Int J Adv Manuf Technol (2013) 67:1217–1230
DOI 10.1007/s00170-012-4560-y

ORIGINAL ARTICLE

Fault level diagnosis for planetary gearboxes using hybrid kernel feature selection and kernel Fisher discriminant analysis

Zhiliang Liu · Jian Qu · Ming J. Zuo · Hong-bing Xu

Received: 22 December 2010 / Accepted: 4 October 2012 / Published online: 20 October 2012

Abstract This paper introduces a hybrid dimension reduction method that combines kernel feature selection and kernel Fisher discriminant analysis (KFDA). In the first stage, a kernel feature selection method is proposed to remove redundant and irrelevant features for two purposes: (1) reducing computation burden of the entire fault diagnosis system and (2) alleviating the impact of irrelevant features on KFDA. In the second stage, KFDA is used to establish a more compact feature subset by extracting a smaller number of features. We use Gaussian radial basis function as the kernel function for the two kernel stages in the proposed method. A parameter selection method for this kernel is proposed to select the optimal values for the proposed method. Experimental results on fault level diagnosis demonstrate that the proposed hybrid dimension reduction method has advantages over other approaches that use feature selection or KFDA separately.

Keywords Fault diagnosis · Dimension reduction · Feature selection · Kernel Fisher discriminant analysis · Planetary gearbox

Z. Liu · H.-b. Xu
School of Automation Engineering,
University of Electronic Science and Technology of China,
Chengdu, People's Republic of China 611731

M. J. Zuo
School of Mechatronics Engineering,
University of Electronic Science and Technology of China,
Chengdu, People's Republic of China 611731

J. Qu · M. J. Zuo (✉)
Department of Mechanical Engineering, University of Alberta,
Edmonton, Canada T6G2G8
e-mail: ming.zuo@ualberta.ca

1 Introduction

Planetary gearboxes are important components of rotating machinery because they can achieve a large transmission ratio in a compact package. Planetary gearboxes therefore have been widely used in mining [1], power [2], aerospace [3], etc.

Gear faults commonly occur during operation in planetary gearboxes since gears experience cyclic stress when gears are meshing. Gear faults are broadly grouped into distributed faults and localized faults [4]. Distributed fault such as wear may decrease transmission accuracy, increase vibration levels, and even lead to catastrophic accidents. Localized fault such as pits may increase transmission errors and worsen the distributed fault. Gear fault diagnosis is valuable for decreasing economic costs and increasing operation safety of planetary gearboxes. If faults are detected in time, preventive actions can be taken to avoid severe consequences, such as a system breakdown and even injuries of workers [5, 6].

In recent years, major efforts have been made in fault detection and fault mode identification; however, few papers are reported on fault level diagnosis, which comes with more challenges. Feng et al. [7] proposed a regularization dimension technique to make vibration signals increase monotonically with respect to gear fault levels. Combet et al. [8] proposed an optimal denoising filter based on spectral kurtosis and used this method to detect small tooth surface pitting in a two-stage helical reduction gearbox. Öztürk et al. [9] used a scalogram and its mean frequency variation for detecting and monitoring fault levels of gear pitting. Loutridis [10] examined an exponent empirical histogram for detecting early-stage damage and further estimating damage magnitude. Loutridis [11] later proposed energy-based features for diagnosis and prediction of gear crack magnitude. However, the methods in Refs. [7–11] require expert knowledge and cannot identify fault levels automatically. Moreover, those methods were proposed for fixed-shaft gearboxes, either horizontal or helical. Compared with fixed-shaft gearboxes, planetary gearboxes are more complex and possess several unique behaviors that are

not found in fixed-shaft gearboxes [12]. For instance, gear mesh frequencies are often completely suppressed, and sidebands are not symmetrical as they are in fixed-shaft gearboxes. Keller et al. [13] modified several standard diagnostic features for planetary gearboxes, but no single feature could detect initial faults in the planetary gearbox under the on-aircraft condition.

Since a fault can be viewed as a "pattern" [14], classification approaches are used to fuse multiple features. However, a large number of features may lead to poor fault diagnosis, such as classification accuracy deterioration, increase of model complexity and computation burden, etc. Dimension reduction is therefore required to improve performance of classification-based fault diagnosis methods.

Feature selection and feature extraction are two main categories of dimension reduction techniques. *Feature selection* refers to a process that selects the best n' features out of n features ($n' \leq n$). *Feature extraction* aims generating a set of new features from the original feature space through a functional projection transform. In the field of feature selection, Lei et al. [4] proposed a two-stage feature selection and weighting technique to select fault-sensitive features. The method was used to classify three levels of gear crack, i.e., 0, 25, and 50 %. Zhang et al. [15] proposed a filter-wrapper model that combines multiple effectiveness criteria for fault diagnosis of bearing and stator winding. On the other hand, relatively fewer publications are found in the field of feature extraction for machinery fault diagnosis. Liu et al. [1] proposed a method for determining the optimal number of discriminant directions of Fisher discriminant analysis (FDA). Zhang and Huang [16] applied kernel Fisher discriminant analysis (KFDA) for bearing fault diagnosis.

Although feature selection and feature extraction can both achieve dimension reduction, they have different characteristics. Feature extraction reduces feature dimension for classifiers only, while feature selection can reduce dimensionality for the entire fault diagnosis system. That is, the nonselected features are not computed once feature selection is done. Feature selection is therefore more computationally efficient from the perspective of the entire system. On the other hand, feature extraction can concentrate classification information into few transformed features. Theoretically, it can produce more informative features than feature selection. For example, the best discriminant direction obtained by FDA contains the largest Fisher discriminant ratio. However, conventional dimension reduction techniques usually treat feature selection and feature extraction separately, which cannot combine their advantages.

In this paper, we propose a hybrid dimension reduction method that combines kernel feature selection and KFDA for fault level diagnosis. FDA is a well-known feature extraction technique in machinery fault diagnosis, but it fails to deal with nonlinear features. KFDA is a kernel extension of FDA. KFDA has two characteristics: (1) it can reduce feature dimension to $L-1$, where L is the number of classes and (2) it can recognize nonlinear and relevant features through the so-called kernel methods. Since KFDA is easily affected by irrelevant features [17, 18], we propose a kernel feature selection method to remove irrelevant features prior to KFDA. We use kernel methods in the proposed feature selection to deal with nonlinear and relevant features that could improve the performance of the proposed method. The experimental results show that this kind of combination of kernel feature selection and KFDA could outperform approaches that use either feature selection only or KFDA only.

The rest of the paper is organized as follows. Section 2 introduces the proposed hybrid dimension reduction method. Section 3 describes pitting damage experiments. Section 4 applies the proposed method to fault level diagnosis for planetary gearboxes. Conclusions are made in Section 5.

2 The proposed method

The proposed method is based on kernel methods for dealing with nonlinear features. Kernel methods map data in the original feature space into the kernel space without ever knowing the mapping function Φ explicitly. Kernel functions define inner product spaces (Hilbert spaces) in the following way:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = < \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >, \quad (1)$$

where $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ returns kernel function value, $<\cdot,\cdot>$ is the dot product operation, and $\Phi(\mathbf{x})$ is the instance in the kernel space corresponding to x in the original feature space by the feature mapping Φ.

The proposed dimension reduction method is illustrated in Fig. 1. It consists of two stages. In the first stage, a kernel feature selection method is proposed to reduce dimension first by removing redundant and irrelevant features. It selects n' features out of n in the original feature set to establish a reduced feature space for the next stage. Physical meanings of features are not changed after feature selection. Two benefits in this stage are that (1) it cuts the computation budget for the entire fault diagnosis system and (2) it shortens the impact of irrelevant features on the following feature extraction. In the second stage, KFDA is employed as the feature extraction method to further reduce feature dimension, say from n' to n''. KFDA in this stage also offers two benefits: (1) it can produce fewer features for classifiers, i.e., L–1 and (2) it can obtain more informative features that could improve accuracy of classification. It is worth pointing out that features generated by KFDA lose their physical meanings due to a functional transformation made in KFDA. We use the Gaussian radial basis function (RBF)

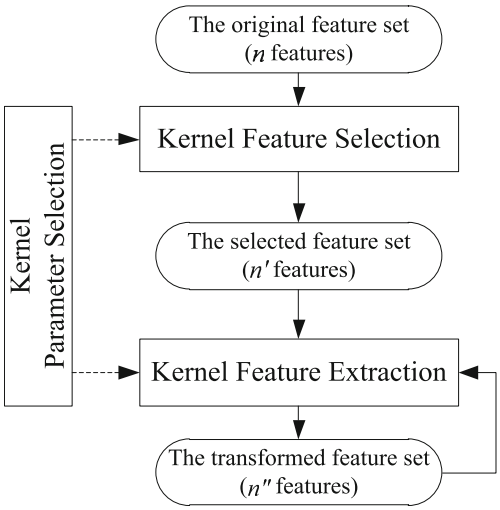

Fig. 1 Scheme of the proposed kernel dimension reduction algorithm

as the kernel function for the proposed kernel feature selection and KFDA. A kernel parameter selection method for this kernel is proposed in Section 2.3.

