A Method of Automatic Feature Extraction from Massive Vibration Signals of Machines

Feng Jia, Yaguo Lei*, Saibo Xing, Jing Lin
State Key Laboratory for Manufacturing Systems Engineering
Xi'an Jiaotong University
Xi'an 710049, China
*Corresponding author: yaguolei@mail.xjtu.edu.cn (Yaguo Lei)

***Abstract*—In the studies of intelligent fault diagnosis of machines, lots of effort goes into designing effective feature extraction algorithms. Such processes would consume plenty of human labor, especially when dealing with massive vibration signals. So it is interesting to automatically extract features using machine learning techniques, instead of manually extracting them. To deal with the problem, this paper presents a new automatic feature extraction method of machines. The proposed method first learns features from the vibration signals by *K*-means, and then maps the learned features into a salient low-dimensional feature space using *t*-distributed stochastic neighbor embedding (*t*-SNE). Through the feature extraction results of a bearing dataset, it is verified that the proposed method is able to effectively learn the features from the raw vibration signals and is superior to the manual features like time-domain features and wavelet features. Therefore, the proposed method has potential to be a tool in the automatic data mining of intelligent fault diagnosis.**

Keywords—intelligent fault diagnosis; automatic feature extraction; k-means; t-SNE

I. INTRODUCTION

Machine learning techniques have been widely applied to mechanical fault diagnosis since it may replace diagnostic specialists to efficiently process massive collected signals and automatically determine the health conditions of machines [1]. The diagnosis based on the machine learning techniques, such as artificial neural networks, support vector machine (SVM), and clustering, is called intelligent fault diagnosis in the field of mechanical fault diagnosis [2]. Typically, there are two main steps in intelligent fault diagnosis: feature extraction and selection, and fault recognition [3-5]. Feature extraction and selection aim to extract representative features from vibration signals based on signal processing techniques, and select salient features so as to reduce the feature dimension (feature selection is beneficial but not essential). Fault recognition is able to use machine learning techniques to automatically determine mechanical health conditions based on the extracted and selected features. Currently, a good deal of effort on intelligent fault diagnosis methods has been taken to improve the diagnosis accuracies of machines, and good results have been achieved. Amar et al. [6] developed a vibration spectrum imaging feature enhancement procedure to obtain robust features from vibration signals and applied an artificial neural network to diagnose the bearing faults. Lei et al. [7] designed two features, i.e. accumulative amplitudes of carrier orders and energy ratio based on difference spectra, to characterize the health conditions of planetary gearboxes and used relevance vector machine to classify these health conditions. Kang et al. [8] proposed a fault diagnosis approach for low-speed bearings in which wavelet-based fault features were used to represent diverse symptoms of bearing defects and SVM is applied to classify the health conditions using the sensitive features selected by kernel discriminative feature analysis. Dual tree complex wavelet transform was applied to extract features by Seshadrinath et al. [9] and probabilistic neural network was used to classify the health conditions of Induction Machines based on the features selected by principal components analysis (PCA). Zhang et al. [10] presented a method based on ensemble empirical mode decomposition and permutation entropy for extracting features and employed SVM for fault recognition of bearings.

Through the literature review, it is noticed that plenty of the effort in intelligent fault diagnosis goes into designing effective feature extraction algorithms since traditional machine learning techniques cannot extract the representative information from these signals. Based on the designed algorithms, salient features could be extracted from raw vibration signals. The feature designing processes take advantage of human knowledge in the characteristics of different mechanical signals and signal processing techniques. But they would consume plenty of human labor, especially when dealing with massive vibration signals. So a new challenging problem in intelligent fault diagnosis is how to automatically extract features using advanced machine learning techniques, instead of extracting them manually. To deal with this problem, a potential way may be unsupervised feature learning [11]. As a new achievement in machine learning, unsupervised feature learning has been used in lots of fields, such as image classification, object recognition, speech recognition and computer vision.

