2479. An automatic feature extraction method and its application in fault diagnosis

Jinrui Wang[1], Shunming Li[2], Xingxing Jiang[3], Chun Cheng[4]

[1, 2, 3, 4]College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, P. R. China

[3]School of Urban Rail Transportation, Soochow University, Suzhou 215137, P. R. China

[2]Corresponding author

E-mail: [1]*wangjr33@163.com*, [2]*smli@nuaa.edu.cn*, [3]*jiangxx@suda.edu.cn*, [4]*chengchun987@163.com*

Check for updates

Abstract. The main challenge of fault diagnosis is to extract excellent fault feature, but these methods usually depend on the manpower and prior knowledge. It is desirable to automatically extract useful feature from input data in an unsupervised way. Hence, an automatic feature extraction method is presented in this paper. The proposed method first captures fault feature from the raw vibration signal by sparse filtering. Considering that the learned feature is high-dimensional data which cannot achieve visualization, t-distributed stochastic neighbor embedding (t-SNE) is further selected as the dimensionality reduction tool to map the learned feature into a three-dimensional feature vector. Consequently, the effectiveness of the proposed method is verified using gearbox and bearing experimental datas. The classification results show that the hybrid method of sparse filtering and t-SNE can well extract discriminative information from the raw vibration signal and can clearly distinguish different fault types. Through comparison analysis, it is also validated that the proposed method is superior to the other methods.

Keywords: fault diagnosis, automatic feature extraction, sparse filtering, t-SNE.

1. Introduction

As the most essential system in rotating machinery, gear and bearing play a major role to keep the entire machine operating normally. Serious faults of gear and bearing may lead to catastrophic consequences and cause enormous economic losses. So, condition monitoring and fault diagnosis have attracted broad attention in reducing undesirable casualties and minimizing production loss [1].

The feature extracted from vibration signal is commonly used to detect faults in machines [2]. And the more meaningful feature can enhance the identification accuracy. However, complex structures and noises in the observed signal make it difficult to extract the effective feature. For this reason, Different kinds of signal processing methods have been performed such as time-domain analysis [3], frequency transform [4], time-frequency analysis [5, 6] and envelope demodulation [7]. Among the conventional methods, researchers have spent a large amount of time on feature extraction and selection which are complicated and longstanding tasks. With the research of machine learning, neural network [8, 9] has attracted more and more attention. It can automatically learn high-dimensional feature from the signal by the hidden layers, but it still requires lots of label data.

As a viable alternative to manually design feature representations, unsupervised feature learning has been successfully implemented to extract good characteristics in many image [10], video [11] and audio [12] tasks. However, many current unsupervised feature learning algorithms are challenging to implement because they need to turn the various hyperparameters. If the hyperparameters are set improperly, it will produce a great impact on the diagnosis accuracy [13]. These algorithms include sparse RBMs [14], sparse autoencoders [15], sparse coding [16], independent component analysis (ICA) [17] and others. A comparison of the tunable hyperparameters in these algorithms are shown in Table 1. For instance, Sparse RBMs has up to half a dozen hyperparameters which makes it difficult to tune and monitor convergence. ICA has just one tunable hyperparameter, but it scales poorly to large inputs or large sets of features [18].

Ngiam et al. [19] proposed an unsupervised feature learning network named sparse filtering. It only focuses on optimizing the sparsity of the learned representations and ignores the problem of learning the data distribution. It also scales excellently with the dimension of the input. Only the number of features needs to set, so it is extremely simple to tune and easy to implement by a few lines of MATLAB code. Meanwhile in Ref. [19], the author adopted it on image recognition and phone classification, which generated the state-of-the-art performance.

Table 1. Tunable hyperparameters in various algorithm

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

Because of its simplicity and performance, sparse filtering is proposed to solve fault diagnosis of rotating machines in this paper. However, the dimension of the learned feature is too high so that visualization is difficult. So, it is necessary to select a suitable dimensionality reduction tool to embed the learned feature into a low-dimensional space. Maaten et al. [20] introduced an nonlinear dimensionality reduction technique called t-distributed stochastic neighbor embedding (t-SNE), which is more effective in creating a single map at different scales than other techniques. Then the reduced features of different fault types can achieve visualization in a scatter plot.

This paper is organized as follows. Section 2 briefly introduces the algorithms of sparse filtering and t-SNE. Section 3 is dedicated to detail the content of the proposed method. In Sections 4, the diagnosis cases of gearbox and bearing datasets are adopted to validate the effectiveness of the proposed method. Furthermore, the superiority of the proposed method is exhibited by comparing with other methods. Finally, the conclusion is drawn in Section 5.

2. Theoretical background

2.1. Sparse filtering

Sparse filtering is a simple unsupervised two-layer network which aims to optimize the sparsity of the learned features but not attempts to model the data distribution [19]. The method learns the excellent features in an unsupervised way which possesses three principles:

(1) Population sparsity: each sample should be sparsity, which means each sample is represented by just a few activated elements.

(2) Lifetime sparsity: each feature should be sparsity, which means each feature allows to be activated only for a small number of samples.

