electronics

Article

A New Fault Diagnosis Method of Bearings Based on Structural Feature Selection

Wentao Mao [1,2,*,†], Liyun Wang [1,†] and Naiqin Feng [1]

[1] School of Information Engineering, Zhengzhou University of Industrial Technology, Zhengzhou 451100, China; zzwly0428@163.com (L.W.); fengnaiqin@163.com (N.F.)

[2] School of Mechanics and Civil & Architecture, Northwestern Polytechnical University, Xi'an 710129, China

* Correspondence: maowt.mail@gmail.com; Tel.: +86-177-3735-3087

† These authors contributed equally to this work.

Received: 15 October 2019; Accepted: 22 November 2019; Published: 25 November 2019

Abstract: By using signal processing and statistical analysis methods simultaneously, many heterogeneous features can be produced to describe the bearings fault with more comprehensive and discriminant information. At same time, there may exist redundant or irrelevant information which will instead reduce the diagnosis performance. To solve this problem, it is necessary to conduct feature selection which tries to choose the most typical and discriminant features by evaluating their effect on fault status. However, if the structural relationship between features has not been considered well, some similar or redundant features are still probably chosen, which would introduce bias into the final diagnosis model. In this paper, a new fault diagnosis method of bearings based on structural feature selection is proposed to solve the aforementioned problem. Obeying the hypothesis that the features with strong relatedness have close coefficient distance, the proposed method aims to improve diagnosis performance via determining group structure in fault features. First, a new feature selection strategy is proposed by introducing a group identification matrix. Using this matrix, two evaluation criteria about intra-group feature correlation and inter-group feature difference are constructed by means of coefficient's distance. Consequently, we get a multi-objective 0–1 integer programming problem by minimizing intra-group distance and maximizing inter-group distance simultaneously. Second, we use the multi-objective particle swarm optimization algorithm to solve this problem, and then determine the optimal group structure of features adaptively. Finally, a diagnosis model can be trained by support vector machine on the typical features extracted from these groups. Experimental results on four UCI datasets show the effectiveness of the proposed group feature selection strategy. Moreover, the experimental results on two bearing datasets (i.e., CWRU and IMS datasets) demonstrate that the proposed method can identify the inherent group structure in fault features, and then has better diagnosis performance compared with several state-of-the-art methods.

Keywords: fault diagnosis; bearing; structural information; support vector machine; heterogeneous feature; feature selection

1. Introduction

As an important unit of common rotating machinery, rolling bearings easily fall into different kinds of faults under complex working condition such as long-term heavy load and strong impact, etc. Faulty bearings will lead to the performance deterioration of whole machinery, so fault diagnosis for bearings always plays a vital role in health management of machinery. In recent decades, fault diagnosis has received much attention from academic researchers and engineers [1,2]. With quick development of artificial intelligence in the past decade, machine learning-based fault diagnosis methods have shown their comparative performance in health monitoring for rotating machinery. In these methods, two key issues are generally included: feature extraction and model construction [3,4]. For feature extraction, representative features of vibration signals need to be extracted by using statistical analysis or following failure mechanism. Heterogeneous features can be calculated in time/frequency/time-frequency domain [5]. There are a large number of features such as Kurtosis value [6], wavelet packet transform (WPT) [7], empirical mode decomposition (EMD) [8], Garch model [9] and so on which are applied to bearing fault diagnosis. These features have different characteristics. For instance, Kurtosis value is sensitive to early fault, while WPT and EMD can effectively decompose non-stationary signal. Moreover, entropy theory has also been introduced to extract discriminant features [10] Based on these features, some machine learning algorithms, e.g., support vector machines (SVMs) [11,12], decision tree [13] and artificial neural network [14], have been introduced to establish diagnosis model. Despite different operational principles, these algorithms all build classification model on the extracted features of different health conditions. In very recent years, we also observe deep learning techniques [15–17] have been successfully introduced to solve the fault diagnosis problem. Different kinds of deep neural networks such as stacked auto-encoder [18] and convolutional neural network [19] have been proved promising in solving the problem of fault diagnosis. In contrast to the hand-crafted features listed above, deep learning techniques are capable of extracting features adaptively from raw signals and directly output the diagnosis result. However, as running on deep neural networks, deep learning techniques tend to perform well on sufficient training data, which is too restrict for many real-world applications. Even these techniques would extract features from small-scale samples (by avoiding over-fitting), these features still need to be analyzed and refined again for achieving better discriminant ability.

In this paper, we turn around our focus from the deep learning-based fault diagnosis methods back to the diagnosis method on small-scale samples, as mass of bearings fault data are not easy to collect in many real-world applications. From the discussion above, features of fault statuses play a vital role to establish the decision model of fault diagnosis. Then we hope to find an effective way on feature level to improve the generalization performance of diagnosis model on unseen data. We observe that different signal processing techniques, even including some deep learning techniques, can only provide limited information to describe bearings fault. Different features have the different representative ability. For example, as a fourth-order index to measure stochastic signal, Kurtosis is rather sensitive to the incipient fault [8]. Moreover, EMD contributes to decomposing non-stationary signal into a collection of intrinsic mode functions with a trend [4]. Then EMD can choose proper components of intrinsic mode function for statistical analysis. As an example, we draw four statistical features (Kurtosis, RMS, frequency spectrum, wave factor) of total degradation process of one same bearing from CWRU dataset (please refer to the section of experiment), as shown in Figure 1. It is visually obvious in Figure 1 that Kurtosis goes up in the early fault stage, while frequency spectrum shows better tendency in the fast degradation stage). Figure 1 indicates that different features have different working areas on this signal, which just shows the necessity of heterogeneous feature selection.