Most of the unsupervised feature learning techniques attempt to model the good distribution of raw data without labels, such as sparse autoencoders, sparse coding and restricted Boltzmann machines. These methods may provide the good feature representations of mechanical signals. However, they often require the tuning of various hyper-parameters to perform well and there are not many principles for how to tune them, which is a great challenge. For example, the hyper-parameters of the sparse autoencoders [12] needing to be tuned are: the number of features, learning rate, momentum, sparsity penalty and weight decay. If these hyper-parameters are not set appropriately, the learned features may result in poor diagnosis accuracy. Recently, Coates et al. [13] found that *K*-means clustering can be an effective unsupervised feature learning method, rivaling the state-of-the-art performance in image classification. Inspired by their work, we explore the application of *K*-means clustering for feature extraction in mechanical intelligent fault diagnosis. However, the features extracted by *K*-means are high-dimensional, containing irrelevant or redundant information. So dimensionality reduction techniques are needed to obtain a compact representation from the *K*-means features so as to capture the discriminative information. Among dimensionality reduction techniques, *t*-distributed stochastic neighbor embedding (*t*-SNE) developed by Maaten et al. [14] is an unsupervised nonlinear technique that is particularly suitable for mapping high-dimensional features into a feature vector of two or three dimensions. So the reduced features can be visualized in a scatter plot and present the characteristic pattern of the high-dimensional features.

Based on *K*-means and *t*-SNE, this paper proposes a new feature extraction method for intelligent fault diagnosis of machines. In the method, first, *K*-means is used to learn features from the raw vibration signals. Then, *t*-SNE is employed to reduce the dimensionality of the learned features. Since *K*-means and *t*-SNE are both unsupervised techniques, the proposed method releases us from the tasks of designing feature extraction algorithms and may reduce the human labor in intelligent fault diagnosis.

The rest of this paper is organized as follows. In Section 2, the algorithms of *K*-means and *t*-SNE are briefly described. Section 3 details the proposed method for automatic feature extraction. In Section 4, a bearing dataset containing massive samples is used to verify the effectiveness of the proposed method. Moreover, the proposed method is compared with manual feature extraction methods. The comparison results show the superiority of the proposed method. Conclusions are drawn in Section 5.

II. Theoretical Background

A. K-means Algorithm

K-means is a classical unsupervised clustering algorithm in intelligent fault diagnosis of machinery. It aims to partition *N* samples into *K* clusters where each sample belongs to the cluster with the nearest mean. Let $\{\boldsymbol{x}_i\}_{i=1}^{N}$ be the set of *N* samples to be clustered into the set of *K* clusters $\{c_k\}_{k=1}^{K}$ where $\boldsymbol{x}_i \in \Re^{M\times 1}$. Suppose that $\boldsymbol{u}_k$ is the mean of cluster c_k. The squared error between $\boldsymbol{u}_k$ and the points in cluster c_k is defined as

$$J(c_k) = \sum_{\boldsymbol{x}_i \in c_k} \left\| \boldsymbol{x}_i - \boldsymbol{u}_k \right\|^2. \tag{1}$$

The goal of *K*-means is to minimize (1) over all *K* clusters, so the squared error over *K* clusters is defined as

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in c_k} \left\| \boldsymbol{x}_i - \boldsymbol{u}_k \right\|^2. \tag{2}$$

Through minimizing this objective function, *K*-means can converge to its optimum. Based on this classic algorithm, Coates et al. [15] viewed *K*-means as a way of constructing a "dictionary" of *K* vectors so as to map the sample $\boldsymbol{x}_i$ to a code vector. So the modified objective function of *K*-means algorithm is optimized as follows.

$$\begin{aligned} &\underset{\boldsymbol{D},\boldsymbol{s}}{\text{minimize}} \ \sum_i \left\| \boldsymbol{D}\boldsymbol{s}_i - \boldsymbol{x}_i \right\|^2 \\ &\text{subject to } \left\| \boldsymbol{s}_i \right\|_0 \le 1, \\ &\qquad\qquad \left\| \boldsymbol{D}_k \right\|_2 = 1 \end{aligned} \tag{3}$$

where $\boldsymbol{s}_i \in \Re^{K\times 1}$ is a code vector for the sample $\boldsymbol{x}_i$, and $\boldsymbol{D}_k$ is the *k*th column of the dictionary $\boldsymbol{D} \in \Re^{M\times K}$. It is noted that $\left\| \boldsymbol{t} \right\|_p$ is ℓp-norm taking the form $\sqrt[p]{\left|t_1\right|^p + \cdots + \left|t_n\right|^p}$ supposing $\boldsymbol{t} = [t_1, t_2, \cdots, t_n]$. Actually, the trained $\boldsymbol{D}_k$ is the *k*th centroid of cluster c_k in the traditional view of *K*-means and $\boldsymbol{s}_i$ indicates which cluster $\boldsymbol{x}_i$ should belong to. The *k*th feature of $\boldsymbol{x}_i$ is calculated by a non-linear mapping function:

