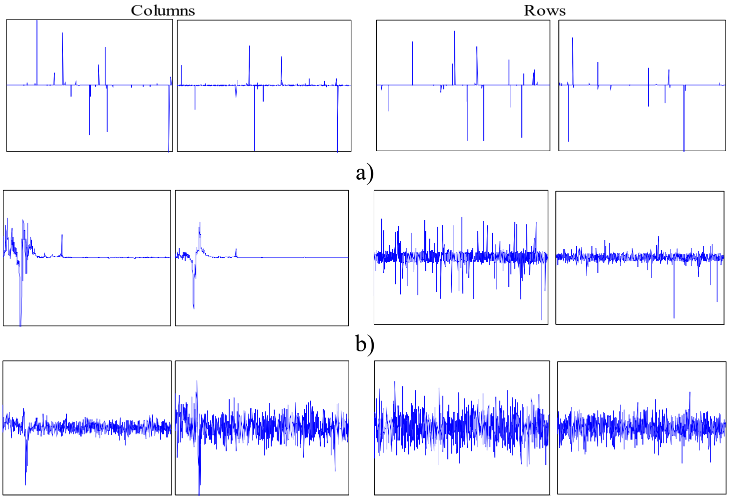


Fig. 8. Columns and rows of weight matrices learned from the bearing dataset:
a) L1SF-LOG, b) L2SF-LOG, c) SF-LOG

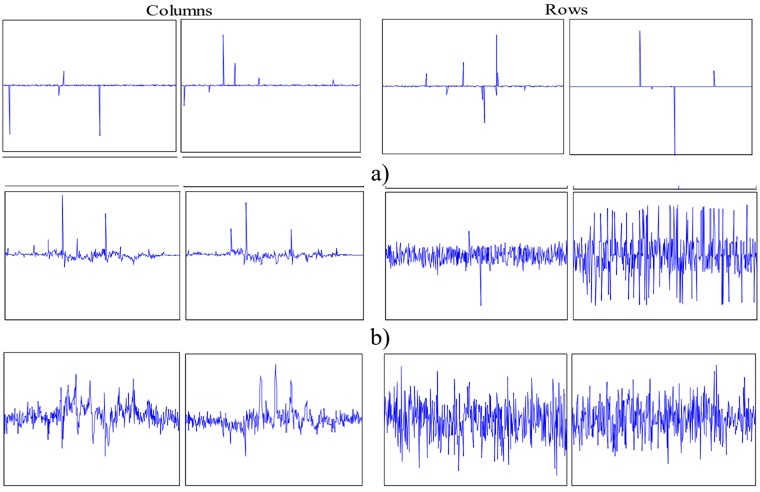
We regard each column of $\mathbf{W}$ as a filter [2] and the features are extracted by the inner product of weight matrix columns and sample $\mathbf{x}^i$. When samples are fixed, we want the filters to extract discriminative features and do not want these features to be extracted repeatedly, which means that the sparsity of the weight matrix should be ensured in a certain degree. Heuristically, as sparse filtering achieves the sparsity of the feature matrix by implementing L1 norm in the objective function, we utilize L1 regularization to constrain the sparsity of the weight matrix. Similarly, when the samples are frequency spectra, we propose defining the population sparsity and lifetime sparsity of weight matrix as follows:

(1) Population sparsity, which refers that the columns of the weight matrix have very narrow passing bandwidths. It should be noticed that different from common filters, each column of the

weight matrix can have several main lobes, which means that it can fuse several weakly discriminative features by composing them into a prominent one.

(2) Lifetime sparsity, which refers that the rows of the weight matrix should be sparsely activated, which means the same frequency band will not be filtered repeatedly. It prevents the discriminative features from being extracted by many times and reduces the information redundancy in feature representation. As a result, more discriminative features are mined.

To further present the effectiveness of sparsity in the weight matrix, typical rows and columns of weight matrices learned from L1SF-LOG, L2SF-LOG, SF-LOG are presented in Fig. 8 and Fig. 9. Firstly, it can be seen that all the columns of weight matrix learned by each network serve as band-width filters for frequency-domain samples and the ones learned by L1SF-LOG have the best frequency-domain filtering property. By comparison, it can be concluded that L1SF-LOG holds this property because its weight matrix is constrained by L1 regularization. Secondly, it also presents that the columns and rows of weight matrices learned by L1SF-LOG are all highly sparse, which means that the weight matrix has the population sparsity and lifetime sparsity defined above. Therefore, it is validated that L1 regularization is effective in constraining the sparsity of the weight matrix in sparse filtering. Furthermore, it can also be found that the columns of weight matrix learned by L2SF-LOG are smoother than SF-LOG.



c)

Fig. 9. Columns and rows of weight matrices learned from the gearbox dataset:
a) L1SF-LOG, b) L2SF-LOG, c) SF-LOG

Weiwei Qian Conceived and designed the work that led to the submission and wrote the submission. Shunming Li and Xingxing Jiang revised the manuscript. Jinrui Wang and Zhenghui An helped perform the analysis with discussions.

