Feature selection for damage degree classification of planetary gearboxes using support vector machine

J Qu, Z Liu, M J Zuo and H-Z Huang

Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 2011 225: 2250 originally published online 27 June 2011
DOI: 10.1177/0954406211404853

Feature selection for damage degree classification of planetary gearboxes using support vector machine

J Qu[1], **Z Liu**[1,2], **M J Zuo**[1]*, and **H-Z Huang**[3]
[1]Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta, Canada
[2]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[3]School of Mechatronics Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

The manuscript was received on 25 June 2010 and was accepted after revision for publication on 4 March 2011.

DOI: 10.1177/0954406211404853

Abstract: Feature selection is an effective way of improving classification, reducing feature dimension, and speeding up computation. This work studies a reported support vector machine (SVM) based method of feature selection. Our results reveal discrepancies in both its feature ranking and feature selection schemes. Modifications are thus made on which our SVM-based method of feature selection is proposed. Using the weighting fusion technique and the one-against-all approach, our binary model has been extensively updated for multi-class classification problems. Three benchmark datasets are employed to demonstrate the performance of the proposed method. The multi-class model of the proposed method is also used for feature selection in planetary gear damage degree classification. The results of all datasets exhibit the consistently effective classification made possible by the proposed method.

Keywords: feature selection, support vector machine, damage degree classification, planetary gearbox

1 INTRODUCTION

Planetary gearboxes are a key component in certain rotating machinery used in automotive, aerospace, and industrial applications. During operation, the planetary gears experience cyclic stress which ultimately results in gear damage such as pits and cracks. This damage will progress until a failure occurs. For this reason, non-intrusive measurement of gear damage for classification of its degree and rate of growth provides users with the information needed in order to schedule preventive actions. In this way, serious consequences such as system breakdowns, injuries, and fatalities can be prevented.

Feature selection is an effective way of achieving good damage degree classification. The goal of feature selection is to eliminate irrelevant and redundant features to enhance the generalization ability of a given classifier [**1**]. Other advantages of feature selection include reducing feature dimension and speeding up classification computation. The applications of feature selection in damage detection, damage mode classification, and damage degree classification have been reported in many published papers. Malhi and Gao [**2**] developed a principal component analysis based method of feature selection for detecting bearing damage. Fei *et al.* [**3**] used genetic algorithm and support vector machine (SVM) to select useful features for classifying the damage mode of seafloor petroleum pipelines. Qu and Zuo [**4**] classified degree of damage in slurry pump system impellers using the backward feature selection algorithm to process data.

SVM is a relatively new machine learning method based on Vapnik–Chervonenkis theory [**5**], which recently emerged as a general mathematical framework for estimating dependency from finite samples.

*Corresponding author: Department of Mechanical Engineering, University of Alberta, 4-9 Mechanical Engineering Building, Edmonton, Canada.
email: ming.zuo@ualberta.ca

The performance of SVM classification has been studied by many researchers. Widodo and Yang [**6**] summarized the applications of SVM in fault diagnosis of rotating machinery. Samanta [**7**] compared the performance of artificial neural network and SVM in classifying the damage mode of fixed shaft gearboxes. Khawaja *et al.* [**8**] developed a one-class classifier for detecting crack on planetary gear plates using the least squares SVM.

Recently, Gualdrón *et al.* [**9**] reported a feature selection method for multi-sensor systems based on the norm of the weight vector of SVM classification. The method has been studied and the drawbacks are found in both the measure for feature ranking and the scheme for feature selection. The measure for feature ranking of the reported method cannot effectively evaluate the importance of features to classification. The modification is made on the measure based on which our new feature ranking is developed. The feature selection scheme of the reported method relies considerably on the quality of its feature rank. A feature selection scheme which recursively eliminates useless features based on updated rank is developed. The performance of our proposed feature selection method is studied using three benchmark datasets and one planetary gear damage dataset. The remaining parts of this article are organized as follows. The basics of SVM classification are presented in Section 2. The reported SVM-based feature selection method is reviewed and the proposed method for feature selection is presented in Section 3. The demonstration of the proposed method based on three benchmark datasets is given in Section 4. The applications of the proposed method in feature selection for classifying the degree of planetary gear damage is presented in Section 5. The feature subsets obtained in Section 5 are further studied in Section 6. The conclusions are drawn in Section 7.

2 SUPPORT VECTOR MACHINE

This section covers the fundamentals of SVM, presented in a way similar to that used in Qu and Zuo [**4**]. Suppose that the training data, $\{\boldsymbol{x}_1, y_1; \boldsymbol{x}_2, y_2; \ldots; \boldsymbol{x}_M, y_M\}$, $\boldsymbol{x}_i \in \boldsymbol{R}^n$, have binary classes of $y_i \in \{1,-1\}$ $(i = 1, 2, \ldots, M)$. The labels of 1 and -1 represent the two classes. A separating plane (SP) in the input space can be expressed as follows, if the training data are linearly separable

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{1}$$

where $\boldsymbol{w} \in \boldsymbol{R}^n$ is a weight vector, b is a scalar, and T means the transpose operator. The parameters of $\boldsymbol{w}$ and b which define the location of SP are determined during the training process.

When the training data are non-linearly separable, equation (1) is no longer applicable. In SVM theory, one can introduce a mapping function, $\Phi(\cdot)$, which projects the original feature space onto a high-dimension feature space in which the training data can be linearly separated again. Figure 1 shows an example of such mapping.

The SP for non-linearly separable data can still be expressed in a linear form but only with the addition of the mapping, $\Phi(\cdot)$

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}) + b = 0 \tag{2}$$

A distinct SP should satisfy the following constraints

$$y_i f(\boldsymbol{x}_i) = y_i(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i) + b) \geq 1, i = 1, 2, \ldots, M \tag{3}$$

where an underlying constraint applied is that the predicted $f(\boldsymbol{x})$ value of data point $\boldsymbol{x}$ should have the same sign as its virtual class label.

Figure 2 is an example of a linearly separable classification problem in a two-dimensional feature

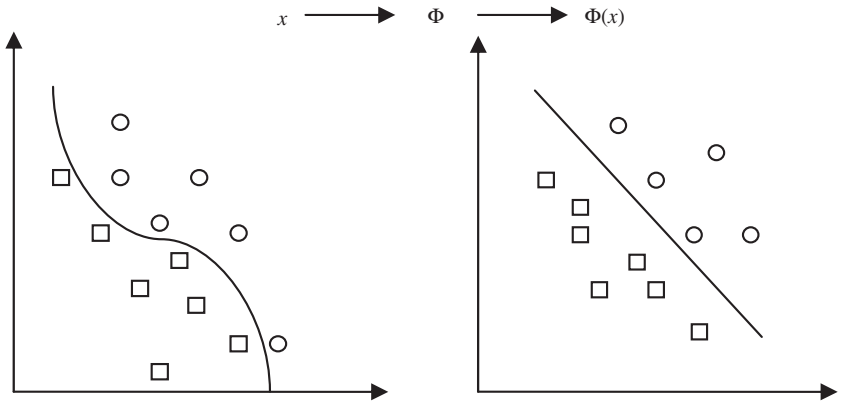

Fig. 1 An example of feature mapping enabling linear data separation (adapted from [**10**])

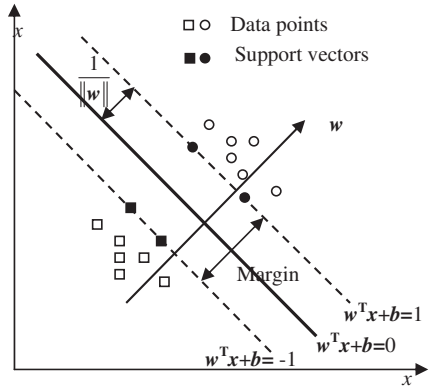

Fig. 2 An example of a linearly separable classification problem in $\boldsymbol{R}^2$

space where two classes of data points are labelled −1 (squares) and +1 (circles), respectively. The SP represented by the solid straight line merely relies on the solid squares and the solid circles called support vectors. Two parallel planes which cross the support vectors of each class determine the margin; this can be computed by

$$\text{Margin} = \left|[(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b-1) - (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b+1)]/\|\boldsymbol{w}\|\right| = 2/\|\boldsymbol{w}\|. \tag{4}$$

SVM searches for the SP that provides the largest margin. Basically, the larger the margin, the better the separation of data points. For this reason, one can use the margin value to quantify the importance to the classification result of a certain feature subset. This observation is the essential theoretical support for the feature selection method proposed in this article. It is explained in detail in Section 3.