(3) High dispersal: each feature should have similar statistical properties.

The architecture of sparse filtering is shown in Fig. 1. The collected raw vibration signal is directly used as the input data. Firstly, the vibration signal is separated into M samples to compose a training set $\{\mathbf{x}^i\}_{i=1}^M$, where $\mathbf{x}^i \in \Re^{N \times 1}$ is a training sample contains N data points. Then, the training set is used to train the sparse filtering model so as to obtain a weight matrix $\mathbf{W} \in R^{N \times L}$. Finally, each sample is mapped into a feature vector $\mathbf{f}^i \in \Re^{L \times 1}$ by the weight matrix $\mathbf{W}$. For sparse filtering, an activation function is needed for calculating the nonlinear features. In our experiment, the soft-absolute function is adopted as the activation function and the features of each sample can be calculated as follows:

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

where $\mathbf{f}_j^i$ is the jth feature value corresponding to rows in the ith column, $\varepsilon = 10^{-8}$.

The feature matrix is comprised by the features $\mathbf{f}_j^i$. Firstly, each row is normalized to be equally active by its $\ell 2$-norm:

$$\tilde{\mathbf{f}}_j = \frac{\mathbf{f}_j}{||\mathbf{f}_j||_2}. \tag{2}$$

Then, each column (or each sample) is normalized by its $\ell 2$-norm. As a result, each feature is constrained to lie on the unit $\ell 2$-ball:

$$\hat{\mathbf{f}}^i = \frac{\tilde{\mathbf{f}}^i}{||\tilde{\mathbf{f}}^i||_2}. \tag{3}$$

At last, the normalized features are optimized for sparseness using the $\ell 1$ penalty. For the training set $\{\mathbf{x}^i\}_{i=1}^M$, the sparse filtering objective is shown as follows:

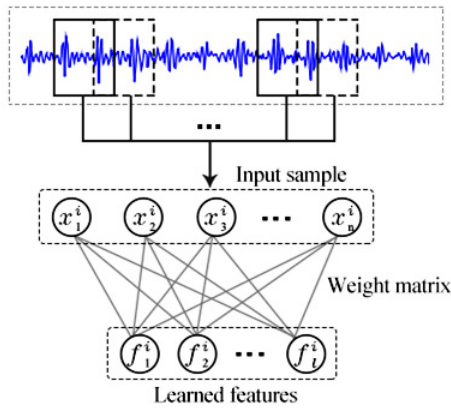$$\underset{W}{\text{minimize}} \sum_{i=1}^M ||\hat{\mathbf{f}}^i||_1. \tag{4}$$

Fig. 1. Architecture of sparse filtering

2.2. T-SNE

As a nonlinear dimensionality reduction technique, t-SNE is extremely suited for embedding the high-dimensional dataset into an s-dimensional vector (typical values for s are 2 or 3) [21]. So, each object can be represented by a point in the scatter plot. To this end, t-SNE determines the joint probabilities p_{ij} that measure the pairwise similarity between features $\mathbf{f}_i$ and $\mathbf{f}_j$ by symmetrizing two conditional probabilities as follows:

$$p_{j|i} = \frac{\exp\left(-||\mathbf{f}_i - \mathbf{f}_j||^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp(-||\mathbf{f}_i - \mathbf{f}_k||^2 / 2\sigma_i^2)}, \tag{5}$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \tag{6}$$

where σ_i^2 is the variance of the Gaussian, which is centered on feature $\mathbf{f}_i$ and is determined by the way that the perplexity of the conditional distribution equals to a predefined perplexity. As a result, σ_i tends to be smaller in the data space with a higher data density than a lower data density. So,

for each input object, the optimal value of σ_i can be found using a simple binary search.

In the low-dimensional space, the similarities between two features $\mathbf{y}_i$ and $\mathbf{y}_j$ (i.e., the mapped features of $\mathbf{f}_i$ and $\mathbf{f}_j$) are measured using a normalized heavy-tailed kernel. Specifically, the joint probabilities q_{ij} between $\mathbf{y}_i$ and $\mathbf{y}_j$ is computed as a normalized Student-t distribution:

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}. \tag{7}$$

The heavy tails of the normalized Student-t distribution can make the input objects $\mathbf{f}_i$ and $\mathbf{f}_j$ to be modeled far apart by $\mathbf{y}_i$ and $\mathbf{y}_j$. And it creates more space to accurately model the small pairwise distance. The locations of the embedding points $\mathbf{y}_k$ are computed by minimizing the KL divergence between the joint distributions q_{ij} and p_{ij}:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right), \tag{8}$$

where P and Q are the matrix formation of p_{ij} and q_{ij}. Then $\mathbf{f}_k$ and $\mathbf{y}_k$ are becoming more and more similar with each other. That is to say, $\mathbf{y}_k$ could represent the characteristics of $\mathbf{f}_k$.