2.1 Kernel feature selection

In this section, we propose a feature selection method that takes into account not only feature effectiveness but also feature correlation. Feature effectiveness is measured by class separability in the kernel space. Class separability is a classic concept that describes how instances scatter in a feature space [19]. It usually considers two principles: (1) instances from the same class should be as similar as possible and (2) instances from different classes should be as different as possible.

Within-class separability (denoted by W) and between-class separability (denoted by B) are introduced to quantify the two principles above, respectively. Class separability can be measured by either distance similarity or cosine similarity. Since distance similarity is less effective in the high-dimensional space [20], we adopt cosine similarity to measure within-class separability and between-class separability. Cosine similarity in the kernel space is computed by

$$\cos\theta\left(\Phi(\mathbf{x}_i),\Phi\left(\mathbf{x}_j\right)\right)=\frac{\kappa\left(\mathbf{x}_i,\mathbf{x}_j\right)}{\sqrt{\kappa\left(\mathbf{x}_i,\mathbf{x}_i\right)}\sqrt{\kappa\left(\mathbf{x}_j,\mathbf{x}_j\right)}}, \tag{2}$$

where θ is the angle between two instances.

Within-class separability can be estimated by the average cosine similarity of instances from the same class. Between-class separability can be estimated by the average cosine similarity of instances from different classes. That is, W and B are estimated by:

$$W=-\frac{1}{\sum_{i=1}^{L}N_i^2}\sum_{i=1}^{L}\sum_{t=1}^{N_i}\sum_{k=1}^{N_i}\cos\theta\left(\Phi\left(\mathbf{x}_t^{(i)}\right),\Phi\left(\mathbf{x}_k^{(i)}\right)\right), \tag{3}$$

$$B=-\frac{1}{\sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L}N_iN_j}\sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L}\sum_{t=1}^{N_i}\sum_{k=1}^{N_j}\cos\theta\left(\Phi\left(\mathbf{x}_t^{(i)}\right),\Phi\left(\mathbf{x}_k^{(j)}\right)\right), \tag{4}$$

where $\mathbf{x}^{(i)}$ is an instance from the ith class, and N_i is the number of instances from class i.

The objective function of class separability is defined as

$$J=B-W. \tag{5}$$

By this definition, a large score of class separability means large between-class separability but small within-class separability. Following the suggestion in Ref. [21], we evaluate the effectiveness of the ith feature by the change (δ_i) in class separability, that is,

$$\delta_i=J-J_i, \tag{6}$$

where J is the score of class separability using all features and J_i is the score of class separability after removing the ith feature.

We also consider the correlation between features by the criterion of the Pearson correlation coefficient, which is an effective measure of how well two features vary jointly [22]. The Pearson correlation coefficient is defined as follows:

$$\rho(\mathbf{x},\mathbf{z})=\frac{\sum_{i=1}^{N}(x_i-\overline{x})(z_i-\overline{z})}{\sqrt{\sum_{i=1}^{N}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{N}(z_i-\overline{z})^2}}, \tag{7}$$

where $\mathbf{x}=[x_1, x_2, \ldots, x_N]^{\mathrm{T}}$ and $\mathbf{z}=[z_1, z_2, \ldots, z_N]^{\mathrm{T}}$ are two feature vectors that contain a serial values of observations, and N is the number of instances in a training subset.

The Pearson correlation coefficient has a range of [−1, +1], where +1 and −1 indicate a perfect fit to a positive or negative linear relationship between $\mathbf{x}$ and $\mathbf{z}$, respectively. A value close to +1 or −1 indicates a high degree of correlation between $\mathbf{x}$ and $\mathbf{z}$. A value close to zero indicates a poor fit to a linear model between $\mathbf{x}$ and $\mathbf{z}$. We take an absolute value for the Pearson correlation coefficient so that a large value means a strong correlation, and the value is within a range of [0, +1].

Prior to fusion of feature effectiveness and feature correlation, we scale feature effectiveness in Eq. (6) within the

same range as the Pearson correlation coefficient by the following normalization approach [23]:

$$c_i = \frac{1}{1+\exp(-b_i)}, b_i = \frac{\delta_i - \mu}{\sigma}, \tag{8}$$

where μ and σ are mean and standard derivation of feature effectiveness scores $\{\delta_1, \delta_2, \ldots, \delta_n\}$, respectively.

The ad hoc technique [23] is employed to combine feature effectiveness and feature correlation. Mathematically, feature ranking is achieved by the following equations:

$$i_m = \begin{cases} \arg\max_j c_j, & m = 1 \\ \arg\max_{j \neq i_r} \left\{ \alpha_1 c_j - \alpha_2 \frac{1}{m-1} \sum_{r=1}^{m-1} \left| \rho\left(\mathbf{x}_{i_r}, \mathbf{x}_j\right) \right| \right\}, & m \geq 2 \end{cases}, \tag{9}$$

where i_m is the index of a feature in the mth place with the corresponding score JR_m, e.g., i_1=2 means that the second feature (F2) is the top feature. Two parameters α_1 and α_2 are used to balance the effectiveness term and the correlation term. A relatively larger parameter α_1 emphasizes the effectiveness term more. A relatively larger parameter α_2 weights the correlation term more heavily and thus could produce a feature subset with less redundancy.

By the proposed feature ranking method, the top n' features are selected to establish a reduced feature space that is further processed by KFDA in the next section.

2.2 Kernel feature extraction

In this section, KFDA [24] is conducted on the reduce feature space generated in Section 2.1. KFDA is a kernel extension of FDA that is a representative technique of feature extraction. Although the proposed kernel feature selection can recognize nonlinear and relevant features, those features cannot show their effectiveness in FDA because FDA fails to recognize nonlinear features. Instead, KFDA can handle nonlinear features selected by the kernel feature selection method.

We introduce KFDA as follows. Let us define the between-class scatter matrix ($\mathbf{S}_b$), the within-class scatter matrix ($\mathbf{S}_w$), and the total scatter matrix ($\mathbf{S}_t$), respectively, as follows:

$$\mathbf{S}_b^{\Phi} = \sum_{i=1}^{L} N_i \left(\mu_i^{\Phi} - \mu_0^{\Phi}\right)\left(\mu_i^{\Phi} - \mu_0^{\Phi}\right)^{\mathrm{T}}, \tag{10}$$

$$\mathbf{S}_w^{\Phi} = \sum_{i=1}^{L} \sum_{j=1}^{N_i} \left(\Phi\left(\mathbf{x}_j^{(i)}\right) - \mu_i^{\Phi}\right)\left(\Phi\left(\mathbf{x}_j^{(i)}\right) - \mu_i^{\Phi}\right)^{\mathrm{T}}, \tag{11}$$

$$\mathbf{S}_t^{\Phi} = \sum_{i=1}^{L} \sum_{j=1}^{N_i} \left(\Phi\left(\mathbf{x}_j^{(i)}\right) - \mu_0^{\Phi}\right)\left(\Phi\left(\mathbf{x}_j^{(i)}\right) - \mu_0^{\Phi}\right)^{\mathrm{T}}, \tag{12}$$

where μ_i is the mean of the ith class, μ_0 is the mean of all points, and $(\cdot)^{\mathrm{T}}$ is the transpose operation. The three scatter matrices are employed to estimate between-class, within-class, and overall variances of a given training subset, respectively. It can be proven that $\mathbf{S}_b^{\Phi} + \mathbf{S}_w^{\Phi} = \mathbf{S}_t^{\Phi}$.