$$f_k = \max\left\{0, \frac{1}{K}\sum_{k=1}^{K} z_k - z_k\right\} \tag{4}$$

where $z_k = \left\| \boldsymbol{x}_i - \boldsymbol{D}_k \right\|_2$ and this function measures the distance between $\boldsymbol{x}_i$ and each centroid. So we can learn *K*-dimensional feature vector from each sample $\boldsymbol{x}_i$ using *K*-means. Since there is only one parameter, *K*-means does not necessarily include the hyper-parameter tuning and easily converges to an optimal solution. *K*-means has been used for unsupervised feature extraction in many fields, like image classification and computer vision, and obtains good performance. In this paper, we use it to learn features from raw vibration signals in an unsupervised manner.

B. T-SNE

T-SNE is an nonlinear dimensionality reduction technique that is used to embed high dimensional features into low dimensional features [14]. Given a set of *N* high-dimensional features $\{\boldsymbol{f}_i\}_{i=1}^{N}$, *t*-SNE first computes probabilities p_{mn} that are proportional to the similarity of features $\boldsymbol{f}_m$ and $\boldsymbol{f}_n$ as follows.

$$p_{n|m} = \frac{\exp(-\left\| \boldsymbol{f}_m - \boldsymbol{f}_n \right\|^2 / 2\sigma_m^2)}{\sum_{i\ne m} \exp(-\left\| \boldsymbol{f}_m - \boldsymbol{f}_i \right\|^2 / 2\sigma_m^2)} \tag{5}$$

$$p_{mn} = \frac{p_{n|m} + p_{m|n}}{2N} \tag{6}$$

The bandwidth of the Gaussian kernels σ_m is set in such a way that the perplexity of the conditional distribution equals a predefined perplexity using a binary search. As a result, the bandwidth is adapted to the density of the features, namely smaller values of σ_m are used in denser parts of the feature space. T-SNE aims to learn a d-dimensional map $\{\boldsymbol{y}_i\}_{i=1}^N$ that reflects the similarities p_{mn} as well as possible where $\boldsymbol{y}_i \in \Re^{d\times 1}$ is a low-dimensional feature. To achieve this goal, t-SNE measures similarities q_{mn} between two features in the map $\boldsymbol{y}_m$ and $\boldsymbol{y}_n$, which is defined as follows.

$$q_{mn} = \frac{\left(1+\|\boldsymbol{y}_m - \boldsymbol{y}_n\|^2\right)^{-1}}{\sum_{i\neq l}\left(1+\|\boldsymbol{y}_i - \boldsymbol{y}_l\|^2\right)^{-1}} \tag{7}$$

A heavy-tailed student-t distribution is used to measure similarities between low-dimensional features in order to allow dissimilar features to be modeled far apart in the map. The locations of the features $\boldsymbol{y}_i$ in the map are determined by minimizing the Kullback-Leibler (KL) divergence of the distribution q_{mn} from the distribution p_{mn}.

$$KL(P\|Q) = \sum_{m\neq n} p_{mn} \log \frac{p_{mn}}{q_{mn}} \tag{8}$$

where P and Q are the matrix form of p_{mn} and q_{mn}. By minimizing (8), the similarities between the high-dimensional feature $\boldsymbol{f}_i$ and the low-dimensional feature $\boldsymbol{y}_i$ are as similar as possible. So $\boldsymbol{y}_i$ could represent the characteristics of $\boldsymbol{f}_i$.

III. The Proposed Method

This section details the proposed method for automatic feature extraction of machinery. As shown in Fig. 1, the method can be described in five steps as follows.

1) Collect signals. The collection of raw vibration signals is the first step in the fault diagnosis process of rolling element bearings. The collected signals compose the sample set $\{\boldsymbol{x}^i\}_{i=1}^N$, where $\boldsymbol{x}^i \in \Re^{D\times 1}$ is the ith sample containing D data points.

2) Train K-means. After setting the parameter K, we randomly sample N_{seg} segments from the training samples and each segment contains M data points. Through minimizing (3), we can obtain the dictionary or the centroid matrix $\boldsymbol{D}$.