3. Proposed framework

This section details how to automatically extract features from mechanical signal. The procedure of our method is displayed in Fig. 2 and it can be described as following:
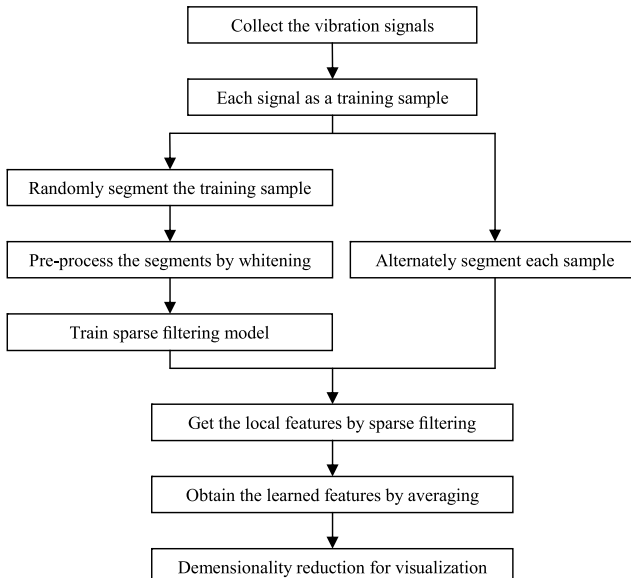

Fig. 2. Flowchart of the proposed method

(1) Collect signals. The vibration signals are collected under different health conditions and are directly adopted as the training samples. We collect Z segments from each sample to compose the training set $\{\mathbf{s}^j\}_{j=1}^Z$ by an overlapped manner, where $\mathbf{s}^j \in \mathfrak{R}^{N_{in} \times 1}$ is the jth segment containing N_{in} data points.

(2) Whitening. The training set $\{\mathbf{s}^j\}_{j=1}^Z$ is written as a matrix formation $\mathbf{S} \in \mathfrak{R}^{N_{in} \times Z}$ and

pre-processed by whitening. In this way, the segments can become less correlated with each other and the convergence rate can get faster [22]. By computing:

$$cov(\mathbf{S}^T) = \mathbf{E}\mathbf{U}\mathbf{E}^T, \tag{9}$$

where $cov(\mathbf{S}^T)$ is the covariance matrix, $\mathbf{E}$ denotes the orthogonal matrix of eigenvectors, and $\mathbf{U}$ is the diagonal matrix of the eigenvalues. Then the whitened training set $\mathbf{S}_w$ can be obtained as follows:

$$\mathbf{S}_w = \mathbf{E}\mathbf{U}^{-1/2}\mathbf{E}^T\mathbf{S}. \tag{10}$$

(3) Train sparse filtering. $\mathbf{S}_w$ is employed to train the sparse filtering model, and then the weight matrix $\mathbf{W}$ is obtained by minimizing Eq. (4).

(4) Calculate the local features. The training sample $\mathbf{x}^i$ is alternately divided into K segments, where $K = N/N_{in}$ And these segments constitute a set $\{\mathbf{x}_k^i\}_{k=1}^K$, where $\mathbf{x}_k^i \in \mathfrak{R}^{N_{in} \times 1}$. Then the local features $\mathbf{f}_k^i \in \mathfrak{R}^{1 \times N_{out}}$ can be calculated from each training sample $\mathbf{x}_k^i$ by the weight matrix $\mathbf{W}$.

(5) Obtain the learned features. The local features $\mathbf{f}_k^i$ are combined into a feature vector $\mathbf{f}^i$ by the method of averaging, and $\mathbf{f}^i$ is the learned feature:

$$\mathbf{f}^i = \left(\frac{1}{K}\sum_{k=1}^K \mathbf{f}_k^i\right)^T. \tag{11}$$

(6) Dimensionality reduction. Since the obtained feature $\mathbf{f}^i$ is still high-dimensional data, t-SNE is adopted to reduce its dimension for visualization.

4. Fault diagnosis using the proposed method

In this section, a gearbox and a bearing experimental datasets are employed to validate the effectiveness of our method. In order to further illustrate the superiority of our method, several commonly used dimensionality reduction tools are adopted to combine with sparse filtering respectively for comparison analysis.

4.1. Case 1. Gearbox experiment verification

4.1.1. Data description

A four-speed motorcycle gearbox [23] is used to collect vibration signals as shown in Fig. 3. Besides the gearbox, there are an electrical motor, a tachometer, a tri-axial accelerometer, a data acquisition system, a load mechanism and four shock absorbers. There are four kinds of gearbox defects under a certain load: normal condition (NC), slight-worn (SW), medium-worn (MW) and broken-tooth (BT), as shown in Fig. 3. The sampling frequency was 16384 Hz. We collect 50 samples from the raw vibration singal of NC and 100 samples from the raw vibration signal of SW, MW and BT, where each sample contains 1200 data points.