An intuitive idea to improve the diagnosis performance is merging as many heterogeneous features as possible to describe the bearings fault with more comprehensive information, named *feature pooling* [7]. To select typical and discriminative features, it is recommended to conduct feature selection which needs to evaluate the effect of features on fault status. Currently, two strategies, i.e., Filter and Wrapper [20], have been widely used in feature selection. The filter methods pick up the intrinsic relevance of the features measured via univariate statistics such as information entropy and rough set, while the Wrapper methods measure the “usefulness” of features based on the classifier performance. Despite impressive performance of these two strategies in many fields, there are still two challenges arising from the field of bearing fault diagnosis:

(1) First, it probably exists redundant information after common feature selection for heterogeneous features. For example, to improve the measurement quality, many engineers used to set up vibration sensors on adjacent places, and this would cause the phenomenon of *cross detection* [7]. The number of features duplicates, but the features are rather redundant. Traditional feature selection

methods may eliminate such duplicated or irrelevant features to some extent. However, it is still necessary to further exploit the relevance of the rest of features to improve the feature representation as well as reduce the bias of diagnosis model.

(2) Second, most current Wrapper methods ignore the inner structure among heterogeneous features. If we can explore the features' inner structure, the most typical features can be determined more easily in an effective subset.

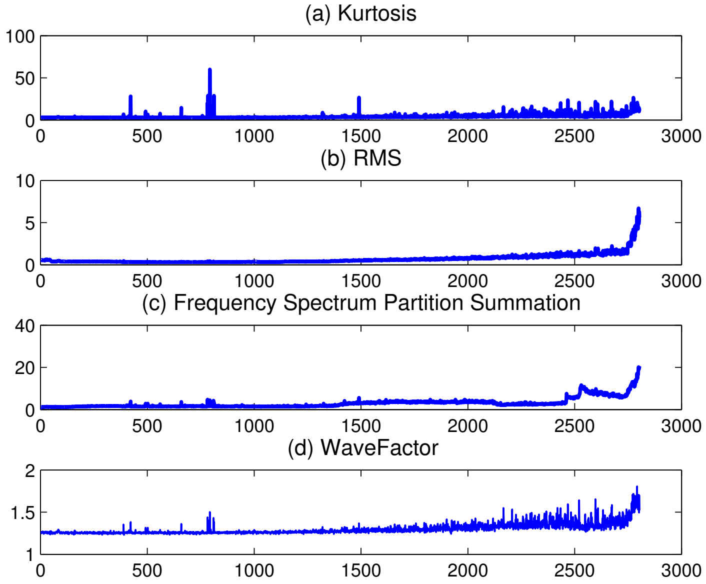


Figure 1. Four statistical features in time domain of total fatigue life of one same bearing.

These two challenges can be put into the problem of *structural feature selection*. Although some related topics like sparse feature selection [21,22] have been studied, the structural selection methods for heterogeneous features are seldom found in the field of bearing fault diagnosis. In this paper, we try to solve this problem. We start with a simple but typical structure: group structure. Obeying the hypothesis that the features with strong relatedness have close coefficient distance, this paper aims to determine the group structure of features while doing feature selection. The basic idea is introducing an identification matrix to indicate the distance between features and then optimizing such matrix to obtain an optimal inner feature structure. From the theoretical aspect, the main contribution of this paper is proposing a new structural feature selection method on heterogeneous features. This method transforms the structural feature selection problem into a multi-objective 0–1 integer programming problem by minimizing intra-group feature correlation and maximizing inter-group feature discrimination at the same time. After solving this optimization problem, the optimal structure can be adaptively determined. From the practical aspect, the main contribution of this paper is proposing a new fault diagnosis method for bearings. In contrast to most of current diagnosis methods, the proposed method can effectively obtain a set of typical fault features for building diagnosis model no matter from traditional statistical features or deep features. To our best knowledge, there are very few studies on structural heterogeneous feature selection for bearing fault diagnosis.

It is worth noting that this method uses SVM as baseline algorithm. SVMs are supervised learning models for classification with structural risk minimization. By using maximum-margin strategy and kernel trick, SVMs perform well on non-linear small-scale data. However, the basic idea listed above can also generalize to some discriminant classifiers such as perception network, linear discriminant analysis, and Logistic regression. Moreover, the proposed structural feature selection method can link to the last layer of a deep neural network. Therefore, the redundant features which are easily generated on small-scale data can be detected and removed. In the experiment section, the effect of structural feature selection is first evaluated on four UCI Machine Learning Repository datasets which are widely used and cited for machine learning research, and then tested on two bearing datasets. We also find that the least absolute shrinkage and selection operator (LASSO) is effective in variable selection by forcing certain coefficients of some variables to be set to zero. As an extension of LASSO, group LASSO [21,22] can identify predefined groups of covariates to be selected into or out of a model together. However, this method relies on L1-norm minimization which is the optimal convex approximate of L0-norm minimization. With special optimization methods (e.g., proximal methods) required to select sparse features, this method could not apply to the existing classification algorithm form.