KFDA seeks the best discriminant direction (denoted by $\mathbf{v}_{opt}$) in the kernel space by maximizing the Fisher discriminant ratio. That is,

$$\mathbf{v}_{opt} = \arg\max_{\mathbf{v}} \frac{\mathbf{v}^{\mathrm{T}} \mathbf{S}_b^{\Phi} \mathbf{v}}{\mathbf{v}^{\mathrm{T}} \mathbf{S}_w^{\Phi} \mathbf{v}}, \tag{13}$$

which can be solved by the following eigenproblem:

$$\mathbf{S}_b^{\Phi} \mathbf{v} = \lambda \mathbf{S}_w^{\Phi} \mathbf{v}. \tag{14}$$

Because the eigenvectors are linear combinations of $\Phi(\mathbf{x}_i)$, there exists coefficient α_i such that

$$\mathbf{v} = \sum_{i=1}^{N} \alpha_i \Phi(\mathbf{x}_i). \tag{15}$$

Let $\boldsymbol{\alpha}=[\alpha_1, \alpha_2, \ldots, \alpha_N]^{\mathrm{T}}$, it can be proven that Eq. (13) is equivalent to:

$$\alpha_{opt} = \arg\max_{\alpha} \frac{\alpha^{\mathrm{T}} \mathbf{KHK} \alpha}{\alpha^{\mathrm{T}} \mathbf{KK} \alpha}, \tag{16}$$

and the corresponding eigenproblem is

$$\mathbf{KHK}\alpha = \lambda \mathbf{KK}\alpha, \tag{17}$$

where $\mathbf{K}$ is the kernel matrix, $\mathbf{K}=\{\kappa(\mathbf{x}_i, \mathbf{x}_j)\}_{N\times N}$, $1 \leq i, j, \leq N$, and each element $h(\mathbf{x}_i, \mathbf{x}_j)$ in $\mathbf{H}$ is given by

$$h\left(\mathbf{x}_i, \mathbf{x}_j\right) = \begin{cases} 1/N_i & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same class} \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

Each eigenvector in Eq. (14) gives a discriminant direction v in the feature space. For a test instance $\mathbf{x}$, we have

$$\mathbf{v}^{\mathrm{T}} \Phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \Phi(\mathbf{x}_i)^{\mathrm{T}} \Phi(\mathbf{x}) = \alpha^{\mathrm{T}} \boldsymbol{\kappa}(:, \mathbf{x}), \tag{19}$$

where $\boldsymbol{\kappa}(:, \mathbf{x})=[\boldsymbol{\kappa}(\mathbf{x}_1, \mathbf{x}), \boldsymbol{\kappa}(\mathbf{x}_2, \mathbf{x}), \ldots, \boldsymbol{\kappa}(\mathbf{x}_N, \mathbf{x})]^{\mathrm{T}}$.

The upper bound of the matrix rank of $\mathbf{S}_b^{\Phi}$ is L–1, so there are at most L–1 eigenvectors corresponding to nonzero eigenvalues [25]. Let $\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{L-1}\}$ be the L–1 eigenvectors of the eigenproblem in Eq. (17) with respect to the nonzero eigenvalues. The transformation matrix $\mathbf{A}$ is an $N\times$ (L–1) matrix and a test instance $\mathbf{x}$ can be mapped into L–1 dimension subspace by:

$$\mathbf{x} \rightarrow \mathbf{z} = \mathbf{A}^{\mathrm{T}} \boldsymbol{\kappa}(:, \mathbf{x}). \tag{20}$$

2.3 Kernel parameter selection

In this section, a kernel parameter selection method is introduced for the proposed hybrid dimension reduction method. We use the Gaussian RBF for both kernel feature selection and KFDA because this kernel is reported to have good generality in many applications [26, 27]. The Gaussian RBF is defined by

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \tag{21}$$

where $\|\cdot\|$ is the two norm of a vector, and σ is the width of features.

The parameter σ is the only parameter in the Gaussian RBF kernel. Parameter selection for this kernel is important to the robustness of the proposed method. Based on the interpretation of kernel methods, kernel parameter results in a kernel space. It is reasonable to believe that a good kernel space comes with a parameter resulting in large class separability. Class separability in Eq. (5) is an implicit function with respect to the kernel parameter. Therefore, we here define an optimal σ as the one that can maximize class separability in Eq. (5). In this sense, kernel parameter selection becomes a one-dimensional optimization problem. The optimal σ is the maximizer of the objective function.

The Gaussian RBF kernel is continuous and twice differentiable with respect to σ. The optimal σ could be found via a one-dimensional search method, e.g., Newton's method [28] with the following iterative equation:

$$\sigma_{k+1} = \sigma_k - \frac{\dot{J}(\sigma)}{\ddot{J}(\sigma)} = \sigma_k - \frac{\dot{B}(\sigma) - \dot{W}(\sigma)}{\ddot{B}(\sigma) - \ddot{W}(\sigma)} \tag{22}$$

where $\dot{J}$ and $\ddot{J}$ are the first derivative and the second derivative of J with respect to σ, respectively; and

$$\dot{B}(\sigma) = -\frac{1}{\sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L} N_i N_j} \sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L}\sum_{t=1}^{N_i}\sum_{k=1}^{N_j}\left[\kappa\left(\mathbf{x}_t^{(i)}, \mathbf{x}_k^{(j)}\right)\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_k^{(j)}\right\|^2/\sigma^3\right],$$

$$\dot{W}(\sigma) = -\frac{1}{\sum_{i=1}^{L} N_i^2}\sum_{i=1}^{L}\sum_{t=1}^{N_i}\sum_{k=1}^{N_i}\left[\kappa\left(\mathbf{x}_t^{(i)}, \mathbf{x}_k^{(i)}\right)\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_k^{(i)}\right\|^2/\sigma^3\right],$$

$$\ddot{B}(\sigma) = -\frac{1}{\sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L} N_i N_j} \sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L}\sum_{t=1}^{N_i}\sum_{k=1}^{N_j}\left[\kappa\left(\mathbf{x}_t^{(i)}, \mathbf{x}_k^{(j)}\right)\left(\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_k^{(j)}\right\|^4 - 3\sigma^2\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_k^{(j)}\right\|^2\right)/\sigma^6\right],$$

$$\ddot{W}(\sigma) = -\frac{1}{\sum_{i=1}^{L} N_i^2}\sum_{i=1}^{L}\sum_{t=1}^{N_i}\sum_{k=1}^{N_i}\left[\kappa\left(\mathbf{x}_t^{(i)}, \mathbf{x}_k^{(i)}\right)\left(\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_k^{(i)}\right\|^4 - 3\sigma^2\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_k^{(i)}\right\|^2\right)/\sigma^6\right].$$

The kernel feature selection and KFDA use the same parameter selection algorithm in different feature spaces, namely, the original feature space (n) and the reduced feature space (n'), respectively.

3 Fault experiments

3.1 Planetary gearbox test rig

Figure 2 shows a planetary gearbox test rig designed to perform fully controlled experiments for developing a reliable diagnostic system. The planetary gearbox test rig has an over-hung floating configuration that can mimic the operation of field mining. It mainly includes a 20-HP drive motor, a one-stage bevel gearbox, a two-stage planetary gearbox, a two-stage speed-up gearbox, and a 40-HP load motor. Table 1 lists the number of teeth and transmission ratio achieved by each stage.

The study object in this paper is the two-stage planetary gearbox diagramed in Fig. 3. The sun gear of the first-stage planetary gearbox is mounted on the right end of shaft 1 with a driven bevel gear mounted on the left end. The first-stage planet gears are mounted on the first-stage carrier that

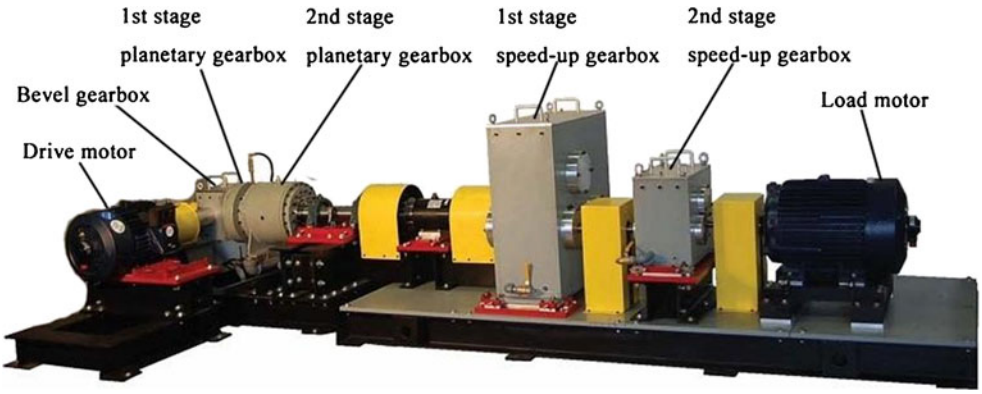

Fig. 2 The planetary gearbox test rig [1]

is connected to shaft 2 with the second-stage sun gear mounted on the other end. The second-stage planet gears are mounted on the second-stage carrier located on output shaft 3. Ring gears of the first and the second stages are mounted on the housing of their stage, respectively.