To acquire the SP which can provide the largest margin, the following optimization model is established

$$\begin{aligned} &\text{Minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{M}\xi_i \\ &\text{Subject to } y_i\,(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b) \geq 1-\xi_i,\ i=1,2,\ldots,M \\ &\xi_i \geq 0,\ i=1,2,\ldots,M, \end{aligned} \tag{5}$$

where C is a positive constant and ξ_i represents the distance between the data point, $\boldsymbol{x}_i$, lying on the false class side and the margin of its virtual class. The slack variables, ξ_i, are adopted to allow some falsely classified data points in the training process which are useful for naturally non-separable data points [**10**]. The optimization model can be solved using the Lagrange multipliers α_i and β_i

$$\begin{aligned} L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{M}\xi_i \\ &\quad - \sum_{i=1}^{M}\alpha_i(y_i(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b)-1+\xi_i) - \sum_{i=1}^{M}\beta_i\xi_i \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_M)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_M)^{\mathrm{T}}$, and $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_M)^{\mathrm{T}}$.

For the optimal solution, the derivatives of the Lagrange function with respect to $\boldsymbol{w}$, b, and $\boldsymbol{\xi}$ should vanish

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \displaystyle\sum_{i=1}^{M}\alpha_i y_i \Phi(\boldsymbol{x}_i), \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \displaystyle\sum_{i=1}^{M}\alpha_i y_i = 0, \\ \dfrac{\partial L}{\partial \boldsymbol{\xi}} = 0 \rightarrow \alpha_i + \beta_i = C,\ i=1,\ldots,M. \end{cases} \tag{7}$$

Incorporating the first line of equation (7) into equation (6) yields

$$L(\alpha) = \sum_{i=1}^{M}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{M}\alpha_i\alpha_j y_i y_j \Phi^{\mathrm{T}}(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j) - b\sum_{i=1}^{M}\alpha_i y_i, \tag{8}$$

Incorporating the second line of equation (7) into equation (8) yields a dual maximization problem

$$\begin{aligned} &\text{Maximize } L(\alpha) = \sum_{i=1}^{M}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{M}\alpha_i\alpha_j y_i y_j \Phi^{\mathrm{T}}(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j), \\ &\text{Subject to } \begin{cases} \displaystyle\sum_{i=1}^{M} y_i\alpha_i = 0, \\ C \geq \alpha_i \geq 0,\ i=1,\ldots,M. \end{cases} \end{aligned} \tag{9}$$

Solving equation (9) yields the coefficients, α_i, which are required to express $\boldsymbol{w}$. Following the Karush–Kuhn–Tucker condition, the products between the dual variables and the constraints should be equal to zero at the optimal solution point

$$\alpha_i[\,y_i(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b)-1] = 0,\ i=1,2,\ldots,M. \tag{10}$$

From equation (10) it can be seen that the support vectors correspond to non-zero α_i. According to equations (7) and (10), the expressions of $\boldsymbol{w}$ and b can be obtained as

$$\boldsymbol{w} = \sum_{i=1}^{M}\alpha_i y_i \Phi(\boldsymbol{x}_i),\ b = \frac{1}{p}\sum_{j=1}^{p}[\,y_j - \boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_j)], \tag{11}$$

where p represents the number of support vectors. There is a value for b only when α_i is non-zero.

The linear decision function can thus be given as

$$i_{\mathrm{f}} = \mathrm{sign}\left(\sum_{i=1}^{M} \alpha_i y_i \Phi^{\mathrm{T}}(\boldsymbol{x}_i)\Phi(\boldsymbol{x}) + b\right) \tag{12}$$

where if i_{f} is positive, the new input data point, $\boldsymbol{x}$, belongs to class 1 ($y_i = 1$) and if i_{f} is negative, $\boldsymbol{x}$ belongs to class 2 ($y_i = -1$).

The function $\Phi(\cdot)$ is, however, usually difficult to compute. Because the mapping functions are in the form of an inner product in equation (12), SVM theory adopts a kernel function, namely $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi^{\mathrm{T}}(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)$, to avoid computing the mapping function explicitly. Any function that satisfies Mercer's theorem [**10**] can be used as a kernel function. More details on kernel functions can be found in Schölkopf and Smola [**11**]. The non-linear decision function can thus be expressed as

$$i_{\mathrm{f}} = \mathrm{sign}\left(\sum_{i=1}^{M} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \tag{13}$$

The SVM classifier is for binary classification problems. Some strategies can be used to adapt SVM for multi-class classification problems. One commonly used strategy is called one-against-all (OAA). The OAA strategy actually divides an N-class classification problem into N binary class classification problems in each of which the data points of one class are labelled $+1$ and the data points of the rest of the classes are labelled -1. This enables N sets of training data to train N SVM classifiers, individually. Ideally, when inputting a new data point into N SVM classifiers, only one classifier will return a positive value in equation (2). For some non-separable data, however, positive values may be returned by multiple classifiers. SVM adopts the following decision function in order to determine which class a data point, $\boldsymbol{x}$, belongs to

$$\begin{aligned}\text{Class of } \boldsymbol{x} &\equiv \arg\max_{j=1,\ldots,N}((\boldsymbol{w}^{(j)})^{\mathrm{T}}\Phi(\boldsymbol{x}) + b^{(j)}) \\ &= \arg\max_{j=1,\ldots,N}\left(\sum_{i=1}^{M} \alpha_i^{(j)} y_i^{(j)} K^{(j)}(\boldsymbol{x}_i, \boldsymbol{x}) + b^{(j)}\right)\end{aligned} \tag{14}$$

where the parameters with the superscript (j) are for the jth SVM classifier.

3 FEATURE SELECTION

This section first introduces an SVM-based feature selection method reported in the literature and then examines its feature ranking and feature selection schemes. Next, modifications are made on both schemes. Classification accuracy is used throughout this article to evaluate classification performance. N_c denotes the number of data that are correctly classified and N_f denotes the number of data that are falsely classified. Classification accuracy is defined as $N_c/(N_c + N_f) \times 100\%$. The reported method is introduced and evaluated in Subsection 3.1 and the proposed method is presented in Subsection 3.2. The multi-class model of the proposed method is developed in Subsection 3.3. Additional details of the proposed method are presented in Subsection 3.4.

3.1 Gualdrón's method

3.1.1 Introduction to Gualdrón's method

Recently, Gualdrón *et al.* [**9**] reported an SVM-based feature selection method for the classification of multi-sensor systems. They stated that the feature whose removal leads to a large variation of the norm of the weight vector, $\|\boldsymbol{w}\|$ as shown in equation (5), is of greater importance for classification. As a result, they proposed the following measure for feature ranking

$$\delta_i = |\|\boldsymbol{w}_0\| - \|\boldsymbol{w}_i\||, \; i = 1, 2, \ldots, L, \tag{15}$$

where L represents the number of features, $\|\boldsymbol{w}_i\|$ represents the norm of the weight vector when the ith feature is removed from the original feature space, and $\|\boldsymbol{w}_0\|$ represents the one when all features are in the original feature space. The features are ranked in accordance with the δ values. The feature having a larger δ value is placed ahead of the one having a smaller δ value. This δ value is hereafter called the variation value.

Gualdrón's method adopts forward selection (FS) as its feature selection scheme. One by one, it adds the top-ranked features to an empty feature set until classification accuracy is not increased in response to adding a particular feature. The resultant feature subset is then considered as the optimal feature subset.

3.1.2 Evaluation of Gualdrón's method

First, Gualdrón's feature ranking scheme is evaluated. As discussed in Section 2, SVM classification seeks the maximal margin value which gives the largest separation of data points. Because $\|\boldsymbol{w}\|$ is inversely proportional to the margin in SVM classification, one expects the value for $\|\boldsymbol{w}\|$ to be as small as possible. It is found that Gualdrón's method of using equation (15) for feature ranking is not appropriate and an example to explain why is given below.