The paper is organized as follows. In Section 2, we review the typical feature extraction, feature selection methods and support vector machine. In Section 3, we introduce an identification matrix to establish the multi-objective 0–1 programming problem, and provide a solution with multi-objective particle swarm optimization. Section 4 is devoted to computer experiments on UCI datasets and two bearing fault diagnosis data sets, followed by a conclusion of the paper in the last section.

2. Background

In this section, we provide the review for some commonly used techniques and algorithms which will be used in our proposed algorithm and experiments.

2.1. Feature Extraction Using EMD and WPT

In recent decades, EMD [4] and WPT [3] have become the promising tool of extracting features from vibration signals, especially from non-stationary signals. The basic idea of EMD is decomposing the given signal into a series of intrinsic mode function (IMF) [23] on the base of local characteristic time scale. Here IMF indicates a simple oscillatory mode as a counterpart to a simple harmonic function. Using the IMFs, the obtained sequence $c_1(t), c_2(t), \cdots, c_n(t)$ covers the bands' component from high frequency to low frequency. The first few IMF components are significant as they contain the most important information of raw signal. In Experiment section, we extract 8 IMF components for the fault signals. After decomposition, we calculate the energy spectrum of each IMF component to build EMD's feature vector, as the following formulation [24]:

$$\begin{aligned} &E_{EMD} = \{E_1, E_2, \cdots, E_n\} \\ &where \quad E_i = \int |c_i(t)|^2 \mathrm{d}t = \sum_{k=1}^{m} |y_{ik}|^2 \end{aligned}$$

Here $y_{ik}\,(i = 0, 1, \cdots, n; k = 0, 1, \cdots, m)$ denotes the amplitude of discrete points in $c_i(t)$.

Like EMD, WPT also can extract energy spectrum as a set of features of bearing vibration signals. WPT is a wavelet transform where the discrete-time signal is passed through more filters than the discrete wavelet transform. Then WPT can offer a richer signal analysis for non-stationary signal. Using WPT, we obtain the decomposition coefficients $\left\{X_d^0, X_d^1, \cdots, X_d^{J-1}\right\}$, where d is decomposition level, and J is the total number of sub-band. In this experiment, we set $d = 3, J = 8$. The energy of each sub-band signal can be calculated by using the following formulation [5]:

$$\begin{aligned} &E_{WPT} = \left\{E_d^0, E_d^1, \cdots, E_d^{J-1}\right\} \\ &where \quad E_d^J = \int \left|X_d^j(t)\right|^2 \mathrm{d}t = \sum_{k=1}^{n} \left|x_{jk}\right|^2 \end{aligned}$$

Here $x_{jk}\,(j = 0, 1, \cdots, J-1; k = 0, 1, \cdots, n)$ denotes the amplitude of discrete points in $X_d^j(t)$. Therefore, the WPT energy spectrum can be obtained after the normalization of E_{WPT} by calculating the proportion of each WPT component in total E_{WPT}.

2.2. Support Vector Classification

SVM is an effective binary classification algorithm on small-scale data. Based on the statistical learning theory, SVM tries to seek structural risk minimization, and then can get good generalization ability on limited observations while avoiding over-fitting [25]. By introducing the kernel trick, SVM effectively solves the problem of "dimension disaster" for the high-dimensional problem. SVM has been a promising tool in the fields of pattern recognition and fault diagnosis, etc. For the sake of better understanding, we provide a sketch map to show the principle of SVM classification in Figure 2. Obviously, SVM seeks the best classification hyper-plane (solid line) with maximum margin between negative and positive classes.

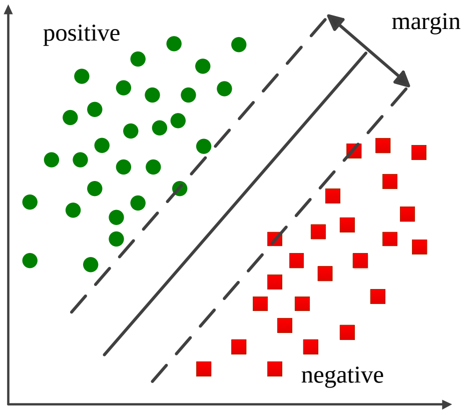


Figure 2. Sketch map of SVM classification.