6. Conclusions

A new feature extraction method called L1-regularized sparse filtering is proposed to extract more discriminative features and improve generalization ability of standard sparse filtering through constraining the sparsity of the weight matrix. Based on FFT, L1-regularized sparse filtering and softmax regression, a novel three-stage fault diagnosis method is developed. Experiments on a bearing dataset and a gearbox dataset demonstrate that L1SF-LOG is more effective in preventing overfitting than L2SF-LOG, SF-LOG and SF-ABS. Meanwhile, the proposed L1SF-LOG can obtain much higher diagnosis accuracy with less training samples and be more stable.

We propose adopting different activation functions in feedforward and weight matrix optimization for considering their different requirements of anti-noise and non-saturating properties. Experiments show that this operation improves the testing accuracy and the diagnosis network stability a lot. Furthermore, we give an interpretation of the weight matrix and define two sparse properties of weight matrix. The significance of the two properties is validated by experiments. It shows that the weight matrix learned by L1SF-LOG has the sparse properties described above, which accounts for its superiority.

Acknowledgements

The research was supported by National Natural Science Foundation of China (51675262) and also supported by the Project of National Key Research and Development Plan of China “New Energy-Saving Environmental Protection Agricultural Engine Development” (2016YFD0700800) and the Advance Research Field Fund Project of China (6140210020102).

References

[1] **Yin S., Li X., Gao H., et al.** Data-based techniques focused on modern industry: an overview. IEEE Transactions on Industrial Electronics, Vol. 62, Issue 1, 2014, p. 657-667.

[2] **Lei Y., Jia F., Lin J., Xing S., Ding S.** An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Transactions on Industrial Electronics, Vol. 63, Issue 5, 2016, p. 31-37.

[3] **Nembhard A. D., Sinha J. K., Yunusa Kaltungo A.** Development of a generic rotating machinery fault diagnosis approach insensitive to machine speed and support type. Journal of Sound and Vibration, Vol. 337, 2015, p. 321-341.

[4] **Qin S. J.** Process data analytics in the era of big data. AIChE Journal, Vol. 60, Issue 9, 2014, p. 3092-3100.

[5] **Frankel F., Reid R.** Big data: distilling meaning from data. Nature, Vol. 455, Issue 7209, 2008, p. 30.

[6] **Gan M., Wang C., Zhu C.** Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings. Mechanical Systems and Signal Processing, Vol. 72-73, Issues 2, 2016, p. 92-104.

[7] **Yan R., Gao R. X., Chen X.** Wavelets for fault diagnosis of rotary machines: a review with applications. Signal Processing, Vol. 96, Issue 5, 2014, p. 1-15.

[8] **Yin J., Wang W., Man Z., Khoo S.** Statistical modeling of gear vibration signals and its application to detecting and diagnosing gear faults. Information Sciences an International Journal, Vol. 259, Issue 3, 2014, p. 295-303.

[9] **Li W., Zhu Z., Jiang F., Zhou G., Chen G.** Fault diagnosis of rotating machinery with a novel statistical feature extraction and evaluation method. Mechanical Systems and Signal Processing, Vol. 50, Issue 51, 2015, p. 414-426.

[10] **Feng Z., Liang M., Chu F.** Recent advances in time-frequency analysis methods for machinery fault diagnosis: a review with application examples. Mechanical Systems and Signal Processing, Vol. 38, Issue 1, 2013, p. 165-205.

[11] **Li Y., Xu M., Wang R., Huang W.** A fault diagnosis scheme for rolling bearing based on local mean decomposition and improved multiscale fuzzy entropy. Journal of Sound and Vibration, Vol. 360, 2016, p. 277-299.

[12] **Ming A. B., Zhang W., Qin Z. Y., Chu F. L.** Envelope calculation of the multi-component signal and its application to the deterministic component cancellation in bearing fault diagnosis. Mechanical Systems and Signal Processing, Vol. 50, Issue 51, 2016, p. 70-100.

[13] **Li Y., Yang Y., Li G., Xu M., Huang W.** A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and MRMR feature selection. Mechanical Systems and Signal Processing, Vol. 91, 2017, p. 295-312.

[14] **Li Y., Li G., Yang Y., Liang X., Xu M.** A fault diagnosis scheme for planetary gearboxes using adaptive multi-scale morphology filter and modified hierarchical permutation entropy. Mechanical Systems and Signal Processing, Vol. 105, 2018, p. 319-337.

[15] **Zhao H., Sun M., Deng W., et al.** A new feature extraction method based on EEMD and multi-scale fuzzy entropy for motor bearing. Entropy, Vol. 19, Issue 1, 2016, p. 14.

[16] **Landi A., Piaggi P., Pioggia G.** Backpropagation-based non linear PCA for biomedical applications. International Conference on Intelligent Systems Design and Applications, Vol. 58, 2009, p. 635-640.