4.1.2. Diagnosis results

In this subsection, we will process the dataset by the proposed method. Firstly, we randomly select 10 % samples from each health condition to train the sparse filtering model. And then we randomly select 50 segments from each sample by using an overlapped manner, where each segment contains 100 data points. These segments are employed to constitute the training set $\{\mathbf{s}^j\}_{j=1}^{50}$, where $\mathbf{s}^j \in \mathfrak{R}^{100 \times 1}$ is the jth segment containing 100 data points. Subsequently, we use

these samples to train the sparse filtering model. There are two tunable feature parameters, i.e., the input and output dimension (N_{in} and N_{out}). Ref. [24] investigated the selection of these two parameters. It randomly selected 10 % of samples to train and the rest to test. The diagnosis results showed that the larger N_{in} is, the more time spends. Considering that the testing accuracy of $N_{in} = 100$ is the highest and the spent time is low, so N_{in} is identified as 100. And the selection of N_{out} is tradeoff between the diagnosis accuracy and the spent time. Since the increasing of the accuracy is not obvious after $N_{out} = 100$, so N_{out} is also identified as 100. After the sparse filtering model is trained and the weight matrix **W** is obtained, we use the rest samples to calculate the learned features by the weight matrix **W**. Each sample is alternately divided into 12 segments, where each segment contains 100 data points. As a result, a 100-dimensional feature vector is obtained for each $\mathbf{s}^j$ as show in Eq. (11). In Fig. 4, it exhibits all the learned feature vectors of each health condition. It can be observed that the dimension of the learned feature is too high so that visualization is difficult. Finally, t-SNE is adopted to embed the learned feature into a three-dimensional feature vector for visualization.
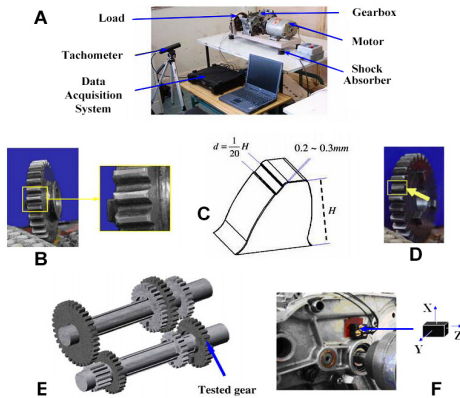

Fig. 3. a) Experimental set up; b) worn teeth; c) worn model;
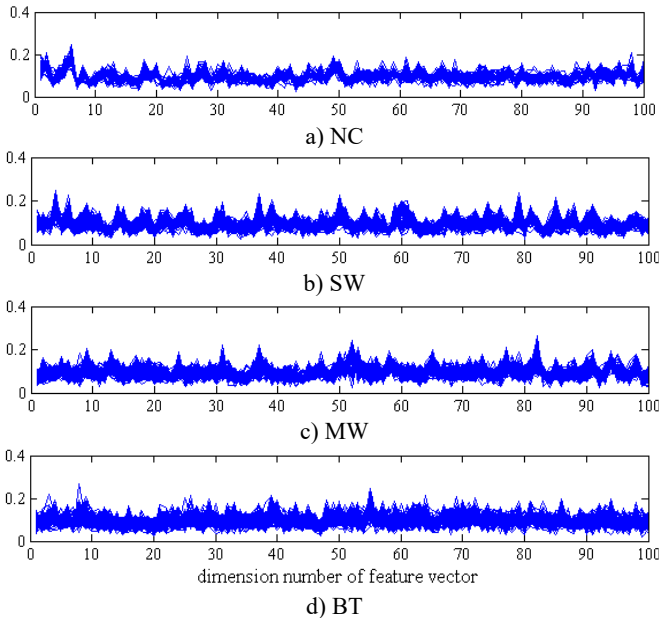d) broken teeth; e) schematic of the gearbox; and f) accelerometer location