Here we take the non-linear SVM as an example. Given a set of independently and identically distributed (*i.i.d*) training samples $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\} \subset \mathbb{R}^d \times \mathbb{R}$, SVM is formulated as minimizing the following functional [26]:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ &s.t.\ y_i\left(\mathbf{w}\cdot\varphi\left(\mathbf{x}_i\right)+b\right) \geq 1-\xi_i \\ &\quad \xi_i \geq 0 \quad i = 1,2,...,N \end{aligned} \tag{1}$$

where C is the regularization parameter, $\varphi(\cdot)$ is the non-linear mapping function. In Equation (1), the first term is the regularization term which is used to prevent over-fitting, and the second term means training error on the available training samples. Minimizing Equation (1) will reach structural risk minimization. By calculating the Karush–Kuhn–Tucker (KKT) condition which permits replacing the primal problem by a dual problem, Equation (1) can be transformed to the following dual form [26]:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i,\mathbf{x}_j\right) - \sum_{i=1}^{N}\alpha_i \\ &s.t. \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\quad C \geq \alpha_i \geq 0, i = 1,2,...,N \end{aligned} \tag{2}$$

where $K\left(\mathbf{x}_i,\mathbf{x}_j\right) = <\varphi\left(\mathbf{x}_i\right), \varphi\left(\mathbf{x}_j\right)>$ is kernel function, α_i means Lagrange multiplier. Please note that in Equation (2), the first term is convex and quadratic on α_i while the second term is convex.

Therefore, Equation (2) is a convex quadratic problem. After solving this problem, the optimal solution $\alpha^* = (\alpha_1{}^*, \alpha_2{}^*, ..., \alpha_N{}^*)^T$ can be obtained. Therefore, we have the following classification model [26]:

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right)$$

where the $sign(\cdot)$ function is:

$$sign(x) = \begin{cases} 1 & if x \geq 0 \\ -1 & if x < 0 \end{cases}$$

2.3. SVM-RFE

As a typical Wrapper feature selection method, SVM-Recursive Feature Elimination (SVM-RFE) [27] has been successfully used to find discriminative relationships and identify the inner patterns within bearing fault datasets. SVM-RFE is an iterative algorithm that works backward from an initial set of features obtained from SVM. At each round, SVM-RFE first fits a linear SVM, and then ranks the features in terms of their weights in the SVM solution, and finally eliminates the feature with the lowest weight. The ranking criterion R_c is the difference between the weight $\|\mathbf{W}\|^2$ of all features and the weight $\|\mathbf{W}^{-p}\|^2$ after eliminating pth feature, as follows [27]:

$$R_c = \left| \|\mathbf{W}\|^2 - \|\mathbf{W}^{-p}\|^2 \right|$$

Please note that $\mathbf{W} = \sum_{i=1}^{m} \alpha_i y_i \varphi(x_i)$ is the model weight of SVM and can be calculated by Equation (1). After substituting $\mathbf{W}$, we have [27]:

$$R_c = \frac{1}{2} \left| \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i,j=1}^{m} \alpha_i{}^{(-p)} \alpha_j^{(-p)} y_i y_j K\left(\mathbf{x}_i^{(-p)}, \mathbf{x}_j^{(-p)}\right) \right|$$

where m is the number of samples, $K(\mathbf{x}_i, \mathbf{x}_j)$ means the kernel function of sample $\mathbf{x}_i$ and $\mathbf{x}_j$, α_i is the Lagrange multiplier of SVM while the superscript $(-p)$ indicates the pth feature is removed.

3. The Proposed Fault Diagnosis Method

In this section, we present the proposed fault diagnosis method based on structural feature selection of heterogeneous fault features. Briefly speaking, we first exploit the inner structure of features and then choose the most representative fault features for building diagnosis model. We adopt the following assumption: the features with strong relatedness have short weights' distance, and then construct a multi-objective 0–1 programming problem by making the feature weights in the same group as close as possible while making the feature weights in the different group as unlike as possible. By solving this problem, the optimal inner structure of fault features can be determined adaptively, and a set of fault features with good discriminant ability can be chosen for model training.

3.1. Model Construction

Given a set of *i.i.d* training samples $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\} \subset \mathbb{R}^d \times \mathbb{R}$, we have the weight coefficient vector $\beta = [\beta_1, \beta_2, ..., \beta_d]^T$ of d-dimensional features by running linear SVM. Suppose that there exists some group structures among the features. Therefore, to conduct structural feature selection, we need to assign all d-dimensional features into G groups with making the weight distance of features in same group as small as possible and the distance in different group as big as possible. Here we first present the expression of intra-group weight distance:

$$IntraDist = \sum_{g=1}^{G} \sum_{i<j}^{m} |\beta_i - \beta_j| \tag{3}$$

where G is the total number of group, m is the number of features in group g. Equation (3) means the sum of the weight distance of each two features in all groups. Obviously, minimizing Equation (3) can prompt features with close weight into a same group.