3) Get the local features of each sample. We first divide each sample $\boldsymbol{x}^i$ into a set of segments $\{\boldsymbol{x}_j^i\}_{j=1}^J$ where $J = D/M$. With the trained $\boldsymbol{D}$ and (4), we can get the feature $\boldsymbol{f}_j^i \in \Re^{1\times K}$ for $\boldsymbol{x}_j^i$, which is the local feature of the sample $\boldsymbol{x}^i$.

4) Obtain the learned feature of each sample. We combine the local features $\boldsymbol{f}_j^i$ into a feature vector $\boldsymbol{f}^i$ and this vector is the learned feature.

$$\boldsymbol{f}^i = \left[\boldsymbol{f}_1^i, \boldsymbol{f}_2^i, \cdots, \boldsymbol{f}_J^i\right]^{\mathrm{T}} \tag{9}$$

5) Reduce the dimension of the learned features. After obtaining the learned features $\boldsymbol{f}^i$, we use t-SNE to reduce its dimension. Through minimizing of the KL divergence in (8), the high-dimensional feature $\boldsymbol{f}_i$ can be mapped into low-dimensional feature $\boldsymbol{y}_i$.

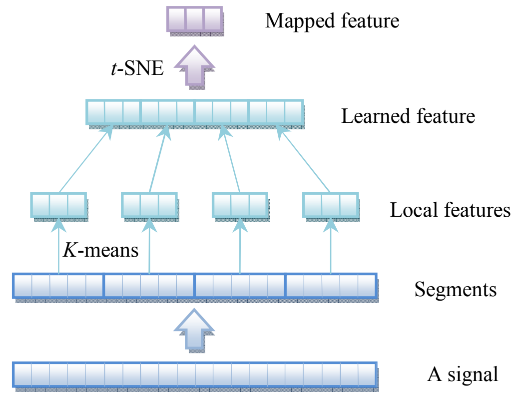

Fig. 1. Illustration of the proposed method.

IV. Fault Features Extraction of Bearings Using the Proposed Method

A. Data Description

The raw vibration signals of bearings were collected from a test bench. It is composed of a hydraulic motor, a hydraulic cylinder, and a hydraulic radial load application system. When tested, the bearings were installed in the hydraulic motor driven system and the hydraulic cylinder was used to add loads. Accelerometers were mounted on the load module to collect vibration signals with a sampling frequency of 12,800 Hz.

The bearing dataset has five different health conditions under two different loads. The five health conditions are: normal condition (NC), roller fault (RF), inner race fault (IF), slight rub fault in the outer race (OF1), serious flaking fault in the outer race (OF2). There are 273 samples for each health condition under one load. Each sample is a vibration signal containing 1,200 data points. So the bearing dataset totally has 2,730 samples.

B. Feature Extraction Result

The proposed method is used to learn the features from the samples of the five bearing health conditions. We first select 40,000 segments randomly from these samples to train K-means and the parameter K equals 25. The dimension of the segments is 100, and each segment actually is a signal containing 100 data ponits. After K-means is trained, we divide each sample into 12 segments alternately. For each segment, 25 dimensional features can be extracted by the trained K-means. So the learned features of each sample is a 300 dimensional feature vector. Using t-SNE, we reduce the dimension of each learned feature mapped into 2 dimensional feature vector, which could exploit the characteristic patterns of the bearing dataset. Finally, the mapped features are obtained by the proposed method, as shown in Fig. 2(a). It can be seen that most features of the same health condition are gathered in the corresponding cluster and most features of the different health conditions are separated. There is only one sample of OF2 clustered as the samples of IF, and there are three samples of NC clustered as the samples of OF1.

To demonstrate the effectiveness of the proposed method, three dimensionality reduction methods, which are commonly used in mechanical fault diagnosis, are employed to process the learned features of the bearing dataset. The three methods are PCA, locality preserving projection (LPP) [16], and Sammon mapping (SM) [17]. The mapped features by PCA for the bearing dataset are shown in Fig. 2(b). It is shown that the mapped features of NC and those of OF1 are mixed with each other. The reason may be that the fault of OF1 is slight, so the signals of the two faults are similar, which makes it difficult to distinguish the two health conditions using PCA. It is noticed that parts of the mapped features of RF are mixed with the mapped features of IF2 in Fig. 2(b). Actually these mapped features are separated in the three-dimensional plot, which indicates that the ability of the proposed method in dimensionality reduction is better than that of PCA. In Fig. 2(c), it displays the mapped features by LPP for the bearing dataset. It can be seen that the mapped features of NC and the mapped features of OF1 are clustered better than the results by PCA, but these mapped features are not well separated as well. Besides, the mapped features of RF are mixed with the mapped features of IF2 in this figure. The mapped features by SM for the bearing dataset are shown in Fig. 2(d). It is seen that the mapped features of each health conditions are separated better than the features mapped by PCA and LPP. However, the mapped features of OF2 are also separated into two clusters because of the two operation conditions under different loads. This is improper for mechanical fault diagnosis since we traditionally regard a health condition under different operation conditions as one class. Through the comparison above, it is verified that the proposed method is able to well exploit the characteristic patterns of the bearing dataset.