Fig. 4. Learned feature vectors of gearbox

The classification result by our method is shown in Fig. 5(a). It is seen that the mapped features of the different types are separated excellently and the features of the same type are gathered together, and the distance between each type is large enough for distinguishing different health conditions. In addition, two similar models of worn gear are separated clearly. To testify the superiority of our method, five common dimensionality reduction tools combined with sparse filtering are adopted to process the gearbox dataset respectively. The five tools are: principal component analysis (PCA) [25], locality preserving projection (LPP) [26], Sammon mapping (SM) [27], linear discriminant analysis (LDA) [28], and stochastic proximity embedding (SPE) [29]. The classification results by the five methods are shown in Fig. 5(b)-(f).
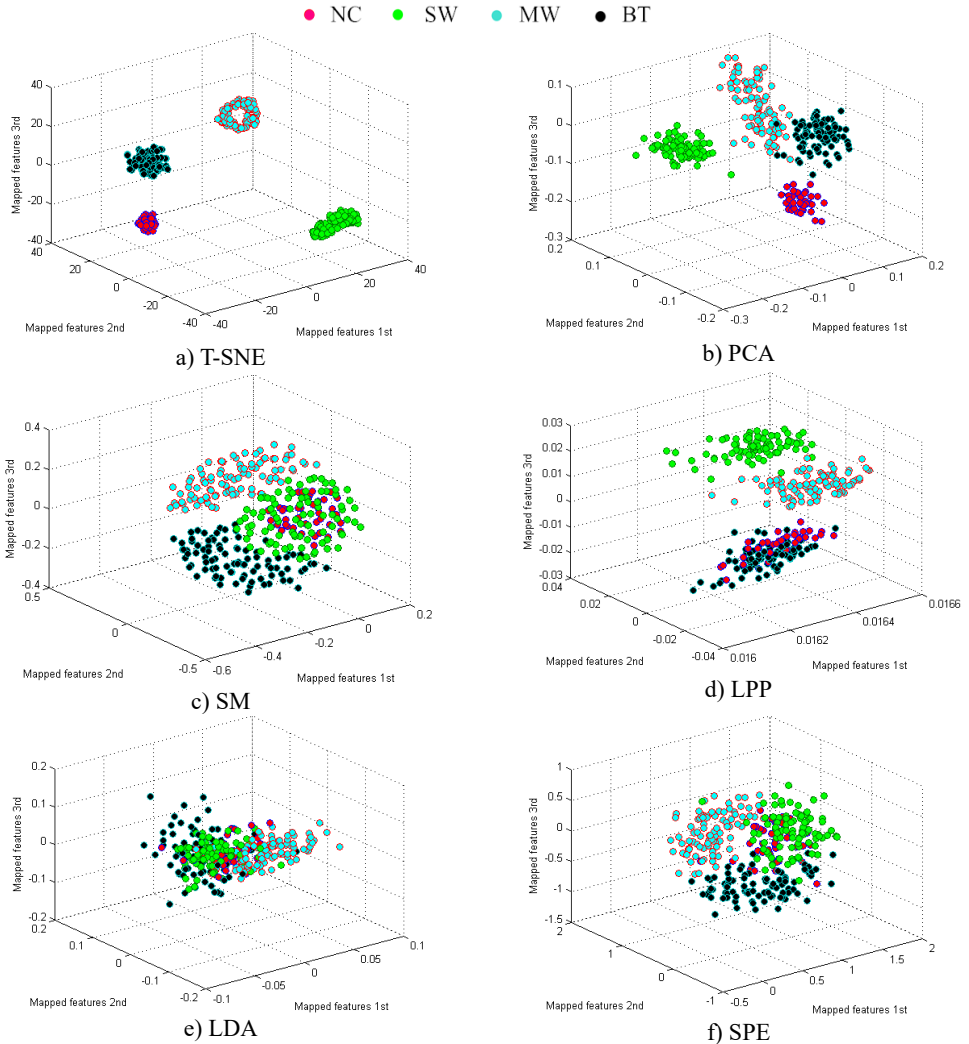

Fig. 5. Visualize features of gearbox

By comparing the results of the six methods, it is easy to see that the classification by t-SNE is the best. The mapped features by PCA are shown in Fig. 5(b), and its classification result is better than the rest methods. But there is still a big gap with the proposed method. For instance, the mapped features of MW are not well clustered and several of them are mixed with BT. In Fig. 5(c) which exhibits the mapped features by SM. In fact, these features are separated from another view to see. But the distance between each type is too small to distinguish. As we noticed

in Fig. 5(d), the mapped features of NC and BT are almost mixed together which show a bad performance by LPP. In Fig. 5(e), it can be observed that the mapped features by LDA are not separated out at all, which shows it cannot be used for classification. And in Fig. 5(f) which displays the mapped features by SPE is also not performed very well. Through the above comparison analysis, it is indicated that our method is the best choice for discriminating different fault types of the gearbox dataset.

Furthermore, to illustrate the advantage of the proposed method. softmax regression [30] is adopted as the classifier to test the accuracies of the mapped features by the six methods. We randomly select 10 % samples from each health condition to train the softmax regression model and the rest to test. The weight decay term λ of softmax regression is 1E-5. To reduce the effects of randomness, 20 trials are carried out for the experiment. The diagnosis results are depicted in Fig. 6. It shows that the diagnosis accuracies of t-SNE and PCA vary a little and the other methods vary greatly. As we can notice that the performance of t-SNE is a little better than PCA. Then the test accuracies are averaged by 20 trials and the results are displayed in Table 2. It can be seen that the average testing accuracy of the proposed method is 99.87 %, which is the best of all. And the accuracies of all the methods are the same as the performances of the visualize features in Fig. 5.

To verify the effectiveness of our method, a bearing dataset is employed in the next section.

Table 2. Average testing accuracy of gearbox using the six methods

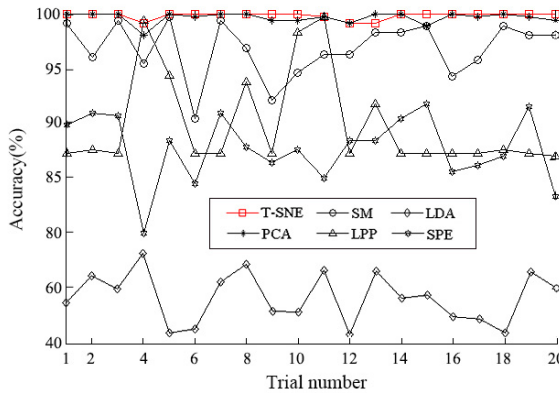| Methods	T-SNE	PCA	SM	LPP	LDA	SPE
Accuracy	99.84 %	99.62 %	96.43 %	88.74 %	56.73 %	86.22 %