Four vibration sensors, including two identical low-sensitivity sensors (denoted by LS1 and LS2) and two identical high-sensitivity sensors (denoted by HS1 and HS2), are mounted on the housing of the planetary gearbox. The frequency response range is 0.3 to 8,000 Hz for the low-sensitivity sensors and 0.2 to 200 Hz for the high-sensitivity sensors. The two sensors LS1 and HS1 are used to monitor vibration signals from the first-stage planetary gearbox. The other two sensors, LS2 and HS2, are used to monitor vibration signals of the second-stage planetary gearbox. Figure 4a shows the locations of the four sensors in the test rig.

3.2 Artificially created pitting damage

Based on stress calculation [29], we found that pitting damages should first occur on the planet gears in the second-stage planetary gearbox. There are four planet gears in the second-stage planetary gearbox. We designed pitting damages on one of the planet gears with four pitting levels: baseline, slight, moderate, and severe. A pit was artificially created as a circular hole with a 3-mm diameter and 0.1-mm depth [30]. The number of pits varies with the pitting levels. The four planet gears associated with different fault levels of the pitting damages are shown in Fig. 5. Profiles of the four pitting levels created in the experiment are described as follows [31]:

1. *The baseline level*—a brand new gear was considered the baseline level (see Fig. 5a);
2. *The slight level*—three holes on a tooth and one hole on each of the two neighboring teeth (see Fig. 5b). The middle pitted tooth has a pitting area of 7.95 %. Each of the two neighboring teeth has a pitting area of 2.65 %. The most pitted tooth of the slight level meets the criterion of level two described in the ASM standard [32], that is, a pitting area within 3 to 10 % of the tooth surface area;
3. *The moderate level*—ten holes on a tooth, three holes on each of the two neighboring teeth, and one hole on each of the next two neighbors (see Fig. 5c). The pitting areas on the five teeth are 2.65, 7.95, 26.5, 7.95, and 2.65 %, respectively. The most pitted tooth of the moderate level meets the criterion of level three described in the ASM standard [32], that is, a pitting area within 15 to 40 %; and
4. *The severe level*—24 holes on one tooth, 10 holes on each of the two neighbor teeth, and 3 holes on each of the next neighboring teeth (see Fig. 5d). The pitting areas of the five teeth are 7.95, 26.5, 63.6, 26.5, and 7.95 %, respectively. The most pitted tooth of the severe level meets the criterion of level four described in the ASM standard [32], that is, a pitting area within 50 to 100 %.

Table 1 Specification of the planetary gearbox test rig

Gears	Bevel		First-stage planetary			Second-stage planetary			First-stage speed-up				Second-stage speed-up			
	BI	BO	S	P	R	S	P	R	LI	SI	LO	SO	LI	SI	LO	SO
Teeth	18	72	28	62 (3)	152	19	31 (4)	81	72	32	80	24	48	18	64	24
Ratio	4.00		6.43			5.26			0.13				0.14			

The number of planet gears is indicated in parentheses

BI bevel gear input, *BO* bevel gear output, *S* sun gear, *P* planet gear, *R* ring gear, *LI* large gear on input shaft, *SI* small gear on input shaft, *LO* large gear on output shaft, *SO* small gear on output shaft

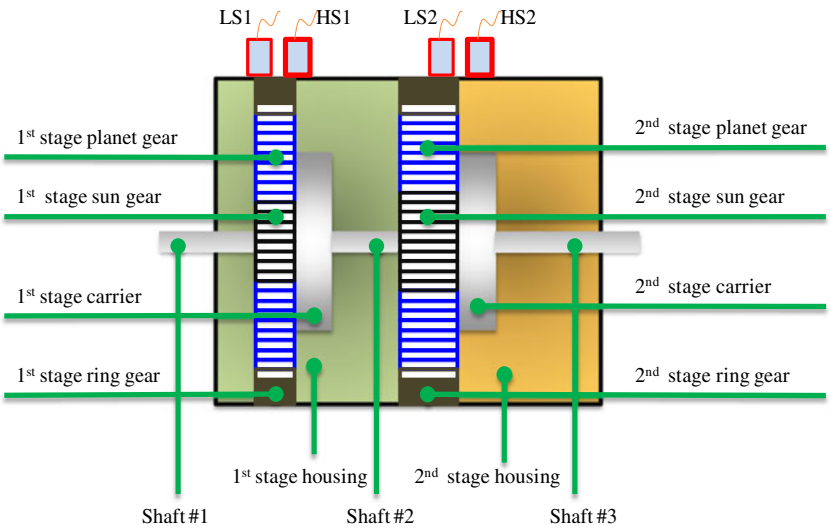

Fig. 3 Structure diagram of the two-stage planetary gearbox (LS1, LS2, HS1, and HS2 are four vibration sensors) [1]

3.3 Data acquisition

We conducted an experiment for each pitting level. The three no-fault planet gears and the one pitted planet gear were mounted in the second-stage planetary gearbox (see Fig. 4b) in one experiment. Four sensors were used to collect vibration signals with a sampling frequency of 10,000 Hz. We considered two load conditions, namely, no-load and load. The amount of 10,000 lb·in torque was applied to the load motor for the load condition. We considered four speed conditions of the drive motor: 300, 600, 900, and 1,200 rpm. A 10-min span of data was recorded under a working combination. These procedures were repeated until four pitting levels were conducted.

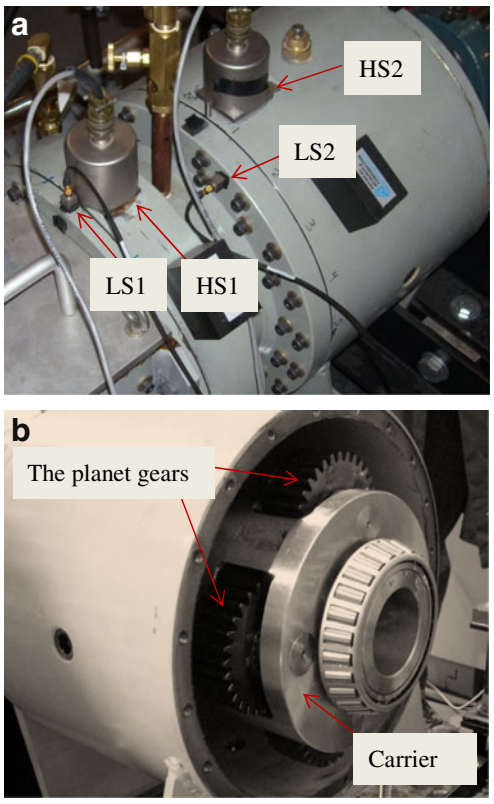

Fig. 4 **a** Location of four accelerometers on the test rig. **b** Location of the planet gears in the second-stage planetary gearbox

4 Fault level diagnosis

We introduce the procedure of pitting level diagnosis as follows. First, raw vibration signals collected from the data

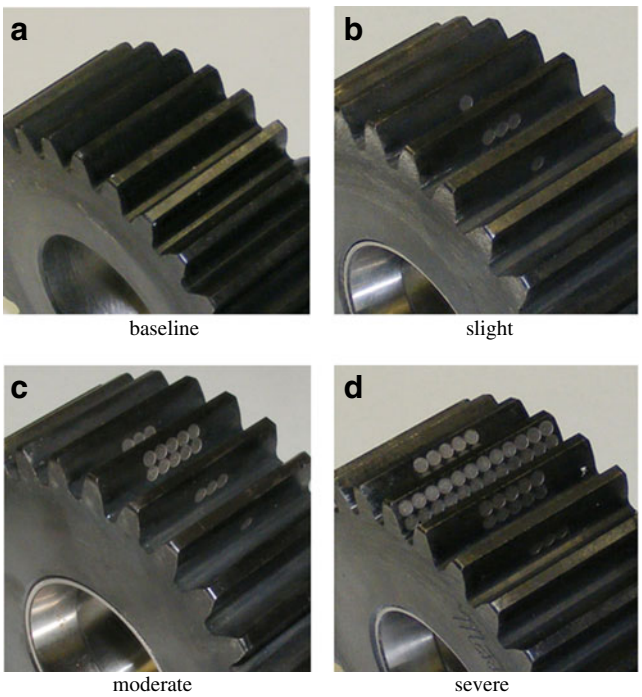

Fig. 5 Planet gears with artificially created pitting damage of four levels

acquisition system have to be pre-processed. That is, the 10-min-long data were split into 20 time records with an equal length that can cover the lowest frequency component of interested, i.e., the second-stage carrier frequency under the drive motor speed of 300 rpm. A bunch of features are calculated on each time record according to their definitions and corresponding signal processing techniques. The original feature space is established for fault diagnosis. Secondly, the proposed dimension reduction method is used to reduce feature dimension. Finally, fault level diagnosis is solved as a classification problem.

