Springer

KSME

Journal of Mechanical Science and Technology 28 (11) (2014) 4441~4454

www.springerlink.com/content/1738-494x
DOI 10.1007/s12206-014-1012-7

Fault diagnosis of rotating machinery based on kernel density estimation and Kullback-Leibler divergence[†]

Fan Zhang, Yu Liu[*], Chujie Chen, Yan-Feng Li and Hong-Zhong Huang

School of Mechanical, Electronic, and Industrial Engineering,
University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China

(Manuscript Received February 6, 2014; Revised July 2, 2014; Accepted July 28, 2014)

Abstract

Based on kernel density estimation (KDE) and Kullback-Leibler divergence (KLID), a new data-driven fault diagnosis method is proposed from a statistical perspective. The ensemble empirical mode decomposition (EEMD) together with the Hilbert transform is employed to extract 95 time- and frequency-domain features from raw and processed signals. The distance-based evaluation approach is used to select a subset of fault-sensitive features by removing the irrelevant features. By utilizing the KDE, the statistical distribution of selected features can be readily estimated without assuming any parametric family of distributions; whereas the KLID is able to quantify the discrepancy between two probability distributions of a selected feature before and after adding a testing sample. An integrated Kullback-Leibler divergence, which aggregates the KLID of all the selected features, is introduced to discriminate various fault modes/damage levels. The effectiveness of the proposed method is demonstrated via the case studies of fault diagnosis for bevel gears and rolling element bearings, respectively. The observations from the case studies show that the proposed method outperforms the support vector machine (SVM)-based and neural network-based fault diagnosis methods in terms of classification accuracy. Additionally, the influences of the number of selected features and the training sample size on the classification performance are examined by a set of comparative studies.

Keywords: Data-driven fault diagnosis; Kernel density estimation; Kullback-Leibler divergence; Ensemble empirical mode decomposition

1. Introduction

Rotating machinery has broad applications in engineering practices, like wind turbines, power generators, aircraft engines, etc. The components, such as bearings, gears, in rotating machinery are usually subjected to undesirable stresses and sudden shocks under which defects or degradations will gradually increase and eventually cause severe damage and unexpected shutdown of the entire system. Unexpected downtime and failures of critical systems oftentimes lead to extra cost due to production delay, unplanned corrective maintenance activities or fatal risk to humans [1]. It is, therefore, of paramount importance to accurately detect the presence of faults as early as possible to avoid the consequence of severe damages caused by faults and also facilitate preventive maintenance planning before the complete failure of engineering systems [2]. On the other hand, rotating machinery and its components oftentimes suffer several fault modes or damage levels, so correctly identifying fault types and/or damage levels can provide engineers comprehensive knowledge of the health status of the monitored system. Hence, developing an effective and reliable fault diagnostic technique to identify various sorts of fault modes and damage levels at their incipient stage becomes extremely important.

In general, existing fault diagnosis methods can be classified into two categories [3]: model-based methods and data-driven methods. Samuel and Pines [4] demonstrated the implementation of these two types of fault diagnosis methods in a helicopter transmission system. Isermann [5] reviewed the development process of model-based method and data-driven method with three applications: DC motors, buses, and diesel engines. Theoretically, the model-based fault diagnosis is to determine a fault in a system by comparing available system measurements with a priori information represented by the system's analytical/mathematical model [6]. Ideally, model-based methods will be very effective if the analytical/mathematical model associated a specific fault can be accurately constructed. In recent years, with the development of intelligent computing technology, data-driven methods have received much attention. In these methods, fault diagnosis is realized by mapping the fault space to the feature space [7]. Put another way, the underlying relationship between features extracted from condition monitoring data and fault

[*] Corresponding author. Tel.: +86 28 61830229, Fax.: +86 28 61830227
E-mail address: yuliu@uestc.edu.cn

[†] Recommended by Associate Editor Cheolung Cheong

modes/damage levels can be learned and constructed solely based on a set of historical data (also called training data). The data-driven methods therefore possess two distinctive advantages over the model-based methods: Fault diagnosis can be carried out automatically by the data-driven methods without requiring too much assistance from engineers, and as opposed to the model-based methods which require professional expertise to make judgments, the data-driven methods do not heavily rely on experience and knowledge from experts [8]. In general, both the model-based and data-driven methods have their own advantages. It is a case-dependent problem to choose one of them or use both. If the system's analytical/mathematical model with respect to a specific fault is readily to be modeled, the model-based methods may be preferred, otherwise the data-driven methods provide an alternative way to reveal the input (extracted features)-output (faults types/ damage levels) relationship.

In most cases, a data-driven diagnosis method consists of five basic elements as shown in Fig. 1 [9]. The raw data, say vibration signals of rotating machinery, collected from condition monitoring program serve as inputs of a data-driven fault diagnosis method. It is followed by the feature extraction, one of the critical steps in a data-driven diagnosis method, to extract a bunch of features from raw signal data [10]. These extracted features are more or less related to the health status of the monitored device. Many advanced signal processing algorithms, such as fast Fourier transform (FFT) [2], empirical mode decomposition (EMD) [11], wavelet transform, Hilbert transform (HT) [12], can be used at this stage to extract a set of features reflecting various types of faults and/or damage levels. Nevertheless, not all the extracted features are sensitive to every type of fault and/or damage level. Eliminating irrelevant features in the original feature set and retaining the most sensitive features related to specific types of faults and/or damage levels can not only significantly reduce the computational cost in ensuing fault classification, but also improve the classification accuracy [13, 14]. Such a task is accomplished in the stage of feature selection. Fault classification in step four can be essentially viewed as a mapping from selected features to specific fault modes/damage levels. The fault modes/damage levels can be therefore identified by the values of selected features. With the development of computational intelligence techniques, many advanced classification methods have been applied to data-driven fault diagnosis [7, 15-17]. Among them, support vector machine [18] and artificial neural network (ANN) [19] are two representative and powerful classification methods, and they have been extensively used in fault diagnosis for rotating machinery [20-22].

It is noteworthy that most reported data-driven fault diagnosis methods for rotating machinery, like SVM-based methods [20], ANN-based methods [22], and K-nearest neighbor methods [23], are seeking an optimal classification hyperplane or a set of the best thresholds and weights to divide all the training/testing samples into difference classes in a sample space via deterministic approaches rather than a probabilistic method. However, due to noises, measurement errors, or other uncontrollable variations, even the training/testing samples belonging to the same fault modes/damage levels may exhibit a certain degree of uncertainty. Such uncertainty among the samples within the same fault modes/damage levels can be quantified by statistic tools. The fault diagnosis can be thereby realized by examining whether the condition monitoring data possess identical statistical characteristics with the training samples. With this idea in mind, we developed a new data-driven fault diagnosis method. Following the familiar framework as shown in Fig. 1, the proposed data-driven fault diagnosis method consists of five steps as depicted in Fig. 2. The ensemble empirical mode decomposition (EEMD) developed by Wu and Huang to alleviate the mode mixing problem of the empirical mode decomposition (EMD) [24, 25], is utilized in this work to extract both time-domain and frequency-domain features from the raw vibration signals of rotating machinery. However, since some of the extracted features may not be sensitive to some specific fault modes/damage levels, the distance-based feature evaluation approach [9] is employed to remove the irrelevant features in this study. Instead of using conventional computational intelligence methods, like SVM and ANN etc., we used two statistical approaches, i.e., the kernel density estimation (KDE) and the Kullback-Leibler divergence (KLID), together to identify fault modes/damage levels from a statistical point of view. The KDE is a nonparametric probability density estimation approach in statistics, and it is able to adaptively fit any data set to a smooth density function without the restriction of distribution type [26, 27]. On the other hand, the KLID (also called information divergence, relative entropy) is a measure of the difference between two probability distributions [28]. By aggregating the KDE and KLID of all the selected features, an integrated Kullback-Leibler divergence, acting as an indicator, is proposed to identify faults modes/damage levels.

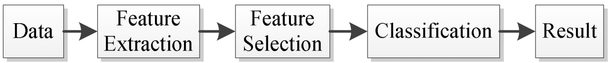


Fig. 1. The basic procedure of data-driven fault diagnosis methods [9].

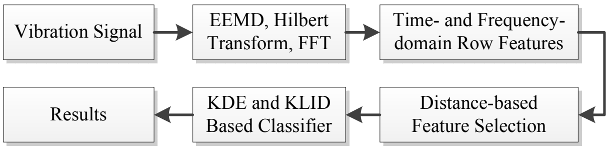

Vibration Signal
EEMD, Hilbert Transform, FFT
Time- and Frequency-domain Row Features
Distance-based Feature Selection
KDE and KLID Based Classifier
Results

Fig. 2. The flow chart of the proposed data-driven fault diagnosis method.