We also deal with the bearing dataset using the manual feature extraction methods. In this work, we use time-domain features combined with wavelet features to represent the characteristics of each health condition. These features are widely applied in bearing fault diagnosis and perform well [18, 19]. In detail, ten time domain features, i.e., mean value, peak

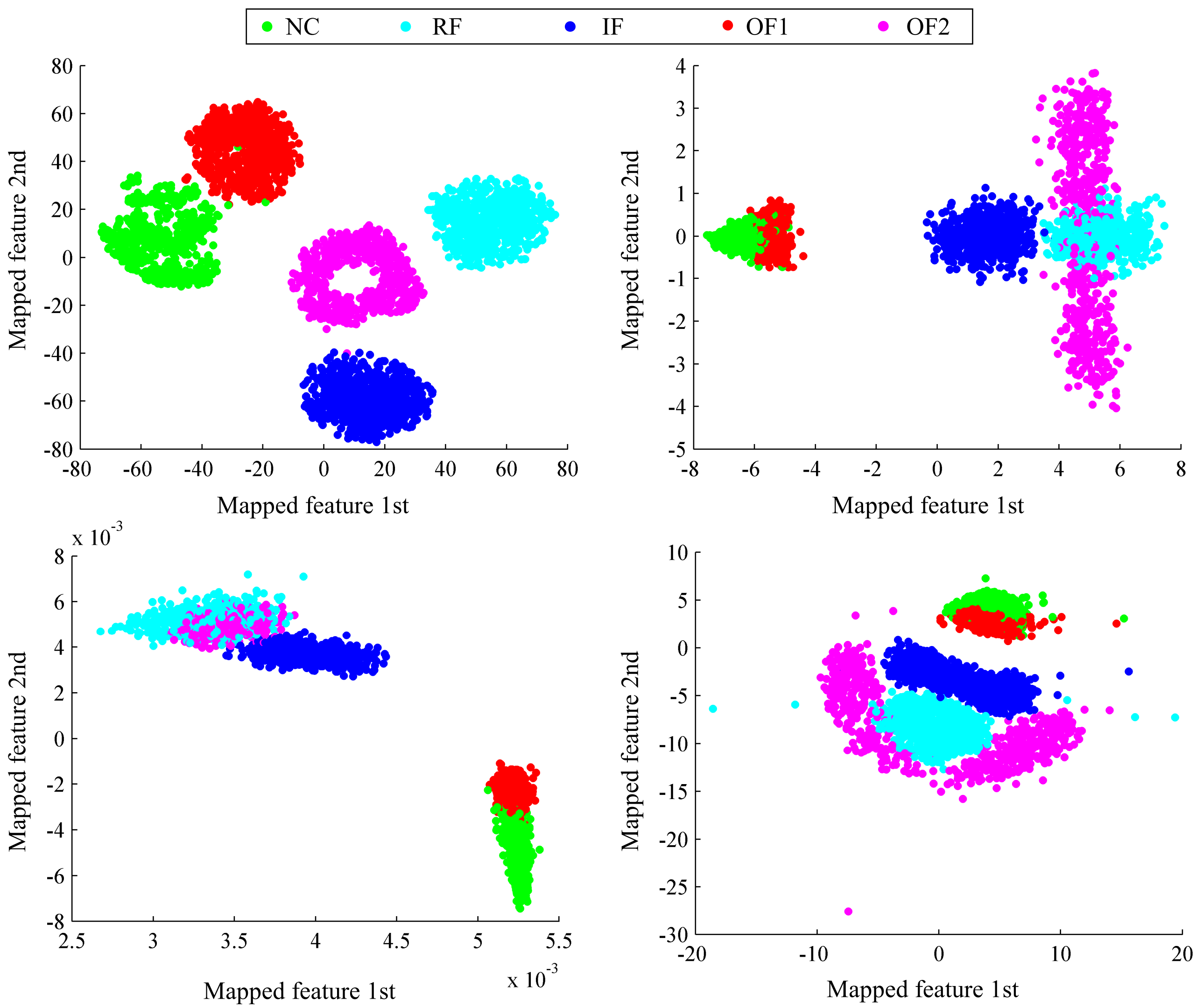

Fig. 2. Visualize features learned from raw vibration signals: (a) the proposed method, (b) *K*-means features and PCA, (c) *K*-means features and LPP and (d) *K*-means features and SM.

value, peak-to-peak value, root mean square, impulse factor, crest factor, waveform factor, clearance, skewness factor and kurtosis factor, and eight wavelet packet energy ratios are selected. So 18 manual features are obtained for each sample of the bearing dataset. To visualize these features, we employ *t*-SNE and PCA to map them into two-dimensional features, respectively. The mapped features are shown in Fig. 3. It can be seen that both the two dimensionality reduction methods fail to represent the characteristic pattern of the bearing dataset in two-dimensional plane. In Fig. 3(b), the mapped features of RF, IF and OF2 are mixed with each other. By comparing the results in Fig. 2(b), it shows that the learned features by *K*-means are more discriminative than those of the manual feature-based method. Through comparison between the results in Fig. 3 and the corresponding results in Fig. 2, the superiority of the proposed method is verified.

V. Conclusions

In this paper, we present a new method to automatically extract features from massive vibration signals. In the method, *K*-means first learns local features from the segments of each signal in an unsupervised way and the learned features are combined by these local features. Then, *t*-SNE mapped the dimensionality of the learned features into a low-dimensional feature vector. Through the experiment results, the superiority and effectiveness of the proposed method are verified.