Equation (3) cannot be minimized directly. To get a feasible minimization target, we introduce group assignment matrix Q whose elements $q_{gi} \in \{0,1\}$ indicate whether the t-th feature is assigned to the group g. Moreover, let $Q_g \in \mathbb{R}^{G \times G}$ be the diagonal matrix whose diagonal elements are q_{gi}. It is clear that $\sum_{g=1}^{G} Q_g = I$, where I stands for the identity matrix. We also set the weight distance matrix $\mathbf{A}$ as:

$$\mathbf{A} = \mathbf{B} - \mathbf{B}^T$$

where $\mathbf{B}$ is square matrix composed by feature's weight vector β, i.e., $\mathbf{B} = [\beta, \cdots, \beta]$. Using the element $a_{ij} \in \mathbf{A}$, we can re-write Equation (3) as:

$$\begin{aligned} IntraDist &= \sum_{g=1}^{G} \sum_{i<j}^{m} |a_{ij}| \\ s.t \quad & a_{ij} \in \mathbf{A} \end{aligned} \tag{4}$$

Because $\mathbf{A} = \sum_{g=1}^{G} Q_g \mathbf{A} Q_g$, we first give the following definition:

Definition 1 (Intra-group weight distance sum)**.**

$$\begin{aligned} IntraDist\left(Q_{\mathrm{g}}\right) &= \sum_{\mathrm{g}=1}^{G} \sum_{i<j}^{m} |a_{ij}| \\ s.t \quad & a_{ij} \in Q_{\mathrm{g}} \mathbf{A} Q_{\mathrm{g}} \\ & \sum_{\mathrm{g}=1}^{G} Q_{\mathrm{g}} = I \end{aligned} \tag{5}$$

The summation constraint guarantees that each feature can be assigned to one and only one group. Obviously, if $G = 1$*, Equation (5) equals a standard formulation of weight distance. It is clear that the smaller the value of f is, the better the grouping performance is.*

Similarly, to measure the inter-group distance, we have the following formulation:

$$InterDist = \sum_{\mathrm{g}=1}^{k} \sum_{i=1}^{m} \sum_{j=1}^{n} |\beta_i - \beta_j| \tag{6}$$

where $k = \left\lceil \frac{G}{2} \right\rceil$*, m is the number of features in group* g*, n is the number of features which are not in the group* g*. Equation (6) indicates the sum of inter-group feature distance. Obviously, maximizing Equation (6) can prompt features in different groups to be farther away from each other, or say, to be more divergent.*

Similarly, we use group assignment matrix Q to express the inter-group feature divergence. We can re-write Equation (6) as:

$$\begin{aligned} InterDist &= \sum_{\mathrm{g}=1}^{G} \sum_{i<j}^{m} \left| b_{ij} \right| \\ s.t\ \ b_{ij} &\in \mathbf{B_g} \\ \mathbf{B_g} &= \mathbf{A} - \mathbf{M_g} - \mathbf{N_g} \end{aligned} \tag{7}$$

where $\mathbf{M_g} = Q_{\mathrm{g}}\mathbf{A}Q_{\mathrm{g}}$ *is intra-group weight distance matrix for group* g, $\mathbf{N_g} = \left(\mathbf{E} - Q_{\mathrm{g}}\right)\mathbf{A}\left(\mathbf{E} - Q_{\mathrm{g}}\right)$ *is beyond-group weight distance matrix for the group* g*. Consequently,* $\mathbf{B_g}$ *means the inter-group divergence matrix for the group* g*. According to the matrix theory, we have:*

$$\begin{aligned} \mathbf{B_g} \;=\; & \mathbf{A} - \mathbf{M_g} - \mathbf{N_g} \\ & = \mathbf{A} - Q_{\mathrm{g}}\mathbf{A}Q_{\mathrm{g}} - \left(\mathbf{E} - Q_{\mathrm{g}}\right)\mathbf{A}\left(\mathbf{E} - Q_{\mathrm{g}}\right) \\ & = \mathbf{A} - Q_{\mathrm{g}}\mathbf{A}Q_{\mathrm{g}} - \mathbf{A} + \mathbf{A}Q_{\mathrm{g}} + Q_{\mathrm{g}}\mathbf{A} - Q_{\mathrm{g}}\mathbf{A}Q_{\mathrm{g}} \\ & = \mathbf{A}Q_{\mathrm{g}} + Q_{\mathrm{g}}\mathbf{A} - \mathbf{2}Q_{\mathrm{g}}\mathbf{A}Q_{\mathrm{g}} \end{aligned} \tag{8}$$

Based on the derivation above, we give the following definition:

Definition 2 (Inter-group weight distance sum)**.**

$$\begin{aligned} InterDist\left(Q_{\mathrm{g}}\right) &= \sum_{\mathrm{g}=1}^{G} \sum_{i<j}^{m} \left| b_{ij} \right| \\ s.t\ \ b_{ij} &\in \left(\mathbf{A}Q_{\mathrm{g}} + Q_{\mathrm{g}}\mathbf{A} - \mathbf{2}Q_{\mathrm{g}}\mathbf{A}Q_{\mathrm{g}}\right) \\ \sum_{\mathrm{g}=1}^{G} Q_{\mathrm{g}} &= I \end{aligned} \tag{9}$$

Equation (9) indicates the distance sum of each feature to any other features in different groups. Obviously, maximizing Equation (9) will assign the features with divergent weights into different groups.