The rest of this paper is organized as follows: Sec. 2 introduces the specific feature extraction and feature selection methods used in our study. The principles of the KDE and KLID are reviewed in Sec. 3, followed by an elaboration of the new data-driven fault diagnosis method. The effectiveness of the proposed method is demonstrated in two case studies of fault diagnosis for bevel gears and rolling element bearings in

Sec. 4, along with a set of comparative studies. Sec. 5 is a brief closure.

2. Feature extraction and selection

2.1 EEMD method

The ensemble empirical mode decomposition (EEMD) is an improved empirical mode decomposition (EMD) to alleviate the mode mixing problem. The EEMD defines the true intrinsic mode function (IMF) components as the mean of an ensemble of trials. Each trial consists of decomposition results of signals which are artificially added by a white noise with finite amplitude [24]. By using the statistical properties of white noise, the EEMD method makes the original EMD a more effective self-adaptive dyadic filter bank in signal processing.

Based on the principle and observations of white noises [24], applying the EEMD on signal $x(t)$ follows four basic steps [25]:

Step 1: Initialize the number of ensembles M and the amplitudes of white noises to be added, and set the index of trial $m = 1$.

Step 2: Perform the mth trial on the white noise-added signals. It contains three procedures:

(a) Add the white noise series with a given amplitude to the investigated signals by:

$$x_m(t) = x(t) + n_m(t) , \tag{1}$$

where $n_m(t)$ is the mth white noise series to be added, and $x_m(t)$ represents the noise-added signals of the mth trial.

(b) Decompose the noise-added signal $x_m(t)$ into I IMFs $c_{i,m}$ $(i = 1,2,\cdots,I)$ using the EMD method developed in Ref. [11], where $c_{i,m}$ denotes the ith IMF of the mth trial, and I is the total number of IMFs.

(c) If $m \ge M$, go to Step 3; otherwise go back to Step 2 and set $m = m+1$, but the white noise series added to the studied signal is different in the next iteration.

Step 3: Compute the ensemble mean c_i of the M trials for each IMF by:

$$c_i = \frac{1}{M}\sum_{m=1}^{M} c_{i,m}\ (i = 1,2,\cdots,I, m = 1,2,\cdots,M) . \tag{2}$$

Step 4: Treat the mean $c_i (i = 1,2,\cdots,I)$ of each of the I IMFs as the final IMFs.

After performing the EEMD, we can get a collection of I IMFs $c_i (i = 1,2,\cdots,I)$ and a residual signal r_I indicating the mean trend of $x(t)$. Each IMF represents an oscillating signal which is much simpler than the raw signal $x(t)$. In essence, the raw signal could be reproduced by the IMFs and the residual signal as follows:

$$x(t) = \sum_{i=1}^{I} c_i + r_I , \tag{3}$$

where the IMFs $c_1, c_2, \cdots, c_I$ contain a wide range of frequency bands from high to low. Also, the EEMD method is self-adaptive as different IMF components will change with the raw signal.

Note that the amplitude of the white noise to be added and the number of ensemble are two crucial parameters in the EEMD method. The relationship of the amplitude of the white noise to be added and the number of ensemble M can be formulated as [25]:

$$e_{\text{std}} = \frac{a}{\sqrt{M}} , \tag{4}$$

where a is the amplitude of the added white noise; e_{std} is the standard deviation of errors. Based on Eq. (4), a smaller a may lead to a smaller standard deviation of errors. Nevertheless, if the amplitude of the added noise is very small, it may not cause the change of extrema on which the EEMD method relies. On the other hand, increasing the number of ensemble M will also cause the reduction of the standard deviation of errors. Generally, an ensemble number of a few hundred will lead to an exact result, and the amplitude of the white noise is oftentimes set to be 0.1~0.4 times of the standard deviation of the investigated signal [24].

2.2 Feature extraction

Features extracted from the raw signals serve as the inputs of classifier of data-driven fault diagnosis methods. A large multitude of features including time- and frequency-domain features, can be extracted from raw and processed signals. Utilizing both time- and frequency-domain features in the data-driven fault diagnosis methods can comprehensively reflect the time- and frequency- distributions of signals. In our study, we define 19 time- and frequency-domain features and extract them from the raw signals and the signals preprocessed by the EEMD and the Hilbert transform [29].

Time-domain features we chose include both dimensional and dimensionless features. The dimensional feature may significantly vary with respect to the load imposed on rotating machinery and the rotary speed. The dimensionless feature, on the other hand, is insensitive to the load and rotary speed and can directly reflect the characteristics of faulty rotating machinery. Nine time-domain features ($p_{t1} \sim p_{t9}$) defined in our work are tabulated in Table 1. The features p_{t2} and p_{t3} reflect the vibration amplitude and energy in time domain; whereas features p_{t1} and $p_{t4} \sim p_{t9}$ characterize the distribution of signals in time domain [12]. These nine features are exacted for the raw signals and the first four IMFs. Thus, we can get 5×9 time-domain features in total.

In most cases, the frequency spectrum of signals and its distribution may change when a fault occurs in rotating machinery. The energy of some frequency components which are related to a specific fault will increase. Here we define ten frequency-domain features (p_{f1}~p_{f10}) as tabulated in Table 1. The fea-

Table 1. The time- and frequency- domain features.

Time domain features		Frequency domain features	
$p_{t1}=\sqrt{\frac{\sum_{n=1}^{N}\left(x(n)-1/N\sum_{n=1}^{N}x(n)\right)^2}{N-1}}$	$p_{t2}=\left(\frac{\sum_{n=1}^{N}\sqrt{\lvert x(n)\rvert}}{N}\right)^2$	$p_{f1}=\frac{\sum_{k=1}^{K}s(k)}{K}$	$p_{f2}=\frac{\sum_{k=1}^{K}\left(s(k)-p_{f1}\right)^2}{N-1}$
$p_{t3}=\sqrt{\frac{\sum_{n=1}^{N}\left(x(n)\right)^2}{N}}$	$p_{t4}=\frac{\sum_{n=1}^{N}\left(x(n)-1/N\sum_{n=1}^{N}x(n)\right)^3}{(N-1)p_{t1}^3}$	$p_{f3}=\frac{\sum_{k=1}^{K}f_k s(k)}{\sum_{k=1}^{K}s(k)}$	$p_{f4}=\sqrt{\frac{\sum_{k=1}^{K}\left(f_k-p_{f3}\right)^2 s(k)}{K}}$
$p_{t5}=\frac{\sum_{n=1}^{N}\left(x(n)-1/N\sum_{n=1}^{N}x(n)\right)^4}{(N-1)p_{t1}^4}$	$p_{t6}=\frac{\max\lvert x(n)\rvert}{p_{t3}}$	$p_{f5}=\sqrt{\frac{\sum_{k=1}^{K}f_k^2 s(k)}{\sum_{k=1}^{K}s(k)}}$	$p_{f6}=\sqrt{\frac{\sum_{k=1}^{K}f_k^4 s(k)}{\sum_{k=1}^{K}f_k^2 s(k)}}$
$p_{t7}=\frac{\max\lvert x(n)\rvert}{p_{t2}}$	$p_{t8}=\frac{p_{t3}}{\frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert}$	$p_{f7}=\frac{\sum_{k=1}^{K}f_k^2 s(k)}{\sqrt{\sum_{k=1}^{K}s(k)\sum_{k=1}^{K}f_k^4 s(k)}}$	$p_{f8}=\frac{p_{f4}}{p_{f3}}$
$p_{t9}=\frac{\max\lvert x(n)\rvert}{\frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert}$		$p_{f9}=\frac{\sum_{k=1}^{K}\left(f_k-p_{f3}\right)^3 s(k)}{Kp_{f4}^3}$	$p_{f10}=\frac{\sum_{k=1}^{K}\left(f_k-p_{f3}\right)^4 s(k)}{Kp_{f4}^4}$
where $x(n)$ ($n=1,2,\cdots N$) is the time series of signals; N is the number of data points.		where $s(k)$ is a spectrum of $x(n)$, ($k=1,2,\cdots K$); K is the number of spectrum lines; f_k is the frequency value of the *k*th spectrum line.	

ture p_{f1} indicates the vibration energy. The features p_{f2}, p_{f4}, and $p_{f8}\sim p_{f10}$ reflect the degree of concentration of the spectrum. The features p_{f3}, $p_{f5}\sim p_{f7}$ represent the main frequency bands of signals [12]. Ten frequency-domain features are extracted from the Fourier spectrum of the raw signals, and the other 40 frequency-domain features are extracted from the Hilbert envelope of the first four IMFs of the raw signal. In total, we can acquire 50 frequency-domain features.

Even though we only chose a set of 95 features (45 time-domain features and 50 frequency-domain features) in our study, other features extracted by different equations or from different processed signals (say higher order IMFs) can be also included in the feature set.