4.1 Signal processing and feature calculation

Feature calculation is crucial to the success of classification-based fault diagnosis, so we would like to make sure that all relevant features of candidate are already included in the original feature pool.

Unlike fixed-shaft gearboxes, sidebands of planetary gearboxes appear at integer multiples of planet passing frequency, and the sideband with the largest magnitude is the closest one to gear mesh frequency (GMF) [12]. In this paper, the upper sideband of the first-order sidebands is defined as the sideband closest to the GMF from the right-hand side, and the lower sideband is defined as the sideband closest to the GMF from the left-hand side. The regular mesh components (RMCs) are defined as the fundamental shaft frequency with its first-order sidebands.

Feature calculation usually comes with signal processing techniques. Five types of signals in signal processing in Fig. 6 are raw signals (RAW), residual signals (RES), difference signals (DIFF), band-pass mesh signals (BPM), and frequency spectrum signals (FS). RAW is derived as the collected vibration signals minus their mean. FS is the frequency spectrum of RAW. DIFF is RAW excluding RMCs. RES is the same as DIFF except that it includes the first-order sidebands. BPM is the band-pass mesh signal obtained by using a band-pass filter filtering around the first-order sidebands. RAW, DIFF, RES, BPM, and FS are denoted by $x(n)$, $d(n)$, $r(n)$, $b(n)$, and $X(k)$, respectively.

We compute 34 features based on the five signals according to their definitions in Table 2. The 34 features consist of 26 features in the time domain, 4 features in the frequency domain, and 4 features modified for planetary gearboxes. Features F1 to F16 are time domain features commonly used for fault diagnosis of generic systems. Features F17 to F26 are proposed for gear damage detection of fixed-shaft gearboxes. Features F27 to F30 are features of frequency domain. Features F31 to F34 are developed for fault diagnosis of planetary gearboxes. Details about feature calculation are provided in [13, 33–36].

The 34 features are extracted from one sensor, and 136 features are extracted from four sensors. To label them uniquely, the 136 features are numbered thus: F1–F34 from LS1, F35–F68 from LS2, F69–F102 from HS1, and F103–F136 from HS2. For features from one sensor, the definitions follow the same order in Table 2. For example, F35 from LS2 is the first feature in LS2, i.e., the maximum value. Finally, *the pitting damage dataset* is established by 640 samples × 136 features.

4.2 Application of the proposed method

In this section, we compare the proposed method with another reported method in Ref. [4] for fault level diagnosis. The method in Ref. [4] is called "Lei's method" in the following discussion. *k*-nearest neighbor is employed as the classifier to assess performance of the two methods.

In Section 4.1, the pitting damage dataset is established by 640 instances × 136 features after feature calculation. The proposed hybrid dimension reduction method is applied to this dataset. Since feature effectiveness is scaled into the same range as feature correlation, we treat the feature effectiveness term and the feature correlation term identically in the feature selection by specifying the same value to the two parameter α_1 and α_2.

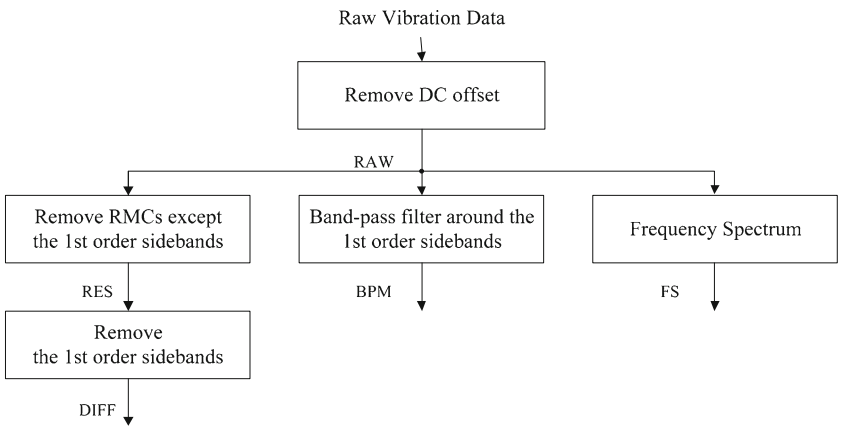

Fig. 6 Processing flow of vibration signals for feature calculation

Table 2 Definitions of the 34 features

No.	Name	Definition	No.	Name	Definition
F1	Maximum value	$\max x(n)$	F2	Minimum value	$\min x(n)$
F3	Average absolute value	$\frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert$	F4	Peak to peak	F1−F2
F5	Variance	$\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^2$	F6	Standard deviation	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^2}$
F7	Skewness	$\frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^3}{\mathrm{F6}^3}$	F8	Kurtosis	$\frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^4}{\mathrm{F5}^2}$
F9	Root mean square (RMS)	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}x(n)^2}$	F10	Crest factor	F1/F9
F11	Clearance factor	$\frac{\mathrm{F1}}{\frac{1}{N}\sum_{n=1}^{N}x(n)^2}$	F12	Impulse factor	F1/F3
F13	Shape factor	F9/F3	F14	Delta RMS	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}x_m(n)^2}-\sqrt{\frac{1}{N}\sum_{n=1}^{N}x_{m-1}(n)^2}$
F15	Energy ratio	$\frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^2}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^2}}$	F16	Energy operator	$\frac{\frac{1}{N}\sum_{n=1}^{N}(\Delta x(n)-\Delta\overline{x})^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(\Delta x(n)-\Delta\overline{x})^2\right)^2}$
F17	NA4	$\frac{\frac{1}{N}\sum_{n=1}^{N}(r(n)-\overline{r})^4}{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{n=1}^{N}(r_m(n)-\overline{r}_m)^2\right)\right)^2}$	F18	NA4*	$\frac{\frac{1}{N}\sum_{n=1}^{N}(r(n)-\overline{r})^4}{\left(\frac{1}{M'}\sum_{m=1}^{M'}\frac{1}{N}\sum_{n=1}^{N}(r_m(n)-\overline{r}_m)^2\right)^2}$
F19	FM4	$\frac{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^4}{\left(\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^2\right)^2}$	F20	FM4*	$\frac{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^4}{\left(\frac{1}{M'}\sum_{m=1}^{M'}\left(\frac{1}{N}\sum_{n=1}^{N}\left(d_m(n)-\overline{d}_m\right)^2\right)\right)^2}$
F21	M6A	$\frac{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^6}{\left(\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^2\right)^3}$	F22	M6A*	$\frac{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^6}{\left(\frac{1}{M'}\sum_{m=1}^{M'}\left(\frac{1}{N}\sum_{n=1}^{N}\left(d_m(n)-\overline{d}_m\right)^2\right)\right)^3}$
F23	M8A	$\frac{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^8}{\left(\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^2\right)^4}$	F24	M8A*	$\frac{\frac{1}{N}\sum_{n=1}^{N}\left(d(n)-\overline{d}\right)^8}{\left(\frac{1}{M'}\sum_{m=1}^{M'}\left(\frac{1}{N}\sum_{n=1}^{N}\left(d_m(n)-\overline{d}_m\right)^2\right)\right)^4}$
F25	NB4	$\frac{\frac{1}{N}\sum_{n=1}^{N}(e(n)-\overline{e})^4}{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{n=1}^{N}(e_m(n)-\overline{e}_m)^2\right)\right)^2}$	F26	NB4*	$\frac{\frac{1}{N}\sum_{n=1}^{N}(e(n)-\overline{e})^4}{\left(\frac{1}{M'}\sum_{m=1}^{M'}\left(\frac{1}{N}\sum_{n=1}^{N}(e_m(n)-\overline{e}_m)^2\right)\right)^2}$
F27	Mean frequency	$\frac{1}{K}\sum_{k=1}^{K}X(k)$	F28	Frequency center	$\frac{\sum_{k=1}^{K}(f(k)\times X(k))}{\sum_{k=1}^{K}X(k)}$
F29	Root mean square frequency		F30	Standard deviation frequency	