Fig. 6. Diagnosis results of 20 trials of gearbox using the six methods

4.2. Case 2. Bearing experiment verification

4.2.1. Data description

In this section, the motor bearing experimental data supplied by Case Western Reserve University [30] is employed to test the effectiveness of our method. The ball bearing was installed in the driven end of an induction motor and the experimental set-up is shown in Fig. 7. Besides the induction motor, there are a dynamometer, a load mechanism and a tri-axial accelerometer. The dataset contains four bearing health conditions under a certain load: normal condition (NC), inner race fault (IF), roller fault (RF) and outer race fault (OF). The sampling frequency was 12 kHz. We collect 100 samples from each health condition, where each sample contains 1200 data points.
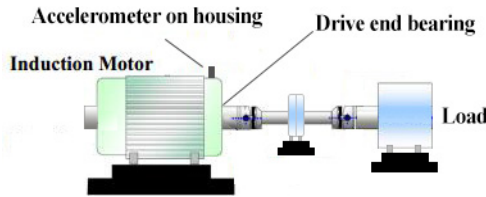
Fig. 7. A schematic of the experimental system

4.2.2. Diagnosis results

The process procedure of each sample is the same as the above experiment. We also randomly select 10 % samples from each health condition to train the sparse filtering model and then use the rest samples to calculate the learned features. The learned feature vectors of each health condition by sparse filtering are shown in Fig. 8. And the classification result by our method is shown in Fig. 9(a). It is seen that t-SNE also shows its excellent ability in clustering, the features of each fault type are clustered into a ball and can be easily distinguished.

a) NC

b) IF

c) RF

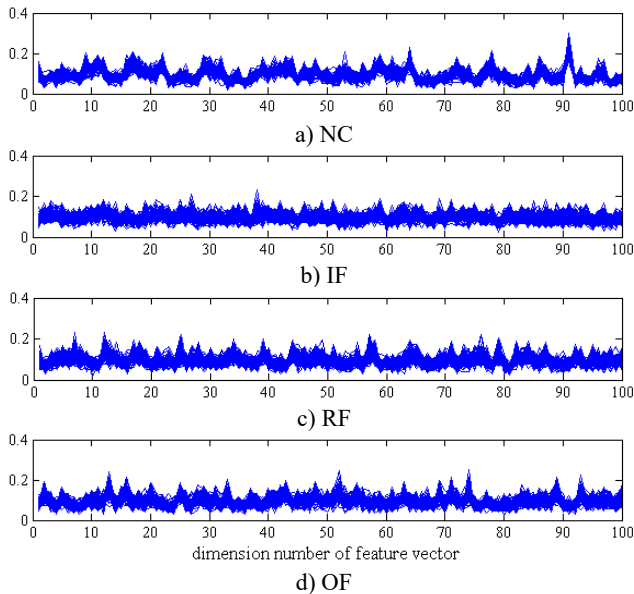dimension number of feature vector

d) OF

Fig. 8. Learned feature vectors of bearing

Then we use the other five methods to process the bearing dataset as above. The results are shown in Fig. 9(b)-(f). By comparing the results of the six methods, it is also easy to see that the classification by t-SNE is the best. In Fig. 9(b), although the mapped features by PCA are well performed, several features of each type are not well gathered which is not good as t-SNE. The mapped features by SM are shown in Fig. 9(c), it is noticed that the distance between each type is still too small for distinction. In Fig. 9(d), it shows the mapped features by LPP. It can be seen that parts of the mapped features of NC, IF and RF are mixed together. In Fig. 9(d), which display the mapped features by LPP. It can be observed that most of the mapped features are mixed together. And the mapped features by SPE are shown in Fig. 9(f) which also has several features mixed with each other.

Finally, we employ softmax regression to test the accuracies of the six method, and the process is the same as above. The diagnosis results are showed in Fig. 10 and the average testing accuracies are displayed in Table 3. It is exhibited that the accuracies of t-SNE and PCA are all 100 %, but we should note that the visualization of PCA in Fig. 8(b) is not so well as t-SNE in Fig. 8(a). And