Please note that Equation (5) indicates the intra-group feature relatedness while Equation (9) indicates the inter-group feature divergence. To exploit the inner structure based on feature distance, we need to minimize Equation (5) and maximize Equation (9) simultaneously. The final target is seeking a series of Q_{g}*, named* Q^**, to make the intra-group weight distance sum as small as possible while inter-group weight distance sum as large as possible, i.e.,*

$$Q^* = \mathrm{argmin}\left(IntraDist\left(Q_{\mathrm{g}}\right), -InterDist\left(Q_{\mathrm{g}}\right)\right) \tag{10}$$

3.2. Solving Method

Considering the group assignment matrix Q, Equation (10) is a multi-objective 0–1 programming problem. As seeking the optimal solution for this kind of problem is NP-hard (i.e., a deterministic algorithm to solve it cannot be found in polynomial time), we choose an evolutionary algorithm to seek a numerically approximate solution. Taking the 0–1 programming into account, we can improve the traditional multi-objective optimization algorithm by adjusting its search space. Compared to single-objective optimization, the multi-objective optimization problem involves more than one objective function to be optimized simultaneously, and seeks one or more solutions (generally denoted as Pareto-optimal) to make each objective reach optimum [28]. In this case, the objective functions are generally conflicting, which results in no single solution that simultaneously optimizes each objective.

For the sake of better understanding, we first provide some general concepts. The multi-objective optimization problem can be expressed as follows [28]:

$$\begin{aligned} \min \ & z = f(x) = (f_1(x), f_2(x), \cdots, f_q(x)) \\ s.t. \ & g_i(x) \leq 0, i = 1, 2, \cdots, m \end{aligned}$$

where $x \in \mathbb{R}^n$ is the solution belonging to the feasible region $S = \{x \in \mathbb{R}^n | g_i(x) \leq 0, i = 1, 2, \cdots, m\}$.

In mathematical terms, a feasible solution x_1 is regarded to dominate another solution x_2, written as $x_1 \prec x_2$, if:

(1) $f_i(x_1) \leq f_i(x_2)$ for all indices $i \in \{1, 2, \cdots, q\}$ and

(2) $f_j(x_1) < f_j(x_2)$ for at least one index $i \in \{1, 2, \cdots, q\}$

Due to conflicting objectives, there exist several Pareto-optimal solutions which construct the Pareto front. A solution is called nondominated, Pareto-optimal or noninferior, if it cannot be dominated by any other solutions. In another word, none of the objective functions can be improved in value without degrading some of the other objective values.

Because of the insensitivity to the shape and continuity of Pareto front, the evolutionary algorithm can approximate the non-convex or non-continuous optimal front well, so it is suitable to solve the multi-objective optimization problem. In recent decades, several nature-inspired evolutionary algorithms have been proved very efficient in solving multi-objective problems. Within these algorithms, particle swarm optimization(PSO) [29], proposed by J. Kennedy and R. Eberhart, tries to find an optimal solution by mimicking the social behavior of birds flock. Compared with other evolutionary algorithms, PSO has its advantages such as few parameters, faster convergence rates and so on. Moreover, due to its simple structure, PSO has been successfully developed for solving the multi-objective optimization problem. Therefore, in this work we apply the multi-objective PSO, named MOPSO [30], to solve Equation (10). Please note that this work just pays the emphasis on the application of MOPSO rather than the development of MOPSO algorithm.

The basic idea of classical PSO and MOPSO refers to the literature [29,30]. Here we provide the main step of application of MOPSO in this work, as follows. The brief flowchart of MOPSO can also be found in Figure 3.

Step 1. Initialize the swarm size and number of generations. Initialize the present location and fitness value of each particle by generating randomly the group assignment matrix Q for each group. The intra-group weight distance sum shown in Equation (5) and the minus of inter-group weight distance sum shown in Equation (9) are taken as the fitness function to evaluate the goodness of the solution. Please note that the fitness value from Equation (9) must be added minus, as shown in Equation (10).

Step 2. Update the noninferior archive by adding the particles which are nondominated by others.

Step 3. Update the particles' velocity and position by following the best particle in the swarm (*gBest*) and best individual particle (*xBest*). Here *gBest* is randomly selected from the noninferior archive.

Step 4. Calculate the fitness value of new particles. Update *xBest* by checking the domination relationship between the current *xBest* and new particles.

Step 5. Update the noninferior archive by merging *xBest* into the current noninferior archive with domination relationship checked.

Step 6. Go to Step 3 or stop if a stop criterion is satisfied.

In the framework described above, three tricks in solving Equation (10) need to be highlighted:

(1) In the initialization of particles, the dimension of a particle is equal to the number of features, i.e., the ith particle's position is $X_i = [x_{i,1}, x_{i,2}, ..., x_{i,d}]$. To assign the jth feature to group m, the value of $x_{i,j}$ should be m ($m \in \{1, 2, \cdots, G\}$), where G is the pre-assigned group amount. Through this method,

every particle gets one-to-one correspondence to the group assignment matrix Q.