Table 2 (continued)

No.	Name	Definition	No.	Name	Definition
		$\sqrt{\frac{\sum_{k=1}^{K}\left(f(k)^2 \cdot X(k)\right)}{\sum_{k=1}^{K} X(k)}}$			$\sqrt{\frac{\sum_{k=1}^{K}\left((f(k)-\mathrm{F28})^2 \cdot X(k)\right)}{\sum_{k=1}^{K} X(k)}}$
F31	Largest sideband amplitude	$\max\left(X\left(k^*\right)\right)$	F32	FM0	$\frac{\mathrm{F4}}{\sum X(k^*)}$
F33	Sideband index	$x_{\mathrm{SI}} = \frac{\sum X(k^*)}{2}$	F34	Sideband level factor	$x_{\mathrm{SLF}} = \frac{\sum X(k^*)}{\mathrm{F6}}$

$\Delta x(n)$ is obtained piecewise. For non-endpoints, it is obtained by squared $x(n)$ minus the product of the data points of $x(n-1)$ and $x(n+1)$. For endpoints, the data point of $x(n)$ is looped around. The bar notation represents the mean, e.g., $\bar{x}$ represents the mean of $x(n)$

$X(k)$ the kth measure of the frequency spectrum of $x(n)$, (fk) the frequency value of the kth spectrum line, where $k=1, 2,\ldots, K$, $x_m(n)$ RAW for the mth time record, $r_m(n)$ RES for the mth time record, $d_m(n)$ DIFF for the mth time record, M the total number of time records up to the present, M' the total number of time records in which the gearbox is "healthy" (see Ref. [36] for details of estimating variance for a "healthy" gearbox), $e(n)$ the envelope of the current time record expressed as $e(n)=|b(n)+j\times H(b(n))|$ (where $H(b(n))$ represents the Hilbert transform of $b(n)$ and $e_m(n)$ represents the envelope of the mth time record signal), k^* the index of the first-order sidebands

Kernel parameter selection has to be conducted first for kernel feature selection and KFDA. The proposed parameter selection method for the Gaussian RBF kernel is illustrated in Fig. 7. The optimal σ determined by Newton's method is marked by red circles. The top sub-figure in Fig. 7 shows that the optimal σ is equal to 7.0781 for the kernel feature selection method that handles the original feature space with 136 features. With the optimal σ, the features are sequenced from first to 136th according to their combined scores. We select n' top features ($n' \leq 136$) for the next stage of KFDA, e.g., $n'=68$ [37]. The bottom sub-figure in Fig. 7 shows that the optimal σ is equal to 3.7248 for KFDA that deals with the reduced feature space with the top 68 features.

Figure 8 shows scores of feature effectiveness (the top sub-figure), normalized scores of feature effectiveness (the middle sub-figure), and combined scores (the bottom sub-figure), respectively. Features are sequenced according to their combined scores. The red dashed line in the bottom sub-figure is the threshold for selecting half of the top features. The top two features (F7 and F20) in the proposed kernel feature selection are highlighted in Fig. 8, and they are plotted in Fig. 9. Although F7 and F20 are the top two features, instances from different classes overlap a lot. Figure 10 shows that with the same number of feature, the top two features obtained by KFDA can produce more classification information.

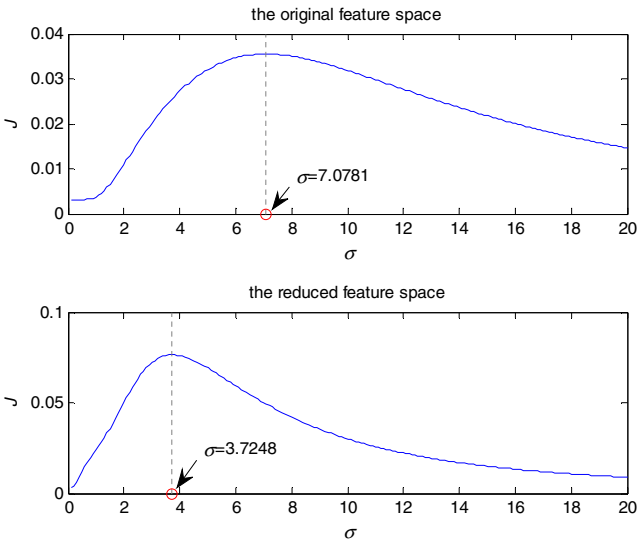

Fig. 7 Class separability with respect to σ on the original feature space (the *top one*) and the reduced feature space (the *bottom one*)

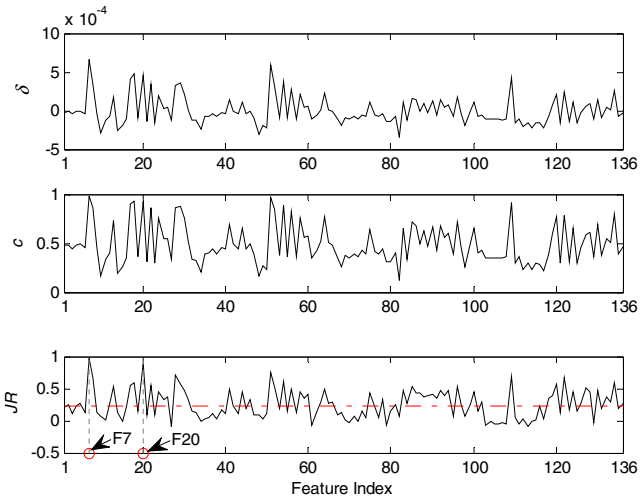

Fig. 8 Scores of the 136 features obtained in the proposed feature selection method (δ scores of feature effectiveness, c normalized scores of feature effectiveness, JR combined scores)

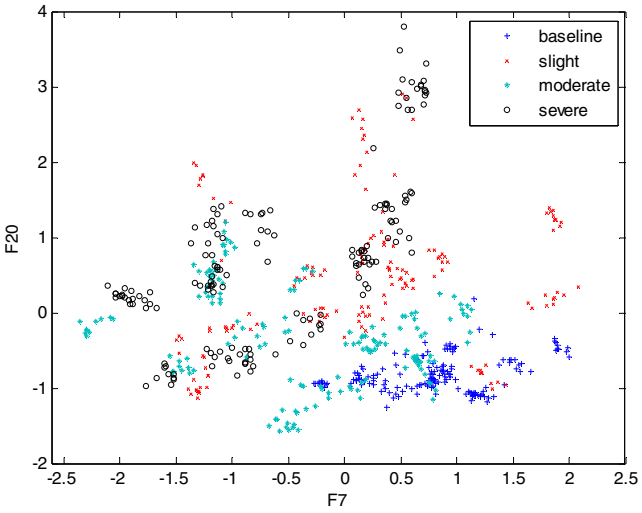

Fig. 9 Scatter plot on the two features: F7 and F20

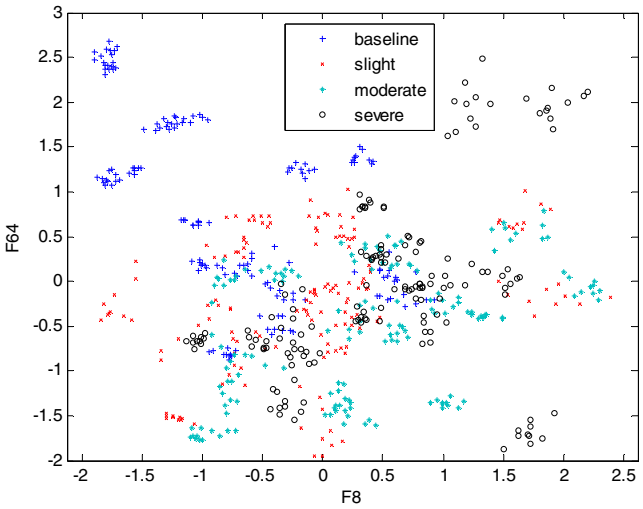

Fig. 11 Scatter plot with the top two features in Lei's method

In KFDA, we can obtain three (4 classes−1) transformed features corresponding to three discriminant directions. For illustration of the proposed method, we use top $n'=68$ features for the reduced feature space. The top two features obtained from KFDA are plotted in Fig. 10. The top two features from Lei's method are plotted in Fig. 11 for comparison. In Fig. 10, class boundaries are clearly identified for four pitting levels. The top two features in Lei's method, i.e., F8 and F64, are hard to distinguish class boundaries, which is similar to the case in Fig. 9. This demonstrates that KFDA can produce more informative individual features than feature selection does. This is not surprising because KFDA optimizes class separability when generating the new features. As stated earlier, the top two features shown in Fig. 10 lost their physical meanings.