Suppose that the original feature space has $n\,(n > 2)$ features. Features C and D are removed from the original feature space, obtaining the resultant $\|\boldsymbol{w}_{\mathrm{C}}\|$ and

$\|\boldsymbol{w}_D\|$, respectively. The $\|\boldsymbol{w}_0\|$ is obtained with the whole original feature space. Now, four scenarios are considered and the performance of Gualdrón's method for these four scenarios is surmmarized. For the sake of brevity, only one scenario arbitrarily selected is explained in detail. The results of other scenarios are summarized in a table, because they follow exactly the same rationale.

Scenario: $\|\boldsymbol{w}_C\| = \|\boldsymbol{w}_0\| + \delta_C$ and $\|\boldsymbol{w}_D\| = \|\boldsymbol{w}_0\| + \delta_D$ ($\delta_C > 0$ and $\delta_D < 0$)

Observations: The equality $\|\boldsymbol{w}_C\| = \|\boldsymbol{w}_0\| + \delta_C$, $(\delta_C > 0)$ indicates that removing feature C increases the $\|\boldsymbol{w}\|$ value (reduces the margin); this means the removal of feature C impairs the classification; therefore, feature C is important to the classification. Inversely, the equality $\|\boldsymbol{w}_D\| = \|\boldsymbol{w}_0\| + \delta_D$, $(\delta_D < 0)$ indicates that removing feature D reduces the $\|\boldsymbol{w}\|$ value (increases the margin); because its removal enhances the classification, feature D is harmful to the classification. It can thus be concluded that, for this scenario, no matter what the absolute values are for δ_C and δ_D, feature C is more important than feature D; this is denoted as $I_C > I_D$ where I represents the importance to classification of a particular feature.

Solution of Gualdrón's method: Gualdrón's method is studied under three conditions, Condition 1: $|\delta_C| < |\delta_D|$, Condition 2: $|\delta_C| = |\delta_D|$, and Condition 3: $|\delta_C| > |\delta_D|$. Using Gualdrón's equation (15), the following results are obtained. For Condition 1, feature D has a larger variation value than does feature C; therefore, Gualdrón's method would conclude that feature D has greater importance to classification than does feature C, i.e. $I_C < I_D$. For Condition 2, features C and D have the same variation value; therefore, Gualdrón's method would conclude that they have the same importance, i.e. $I_C = I_D$. For Condition 3, feature C has a larger variation value than does feature D; therefore, Gualdrón's method would conclude that feature C has greater importance to classification than does feature D, i.e. $I_C > I_D$.

It can be seen that Gualdrón's method fails to comply with the observations outlined earlier under Conditions 1 and 2. The results of other scenarios and conditions are given in Table 1. It is noted that Gualdrón's method fails to give correct conclusions in 6 (bold cells) out of 12 cases; so improvement is needed.

Second, the feature selection scheme is evaluated. Gualdrón's method adopted the FS which selects features based on only one feature rank. This operation may work when the features are independent of each other and a perfect rank is available. However, this is not always the case for practice; hence, the feature selection scheme also needs to be improved.

3.2 The proposed method

In accordance with Table 1, a measure for feature ranking is proposed, which is actually a transform of equation (15) achieved by removing the absolute sign and making the $\|\boldsymbol{w}_i\|$ be the minuend; this is given as

$$\delta_i = \|\boldsymbol{w}_i\| - \|\boldsymbol{w}_0\|, i = 1, 2, \ldots, L. \tag{16}$$

The measure of equation (16) allows for differentiating the importance of features based on the δ values. The smaller the δ value, the less important the corresponding feature. The observation results, as given in Table 1, can be easily obtained using equation (16), since the following causal relations are always applicable: if $\delta_C > \delta_D$, then $I_C > I_D$; if $\delta_C = \delta_D$, then $I_C = I_D$; and if $\delta_C < \delta_D$, then $I_C < I_D$. According to the δ values, the features can be ranked for feature selection.

FS and backward selection (BS) are two commonly used feature selection schemes. FS adds useful features to an empty feature set. BS eliminates useless features from the original feature set. As discussed, FS relies more on a perfect rank. In contrast, BS relies less on rank quality. Though BS may leave some useless features in the final feature subset, most useful features are able to be reserved. BS is more robust than FS in terms of acquiring good classification results, because it is usually difficult to ensure a perfect rank.

The BS scheme is adopted in the proposed method. The classification accuracy is employed to determine whether a particular feature should be removed. Unlike Gualdrón's FS where the feature ranking is

Table 1 Evaluation of Gualdrón's method

Conditions		$\|\delta_C\| > \|\delta_D\|$		$\|\delta_C\| = \|\delta_D\|$		$\|\delta_C\| < \|\delta_D\|$	
Scenarios		Observations	Gualdrón's method	Obsevations	Gualdrón's method	Observations	Gualdrón's method
$\|\boldsymbol{w}_C\| = \|\boldsymbol{w}_0\| + \delta_C$ and $\|\boldsymbol{w}_D\| = \|\boldsymbol{w}_0\| + \delta_D$	$\delta_C > 0, \delta_D > 0$	$I_C > I_D$	$I_C > I_D$	$I_C = I_D$	$I_C = I_D$	$I_C < I_D$	$I_C < I_D$
	$\delta_C > 0, \delta_D < 0$	$I_C > I_D$	$I_C > I_D$	$\boldsymbol{I_C > I_D}$	$\boldsymbol{I_C = I_D}$	$\boldsymbol{I_C > I_D}$	$\boldsymbol{I_C < I_D}$
	$\delta_C < 0, \delta_D > 0$	$\boldsymbol{I_C < I_D}$	$\boldsymbol{I_C > I_D}$	$\boldsymbol{I_C < I_D}$	$\boldsymbol{I_C = I_D}$	$I_C < I_D$	$I_C < I_D$
	$\delta_C < 0, \delta_D < 0$	$\boldsymbol{I_C < I_D}$	$\boldsymbol{I_C > I_D}$	$I_C = I_D$	$I_C = I_D$	$\boldsymbol{I_C > I_D}$	$\boldsymbol{I_C < I_D}$

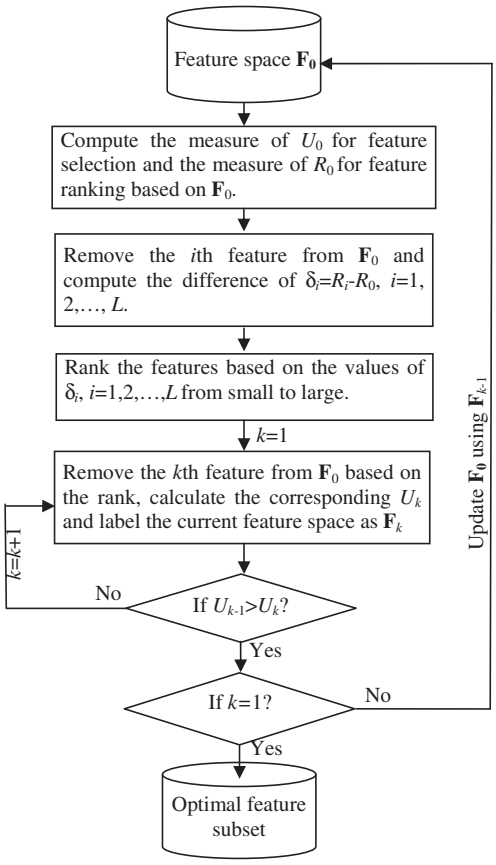

Fig. 3 The flow chart of the proposed feature selection

conducted only once, the proposed method re-ranks the features when no more features can be removed from the current feature space. As a result, BS will be recursively implemented until removing the first top-ranked feature decreases classification accuracy. This recursive backward selection (RBS) allows irrelevant and redundant features remaining in the reduced feature space to have multiple chances of being removed. There is, however, a side effect: it may take more computational time.

Figure 3 shows the flow chart of the proposed method where the measures of U and R represent classification accuracy and $\|\boldsymbol{w}\|$, respectively. The flow chart is introduced in detail in Subsection 3.3.