the accuracies of the rest methods are also the same as the performances in Fig. 9. Through the above analysis, it is also concluded that the proposed method delivers the optimal performance for the bearing dataset.
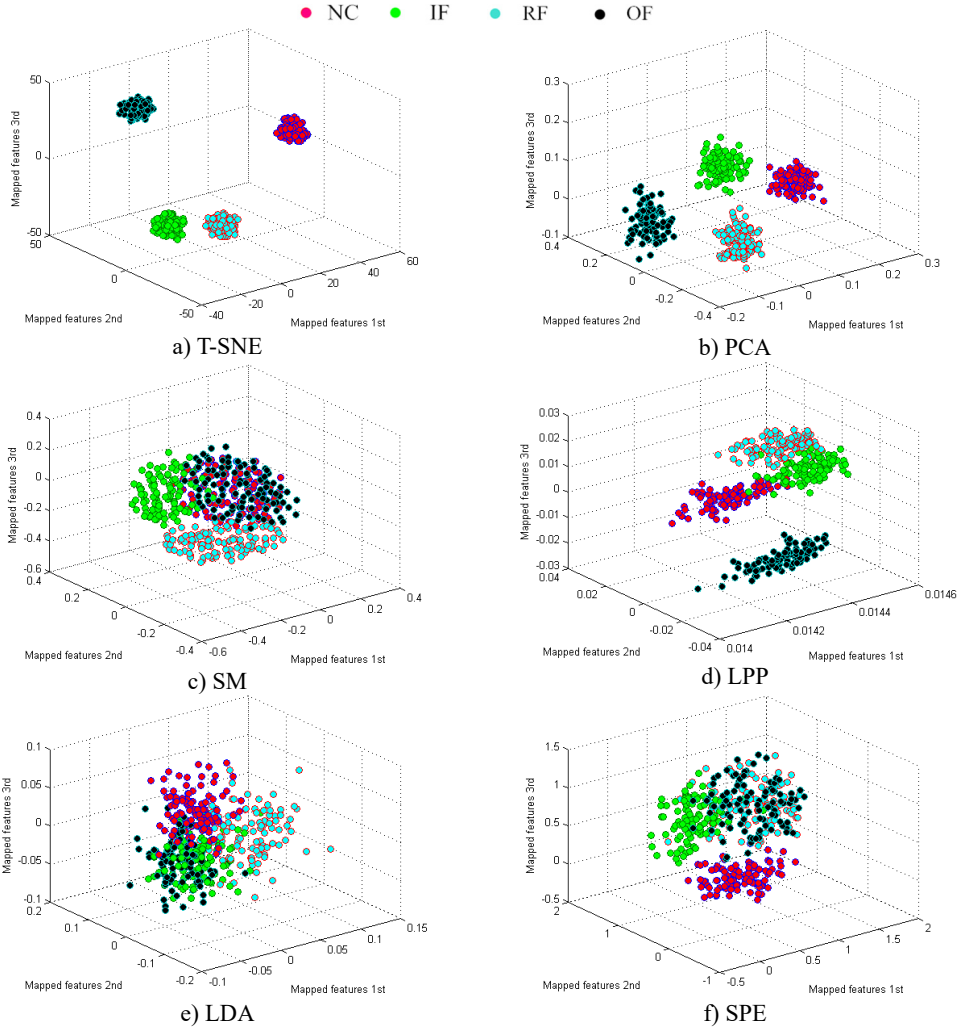

Fig. 9. Visualize features of bearing

Table 3. Average testing accuracy of bearing using the six methods

Methods	T-SNE	PCA	SM	LPP	LDA	SPE
Accuracy	100 %	100 %	99.22 %	95.97 %	63.56 %	92.64 %

5. Conclusions

An automatic feature extraction method was proposed and implemented to pinpoint the mechanical faults in this paper. In the method, sparse filtering and t-SNE were combined to determine the health conditions in an unsupervised way. By the both of gearbox and bearing experimental cases, it is demonstrated that the proposed method has the strong ability in feature extraction and classification. And also, the proposed method showed its superiority by comparing with other methods.
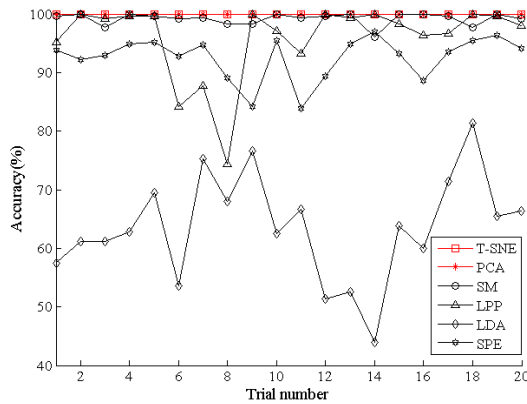
Fig. 10. Diagnosis results of 20 trials of bearing using the six methods

This paper provides the main contributions as follows. Firstly, the proposed method could adaptively learn the features from the raw vibration signals in an intelligent way, which makes it less dependent on the manpower and prior knowledge. Secondly, t-SNE is selected as the dimensionality reduction tool to achieve visualization, which makes the diagnosis results more intuitive. So, the hybrid method of sparse filtering and t-SNE could be an effective automatic feature extraction method for fault diagnosis. At the same time, the proposed method can serve as a reference for fault diagnosis of some other rotating machines.

Acknowledgements

References

[1] **Jiang X., Li S., Wang Y.** A novel method for self-adaptive feature extraction using scaling crossover characteristics of signals and combining with LS-SVM for multi-fault diagnosis of gearbox. Journal of Vibroengineering, Vol. 17, Issue 4, 2015, p. 1861-1878.

[2] **Li Y., Liang X., Xu M., et al.** Early fault feature extraction of rolling bearing based on ICD and tunable Q-factor wavelet transform. Mechanical Systems and Signal Processing, Vol. 86, 2017, p. 204-223.