In the following discussion, we demonstrate the performance of the proposed method and Lei's method from three aspects: the number of selected features (n'), the parameter k in k-NN, and the training ratio.

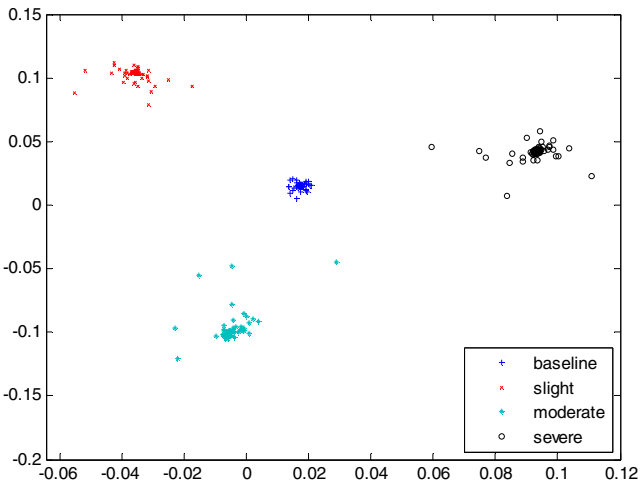

Fig. 10 Scatter plot with the top two features in the proposed method

We first analyze the impact of the number of selected features on the two methods. Following the sequence determined by feature selection, we start with an empty feature set and add one feature at a time until all 136 features are involved in the feature set. Classification accuracy of fivefold cross-validation is computed once a new feature is added. The training ratio of fivefold cross-validation is 0.8. The value of k in k-NN is specified by three. Figure 12 shows classification accuracy with respect to n'. From Fig. 12, the proposed method initially obtains the highest accuracy around $n'=24$. Lei's method initially obtains the highest accuracy around $n'=60$. In addition, the highest accuracy in the proposed method (99.99 %) is slightly larger than the highest one in Lei's method (99.95 %). The proposed feature selection method can reduce feature dimension more and obtain higher classification accuracy than Lei's method.

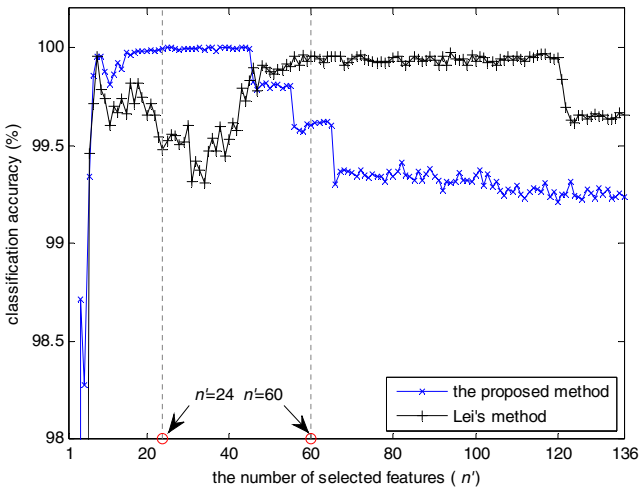

Fig. 12 Classification accuracy with respect to the number of selected features

Performance of the proposed method decreases significantly if $n'>43$. This shows that KFDA is sensitive to irrelevant features as reported in [17, 18] because the features in the latter order contain more noise. If these features are involved in the feature subset, they could misguide KFDA to obtain good discriminant directions. Performance of the proposed method thus deteriorates considerable with the increase of $n'>43$. Lei's method also shows decreased accuracy if $n'>120$. The observations about the proposed method and Lei's method support our claims in Section 1, that is, performance deterioration with a large feature size. Dimension reduction is thus necessary to do for classification-based fault diagnosis. If $n'=136$, the proposed method degenerates KFDA that is used for bearing fault diagnosis in [16]. Classification accuracy in this case decreases substantially. This implies that the proposed feature selection method can improve performance of the proposed method, and the way of combining kernel feature selection and KFDA outperform KFDA alone.

From Fig. 12, we specify n' by 24 and 60 for the proposed method and Lei's method, respectively as the two selections can achieve best performance for the two methods. We use the two selections for the following analysis of the value of k and the training ratio.

We analyze the two methods on the value of k that is the number of nearest neighbors for decision making in k-NN as follows. Classification accuracy of fivefold cross-validation is computed for a set of k values ranging from one to 20. Classification accuracy is shown in Fig. 13. The proposed method keeps good generality for the set of k values, while performance of Lei's method degrades with increase of k values. With increase of k, more training instances close to class boundaries may be used for decision-making in k-NN. Class boundaries in the reduced feature space obtained by Lei's method are not as clearly identified as those of the proposed method. The performance of Lei's method thus decreases. The proposed method uses KFDA to pursue discriminant directions that have large between-class variance and small within-class variance. The proposed method thus retains robustness with increase of k. Figures 9 and 10 support this explanation. The instances on the top two features F7 and F20 overlap a lot for the four pitting levels in Fig. 9, while class boundaries are clearly identified with the top two features in KFDA. In addition, from the perspective of k-NN, the proposed method reduces dimension from 136 to three, and Lei's method reduces dimension from 136 to 60. The proposed method achieves a higher reduction ratio than Lei's method.

Finally, we analyze the two methods with respect to training ratios. The value of k in k-NN is specified by three. Classification accuracy is obtained with respect to each training ratio in a set of {1/10, 1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2, 2/3, 3/4, 4/5, 5/6, 6/7, 7/8, 8/9, and 9/10}. Results are shown in Fig. 14. Generally speaking, the proposed method outperforms Lei's method. In particular, the advantage of the proposed method is much more obvious with small training ratios, i.e., the training ratio of <0.5. For example, if the training ratio is 0.1, classification accuracy of the proposed method is 90 %, which is much higher than 77 % obtained by Lei's method. With small training ratios, classifiers (e.g., k-NN) are insufficiently trained because small size training sets usually contain partial information. Classification accuracy thus decreases for the two methods. On the other hand, the proposed method can extract more representative classification information than Lei's method to prevent k-NN from deterioration.

5 Conclusions

This paper introduces a hybrid kernel dimension reduction method for fault level diagnosis. The proposed hybrid method consists of two sequential stages: kernel feature selection