3.3 The multi-class model of the proposed method

Gualdrón's paper does not mention the explicit model of its feature selection method for multi-class classification problems. In this subsection, the binary model of the proposed method is extended for multi-class classification problems. The OAA approach described in Section 3 is adopted for SVM multi-class classification. When using the OAA approach, N binary SVM classification models are established. The equal weighting fusion technique [**12**] is employed to address these binary models. As a consequence, the norms of the weight vectors of N binary models are equally weighted to determine the rank of the features in the multi-class model. Figure 3 is still applicable; so the measure of U is identically defined as classification accuracy, but the measure of R is redefined to adapt to the multi-class cases given below.

As shown in Fig. 3, first of all, the U_0 (classification accuracy) is calculated based on the feature space $\mathbf{F_0}$. The value of $R_{0,j}$ ($\|\boldsymbol{w}_{0,j}\|$), $j=1,2,\ldots,N$ is also calculated; it is returned by the jth SVM model with all features in the $\mathbf{F_0}$ being used. Next, the impact of removing the ith ($i=1,2,\ldots,L$) feature on the jth ($j=1,2,\ldots,N$) SVM model represented by $\delta_{i,j}=\|\boldsymbol{w}_{i,j}\|-\|\boldsymbol{w}_{0,j}\|$, $i=1,2,\ldots,L$ and $j=1,2,\ldots,N$ is calculated, yielding an impact matrix

$$\begin{bmatrix} \delta_{1,1} & \delta_{2,1} & \cdots & \delta_{L,1} \\ \delta_{1,2} & \delta_{2,2} & \cdots & \delta_{L,2} \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{1,N} & \delta_{2,N} & \cdots & \delta_{L,N} \end{bmatrix} \tag{17}$$

The impact the removal of each feature has on N binary SVM classifications are quantitatively given in the N columns of equation (17). Because there is often no prior knowledge regarding which SVM model should be preferred in classifying a particular data point, the equal weighting technique is used to evaluate the overall impact of removing each feature. For the ith feature, this is given by $\Delta_i=\frac{1}{N}\sum_{j=1}^{N}\delta_{i,j}$, $i=1,2,\ldots,L$ and $j=1,2,\ldots,N$. The features are then ranked according to Δ values. The feature corresponding to the smallest Δ value has the top rank and the one corresponding to the largest Δ value has the lowest rank. Next, the features are removed one at a time starting from the top-ranked feature until removing a feature decreases the classification accuracy. The feature space, $\mathbf{F_0}$, is then updated by eliminating the removed features. The above procedure is repeated until the classification accuracy is decreased by the removal of the top-ranked feature (the least useful feature). The optimal feature subset is thus obtained using the original feature space without the features removed.

3.4 Additional details of the proposed method

3.4.1 Validation methods

The proposed method adopts K-fold cross-validation to overcome over-fitting of data. The K-fold cross-validation splits training data into K disjoint

subsets (folds). Classification accuracy is computed in which $K-1$ folds are used as training data and the remaining one as validation data. When each fold has been used as validation data once, the resultant classification accuracy values are averaged. The average is then used in the proposed method.

3.4.2 Feature ranking strategy

At the feature ranking step of the proposed method, it may occur that several features have the same δ value for binary class cases and the same Δ value for multiclass cases. For the binary class cases, the rankings of these features are arbitrarily selected. For the multiclass cases, the following strategy is taken. Suppose that features H and G have the same Δ value. The number of $\delta_{\mathrm{H},j} > \delta_{\mathrm{G},j}$, $j = 1, 2, \ldots, N$ is counted and it is denoted by n_{com}. If $n_{\mathrm{com}} > N/2$, feature G will be placed ahead of feature H in the rank. If $n_{\mathrm{com}} < N/2$, feature H will be placed ahead of feature G. Otherwise, the features are ranked arbitrarily. When more than two features have the same Δ value, this strategy is used for each pair of features.

4 EXPERIMENTS USING BENCHMARK DATA

4.1 Benchmark datasets

This section examines the performance of the feature selection method that is proposed. Three benchmark datasets are used; all are from the UCI Machine Learning Repository [**13**].

1. The sonar dataset contains 208 observations on 61 variables. The first 60 represent the energy within a particular frequency band, integrated over a certain period of time. The last column contains the class labels. There are two classes, 'R' if the object is a rock and 'M' if the object is a mine (metal cylinder).
2. The breast cancer dataset contains 569 samples of which 357 belong to benign (B) and 212 samples belong to malignant (M). The dataset includes 32 attributes with the ID number and the outcome, benign and malignant. There are 30 real-value features which are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image.
3. The Parkinson dataset is composed of a range of biomedical voice measurements from 31 people, 23 of whom have Parkinson's disease. Each attribute is a particular voice measurement, and there are 195 voice recordings from these individuals.

4.2 Preliminaries of the experiments

The three datasets are pre-arranged identically for the experiments. First, the original dataset is split into training dataset, validation dataset, and testing dataset. The testing dataset will be used to evaluate feature selection methods but will not be involved in the training process. Feature selection is conducted using the training and validation datasets. Once the optimal feature subset is selected, the whole dataset is updated with the selected features. Then, the training and validation datasets are integrated and used to train the SVM model.

The proposed method is compared with Gualdrón's method using the three benchmark datasets. In order to demonstrate the effectiveness of both the proposed feature ranking scheme and the proposed feature selection scheme, two proposed methods are considered: the proposed feature ranking + Gualdrón's feature selection (PFR + GFS) and the proposed feature ranking + the proposed feature selection (PFR + PFS). The comparison of Gualdrón's method and PFR + GFS reveals the effectiveness of the proposed feature ranking versus Gualdrón's feature ranking. The comparison of PFR + GFS and PFR + PFS reveals the effectiveness of the proposed RBS versus FS. In addition, SVM using all features is used as a baseline in relation to which the capabilities of the three methods of improving classification performance can be assessed.

4.3 Experiment results

Since the aim of this work is to study the effectiveness of the proposed method, the degree of improvement achieved using the proposed method is of utmost interest. With that in mind, our focus is not on optimizing the parameters of the experiments in order to obtain the highest classification accuracy. Instead, the parameters are equally selected for every dataset and, most importantly, for every method to be assessed in order to eliminate any deviations caused by factors other than the feature selection methods. The parameters for SVM are $C = 100$ and Gaussian kernel with a width parameter of one. To remove any unexpected singularities due to a particular dataset used, the results of each benchmark dataset are averaged over 30 trials.

Table 2 presents the results of the sonar dataset. With regard to classification accuracy, it is found that Gualdrón's method provided a value even smaller than the baseline value. In contrast, PFR + GFS increased the baseline value by about 3%. This value was further increased by 9%, when PFR + PFS

Table 2 The results for the sonar dataset

Methods	Classification accuracy (%)			Number of features selected		CPU time of feature selection (s)
	Mean	Standard	PBP	Mean	Standard	Mean
SVM using all features	77.55	2.26	0	60	0	—
Gualdrón's method	74.63	7.46	40	3	1.36	8.4729
The proposed method						
PFR + GFS	80.44	9.15	53.3	3	1.38	8.4328
PFR + PFS	89.58	4.95	100	20	7.76	20.819

was used. Similar results can also be observed from the column of 'number of features selected', the values for which are all rounded to their nearest integer. Though with the identical value of three, PFR + GFS provided classification accuracy greater than Gualdrón's method by about 6%. This shows that the proposed feature ranking possesses better capability of detecting useful features.

An index called the percentage of better performance (PBP) is also used to evaluate the robustness of feature selection methods. The PBP is calculated by dividing the number of trials where the classification accuracy is improved using a feature selection method by the total number of trials. It can be seen that PFR + PFS improves the classification accuracy for every trial (PBP = 100%); in contrast, Gualdrón's method has a PBP of 40% and PFR + GFS a PBP of 53.33%, indicating a much lower robustness. Figure 4 shows the classification accuracy of each method for all 30 trials.

It is noticed that PFR + PFS provides a larger optimal feature subset and consumes twice the CPU time of the other two methods. This is a side effect of the RBS; the larger the gain in classification accuracy, the more expensive the computations.

Table 3 presents the results of the breast cancer dataset. Basically, the three methods performed much as they did in the sonar dataset. Figure 5 shows the performance of each method over 30 trials, giving the four curves that are basically separated from each other. The curve of PFR + PFS is above all the others, revealing its superior robustness. The curve of Gualdrón's method is at the bottom, lying even below the baseline curve.