2.3 Distance-based feature selection

Nevertheless, not all the extracted features have equal contributions to faults/damage levels classification. Only some of these features are sensitive to the change of health condition of rotating machinery. The features which are insensitive to the occurrence of faults or damage levels are called irrelevant features. Removing these irrelevant features before conducting classification not only enhances the accuracy of fault diagnosis but also improves the computational efficiency of classification algorithms [13]. In this work, the distance-based evaluation approach proposed in Ref. [9] is used to choose some of the most effective features from the entire feature set, i.e. 95 features. The distance-based evaluation approach is one of the most popular feature selection methods in parallel with the Pearson correlation coefficient, Fisher discriminant ratio (FDR), and information gain, etc. [30] The basic idea behind the distance-based evaluation approach is that when samples are characterized by features, a smaller distance among samples within the same class is better, and a greater distance between different classes is more favorable. Ranking features by the distance-based evaluation approach consists of four steps [9]:

Step 1: Evaluate the average distance of the *j*th feature of training samples belonging to the *c*th class. It can be computed by:

$$\begin{cases} d_{c,j}=\dfrac{1}{M_c\times\left(M_c-1\right)}\sum_{l=1}^{M_c}\sum_{\substack{m=1\\ m\neq l}}^{M_c}\left|q_{m,c,j}-q_{l,c,j}\right| \\ j=1,2,\cdots,J;c=1,2,\cdots,C \end{cases}, \tag{5}$$

where M_c stands for the number of samples belonging to the *c*th class; *J* is the size of a feature set; $q_{m,c,j}$ is the value of the *j*th feature of the *m*th sample in the *c*th class. The average distance $d_j^{(w)}$ of the *j*th feature belonging to all the *C* classes is given by:

$$d_j^{(w)}=\frac{1}{C}\sum_{c=1}^{C}d_{c,j}\ . \tag{6}$$

Step 2: Compute the average value of the *j*th feature of the M_c samples in the *c*th class by:

$$u_{c,j} = \frac{1}{M_c}\sum_{m=1}^{M_c} q_{m,c,j} \ , \tag{7}$$

and then evaluate the average distance $d_j^{(b)}$ of the *C* different classes.

$$d_j^{(b)} = \frac{1}{C\times\left(C-1\right)}\sum_{c=1}^{C}\sum_{\substack{e=1\\ e\neq c}}^{C}\left|u_{e,j}-u_{c,j}\right| . \tag{8}$$

where *c* and *e* represent the indices of two different classes.

Step 3: Assess the effectiveness factor of the *j*th feature by:

$$\alpha_j = \frac{d_j^{(b)}}{d_j^{(w)}} . \tag{9}$$

Step 4: Rank all the features by the value of effectiveness factor α_j. The feature with a greater effectiveness factor is preferred.

After ranking all the features, it is necessary to determine how many features should be selected from the feature set. In general, two criteria can be used here: (1) Sequentially adding the feature to the classifier from the one with the greatest effectiveness factor until the accuracy of classification reaches a pre-set threshold, say 95%~100%. (2) The number of selected features is sufficient if adding extra features cannot lead to any increment of accuracy.

3. Fault classification based on KDE and KLID

3.1 Kernel density estimation

The kernel density estimation (KDE) was originally proposed by Rosenblatt and Parzen [26, 27], and it is also called the Parzen-Rosenblatt window method in some fields such as signal processing and econometrics. In statistics, the KDE is a non-parametric technique to estimate the probability density function of a data set, and it originates from the empirical probability density function.

Let $X_1, X_2, \cdots, X_n$ represent *n* independent and identically distributed (i.i.d.) random samples from a random quantity X with an unknown probability density function $f(x)$. The kernel density function is defined as [31]:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{10}$$

where $K(\bullet)$, a symmetric function with integration equal to 1, is the kernel function. The kernel function may not be necessarily a position function, but has to guarantee $\hat{f}_h(x)$ satisfies the basic requirement of a probability density function; *h*>0 is the bandwidth. Many different types of kernel functions have been proposed in Ref. [32]; some commonly used kernel functions are presented in Table 2.

Table 2. The commonly used kernel functions.

Types	Formula
Uniform (or box)	$\frac{1}{2}I\left(\lvert u\rvert\le 1\right)$
Triangle	$\left(1-\lvert u\rvert\right)I\left(\lvert u\rvert\le 1\right)$
Epanechnikov	$\frac{3}{4}\left(1-u^2\right)I\left(\lvert u\rvert\le 1\right)$
Quaritic	$\frac{15}{16}\left(1-u^2\right)^2 I\left(\lvert u\rvert\le 1\right)$
Triweight	$\frac{35}{32}\left(1-u^2\right)^3 I\left(\lvert u\rvert\le 1\right)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$
Cosinus	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)I\left(\lvert u\rvert\le 1\right)$

In Table 2, $I(\bullet)$ is an indicator function, i.e., if $|u|\le 1$ is true, $I\left(|u|\le 1\right)=1$; otherwise $I\left(|u|\le 1\right)=0$. Among these functions, the Gaussian kernel function possesses many mathematical properties, such as centrality and gradual decay, and has been broadly adopted.

The bandwidth h of the kernel function is an arbitrary parameter which exhibits a strong influence on the smoothness of $\hat{f}_h(x)$. A larger *h* indicates that a greater region of samples around the center point x influences the probability density estimation. That is, the difference of function values between the points closer to the center point *x* and points farther to *x* is smaller, and $\hat{f}_h(x)$ will be a smooth curve; otherwise, $\hat{f}_h(x)$ will be an unsmoothed broken line. However, a smaller *h* indicates that a narrower region of samples is taken into account when estimating the probability density at any place. Thereby, choosing a proper value for the bandwidth *h* is of great importance in the KDE. The most common optimality criterion to select the bandwidth is the mean integrated squared error (MISE) defined as [31]:

$$MISE\left(h\right) = E\int\left(\hat{f}_h\left(x\right) - f\left(x\right)\right)^2 dx . \tag{11}$$

Under weak assumptions on $f(\bullet)$ and $K(\bullet)$ [26, 27], one can get:

$$MISE\left(h\right) = AMISE\left(h\right) + o\left(\frac{1}{nh} + h^4\right), \tag{12}$$

where $o(\bullet)$ is infinitesimal. The *AMISE* is the asymptotic MISE defined as:

$$AMISE\left(h\right) = \frac{R\left(K\left(\bullet\right)\right)}{nh} + \frac{1}{4}m_2\left(K\left(\bullet\right)\right)^2 h^4 R\left(f''\left(\bullet\right)\right), \tag{13}$$

where $R(g)=\int g(x)^2 dx$; $m_2(K)=\int x^2 K(x)dx$; $f''(\bullet)$ is the second-order derivative of $f(\bullet)$; *n* is the total number of samples. The following differential equation can be used to seek the minimal value of the *AMISE* as:

$$\frac{\partial}{\partial h} AMISE(h) = -\frac{R(K(\bullet))}{nh^2} + m_2(K(\bullet))^2 h^3 R(f''(\bullet)) = 0\,. \tag{14}$$

Thus, the minimal value of h is:

$$h^*_{AMISE} = \frac{R(K(\bullet))^{1/5}}{m_2(K(\bullet))^{2/5} R(f''(\bullet))^{1/5} n^{1/5}}\,. \tag{15}$$

Apparently, the above equation cannot be used directly since it is implicit and contains the unknown density function $f(\bullet)$ or its second-order derivative $f''(\bullet)$. In many engineering applications, if the Gaussian basis function is used to approximate univariate data, the underlying density to be estimated is also Gaussian. In such case, based on Eq. (16), one can get the optimal value of *h* as:

$$h^*_{AMISE} = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}\,, \tag{16}$$

where $\hat{\sigma}$ is the standard deviation of samples. Such approximation is named the Gaussian approximation, and it is employed in our work.

3.2 Kullback-Leibler divergence

The Kullback-Leibler divergence (KLID) was first introduced in 1951 [28], and has been applied to quantify the difference of two distributions. For two discrete probability distributions *P* and *Q*, the KLID of *Q* from *P* is written as:

$$D_{KL}(P \,||\, Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i)\,. \tag{17}$$

In essence, Eq. (17) is the expectation of the logarithmic difference between the probabilities *P* and *Q*, and the expectation is taken by the probability *P*. The KLID is valid if the integration of *P* and *Q* are both equal to 1. If $Q(i)=0$, then $P(i)=0$ for all *i*. For the case where $P(i)=0$ and $P(i)/Q(i)=0$, $\ln(P(i)/Q(i))P(i)=0$ since $\lim_{x\to 0} x\ln(x)=0$.

Although the KLID is usually interpreted as a distance between two probability distributions, it does not completely satisfy the properties of distance measure, such as symmetry and triangle inequality. For example, the KLID of *P* from *Q* is generally not the same as the KLID of *Q* from *P*. On the other hand, the KLID is always non-negative. Based on the Gibbs' inequality, $D_{KL}(P \,||\, Q)=0$ if and only if $P=Q$ holds almost everywhere. Based on the definition of the KLID, a smaller value of $D_{KL}(P \,||\, Q)$ implies a higher similarity between the distributions of *P* and *Q*.