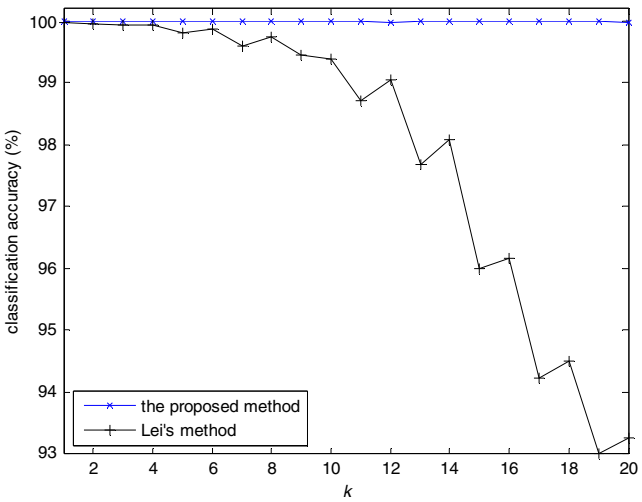

Fig. 13 Classification accuracy with respect to k

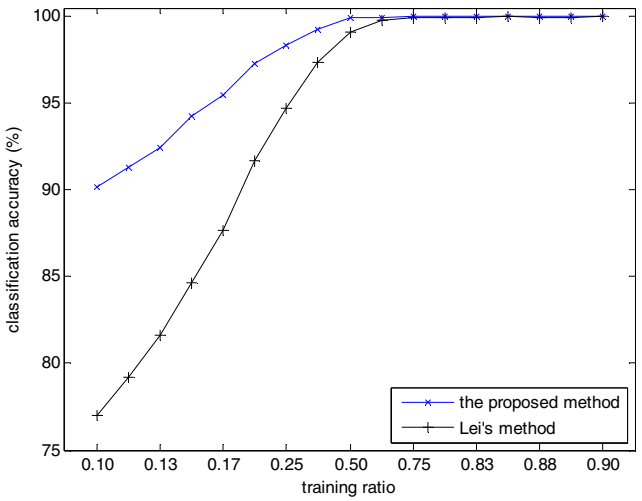

Fig. 14 Classification accuracy with respect to different training ratios

and KFDA. The method of taking an inverse order of kernel feature selection and KFDA fails to combine the advantages of the two types of dimension reduction techniques. The kernel feature selection method is proposed to evaluate features from two aspects: feature effectiveness and feature correlation. Feature effectiveness is measured by class separability in the kernel space, and feature correlation is measured by the Pearson correlation coefficient. A kernel parameter selection method is presented for the Gaussian RBF to determine the optimal values for the kernel feature selection and KFDA. The proposed method is applied to fault level diagnosis with four pitting levels on planet gears, i.e., baseline, slight, moderate, and severe. The experimental results demonstrate that the proposed hybrid dimension reduction method outperforms Lei's method and KFDA.

The following observations are made about the proposed method:

1. The proposed method implicitly assumes normal distribution in the kernel space. In the kernel feature selection method, within-class separability and between-class separability are estimated by average cosine similarity. If the assumption is not satisfied, within-class separability and between-class separability are not properly estimated. The kernel feature selection thus fails to recognize relevant features. On the other hand, normal distribution in the original feature space is made in FDA. KFDA accordingly assumes normal distribution in the kernel space because KFDA is actually the conventional FDA in the kernel space. If the assumption is not satisfied, within-class scatter matrix and between-class scatter matrix cannot reflect variance in a dataset. KFDA is thus misguided for finding the optimal discriminant directions.
2. The proposed method saves computation time for fault diagnosis because (1) it can permanently remove redundant and irrelevant features by feature selection and (2) it can produce a smaller number of individual features with more compact classification information by KFDA.

Acknowledgments The Natural Sciences and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC) supported this research. We highly appreciate the valuable comments of the anonymous reviewers.

References

1. Liu Z, Qu J, Zuo MJ, Xu H (2011) Classification for gear damage levels of planetary gearboxes. In: Proceedings of IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Ottawa
2. Hameed Z, Hong YS, Cho YM, Ahn SH, Song CK (2009) Condition monitoring and fault detection of wind turbines and related algorithms: a review. Renew Sustain Energ Rev 13(1):1–39
3. Samuel PD, Pines DJ (2005) A review of vibration-based techniques for helicopter transmission diagnostics. J Sound Vib 282(1–2):475–508
4. Lei Y, Zuo MJ (2009) Gear crack level identification based on weighted k nearest neighbor classification algorithm. Mech Syst Signal Process 23(5):1535–1547
5. Lei Y, He Z, Zi Y, Hu Q (2008) Fault diagnosis of rotating machinery based on a new hybrid clustering algorithm. Int J Adv Manuf Technol 35:968–977
6. Li R-Q, Chen J, Wu X, Alugongo AA (2005) Fault diagnosis of rotating machinery based on SVD, FCM and RST. Int J Adv Manuf Technol 27:128–135
7. Feng Z, Zuo MJ, Chu F (2010) Application of regularization dimension to gear damage assessment. Mech Syst Signal Process 24(4):1081–1098
8. Combet F, Gelman L (2009) Optimal filtering of gear signals for early damage detection based on the spectral kurtosis. Mech Syst Signal Process 23(3):652–668
9. Öztürk H, Sabuncu M (2008) Early detection of pitting damage in gears using mean frequency of scalogram. J Vib Control 14 (4):469–484
10. Loutridis SJ (2008) Self-similarity in vibration time series: application to gear fault diagnostics. Trans ASME J Vib Acoust 130:0310041–0310049
11. Loutridis SJ (2006) Instantaneous energy density as a feature for gear fault detection. Mech Syst Signal Process 20(5):1239–1253
12. Inalpolat M, Kahraman A (2009) A theoretical and experimental investigation of modulation sidebands of planetary gear sets. J Sound Vib 323(3–5):677–696
13. Keller JA, Grabill P (2003) Vibration monitoring of UH-60A main transmission planetary carrier fault. American Helicopter Society 59th Annual Forum, Phoenix
14. Jiang Q, Jia M, Hu J, Xu F (2009) Machinery fault diagnosis using supervised manifold learning. Mech Syst Signal Process 23 (7):2301–2311
15. Zhang K, Li Y, Scarf P, Ball A (2011) Feature selection for high-dimensional machinery fault diagnosis data using multiple models and radial basis function networks. Neurocomputing 74(17):2941–2952
16. Zhang J-F, Huang Z-C (2005) Kernel Fisher discriminant analysis for bearing fault diagnosis. In: Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Wuhan
17. Dundar MM, Fung G, Bi J, Sandilya S, Rao B (2005) Sparse fisher discriminant analysis for computer aided detection. In: Proceedings of SIAM International Conference on Data Mining, Newport Beach
18. Sierra A (2002) High-order Fisher's discriminant analysis. Pattern Recogn 35(6):1291–1302
19. Wang L (2008) Feature selection with kernel class separability. IEEE Trans Pattern Anal Mach Intell 30(9):1534–1546
20. Apaydin T, Ferhatosmanoglu H (2006) Access structures for angular similarity queries. IEEE Trans Knowl Data Eng 18 (11):1512–1525
21. Guyon I, Weston J, Barnhill S (2002) Gene selection for cancer classification using support vector machine. Mach Learn 46(1–3):389–422
22. Zhao X, Zuo MJ, Liu Z, Hoseini M (2012) Diagnosis of artificially created surface damage levels of planet gear teeth using ordinal ranking. Measurement (in press)
23. Theodoridis S, Koutroumbas K (2008) Pattern Recogn, 4th edn. Academic, Amsterdam
24. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR (1999) Fisher discriminant analysis with kernels. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop
25. Cai D, He X, Han J (2008) SRDA: An efficient algorithm for large-scale discriminant analysis. IEEE Trans Knowl Data Eng 20 (1):1–12

26. Xu Z, Dai M, Meng D (2009) Fast and efficient strategies for model selection of Gaussian support vector machine. IEEE Trans Syst Man Cybern B Cybern 39(5):1292–1307
27. Liu Z, Zuo MJ, Xu H (2011) A Gaussian radial basis function based feature selection algorithm. In: Proceedings of IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Ottawa
28. Chong EKP, Żak SH (2008) An Introduction to optimization, 3rd edn. Wiley, Hoboken
29. Zhou X, Lei Y, Wolfe D, Zuo MJ (2008) Gear stress calculation of the planetary gearbox system (Chinese standard). Technical report, Department of Mechanical Engineering, University of Alberta, Edmonton
30. Hoseini M, Zuo MJ (2008) A literature survey on the creating and quantifying faults on planetary gearbox. Technical report, Department of Mechanical Engineering, University of Alberta, Edmonton
31. Hoseini M, Zuo MJ (2009) A literature survey on the creating and quantifying faults on planetary gearbox. Technical report, Department of Mechanical Engineering, University of Alberta, Edmonton
32. Author of Publication, A-B-C Phase Diagram, ASM Alloy Phase Diagrams Center, Villars P, Okamoto H, Cenzual K (2006) ASM International, Materials Park, OH. Available from http://www1.asminternational.org/AsmEnterprise/APD
33. Lebold M, McClintic K, Campbell R, Byington C, Maynard K (2000) Review of vibration analysis methods for gearbox diagnostics and prognostics. In: Proceedings of the 54th Meeting of the Society for Machinery Failure Prevention Technology, Virginia Beach
34. Večeř P, Kreidl M, Šmíd R (2005) Condition indicators for gearbox condition monitoring systems. Acta Polytechnica 45(6):35–43
35. Zuo MJ, Li W, Fan XF (2005) Statistical methods for low speed planetary gearbox monitoring. Technical report, Department of Mechanical Engineering, University of Alberta, Edmonton
36. Decker HJ (2002) Crack detection for aerospace quality spur gears. NASA/TM-2002-211492, ARL-TR-2682
37. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20 (8):832–844