Unlike for the sonar dataset, PFR + GFS provided a good classification accuracy of 93.85% and a good PBP of 96.67%. Moreover, it selected fewer features in the final feature subset and used the least CPU time. If the requirement for classification accuracy is not very high, PFR + GFS is suitable for the breast cancer dataset because it provides a good balance among classification accuracy, the number of features selected, and computational time.

Table 4 and Fig. 6 show the results of the Parkinson dataset. It is apparent that Gualdrón's method and PFR + GFS provided almost the same numerical results and their curves almost merged as shown in Fig. 6. This is not surprising given that the six cases in the column of observations (non-bold cells) are fulfilled in Table 1. The two feature ranking schemes were studied and it was found that the δ values of equation (16) are positive for most of features except five with nearly zero δ values. These observations satisfy the scenario of the first row and the conditions of the first and the second columns in Table 1, from which it can be concluded that most of the features in the Parkinson dataset are useful and the rest are redundant.

This conclusion accords with the results of the row of 'SVM using all features' where a high classification accuracy of 94.43% is obtained without using a feature selection method. Although a reasonable rank is obtained, FS, however, failed to select as many useful features as possible, whereas the proposed RBS selected a subset containing 11 features and achieved an even higher classification accuracy of 96.15%.

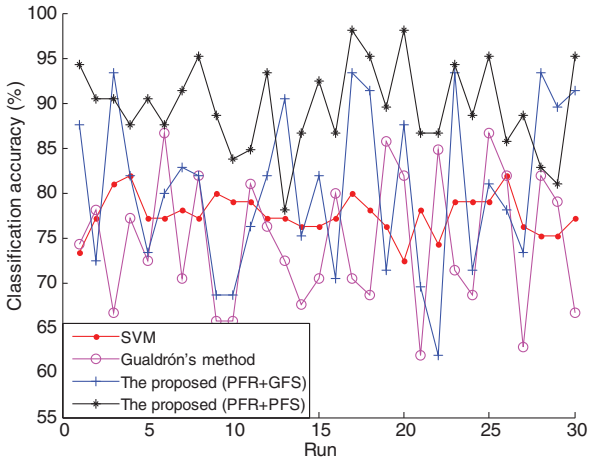

Fig. 4 The results of classification accuracy for the sonar dataset

Table 3 The results for the breast cancer dataset

Methods	Classification accuracy (%)			Number of features selected		CPU time of feature selection (s)
	Mean	Standard	PBP	Mean	Standard	Mean
SVM using all features	88.96	1.79	0	30	0	—
Gualdrón's method	79.72	6.98	3.33	2	0.76	182.96
The proposed method						
PFR + GFS	93.85	2.64	96.67	4	1.43	178.15
PFR + PFS	99.32	0.43	100	22	4.34	364.36

5 APPLICATION OF THE PROPOSED METHOD IN PLANETARY GEARBOXES

This section presents the problem of feature selection for classifying damage degree in planet gears. Descriptions of our test rig and how our experiments were conducted are given in Subsection 5.1. The features to be studied for the given problem are introduced in Subsection 5.2. The datasets used are established and the results are analysed in Subsection 5.3.

5.1 Test rig and experiment conduction

The test rig shown in Fig. 7 was designed to fully enable performing controlled experiments for developing a reliable diagnostic system for planetary gearboxes. The planetary gearbox has an over-hung floating configuration that mimics the support used in the field by Syncrude's mining operations. Its main components include one 20-HP drive motor, one stage bevel gearbox, two stages of planetary gearboxes, two stages of speed-up gearboxes, and one 40-HP load motor. Table 5 lists the number of teeth and the speed ratio achieved by each gearbox.

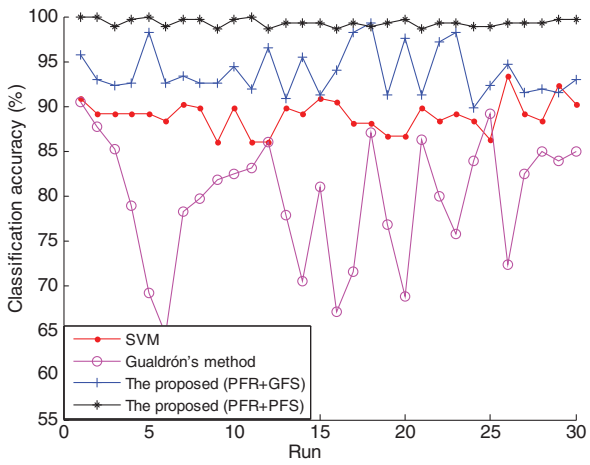

Fig. 5 The results of classification accuracy for the breast cancer dataset

The two-stage planetary gearboxes are our study object. There are four accelerometers located on the housing of the two-stage planetary gearboxes including two identical low sensitivity accelerometers (LS1 and LS2) and two identical high sensitivity accelerometers (HS1 and HS2), as shown in Fig. 8. Figure 9 provides a schematic diagram for the structures of the planetary gearboxes.

The experiments were conducted using planet gears with different degrees of pitting damage. The pitting damage was artificially created on one planet gear of the second stage planetary gearbox. To mimic the pits observed on actual pitted gears, circular holes were created along the pitch line of the gear tooth surface. The number of holes was varied for different degrees of damage. Four damage degrees were considered: baseline, slight, moderate, and severe. A brand new gear was used as the baseline. Figure 10 illustrates the four degrees of damage. More details on the creation of the pitting damage can be found in Hoseini and Zuo [**14**].

There are four planet gears in the second stage planetary gearbox (Table 5). Three normal and one artificially damaged planet gears were installed in the test rig and the experiments were conducted. Upon finishing one experiment, the artificially damaged gear was replaced with the one having another damage degree. This procedure was repeated until all four artificially damaged plant gears were tested.

For each damage degree, experiments were conducted on two separate days. For each day, the load and the drive motor speed were varied. The two load conditions involved were 'no load' and 'load'. For the load condition, torque of 10,000 lb in was applied to the output shaft of the second stage planetary gearbox. Four drive motor speeds were used: 300, 600, 900, and 1200 revolutions per minute (r/min). For each combination of load and speed, vibration data were collected over a 5 min span from each of the four accelerometers. The sampling frequency used was 10,000 Hz. The data were further split into 10 time records of equal length, so that there are 80 (2 days $\times$ 4 damage degrees $\times$ 10 time records) time

Table 4 The results for the Parkinson dataset

Methods	Classification accuracy (%)			Number of features selected		CPU time of feature selection (s)
	Mean	Standard	PBP	Mean	Standard	Mean
SVM using all features	94.43	2.05	0	23	0	—
Gualdrón's method	84.33	7.42	13.33	3	1.66	3.25
The proposed method						
PFR + GFS	84.47	7.44	13.33	3	1.92	3.27
PFR + PFS	96.15	1.65	76.67	11	4.65	7.12

records for each combination of load and speed. The features were extracted from each time record.

5.2 Feature extraction

The features to be extracted for the given damage degree classification problems are introduced in this section. They include the features that are reported for damage detection and damage mode classification of fixed shaft gearboxes [**15–18**]. Due to the unique behaviour of planetary gearboxes, their sidebands are different from those of fixed shaft gearboxes [**19**]. Hence, the features which require information on the sidebands of the planetary gearboxes were modified.

Figure 11 shows the preprocessing of vibration signals for feature extraction. The regular mesh components (RMCs) were defined as the fundamental shaft frequency, its second harmonic, gear meshing frequencies (GMF), its harmonics, and its first-order sidebands. It is reported in Inalpolat and Kahraman [**19**] that unlike the fixed shaft gearbox, the sidebands of planetary gearbox appear at integer multiples of planet passing frequency (the number of planets multiplied by carrier frequency) and the largest sideband is found at the frequency closest to the GMF.