[17] **Hsu C. C., Chen M. C., Chen L. S.** Intelligent ICA–SVM fault detector for non-Gaussian multivariate process monitoring. Expert Systems with Applications, Vol. 37, Issue 4, 2010, p. 3264-3273.

[18] **Romero A., Radeva P., Gatta C.** No more meta-parameter tuning in unsupervised sparse feature learning. Computer Science, 2014.

[19] **Ngiam J., Pang W. K., Chen Z., Bhaskar S., Ng A. Y.** Sparse filtering. International Conference on Neural Information Processing Systems, 2011, p. 1125-1133.

[20] **Cheriyadat A. M.** Unsupervised feature learning for aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. 52, Issue 1, 2013, p. 439-451.

[21] **Chopra P., Yadav S. K.** Erratum to: fault detection and classification by unsupervised feature extraction and dimensionality reduction. Complex and Intelligent Systems, Vol. 1, Issues 1-4, 2015, p. 35-35.

[22] **Liu H., Liu C., Huang Y.** Adaptive feature extraction using sparse coding for machinery fault diagnosis. Mechanical Systems and Signal Processing, Vol. 25, Issue 2, 2011, p. 558-574.

[23] **Ajami A., Daneshvar M.** Data driven approach for fault detection and diagnosis of turbine in thermal power plant using independent component analysis (ICA). International Journal of Electrical Power and Energy Systems, Vol. 43, Issue 1, 2012, p. 728-735.

[24] **Jia F., Lei Y., Lin J., Zhou X., Lu N.** Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mechanical Systems and Signal Processing, Vols. 72-73, 2016, p. 303-315.

[25] **Pellegrini T.** Comparing SVM, Softmax, and Shallow Neural Networks for Eating Condition Classification. INTERSPEECH, 2015.

[26] **Deng W., Yao R., Zhao H., et al.** A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. Soft Computing, Vols. 2-4, 2017, p. 1-18.

[27] **Tibshirani R., Saunders M., Rosset S., Zhu J., Knight K.** Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Vol. 67, Issue 1, 2010, p. 91-108.

[28] **Yang Z., Jin L., Tao D., Zhang S., Zhang X.** Single-layer unsupervised feature learning with l2 regularized sparse filtering. IEEE China Summit and International Conference on Signal and Information Processing, 2014, p. 475-479.

[29] **Deng W., Zhao H., Yang X., et al.** Study on an improved adaptive PSO algorithm for solving multi-objective gate assignment. Applied Soft Computing, Vol. 59, 2017, p. 288-302.

[30] **Deng W., Zhao H., Zou L., et al.** A novel collaborative optimization algorithm in solving complex optimization problems. Soft Computing, Vol. 21, Issue 15, 2016, p. 1-12.

[31] **Jia F., Lei Y., Guo L., Lin J., Xing S.** A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. Neurocomputing, Vol. 272, Issue 10, 2018, p. 619-628.

[32] **Lecun Y., Bengio Y., Hinton G.** Deep learning. Nature, Vol. 521, Issue 7553, 2015, p. 436-444.

[33] **Lecun Y., Ranzato M.** Deep learning tutorial. International Conference on Machine Learning, Citeseer, 2013.

[34] **Lou X., Loparo K. A.** Bearing fault diagnosis based on wavelet transform and fuzzy inference. Mechanical Systems and Signal Processing, Vol. 18, Issue 5, 2004, p. 1077-1095.

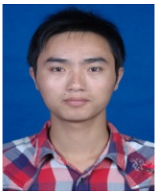
[35] **Rafiee J., Tse P. W., Harifi A., Sadeghi M. H.** A novel technique for selecting mother wavelet function using an intelligent fault diagnosis system. Expert Systems with Applications an International Journal, Vol. 36, Issue 3, 2009, p. 4862-4875.

Weiwei Qian received the B.S. degrees in Jiangsu University of Science and Technology (JUST), Zhenjiang, China, in 2012 and 2016. Now he is a Ph.D. candidate with College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include rotating machinery fault diagnosis and mechanical signal and information processing, machine learning.

Shunming Li received Ph.D. degree in mechanics from Xi’an Jiaotong University, China, in 1988. He is a Professor in Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include noise and vibration analysis and control, signal processing, machine fault diagnosis, sensing and measurement technology, intelligent vehicles.

Jinrui Wang received the B.S. and M.S. degrees in Shandong University of Science and Technology (SDUST), Tsingdao, China, in 2013 and 2015. Now he is a Ph.D. candidate with College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include rotating machinery fault diagnosis and mechanical signal and information processing.

Zhenghui An received the M.S. degrees in University of Jinan, Jinan, China, in 2014 and 2016. Now he is a Ph.D. candidate with College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include rotating machinery fault diagnosis and mechanical signal and information processing.



Xingxing Jiang received the Ph.D. degree in Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. Now he is doing his postdoctoral research in Soochow University, Suzhou, China. His current research interests include rotating machinery fault diagnosis and mechanical signal and information processing.