The symmetry property is very crucial in the classification issue. Therefore, in our work, the symmetrized distance of KLID is used as a measure to quantify the difference/distance between two distributions. The symmetrized distance between the distributions *P* and *Q* is defined as [28]:

$$D_{KL}(P,Q) = \frac{1}{2}\left[D_{KL}(P \,||\, Q) + D_{KL}(Q \,||\, P)\right]. \tag{18}$$

3.3 Fault classification based on KDE and KLID

3.3.1 The proposed fault diagnosis method

In this section, feature extraction, feature selection, kernel density estimation, and Kullback-Leibler divergence introduced earlier will be integrated together to realize fault diagnosis for rotating machinery. Some important symbols to be used hereinafter are explained here:

(1) KD_i^j $(j=1,2,\cdots,n;\ i=1,2,\cdots,C)$ denotes the KDE function of the *j*th feature of the training samples for type i fault. The vector $\mathbf{KD}_i=(KD_i^1,KD_i^2,\cdots,KD_i^n)$ is the KDE function set of all the n selected features of the training samples for type i fault;

(2) TKD_i^j $(j=1,2,\cdots,n;i=1,2,\cdots,C)$ is the KDE function of the *j*th feature of the training samples for type i fault after adding a testing sample. The vector $\mathbf{TKD}_i=(TKD_i^1, TKD_i^2,\cdots,TKD_i^n)$ is the KDE function set of all the n selected features of the training samples for type i fault after a testing sample is added;

(3) KL_i^j $(j=1,2,\cdots,n;\ i=1,2,\cdots,C)$ is the KLID between KD_i^j and TKD_i^j. The vector $\mathbf{KL}_i=(KL_i^1,KL_i^2,\cdots,KL_i^n)$ contains the KLIDs of all the n selected features.

The overall flowchart of the proposed classification method is in Fig. 3. As shown, training sample sets from two types of fault modes (types I and II faults) and one testing sample sets are used here to illustrate the rationale of the proposed method in classifying two fault modes. Nine time-domain features along with ten frequency-domain features are extracted from the raw signal and the first four IMFs obtained by the EEMD, and thus 95 features are acquired. Afterwards, the distance-based evaluation approach is applied to assess all the features and get the effectiveness factor α_j of the *j*th ($j=1,2,\cdots,95$) feature. By sorting all the features from the one with the largest effectiveness factor to the lowest one, the first *n* features are selected from the original feature set and serve as the inputs of the ensuing classifier. It is noted that the value *n* of selected features may have influence on the results, and it will be discussed in Sec. 4.3.1. Therefore, the importance of the *j*th feature to the fault classification is defined as:

$$F_j = \frac{\alpha_j}{\sum_{i=1}^{n} \alpha_i}\,,\ (j=1,2,\cdots,n)\,. \tag{19}$$

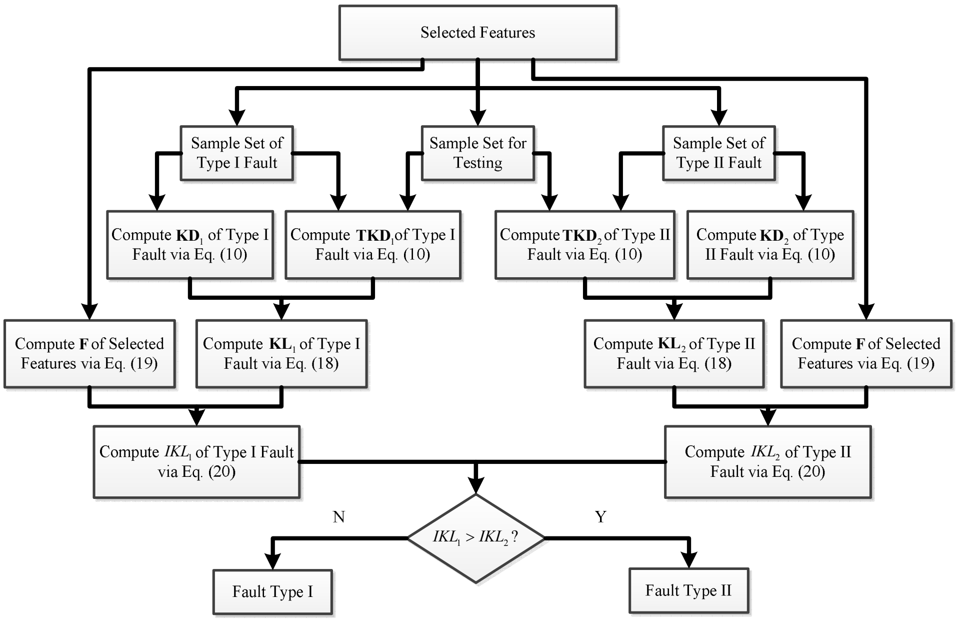

Selected Features
Sample Set of Type I Fault
Sample Set for Testing
Sample Set of Type II Fault
Compute $\mathbf{KD}_1$ of Type I Fault via Eq. (10)
Compute $\mathbf{TKD}_1$ of Type I Fault via Eq. (10)
Compute $\mathbf{TKD}_2$ of Type II Fault via Eq. (10)
Compute $\mathbf{KD}_2$ of Type II Fault via Eq. (10)
Compute $\mathbf{F}$ of Selected Features via Eq. (19)
Compute $\mathbf{KL}_1$ of Type I Fault via Eq. (18)
Compute $\mathbf{KL}_2$ of Type II Fault via Eq. (18)
Compute $\mathbf{F}$ of Selected Features via Eq. (19)
Compute IKL_1 of Type I Fault via Eq. (20)
Compute IKL_2 of Type II Fault via Eq. (20)
N
Y
$IKL_1 > IKL_2$?
Fault Type I
Fault Type II

Fig. 3. The flowchart of the proposed fault diagnosis method based on the KDE and KLID.

The kernel density function is used to characterize the probability density of the selected features of each training set. For example, based on the definition of KD_i^j, one can get KD_1^1 and KD_2^1 for the first feature of type I and type II faults respectively as shown in Fig. 3. One sample from the testing sample set is added into the two training sets, respectively, and the kernel density functions of the first feature of the two new sample sets are estimated as well and denoted as TKD_1^1 and TKD_2^1. In the same manner, KD_1^j, KD_2^j, TKD_1^j, and TKD_2^j ($j=1,2,3,....,n$) for all the selected features can be estimated. It is followed by computing the KL_1^j and KL_2^j, the symmetrized Kullback-Leibler divergences (KLIDs) of KD_1^j and TKD_1^j, KD_2^j and TKD_2^j ($j=1,2,\cdots,n$), via Eq. (18). To get an overall assessment for all the *n* selected features of each sample set, an integrated KLID IKL_i is defined here to aggregate all the symmetrized KLIDs KL_i^j for the type i fault together as:

$$IKL_i = \sum_{j=1}^{n} F_i \times KL_i^j , \tag{20}$$

where $F_j\left(j=1,2,\cdots,n\right)$ is the importance of the *j*th feature computed via Eq. (19), and the vector $\mathbf{F}=(F_1,F_2,\cdots,F_n)$. By using Eq. (20), one can get IKL_1 and IKL_2 for any sample from the testing sample set with respect to the type I and type II faults. A smaller value of IKL_i indicates the testing sample has a greater statistical similarity with the corresponding training sample set, that is, adding the testing sample into the training set leads to a very minor influence on the statistical distribution of the original training sample set. Hence, one can discriminate the fault type of the testing sample. For example, if $IKL_1 > IKL_2$, one can conclude that most likely the fault implied by the testing sample is the type II fault rather than type I fault. Following the same fashion, one can classify all the testing sample sets into one of the two fault types.

In the same manner, the proposed method can be straightforwardly extended to a general case where more than two distinct fault modes/damage levels exist. The fault modes/damage level implied by the testing sample can be identified by finding the smallest integrated KLID over all the fault modes/damage levels.

3.3.2 Numerical example

In this section, the performance of the proposed method will be validated with a numerical example. In Fig. 4, two sample sets each of which has two features following bivariate normal distribution are randomly generated. The mean and covariance of the sample set #1 are $\mu=[2\ 3]^T$ and $C=\left[1,0;0,1\right]$, respectively; whereas the mean and covariance are set to $\mu=[6\ 3]^T$ and $C=\left[1.5,0;0,1.5\right]$, respectively for the sample set #2. Each of sample sets contains 300 pairs of samples. In such case, it may not be easy for the commonly used classification methods, like SVM, to correctly classify all the samples since these two sample sets have an overlapping region. As seen in Fig. 4, the optimal hyperplane (represented by the black dotted line in Fig. 4) solved by the SVM classification method cannot completely separate the two data sets.