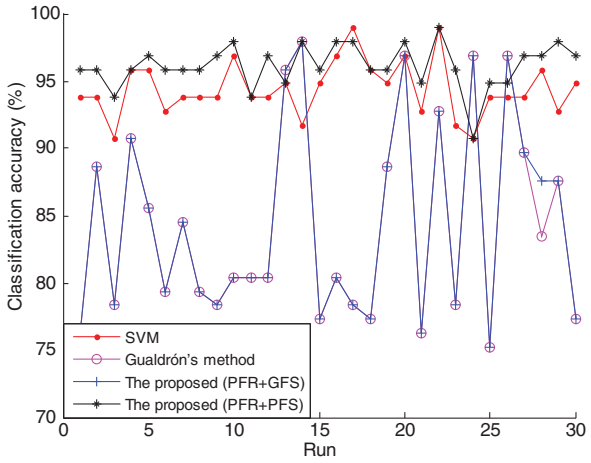

Fig. 6 The results of classification accuracy for the Parkinson dataset

In this article, the first-order sidebands for planetary gearboxes are defined as the lower and the upper sidebands closest to the GMF. Four types of signals were used including raw signals (RAW), residual signals (RES), difference signals (DIFF), and band-pass mesh signals (BPM). RAW denotes the vibration signal subtracted by its mean, DIFF denotes the RAW excluding the RMCs, RES is similar to DIFF but has the first-order sidebands included, and BPM denotes the band-pass mesh signal which is obtained using a band-pass filter filtering around the first-order sidebands. The RAW, DIFF, RES, and BPM are represented, respectively, by $x(t)$, $d(t)$, $r(t)$, and $b(t)$, $t = 1, 2, \ldots, T$ where T is the number of data points in the data series.

Thirty-four features were extracted including 26 time domain features and 8 frequency domain features as listed in Table 6. The time domain features include 16 features (F1–F16) which are commonly used for fault diagnosis of generic systems and 10 advanced features which are exclusively proposed for gear fault detection (F17–F26). Frequency domain features calculated based on sideband values include four (F27–F30) proposed for gear fault detection and four (F31–F34) exclusively developed for planetary gearboxes. More details on these features can be found in references [**15–20**].

5.3 Dataset establishment and results analysis

In the following, a data point is referred as having multiple input dimensions (features) and one output class label. Based on the descriptions in Subsections 5.2 and 5.3, for a particular combination of load and speed, we can establish a dataset with 80 data points, each of which has 136 (34 features $\times$ 4 accelerometers) input features and an output damage degree label. The features are numbered in the following way: features 1–34 are from LS1, features 35–68 from LS2, features 69–102 from HS1, and features 103–136 from HS2. The features for each accelerometer are in the same order as given in Table 6, e.g. features 1, 35, 69, and 103 all correspond to maximal value (F1) in Table 6. There are

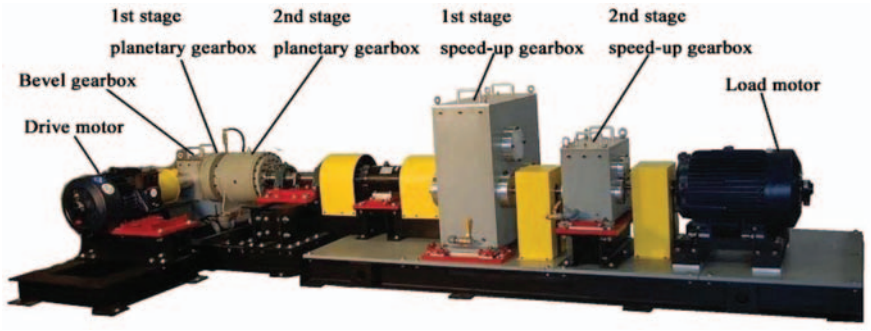

Fig. 7 The view of the test rig

Table 5 The number of teeth for two planetary gearboxes

	Bevel		First planetary			Second planetary			First speed-up				Second speed-up			
Gears	IB	OB	S	P	R	S	P	R	GI	SM	LM	GO	GI	SM	LM	GO
No.	18	72	28	62(3)	152	19	31(4)	81	72	32	80	24	48	18	64	24
Ratio	4↓		6.429↓			5.263↓			3.75↑				7.111↑			

Notes: No., number of gear teeth; IB, input bevel gear; OB, output bevel gear; S, sun gear; P, planet gear; R, ring gear; GI, gear on input shaft; SM, small gear on middle shaft; LM, large gear on middle shaft; GO, gear on output shaft; ↓, speed reduction ratio; and ↑, speed-up ratio. The number of planet gears is given within parentheses.

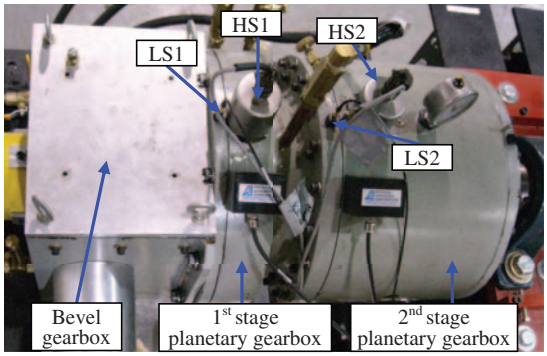

Fig. 8 The actual view of accelerometer locations

Fig. 10 Planet gears with artificially created pitting damage (from left to right corresponding to the damage degrees of severe, moderate, slight, and baseline)