The manual feature-based method requires prior knowledge about the characteristics of the collected signals and diagnostic expertise, whereas the proposed method is simple to perform and is able to learn the features automatically from the raw vibration signals. So the proposed method releases us from the tasks of designing feature extraction algorithms. Since this method is an unsupervised learning method, it could be applied to not just the bearing feature extraction in this paper but also the feature extraction of other machines.

As a tool in the automatic data mining in fault diagnosis, the proposed method could be combined with unsupervised clustering algorithms to automatically determine the health conditions of machines. This could offer a promising way to eliminate the need for human labor in intelligent fault diagnosis, which is definitely beneficial for dealing with massive signals collected. We will focus on this topic in future. Additionally, in the work, we have not investigated the effect of the feature number *K* when applying *K*-means for bearing feature extraction. But in our experiments, the feature extraction results are insensitive to the feature number. We will investigate the selection of *K* systematically in the next work.

Acknowledgment

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] K. Worden, W. J. Staszewski, and J. J. Hensman, “Natural computing for mechanical systems research: A tutorial overview,” *Mech. Syst. Signal Process*., vol. 25, no. 1, pp. 4-111, Jan. 2011.

[2] Y. Lei, Z. He, and Y. Zi, “A new approach to intelligent fault diagnosis of rotating machinery,” *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1593-1600, Nov. 2008.

[3] C. Shen, D. Wang, F. Kong, and W. T. Peter, “Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier,” *Measurement*, vol. 46, no. 4, pp. 1551-1564, May 2013.

[4] Y. Shatnawi, and M. Al-khassaweneh, “Fault diagnosis in internal combustion engines using extension neural network,” *IEEE Trans. Ind. Electron.*, vol. 61, no. 3, pp. 1434-1443, Mar. 2014.

[5] Y. Lei, Z. He, Y. Zi, and Q. Hu, “Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs,” *Mech. Syst. Signal Process*., vol. 21, no. 5, pp. 2280-2294, Jul. 2007.

[6] M. Amar, I. Gondal, and C. Wilson, “Vibration Spectrum Imaging: A Novel Bearing Fault Classification Approach,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 494-502, Jan. 2015.