(2) The velocity of $x_{i,j}$ in kth iteration is updated by [29]:

$$V_{ij}{}^{k+1} = wV_{ij}{}^{k} + c_1 r_1 \left(P_{ij}^{k} - X_{ij}^{k}\right) + c_2 r_2 (P_{gd}^{k} - X_{ij}^{k}) \tag{11}$$

Considering the trick (1), the position of jth feature of ith particle X_{ij}^{k+1} which indicates the group assignment index is then updated by:

$$X_{ij}^{k+1} = \begin{cases} \lceil t \rceil & t \neq 0 \\ \lceil r_3 G \rceil & t = 0 \end{cases} \tag{12}$$

where :

$$t = \left| X_{ij}^{k} + V_{ij}^{k+1} \right| \% G \tag{13}$$

In Equations (11)–(13), V_{ij}^{k+1} is the velocity of jth feature, c_1 and c_2 are values that weight the contribution of the individual and social information, $r_1, r_2, r_3 \in [0, 1]$ are uniformly distributed random numbers, p_{ibest} is the best previous position of the ith particle and g_{best} is the best particle in the swarm, $\lceil t \rceil$ means the top integral function for t.

In our experiment, all particles X are randomly initialized in the range of [0, G], and the corresponding velocities V are initialized to 0. In the searching process, the value of velocity is influenced by w, c_1 and c_2. To guarantee the searching efficiency, we set the value of w linearly decreasing from 1.2 to 0.2, while c_1 and c_2 are both set 0.8. Consequently, the value scope of the velocity V is easily determined, just around [$-X$, X]. We mainly restrict the searching scope of X. If a particle exceeds the bound, we will pull the particle back to the searching scope by adding a random value in reverse direction.

(3) To improve the global search ability, the linear decreasing weight w in Equation (11) is updated by:

$$w = wmax - \frac{k \cdot (wmax - wmin)}{MaxIt} \tag{14}$$

where k is the current iteration number, *MaxIt* is the maximum iteration number, *wmin* and *wmax* are the minimal and maximal value of weight w respectively.

After assigning features into different groups, it needs to choose representative features from each group. In this work, information gain is introduced to conduct this selection. Information gain is an effective tool to evaluate the relatedness between two variables. For a classification problem, the information gain of one feature x_i to the classification label Y is defined as the change in information entropy H only for classification label Y to a state with this feature given, as follows:

$$IG\,(x_i) = H\,(Y) - H\,(Y|x_i) \tag{15}$$

The bigger the information gain is, the larger the relatedness between this feature and classification label is. As a result, we choose the features with the biggest information gain in each group as the representative ones, and finally combine them as a discriminative feature set. Using this feature set, we can apply SVM to construct the fault diagnosis model.

3.3. Method Description

Following the sections above, we give a total description of the proposed method: (1) Extract heterogeneous features of bearing fault, and combine them into a feature pool; (2) Use SVM to calculate the each feature's weight, and build the weight distance matrix; (3) Apply the MOPSO algorithm to determine the optimal grouping structure; (4) Use information gain to choose the most representative feature in each group; (5) Combine the obtained features again into a new feature set, and run SVM with this feature set to construct fault diagnosis model. The key part of the proposed method is determining the inner structure among features to guarantee the feature's discriminant ability. Moreover, as linear SVM has some advantages such as fast speed, simple structure for the extension to the big data problem, we choose it as the baseline algorithm to generate feature vectors.

For better understanding, we provide the flowchart of the proposed fault diagnosis method based on structural feature selection in Figure 3. Furthermore, we provide the pseudocode of the proposed structural feature selection algorithm, as shown in Algorithm 1.