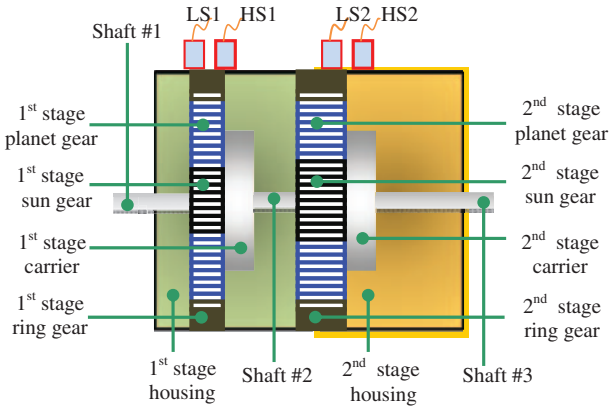

Fig. 9 The diagram of two-stage planetary gearboxes

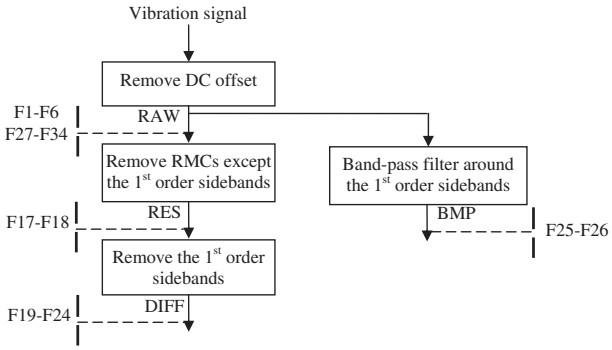

Fig. 11 Processing flow for feature extraction

Table 6 The list of features extracted

	Features	Definition		Features	Definition
F1	Maximal value	$\max(x(t))$	F2	Minimal value	$\min(x(t))$
F3	Average absolute value	$\frac{1}{T}\sum_{t=1}^{T}\lvert x(t)\rvert$	F4	Peak to peak	F1 − F2
F5	Variance	$\frac{1}{T}\sum_{t=1}^{T}(x(t)-\overline{x})^2$	F6	Standard deviation	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x(t)-\overline{x})^2}$
F7	Skewness	$\frac{\frac{1}{T}\sum_{t=1}^{T}(x(t)-\overline{x})^3}{\mathrm{F6}^3}$	F8	Kurtosis	$\frac{\frac{1}{T}\sum_{t=1}^{T}(x(t)-\overline{x})^4}{\mathrm{F5}^2}$
F9	Root mean square (RMS)	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}x(t)^2}$	F10	Crest factor	$\frac{\mathrm{F1}}{\mathrm{F9}}$
F11	Clearance factor	$\frac{\mathrm{F1}}{\frac{1}{T}\sum_{t=1}^{T}x(t)^2}$	F12	Impulse factor	$\frac{\mathrm{F1}}{\mathrm{F3}}$
F13	Shape factor	$\frac{\mathrm{F9}}{\mathrm{F3}}$	F14	Delta RMS	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}x_m(t)^2}-\sqrt{\frac{1}{T}\sum_{t=1}^{T}x_{m-1}(t)^2}$
F15	Energy ratio	$\frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x(t)-\overline{x})^2}}$	F16	Energy operator	$\frac{\frac{1}{T}\sum_{t=1}^{T}(\Delta x(t)-\Delta\overline{x})^4}{\left(\frac{1}{T}\sum_{t=1}^{T}(\Delta x(t)-\Delta\overline{x})^2\right)^2}$
F17	NA4	$\frac{\frac{1}{T}\sum_{t=1}^{T}(r(t)-\overline{r})^4}{\left(\frac{1}{M_n}\sum_{m=1}^{M_n}\left(\frac{1}{T}\sum_{t=1}^{T}(r_m(t)-\overline{r}_m)^2\right)\right)^2}$	F18	NA4*	$\frac{\frac{1}{T}\sum_{t=1}^{T}(r(t)-\overline{r})^4}{\left(\frac{1}{M_h}\sum_{m=1}^{M_h}\frac{1}{T}\sum_{t=1}^{T}(r_m(t)-\overline{r}_m)^2\right)^2}$
F19	FM4	$\frac{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^4}{\left(\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^2\right)^2}$	F20	FM4*	$\frac{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^4}{\left(\frac{1}{M_h}\sum_{m=1}^{M_h}\left(\frac{1}{T}\sum_{t=1}^{T}\left(d_m(t)-\overline{d}_m\right)^2\right)\right)^2}$
F21	M6A	$\frac{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^6}{\left(\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^2\right)^3}$	F22	M6A*	$\frac{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^6}{\left(\frac{1}{M_h}\sum_{m=1}^{M_h}\left(\frac{1}{T}\sum_{t=1}^{T}\left(d_m(t)-\overline{d}_m\right)^2\right)\right)^3}$
F23	M8A	$\frac{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^8}{\left(\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^2\right)^4}$	F24	M8A*	$\frac{\frac{1}{T}\sum_{t=1}^{T}\left(d(t)-\overline{d}\right)^8}{\left(\frac{1}{M_h}\sum_{m=1}^{M_h}\left(\frac{1}{T}\sum_{t=1}^{T}\left(d_m(t)-\overline{d}_m\right)^2\right)\right)^4}$
F25	NB4	$\frac{\frac{1}{T}\sum_{t=1}^{T}(e(t)-\overline{e})^4}{\left(\frac{1}{M_n}\sum_{m=1}^{M_n}\left(\frac{1}{T}\sum_{t=1}^{T}(e_m(t)-\overline{e}_m)^2\right)\right)^2}$	F26	NB4*	$\frac{\frac{1}{T}\sum_{t=1}^{T}(e(t)-\overline{e})^4}{\left(\frac{1}{M_h}\sum_{m=1}^{M_h}\left(\frac{1}{T}\sum_{t=1}^{T}(e_m(t)-\overline{e}_m)^2\right)\right)^2}$
F27	Mean frequency	$\frac{1}{K}\sum_{k=1}^{K}X(k)$	F28	Frequency centre	$\frac{\sum_{k=1}^{K}(f(k)\cdot X(k))}{\sum_{k=1}^{K}X(k)}$

(continued)

Table 6 Continued

	Features	Definition		Features	Definition
F29	RMS frequency	$\sqrt{\frac{\sum_{k=1}^{K}(f(k)^2 \cdot X(k))}{\sum_{k=1}^{K} X(k)}}$	F30	Standard deviation Frequency	$\sqrt{\frac{\sum_{k=1}^{K}\left((f(k)-\mathrm{F28})^2 \cdot X(k)\right)}{\sum_{k=1}^{K} X(k)}}$
F31	Largest sideband amplitude	$\max(X(k^*))$	F32	FM0	$\frac{\mathrm{F4}}{\sum X(k^*)}$
F33	Sideband index	$\frac{\sum X(k^*)}{2}$	F34	Sideband level factor	$\frac{\sum X(k^*)}{\mathrm{F6}}$

Notes: (1) $\Delta x(t)$ is obtained piecewise. For the non-endpoints, it is obtained by the squared $x(t)$ subtracted by the product of the data points of $x(t-1)$ and $x(t+1)$. For the endpoints, the data point of $x(t)$ is looped around. (2) $X(k)$, $k=1,2,\ldots,K$, represents the kth measurement of the frequency spectrum of $x(t)$ and $f(k)$ the frequency value of the kth spectrum line. (3) $x_m(t)$, $r_m(t)$, and $d_m(t)$ represent the RAW, RES, and DIFF of the mth time record, respectively. The bar notation represents the mean, e.g. $\bar{x}$ represents the mean of $x(t)$. M_n represents the total number of the time records up to the present. M_h represents the total number of time records corresponding to the 'healthy' conditions of gearbox. See Decker [**20**] for details of estimating the variance for a gearbox in good condition. (4) $e(t)$ represents the envelope of the current time record expressed as $e(t)=|b(t)+j\cdot H(b(t))|$. $H(b(t))$ represents the Hilbert transform of $b(t)$; $e_m(t)$ represents the envelope of the mth time record signal. (5) k^* represents the index of the first-order sidebands.

Table 7 The results of various combinations of speed and load

	SVM including all features		SVM with the proposed method (feature ranking and feature selection)				
	Classification accuracy (%)		Classification accuracy (%)			number of features selected	
Conditions	Mean	Standard	Mean	Standard	PBP	mean	Standard
300 r/min and no load	63.50	5.82	99.75	1.01	100	2	0.60
300 r/min and load	50.67	9.91	96.58	5.34	100	4	3.93
600 r/min and no load	72.42	8.39	99.58	1.33	100	2	3.38
600 r/min and load	54.67	8.11	97.67	2.86	100	10	8.87
900 r/min and no load	65.83	8.16	100	0	100	1	0.18
900 r/min and load	68.67	9.28	96.25	4.34	100	14	6.78
1200 r/min and no load	59.00	11.83	99.25	1.99	100	2	1.87
1200 r/min and load	64.42	7.90	100	0	100	7	4.89

eight combinations of load and speed; hence, there are eight datasets to be classified.

The given classification problem has multiples classes. The multi-class model of the proposed method is used for feature selection. Since the multi-class model adopts a similar feature ranking scheme and the same feature selection scheme as does the binary model and Section 4 verifies their effectiveness, the application results are directly shown rather than conducting more comparisons to validate it. Nevertheless, the classification accuracy of without using feature selection is still provided as a baseline, which is obtained by an OAA approach-based SVM classification.

The parameters used for SVM classification are as follows. The kernel function uses a Gaussian kernel with a width parameter of one. The parameter C is set at 50. A three-fold cross-validation is used to calculate the classification accuracy. The results averaged over 30 trials are presented in Table 7. For the column of 'SVM using all features', it can be seen that the classification accuracies are quite low for all the conditions. Some are even around 50%, which is unacceptable. Relatively large standard deviations are also observed; these suggest that the data points may not be distributed evenly in the original feature space. Classification accuracy is much improved for all conditions by the use of the proposed feature selection method. The classification accuracy values are all greater than 95% and their variations are reduced. Based on these results, it can be concluded that the proposed feature selection method effectively improves classification performance for the given damage degree classification problem. In the next section, the composition of the selected feature subset is further discussed.

6 DISCUSSION

Studying the resultant feature subsets may mitigate concerns such as which features are most useful for a given gear damage degree classification and which accelerometers are able to provide more useful information. These concerns are addressed in this section. The resultant feature subsets for no load and load conditions are examined in Subsection 6.1.

The usefulness of the accelerometers is discussed in Subsection 6.2.

6.1 Examination of the resultant feature subsets

From Table 7, it can be seen that, given the same speed, the degree of improvement in classification accuracy is greater for no load conditions than for load conditions. In addition, the dimensions of resultant feature subsets are larger for load conditions than for no load conditions; the same is true of their standard deviations. These observations indicate that the classification problem may become more difficult when a load is applied. For this reason, the no load and the load conditions are examined separately in this subsection.