In our study, 50 samples of each set are randomly chosen as the training samples and the remaining 600 samples are treated as the testing samples. The Gaussian function is used as the kernel function for both the SVM-based classification method and our proposed method. The results show that by using the proposed method, one can get a slightly higher (99.7%) classification accuracy than that of the SVM method (97.5%). Only

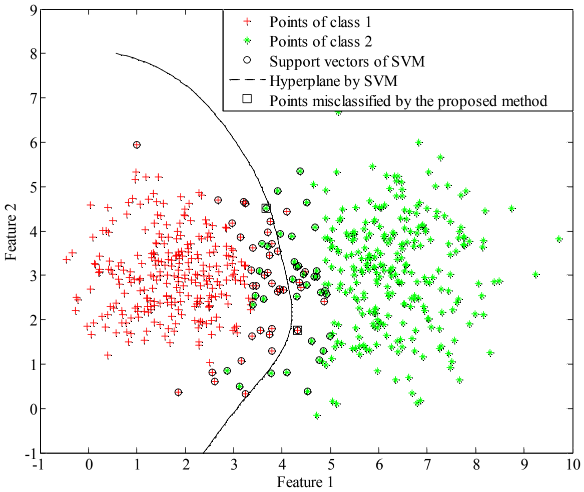


Fig. 4. The classifying results of the SVM-based and proposed methods on the artificial sample sets with an overlapping region.

two testing samples (indicated by the dots with rectangle) are misclassified compared to 15 samples (the green stars on the left of the hyperplane and the red crosses on the right of the hyperplance) by using the SVM method as shown in Fig. 4. The reason is that the proposed method can take into account the statistical characteristics of the sample sets. Even if a testing sample is far from the mean value of the sample sets, it can still correctly identify the class of the testing sample since the testing sample has similar statistical characteristics with the samples in the class. The effectiveness of the proposed method will be further explored via the real fault diagnosis problems in the ensuing section.

4. Applications and result analysis

To validate the effectiveness of the proposed method in terms of rotating machinery fault diagnosis, two case studies of fault diagnosis for bevel gear and rolling element bearing, the critical components of rotating machinery, are presented in this section along with a set of detailed comparative studies.

4.1 Experimental rigs

Case 1: We performed experiments on a machinery fault simulator produced by Spectra Quest, Inc. This equipment is located in the Equipment Reliability and Prognosis and Health Management (ERPHM) Lab at the University of Electronic Science and Technology of China. The experimental test rig and the faulty bevel gears to be tested are shown in Fig. 5. The experimental test rig consists of a motor, a coupling, bearings, two bevel gearboxes (one good right angle gearbox and one worn right angle gearbox), discs, belts, and a shaft. The bevel gearbox is driven by an AC motor and coupled with rub belts. The rotation speed was set to be a constant at 1,800 r/min. Three kinds of faulty gears, i.e., worn gear, gear with missing teeth, and gear with broken tooth, were simulated on the experimental setup. An accelerometer was mounted on the top of the gearbox. Vibration data were collected every 3 min by an eight-channel DAQ, the data sampling rate was 20 kHz and the data length was 4,096 points. Several pieces of time-domain signals and their corresponding spectrum of the normal gear and the faulty gears are shown in Fig. 6.

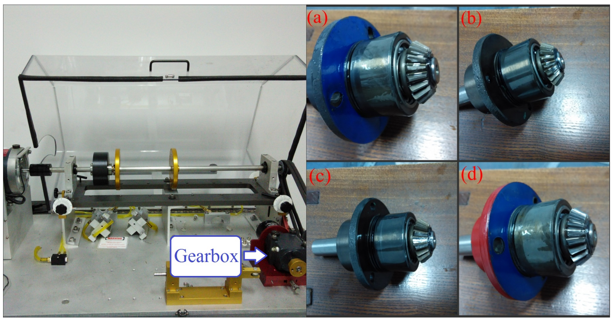


Fig. 5. The experiment rig and the four bevel gears with different damages: (a) normal gear; (b) gear with broken tooth; (c) gear with missing teeth; (d) gear with worn tooth.

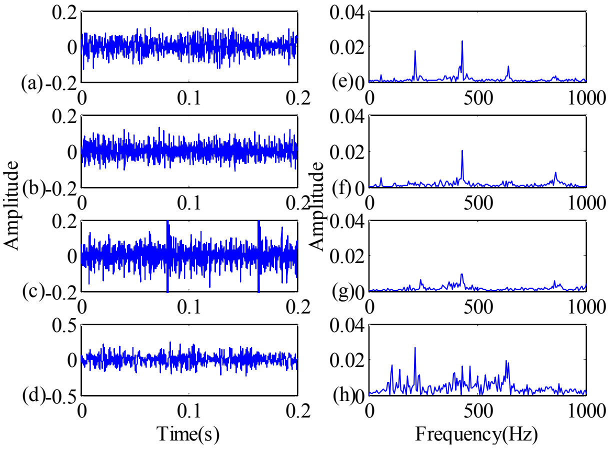


Fig. 6. (a)-(d) are the time-domain signals acquired from the four different gears (normal, broken tooth, missing teeth, and worn tooth); (e)-(h) are the corresponding spectrums.

Case 2: The experimental data come from Case Western Reserve University [33, 34]. The experimental rig is comprised of Reliance Electric 2HP IQPreAlet connected to a dynamometer. The bearings to be examined support the motor shaft. Faults with crack size 0.007, 0.014, 0.021, and 0.028 inches on the drive-end bearing (6205-2RS JEM SKF) were artificially created by the electric discharge machining (EDM). These faults are separately distributed on the inner raceway, rolling element, and outer raceway. Two accelerometers were mounted to collect vibration signals in the experiment. One was placed on the motor housing, and the other one was located on the outer race of the drive-end bearing. Data sampling frequency was 12 kHz and the sampling length was 12 k, rotating speed was fixed at 1750 r/min.

Table 3. The data sets for defect and severity classification.

Data set		Number of training samples	Number of testing samples	Defect size (inch) (training/testing)*	Condition
A		35	35	—	Normal
		35	35	—	Broken tooth
		35	35	—	Missing teeth
		35	35	—	Worn tooth
B	B_1	35	35	0.007/0.021	Inner race
		35	35	0.007/0.021	Ball
	B_2	35	35	0.021/0.007	Inner race
		35	35	0.021/0.007	Ball
C		35 35 35	35 35 35	0.007 0.014 0.021	Inner race

*'-' for the data set A denotes the defect sizes of the training and testing samples are exactly the same but unmeasurable by a physical dimension.

4.2 Experimental scheme and results

The data collected from the aforementioned two experiments are used to validate our proposed method. The data with the same type of defects and severies are randomly divided into training samples and testing samples. The training and testing sample sizes, the places of defects, and the defect sizes in the two case studies are detailed in Table 3. The data set A comes from Case 1, whereas the data sets B and C come from Case 2. The data sets A and B are designed to validate the capability of the proposed method in terms of recognizing the types of defects; whereas the data set C is used to examine whether the proposed method is able to identify the severity of the same type of defect.

The data set A includes 280 data sets for bevel gears with four different operation conditions: normal, gear with broken tooth, gear with missing teeth, and gear with worn tooth. The defect sizes of both training sets and testing sets are exactly the same. Apparently, it can be viewed as a four-class classification problem.

The data set B consists of 280 data samples of the faulty bearings, but with only two types of operation conditions: inner race fault and ball fault. Two subsets B_1 and B_2 are contained in the data set B, each of them have 140 data samples. Two subsets B_1 and B_2 are contained in the data set B, each of them have 140 data samples. The experiment over this data set is carried out to further investigate the robustness and generalization of the proposed method if the fault mode of the training set is the same as the testing set but the defect sizes are distinct. For the subset B_1, 70 samples with the fault detect size of 0.007 inches are treated as the training set, and the rest of 70 samples with the fault detect size of 0.021 inches are the testing samples. The subset B_2 is similar to the subset B_1 except that the training set of the subset B_2 is treated as the testing set of the subset B_1 whereas the testing set of the subset B_2 is the training set of the subset B_1.

Table 4. The selected features for the data set A.