[3] **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, Vol. 259, Issue 3, 2014, p. 295-303.

[4] **Li W., Zhu Z., Jiang F., 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.

[5] **Li Y., Xu M., Wang R., et al.** 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.

[6] **Li Y., Xu M., Wei Y., et al.** An improvement EMD method based on the optimized rational Hermite interpolation approach and its application to gear fault diagnosis. Measurement, Vol. 63, 2015, p. 330-345.

[7] **Ming A., Zhang W., Qin Z., Chu F.** 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, 2015, p. 70-100.

[8] **Monica A.** Analysis of induction motor fault diagnosis with fuzzy neural network. Applied Artificial Intelligence, Vol. 17, Issue 2, 2003, p. 105-133.

[9]   **Su H., Chong K., Kumar R.** Vibration signal analysis for electrical fault detection of induction machine using neural networks. International Symposium on Information Technology Convergence, Vol. 20, Issue 2, 2007, p. 183-194.

[10]   **Yang J., Yu K., Gong Y., Huang T.** Linear spatial pyramid matching using sparse coding for image classification. CVPR, 2009, p. 1794-1801.

[11]   **Le Q., Zou W., Yeung S., Ng A.** Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis. CVPR B, Vol. 42, Issue 7, 2011, p. 3361-3368.

[12]   **Lee H., Largman Y., Pham P., Ng A.** Unsupervised feature learning for audio classification using convolutional deep belief networks. NIPS, 2009, p. 1096-1104.

[13]   **Amar M., Gondal I., Wilson C.** Vibration spectrum imaging: A novel bearing fault classification approach. IEEE Transactions on Industrial Electronics, Vol. 62, Issue 1, 2015, p. 494-502.

[14]   **Cheriyadat A.** Unsupervised feature learning for aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. 52, Issue 1, 2014, p. 439-451.

[15]   **Chopra P, Yadav S.** Fault detection and classification by unsupervised feature extraction and dimensionality reduction. Complex and Intelligent Systems, Vol. 1, Issues 1-4, 2016, p. 1-9.

[16]   **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.

[17]   **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.

[18]   **Hyvarinen A., Hurri J., Patrick O.** Natural Image Statistics: A Probabilistic Approach to Early Computational Vision. Computational Imaging and Vision, Springer, London, 2009.

[19]   **Ngiam J., Chen Z., Bhaskar S., Koh P., Ng A.** Sparse filtering. Proceedings of Neural Information Processing Systems., 2011, p. 1125-1133.

[20]   **Laurens V., Hinton G.** Visualizing Data using t-SNE. Journal of Machine Learning Research, Vol. 9, Issue 2605, 2008, p. 2579-2605.

[21]   **Gisbrecht A., Mokbel B., Hammer B.** Linear basis-function t-SNE for fast nonlinear dimensionality reduction. IEEE International Joint Conference on Neural Networks, Vol. 20, 2012, p. 1-8.

[22]   **Hyvärinen A., Oja E.** Independent component analysis: algorithms and applications. Neural Networks, Vol. 13, Issue 4, 2000, p. 411-430.

[23]   **Jiang X., Li S., Wang Y.** Study on nature of crossover phenomena with application to gearbox fault diagnosis. Mechanical Systems and Signal Processing, Vol. 83, 2016, p. 272-295.

[24]   **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 Power Electronics, Vol. 63, Issue 5, 2016, p. 31-37.

[25]   **Yin S., Steven X., Naik A., et al.** On PCA-based fault diagnosis techniques. IEEE Control and Fault-Tolerant Systems, 2010, p. 179-184.

[26]   **Yu J.** A nonlinear probabilistic method and contribution analysis for machine condition monitoring. Mechanical Systems and Signal Processing, Vol. 37, Issues 1-2, 2013, p. 293-314.

[27]   **Wang X., Zheng Y., Zhao Z., Wang J.** Bearing fault diagnosis based on statistical locally linear embedding. Sensors, Vol. 15, Issue 7, 2015, p. 16225-16247.

[28]   **Prince S., Elder J.** Probabilistic linear discriminant analysis for inferences about identity. IEEE Computer Society, 2007, p. 1-8.

[29]   **Agrafiotis D., Xu H., Zhu F., et al.** Stochastic proximity embedding: methods and applications. Molecular Informatics, Vol. 29, Issue 11, 2010, p. 758-770.

[30]   **Khashei M., Hamadani A., Bijari M.** A novel hybrid classification model of artificial neural networks and multiple linear regression models. Expert Systems with Applications, Vol. 39, Issue 3, 2012, p. 2606-2620.

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

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. Student 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.

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.
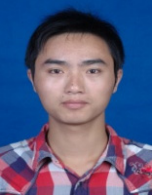
Xingxing Jiang received the B.S. degree in Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2012. Now he is a Ph.D. student with College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include rotating machinery fault diagnosis and mechanical signal and information processing.
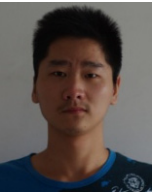
Chun Cheng received the B.S. degree in Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. Now he is a Ph.D. student with College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include vibration analysis and control.