[7] Y. Lei, Z. Liu, X. Wu, N. Li, W. Chen, and J. Lin, “Health condition identification of multi-stage planetary gearboxes using a mRVM-based

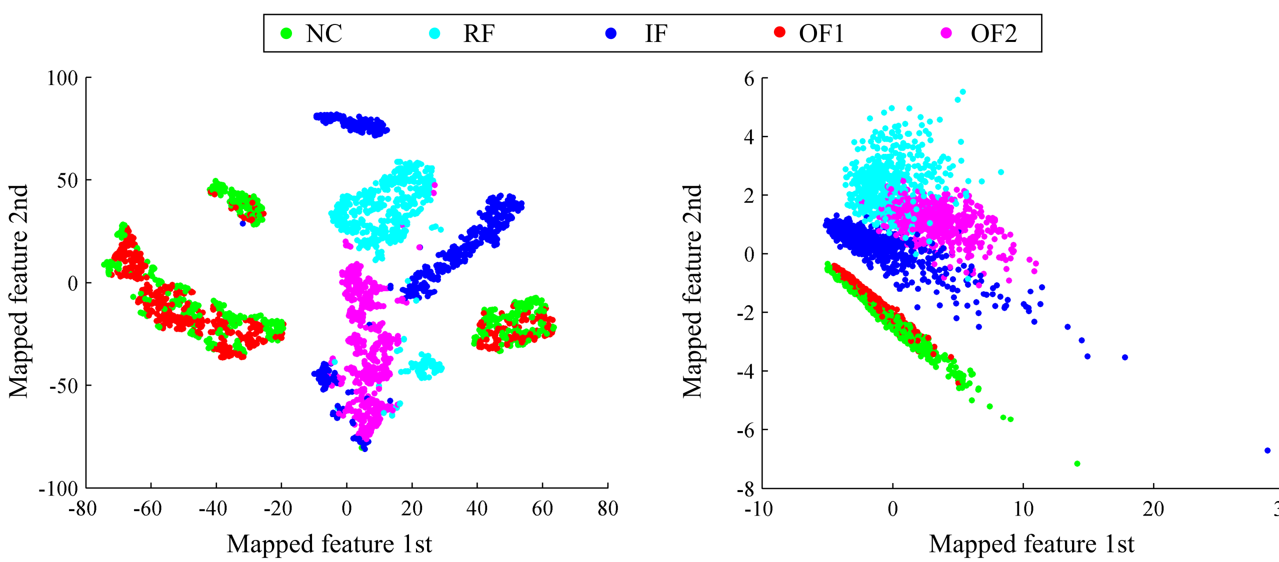

Fig. 3. Visualize the manual features: (a) *t*-SNE, (b) PCA.

method," *Mech. Syst. Signal Process.*, vol. 60, pp. 289-300, Aug. 2015.

[8] M. Kang, J. Kim, J.-M. Kim, A. C. Tan, E. Y. Kim, and B.-K. Choi, "Reliable Fault Diagnosis for Low-Speed Bearings Using Individually Trained Support Vector Machines With Kernel Discriminative Feature Analysis," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2786-2797, May 2015.

[9] J. Seshadrinath, B. Singh, and B. Panigrahi, "Vibration analysis based interturn fault diagnosis in induction machines," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 340-350, Feb. 2014.

[10] X. Zhang, Y. Liang, and J. Zhou, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164-179, Jun. 2015.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[12] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," In *Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 511-516.

[13] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. AISTATS*, 2011, pp. 215-223.

[14] L. Van der Maaten, and G. Hinton, "Visualizing data using t-SNE," *J. Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.

[15] A. Coates, and A. Y. Ng, "Learning feature representations with k-means," *Neural Networks: Tricks of the Trade*, K.-R. Mu ¨ ller, G. Montavon, and G.B. Orr, eds., Springer 2013.

[16] J. Yu, "A nonlinear probabilistic method and contribution analysis for machine condition monitoring," *Mech. Syst. Signal Process.*, vol. 37, no. 1, pp. 293-314, May 2013.

[17] X. Wang, Y. Zheng, Z. Zhao, and J. Wang, "Bearing Fault Diagnosis Based on Statistical Locally Linear Embedding," *Sensors*, vol. 15, no. 7, pp. 16225-16247, Jul. 2015.

[18] Y. Lei, Z. He, and Y. Zi., "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9941- 9948, Aug 2009.

[19] H. Zhou, J. Chen, G. Dong, H. Wang, and H. Yuan, "Bearing fault recognition method based on neighbourhood component analysis and coupled hidden Markov model," *Mech. Syst. Signal Process.*, vol. 66-67, pp. 568-581, Jan. 2016.