Feature ID	14	15	16	27	59
α_j	1.00	0.99	0.96	0.64	0.53
Feature ID	63	26	7	56	8
α_j	0.53	0.53	0.53	0.53	0.51

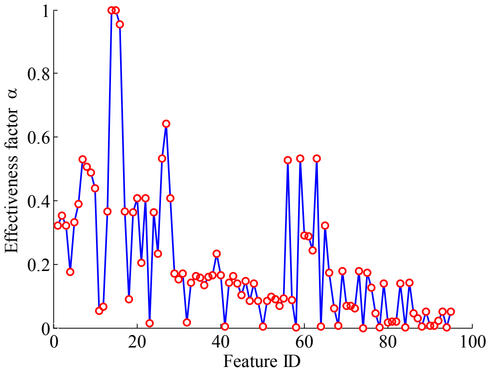
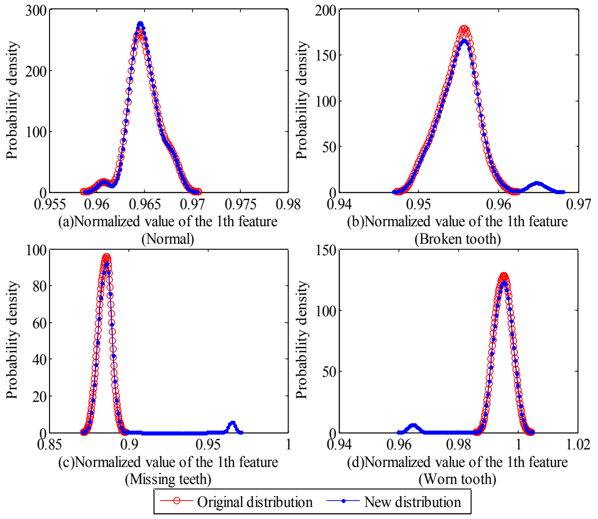


Fig. 7. The effectiveness factor α_j of all the extracted features.

The data set C consists of 210 samples. The data set C is collected from the case where a defect is on the inner race. Three defect sizes, 0.007, 0.021, and 0.028 inches, are contained in these data sets. The purpose of using these three data sets is to validate the effectiveness of the proposed method in identifying the damage levels (defect severity).

We exemplify the implementation of the proposed method to the data set A. 95 features are first extracted from the data set A. The amplitude of the white noise to be added is set to 0.3 and the ensemble number is 100 in EEMD based on the recommended settings in Ref. [25]. The effectiveness factors α_j of all the 95 features computed by the distance evaluation approach are shown in Fig. 7, and the first ten features with the greatest values are listed in Table 4. Note here that the ten selected features may change with different classification problems. Consequently, the probability density functions of the j th feature for the four different training sets can be obtained by the KDE and denoted as KD_i^j ($i=1,2,3,4$) representing bevel gears with normal, broken tooth, missing teeth, and worn tooth conditions, respectively. In the next step, a sample from the testing sample sets is added into the four training sets, respectively, and the probability density functions TKD_i^j ($i=1,2,3,4$) after adding a testing sample can be estimated. The probability density functions of the first feature for the four training sample sets after adding a sample from the one of the four testing sampling sets are shown in Figs. 8-11, respectively.

In Figs. 8-11, the curves with circles represent the original probability distributions of the first feature of the training sets, while the curves with dots are the new probability density functions after adding a testing sample. For example, after a testing sample from the normal condition is added, the two



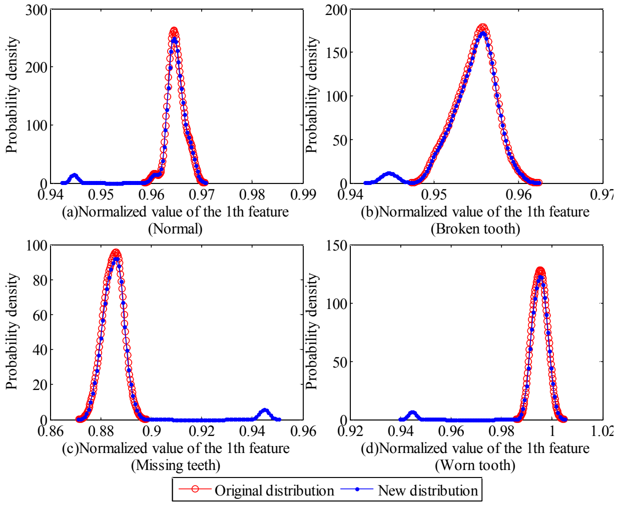
(a)Normalized value of the 1th feature (Normal)

(b)Normalized value of the 1th feature (Broken tooth)

(c)Normalized value of the 1th feature (Missing teeth)

(d)Normalized value of the 1th feature (Worn tooth)

Fig. 8. The original probability densities of the four training sample sets and the corresponding new probability densities after adding a normal testing sample.



(a)Normalized value of the 1th feature (Normal)

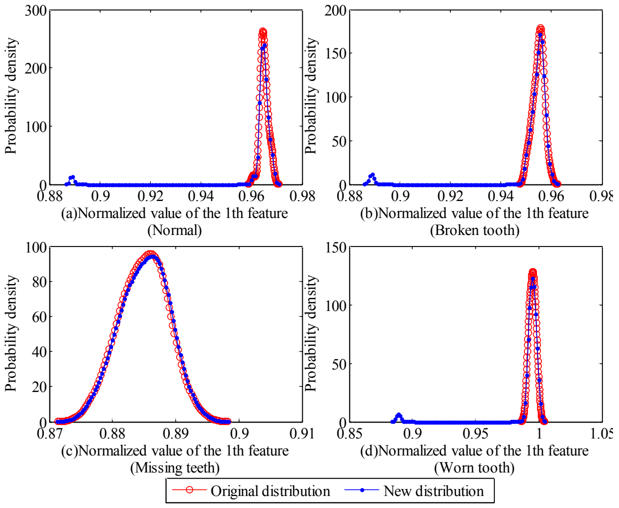
(b)Normalized value of the 1th feature (Broken tooth)

(c)Normalized value of the 1th feature (Missing teeth)

(d)Normalized value of the 1th feature (Worn tooth)

Fig. 9. The original probability densities of the four training sample sets and the corresponding probability densities after adding a testing sample with broken tooth.

probability density functions are almost the same as observed in Fig. 8(a). On the other hand, the probability density functions have a larger discrepancy as seen from Figs. 8(b)-(d) if the testing sample from the normal condition is added to the other three training sample sets. The reason is that the statistical characteristics of the first feature of the testing sample from the normal condition are quite different from these samples from the other three conditions, and it therefore causes a larger change to the probability density functions. In the same manner, as observed in Figs. 9-11, the new sample added to the training sample sets has minor influence on the probability density functions if the conditions of the new sample and training sample sets are the same, otherwise a greater influence can be found.



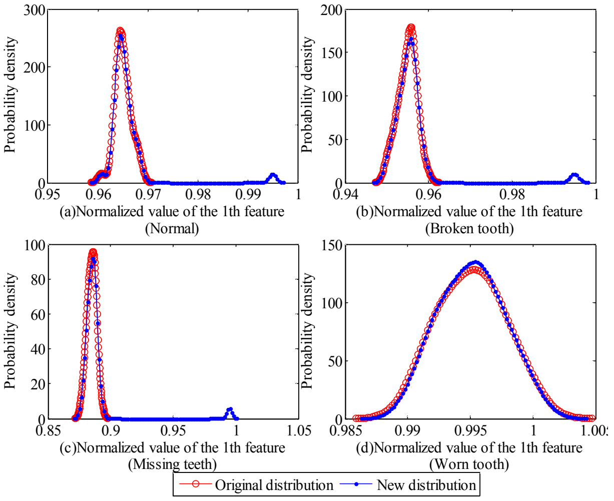
(a)Normalized value of the 1th feature (Normal)

(b)Normalized value of the 1th feature (Broken tooth)

(c)Normalized value of the 1th feature (Missing teeth)

(d)Normalized value of the 1th feature (Worn tooth)

Fig. 10. The original probability densities of the four training sample sets and the corresponding probability densities after adding a testing sample with a missing teeth.



(a)Normalized value of the 1th feature (Normal)

(b)Normalized value of the 1th feature (Broken tooth)

(c)Normalized value of the 1th feature (Missing teeth)

(d)Normalized value of the 1th feature (Worn tooth)

Fig. 11. The original probability densities of the four training sample sets and the corresponding probability densities after adding a testing sample with worn tooth.

In the ensuing step, the KLID is used to quantitatively measure the similarity between the original and the new distributions of the first feature, and the results for the four different conditions are denoted as KL_i^1 ($i = 1,2,3,4$). Following the same fashion, the KLIDs can be computed for all the selected features. The integrated KLIDs IKL_i ($i = 1,2,3,4$) that aggregates the KLIDs of all the selected features is evaluated based on the weights of the 10 selected features via Eq. (20).

The results of classification accuracy (represented by the percentage of correctly identifying the fault modes or defect levels, and a greater value is favorable) for the three data sets are presented in Table 5. To demonstrate the advantages of our proposed method over the conventional data-driven fault diagnosis methods, the results from the SVM-based fault di-

Table 5. The classification accuracy of the three methods*.