6.1.1 Analysis of the resultant feature subsets for no load conditions

The feature subsets of 900 r/min are analysed first. It is found that about one feature on average is selected in this condition. The composition of these subsets over the 30 trials were examined and the results found were quite consistent. Feature 98 was selected for 23 trials, feature 96 for 3 trials, feature 28 for 2 trials, and feature 132 and a combination of features 86 and 90 both for 1 trial. Based on Table 6, it can be seen that apart from features 86 and 90, these are either standard deviation frequency (F30) or frequency centre (F28). Because feature 98 was selected for most of the trials, out of interest, the classification results are plotted using this feature in Fig. 12. It exhibits a good separation of data points for different damage degrees.

For 300 r/min, the resultant feature subsets show two compositions. One contains only feature 96 for 11 trials and the other both features 92 and 126 for 17 trials. For 600 r/min, the resultant feature subsets are a bit various. We find that features 28 and 96 are selected for the most trials, 14 and 10, respectively. These two are both frequency centre (F28) but from different accelerometers. For 1200 r/min, the most frequently selected features are 96 and 98 which are selected for 11 and 10 trials, respectively. Based on the above observations, it can be concluded that the frequency centre (F28) and standard deviation frequency (F30) are the most useful features for no load conditions. To test this conclusion, classification using features 28 (F28 from LS1), 96 (F28 from HS1), and 98 (F30 from HS1) is conducted for the same 30 sets of training and testing data. The resulting classification accuracies are comparable to those obtained using the proposed method.

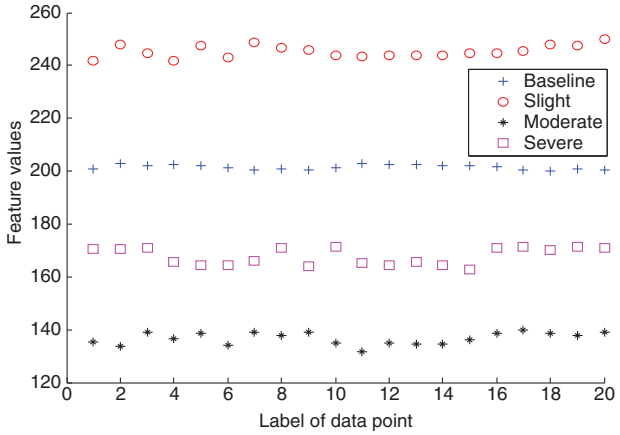

Fig. 12 The classification results using standard deviation frequency of HS1 under 900 r/min and no load condition

6.1.2 Analysis of the resultant feature subsets for load conditions

For each speed, the resultant feature subsets over 30 trials are not as consistent as those from no load conditions. The resultant subsets usually contain more than one feature and display a relatively high variety. However, some features are still found to be appearing in the subsets with high frequencies for all speeds; these include features 29 (F29 from LS1), 30 (F30 from LS1), 96 (F28 from HS1), 129 (F27 from HS2), and 133 (F31 from HS2). It is apparent that these features are all frequency domain features. Similarly, studying the classification results from using these features under load conditions, classification accuracies are obtained, which are not as good as those obtained using the proposed method; they are, however, all over 90%, which is acceptable.

6.2 Analysis of the usefulness of accelerometers

In our test rig, because the four accelerometers are distinct in terms of their sensitivity and location on the gearbox housing, they may provide unequally useful information. Based on the observations of the previous subsection, we find that for no load conditions the useful features are 28, 96, and 98, the first from LS1 and the rest from HS1. For load conditions, features 29, 30, 96, 129, and 133 appear to be most useful. The first two are from LS1, the third is from HS1, and the last two are from HS2. It can be seen that no features are from LS2 under both no load and load conditions. It suggests that LS2 may need to be relocated somewhere more appropriate. In addition, the information from HS1 may be more effective than that from others for no load conditions.

7 CONCLUSIONS

In this study, we propose an SVM-based feature selection method to address both binary-class and multi-class classification problems. The proposed method uses the norm of the weight vector of SVM as a measure for evaluating the importance to classification of a particular feature; as well, it uses a RBS scheme to eliminate useless features through updated feature ranks. The results of three benchmark datasets show this method consistently outperforms its counterparts. This demonstrates that the proposed measure for feature ranking is able to effectively assess the impact of removing features, and the proposed RBS enables the useless features to be removed maximally.

The multi-class model of the proposed method is used for feature selection in damage degree classification of planet gear. The results show the significant effectiveness of the proposed method. Furthermore, our studies on the resultant feature subsets exhibit that the frequency domain features dominate the time domain features, and that two features, frequency centre and standard deviation frequency, are the most useful for the given classification problem under no load conditions.

ACKNOWLEDGEMENT

The authors thank the reviewers for their valuable comments and suggestions.

FUNDING

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

1 **Guyon, I.** and **Elisseeff, A.** An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
2 **Malhi, A.** and **Gao, R. X.** PCA-based feature selection scheme for machine defect classification. *IEEE Trans. Instrum. Meas.*, 2004, **53**(6), 1517–1525.
3 **Fei, C. G., Han, Z. Z.,** and **Liu, Q. K.** Ultrasonic flaw classification of seafloor petroleum transporting pipeline based on chaotic genetic algorithm and SVM. *J. X-ray Sci. Technol.*, 2006, **14**, 1–9.
4 **Qu, J.** and **Zuo, M. J.** Support vector machine based data processing algorithm for wear degree classification of slurry pump systems. *Measurement*, 2010, **43**, 781–791.
5 **Vapnik, V. N.** *Statistical learning theory*, 1998 (John Wiley & Sons, New York).
6 **Widodo, A.** and **Yang, B. S.** Support vector machine in machine condition monitoring and fault diagnosis. *Mech. Syst. Signal Process.*, 2007, **21**, 2560–2574.
7 **Samanta, B.** Gear fault detection using artificial neural networks and support vector machines with genetic algorithms. *Mech. Syst. Signal Process.*, 2004, **18**(3), 625–644.
8 **Khawaja, T. S., Georgoulas, G.,** and **Vachtsevanos, G.** An efficient novelty detector for online fault diagnosis based on least squares support vector machines. In IEEE AUTOTESTCON, Salt Lake City, Utah, 8–11 September 2008.
9 **Gualdrón, O., Brezmes, J., Llobet, E., Amari, A., Vilanova, X., Bouchikhi, B.,** and **Correig, X.** Variable selection for support vector machine based multisensor systems. *Sens. Actuators B*, 2007, **122**, 259–268.
10 **Cristianini, N.** and **Taylor, J. S.** *An introduction to support vector machines and other kernel-based learning methods*, 2000 (Cambridge University Press, Cambridge).
11 **Schölkopf, B.** and **Smola, A.** *Learning with kernels: support vector machines, regularization, and beyond*, 2002 (MIT Press, Cambridge, Massachusetts).
12 **Vachtsevanos, G., Lewis, F. L., Hess, M. R. A.,** and **Wu, B. Q.** *Intelligent fault diagnosis and prognosis for engineering systems*, 2006 (John Wiley & Sons, Inc, New Jersey).
13 **Asuncion, A.** and **Newman, D. J.** UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, California, available from http://www.ics.uci.edu/∼mlearn/MLRepository.html.
14 **Hoseini, M.** and **Zuo, M. J.** A literature survey on the creating and quantifying faults on planetary gearbox, Technical Report, Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta, 2009.
15 **Lei, Y. G.** and **Zuo, M. J.** Gear crack level identification based on weighted K nearest neighbor classification algorithm. *Mech. Syst. Signal Process.*, 2009, **23**(5), 1535–1547.
16 **Lebold, M., McClintic, K., Campbell, R., Byington, C.,** and **Maynard, K.** 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, Virginia, 1–4 May 2000, pp. 623–634.
17 **Večeř, P., Kreidl, M.,** and **Šmíd, R.** Condition indicators for gearbox condition monitoring systems. *Acta Polytech.*, 2005, **45**(6), 35–43.
18 **Zuo, M. J., Li, W.,** and **Fan, X. F.** Statistical methods for low speed planetary gearbox monitoring, Technical Report, Department of Mechanical Engineering, University of Alberta, Edmonton, 29 November 2005.
19 **Inalpolat, M.** and **Kahraman, A.** A theoretical and experimental investigation of modulation sidebands of planetary gear sets. *J. Sound Vib.*, 2009, **323**, 677–696.
20 **Decker, H. J.** Crack detection for aerospace quality spur gears, NASA/TM-2002-211492, ARL-TR-2682, April 2002.