Data set	SVM (%)		BP network (%)		The proposed method (%)	
	Training	Testing	Training	Testing	Training	Testing
A	95.25	92.14	**100**	99.62	**100**	**100**
B	**100**	98.10	89.29	87.14	**100**	**100**
C	**100**	97.86	**100**	96.19	**100**	**100**

*The results with the highest accuracy in each data set are highlighted by the underlined numbers.

agnosis method and the back-propagation (BP) network-based fault diagnosis method are also given and compared. For the SVM-based fault diagnosis method, the parameter σ in SVM is optimized by the grid search method [30, 35]. The number of neurons in the input layer of the BP network-based fault diagnosis method equals to the number of the selected features n. A single hidden layer structure is adopted and the number of neurons in the hidden layer is determined by $\sqrt{n+m}+1$, an empirical formula reported in Ref. [36], where m is the number of the output neurons, the number of all the possible fault modes/damage levels. Since the initial thresholds and weights in a BP network have significant impact on the classification accuracy, the genetic algorithm (GA) is used to determine the initial optimal values of thresholds and weights. The population size of the GA is set to 40 and the maximum number of generations is set to 100. Each individual solution is coded by a 10-bit string, and the generation gap, the crossover rate, the mutation rate are set to be 0.95, 0.7, and 0.01, respectively.

As observed from Table 5, for the data set A, the BP network-based fault diagnosis method possesses a very high (100%) training and testing accuracy (99.62%), whereas the SVM-based fault diagnosis method shows a relatively poor accuracy. Contrary to the data set A, the SVM-based fault diagnosis method has a high training and testing accuracy for the data set B, whereas the BP network-based fault diagnosis method is inferior for the data set B as both the training and testing accuracy are lower than 90%. Both the SVM-based fault diagnosis method and the BP network-based fault diagnosis method have a very high accuracy (greater than 95%) for the data set C. The proposed method outperforms the other two methods on all the three data sets and possesses excellent accuracy (100%).

4.3 Comparative studies for different parameter settings

The number of selected features and the size of training sample sets are two critical parameters for data-driven fault diagnosis methods. The following subsection will examine how the performance of our proposed method may vary with different parameter settings.

4.3.1 The number of selected features

To examine the relation between the number of selected features and the classification accuracy, the numbers of selected features for the gearbox and the bearing fault diagnosis problems are changed from 1 to 95, and the corresponding classification accuracies for the data sets A~C are compared. The number of training samples and the number of testing samples are both set to 35 in this study. The results for the proposed method, SVM-based fault diagnosis method, and BP network-based fault diagnosis on the data sets A~C are plotted in Fig. 12.

As seen from Fig. 12, even if the number of selected features is very small (i.e. less than ten), the proposed method always exhibits the highest accuracy than the other two methods for all the data sets. In addition, the proposed method has the most stable performance with the increase of selected features. The classification accuracy will not decrease or fluctuate with the number of selected features. On the other hand, the SVM-based and BP network-based fault diagnosis methods show inferior performance for the data sets A and B. The accuracy of the SVM-based and BP network-based fault diagnosis methods may not be monotonically increasing with respect to the number of selected features. The SVM-based fault diagnosis method has a very low accuracy for the data set B if the number of selected feature is greater than 30; whereas the oscillating phenomena occur for the BP network-based fault diagnosis method. The reduction of accuracy is also observed for these two methods on the data sets B and C. The observation from Fig. 12 illustrates that the proposed method, which directly considers the importance of the selected features, possesses a more stable performance with respect to the number of selected features than the other two methods.

Using more selected features as the inputs of classification will surely reduce the computational efficiency for all the data-driven fault diagnosis methods. To balance the computational cost and classification accuracy and based on the second criterion for choosing the number of selected features mentioned in Sec. 2.3, the first ten features are selected for our case study in Sec. 4.2.

4.3.2 The size of training sample sets

In this subsection, we also conduct a comparative study to examine whether the size of training samples has any impact on the classification accuracy. The first ten features with the highest value of effectiveness factor α_j are selected as the inputs of the proposed method. The number of testing samples is set to be 35 for all the data sets. The classification accuracies with respect to the number of training samples changing from 1 to 70 are plotted in Figs. 13(a)-(c) for the three data sets respectively.

It is observed from Fig. 13 that the proposed method has a low accuracy when the number of the training samples is less than ten, but the accuracy is significantly improved (100% for all the data sets) when the number of training samples is greater than ten. It is noteworthy that the accuracy curve of the proposed method has a small decrease for the data set B when the number of training samples becomes larger than 56. The reason for this phenomenon is that the new added training

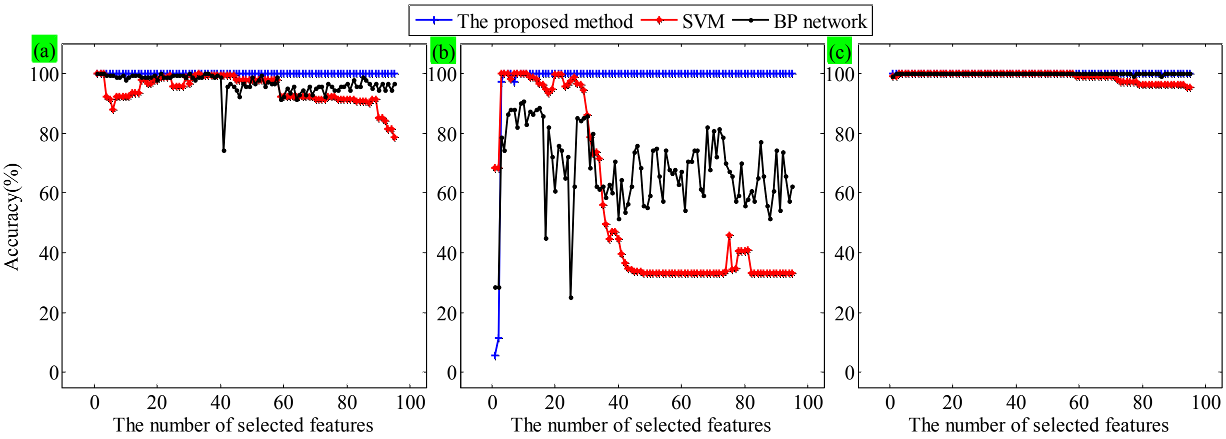


Fig. 12. The number of selected features vs. the classification accuracies of the three methods for the data sets A to C: (a) data set A; (b) data set B; (c) data set C.

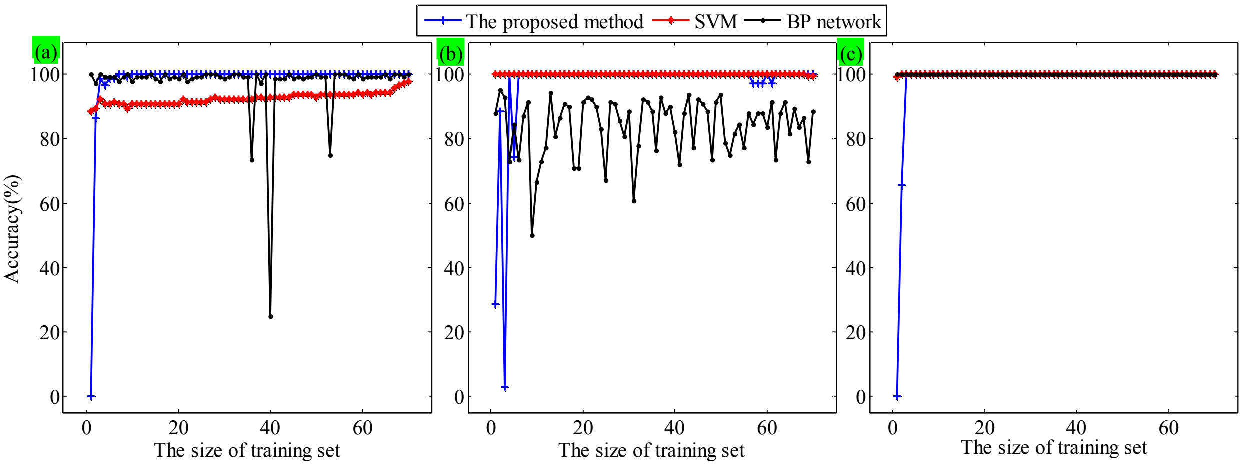


Fig. 13. The size of training set vs. the classification accuracies of the three methods for the data sets A to C: (a) data set A; (b) data set B; (c) data set C.

samples are far from the mean value of the original probability density, that is, these data are outliers and will change the shape of probability density function dramatically, leading to the difficulty in classifying fault modes. For the SVM-based fault diagnosis method, it has an excellent performance (100% accuracy) for the data sets B and C. In addition, the classification accuracy of the SVM-based fault diagnosis method is stable with the increase of training samples size for all the data sets. Nevertheless, the SVM-based fault diagnosis method exhibits relatively low classification accuracy (about 90%) for the data set A regardless of the number of training sample size. For the BP network-based fault diagnosis method, the classification accuracies for the data sets A and C are close or equal to 100% in most cases. However, the classification accuracy shows an oscillating behavior for the data set B.

5. Closure

A new data-driven fault diagnosis method for rotating machinery has been proposed based on kernel density estimation and Kullback-Leibler divergence. The KDE is used to estimate the statistical characteristics of the selected features of the training and testing samples, whereas the KLID is employed to quantitatively measure the similarity between two estimated distributions. With the assistance of the KDE and KLID, the fault modes/damage level is identified by comparing the integrated KLID of the selected features. As demonstrated in fault diagnosis of bevel gears and rolling element bearings, the proposed method has an exceptional performance on faulty pattern recognition and outperforms the conventional SVM-based and BP network-based fault diagnosis methods. Since the proposed method incorporates the statistical characteristics of the samples within one set, it manifests superior classification accuracy and robust performance.

Acknowledgment

The authors greatly acknowledge grant support from the

National Natural Science Foundation of China under contract number 71371042 and the Specialized Research Fund for the Doctoral Program of Higher Education of China under contract number 20110185120014. Comments and suggestions from all reviewers and the editor are very much appreciated.

References

[1] P. Konar and P. Chattopadhyay, Bearing fault detection of induction motor using wavelet and Support vector machines (SVMs), *Applied Soft Computing*, 11 (6) (2011) 4203-4211.

[2] A. K. S. Jardine, D. Lin and D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance, *Mechanical Systems and Signal Processing*, 20 (7) (2006) 1483-1510.

[3] V. Venkatasubramanian et al., A review of process fault detection and diagnosis: Part I: Quantitative model-based methods, *Computers & Chemical Engineering*, 27 (3) (2003) 293-311.

[4] P. D. Samuel and D. J. Pines, A review of vibration-based techniques for helicopter transmission diagnostics, *Journal of Sound and Vibration*, 282 (1) (2005) 475-508.

[5] R. Isermann, Model-based fault-detection and diagnosis-status and applications, *Annual Review in Control*, 29 (1) (2005) 71-85.

[6] J. Chen and R. J. Patton, *Robust model-based fault diagnosis for dynamic systems*, Springer, New York, USA (2012).

[7] A. Heng, S. Zhang, A. C. C. Tan and J. Mathew, Rotating machinery prognostics: State of the art, challenges and opportunities, *Mechanical Systems and Signal Processing*, 23 (3) (2009) 724-739.

[8] B. S. Yang and K. J. Kim, Application of Dempster-Shafer theory in fault diagnosis of induction motors using vibration and current signals, *Mechanical System and Signal Processing*, 20 (2) (2006) 403-420.

[9] B. S. Yang, T. Han and J. L. An, ART-KOHONEN neural network for fault diagnosis of rotating machinery, *Mechanical Systems and Signal Processing*, 18 (3) (2004) 645-657.

[10] L. F. Deng and R. Z. Zhao, Fault feature extraction of a rotor system based on local mean decomposition and Teager energy kurtosis, *Journal of Mechanical Science and Technology*, 28 (4) (2014) 1161-1169.

[11] N. E. Huang et al., The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proceedings of The Royal Society of London*, 454 (1971) (1998) 903-995.

[12] Y. G. Lei, Z. J. He, Y. Y. Zi and Q. Hu, Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs, *Mechanical Systems and Signal Processing*, 21 (5) (2007) 2280-2294.

[13] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3 (2003) 1157-1182.

[14] Y. G. Lei, Z. J. He, Y. Y. Zi and X. Chen, New clustering algorithm-based fault diagnosis using compensation distance evaluation technique, *Mechanical Systems and Signal Processing*, 22 (2) (2008) 419-435.

[15] D. A. Tibaduiza, M. A. Torres-Arredondo and L. E. Mujica et al., A study of two unsupervised data driven statistical methodologies for detecting and classifying damages in structural health monitoring, *Mechanical Systems and Signal Processing*, 41 (1) (2013) 467-484.

[16] Y. G. Lei, Z. J. He and Y. Y. Zi, A new approach to intelligent fault diagnosis of rotating machinery, *Expert Systems with Applications*, 35 (4) (2008) 1593-1600.

[17] Y. Zhang et al., Fault diagnosis of rotating machine by isometric feature mapping, *Journal of Mechanical Science and Technology*, 27 (11) (2013) 3215-3221.

[18] H. Cui, L. Zhang, R. Kang and X. Lan, Research on fault diagnosis for reciprocating compressor valve using information entropy and SVM method, *Journal of Loss Prevention in The Process Industries*, 22 (6) (2009) 864-867.

[19] N. Saravanan and K. I. Ramachandran, Incipient gear box fault diagnosis using discrete wavelet transform (DWT) for feature extraction and classification using artificial neural network (ANN), *Expert Systems with Applications*, 37 (6) (2010) 4168-4181.

[20] A. Widodo and B. S. Yang, Support vector machine in machine condition monitoring and fault diagnosis, *Mechanical Systems and Signal Processing*, 21 (6) (2007) 2560-2574.

[21] Z. L. Liu, M. J. Zuo and H. B. Xu, Feature ranking for support vector machine classification and its application to machinery fault diagnosis, *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 227 (9) (2013) 2077-2089.

[22] J. Rafiee, F. Arvani, A. Harifi and M. H. Sadeghi, Intelligent condition monitoring of a gearbox using artificial neural network, *Mechanical Systems and Signal Processing*, 21 (4) (2007) 1746-1754.

[23] Y. G. Lei and M. J. Zuo, Gear crack level identification based on weighted K nearest neighbor classification algorithm, *Mechanical Systems and Signal Processing*, 23 (5) (2009) 1535-1547.

[24] Z. H. Wu and N. E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, *Advances in Adaptive Data Analysis*, 1 (01) (2009) 1-41.

[25] Y. G. Lei, Z. J. He and Y. Y. Zi, Application of the EEMD method to rotor fault diagnosis of rotating machinery, *Mechanical Systems and Signal Processing*, 23 (4) (2009) 1327-1338.

[26] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, 27 (3) (1956) 832-837.

[27] E. Parzen, On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, 33 (1962) 1065-1076.

[28] S. Kullback and R. A. Leibler, On information and sufficiency, *The Annals of Mathematical Statistics*, 22 (1951) 79-86.

[29] M. Feldman, Hilbert transform in vibration analysis, *Mechanical Systems and Signal Processing*, 25 (3) (2011) 735-802.

[30] Z. L. Liu, M. J. Zuo and H. B. Xu, Fault diagnosis for planetary gearboxes using multi-criterion fusion feature selection framework, *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 227 (9) (2013) 2064-2076.

[31] W. C. Kim, B. U. Park and J. S. Marron, Asymptotically best bandwidth selectors in kernel density estimation, *Statistics & Probability Letters*, 19 (2) (1994) 119-127.

[32] V. A. Epanechnikov, Non-parametric estimation of a multivariate probability density, *Theory of Probability & Its Applications*, 14 (1) (1969) 153-158.

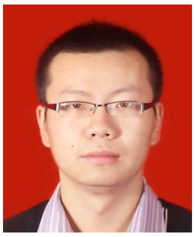
[33] Bearing Data Center, Case Western Reserve University, *Available from: http://www.eecs.cwru.edu/laboratory / Bearing* (accessed on June.10, 2009).

[34] J. B. Yu, Bearing performance degradation assessment using locality preserving projections and Gaussian mixture models, *Mechanical Systems and Signal Processing*, 25 (7) (2011) 2573-2588.

[35] C. W. Hsu, C. C. Chang and C. J. Lin, A practical guide to support vector classification, *Technical Report*, Taipei, Department of Computer Science, National Taiwan University (2010).

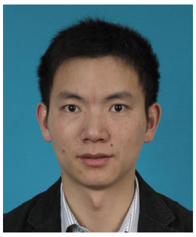
[36] H. Y. Shen et al., Determining the number of BP neural network hidden layer units, *Journal of Tianjin University of Technology*, 24 (5) (2008) 13-15.

Fan Zhang is currently a graduate student in the School of Mechanical, Electronic, and Industrial Engineering, at the University of Electronic Science and Technology of China. His research interests are intelligent fault diagnosis, performance degradation assessment, and signal processing.



Yu Liu is a Professor in the School of Mechanical, Electronic, and Industrial Engineering, at the University of Electronic Science and Technology of China. He received his Ph.D. in Mechatronics Engineering from the University of Electronic Science and Technology of China. His research interests include system reliability modeling and analysis, maintenance decisions, prognostics and health management, and design under uncertainty.

Chujie Chen is currently a graduate student in the School of Mechanical, Electronic, and Industrial Engineering, at the University of Electronic Science and Technology of China. His research interest is reliability modeling and assessment for complex engineering systems.



Yan-Feng Li received his Ph.D. in Mechatronics Engineering from the University of Electronic Science and Technology of China in 2013. He is currently a faculty member of the University of Electronic Science and Technology of China. His research interests include reliability analysis and evaluation of complex systems, dynamic fault tree modeling, Bayesian networks modeling, and probabilistic inference.

Hong-Zhong Huang is a Professor of the School of Mechanical, Electronic, and Industrial Engineering, at the University of Electronic Science and Technology of China. He received the Ph.D. in Reliability Engineering from Shanghai Jiaotong University, China. His current research interests include system reliability analysis, warranty, maintenance planning and optimization, computational intelligence in product design.