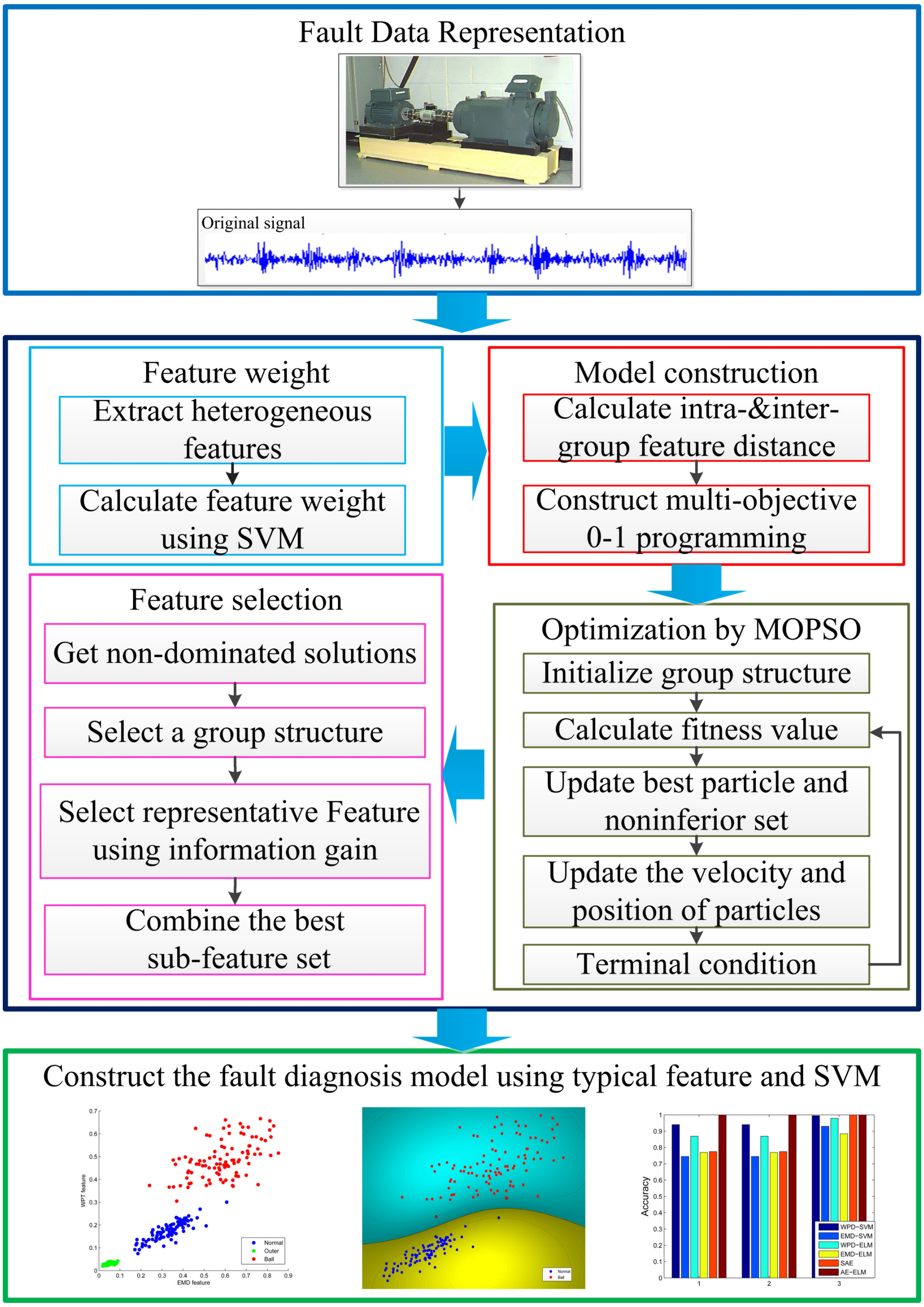

Fault Data Representation
Original signal
Feature weight
Extract heterogeneous features
Calculate feature weight using SVM
Model construction
Calculate intra-&inter-group feature distance
Construct multi-objective 0-1 programming
Feature selection
Get non-dominated solutions
Select a group structure
Select representative Feature using information gain
Combine the best sub-feature set
Optimization by MOPSO
Initialize group structure
Calculate fitness value
Update best particle and noninferior set
Update the velocity and position of particles
Terminal condition
Construct the fault diagnosis model using typical feature and SVM
WPT feature
EMD feature
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Normal
Outer
Ball
Normal
Ball
Accuracy
1
2
3
WPD−SVM
EMD−SVM
WPD−ELM
EMD−ELM
SAE
AE−ELM

Figure 3. Flowchart of the proposed fault diagnosis method based on structural feature selection.

Algorithm 1 The proposed fault diagnosis method based on structural feature selection.

```
Input: Training sample set Xtrain with extracted heterogeneous feature
Output:Representative feature set Rep_Feature
Begin
    Run Linear SVM on Xtrain to calculate feature's weight vector, and then build weight distance matrix $\mathbf{A}$;
    Initialize grouping assignment matrix $Q$;
    Use Equation (5) and (9) to calculate Intra-group weight distance sum and the minus of inter-group weight distance sum, respectively, as initial fitness value.
while $i$ does not reach the maximal iteration number:
    Use Equation (5) and (9) to calculate fitness value;
    Update the positions of best individual particle and best particle in swarm $P_{ij}^{k}$ ,$P_{gd}^{k}$ respectively;
    Update the noninferior set;
    Run the multi-objective PSO, with using Equation (10) to update particle's velocity $V_{ij}^{k+1}$, and using Equation (13) to guarantee the particle's position reaching the search space's requirement;
    Set $i = i + 1$;
end while
Obtain the optimal grouping assignment matrix $Q^{*}$;
Use Equation (15) to choose the feature with biggest information gain from each group, and combine them into a new feature set Rep_Feature;
Train fault diagnosis model by using SVM on Rep_Feature.
End
```

4. Experimental Results

In this section, we use two kinds of data sets to testify the effectiveness of the proposed method. Please note that before using bearing fault data for test, we also introduce four widely used UCI data sets [31] to evaluate the performance of the effect of the proposed structural feature selection method.

In this experiment, we use LibSVM [26] which is a popular open source SVM toolbox to conduct classification of SVM. By implementing the sequential minimal optimization (SMO) algorithm and various model selection algorithms, LibSVM can provide fast and stable prediction results with no need for tuning hyper-parameters repeatedly. For simplicity, the proposed method is called Structural Feature Selection-SVM (SFS-SVM). The proposed method is compared with some typical feature selection algorithms, i.e., Relief [32], SVM-RFE [27], feature selective validation(FSV) [33] and SVM without feature selection. The principle of SVM-RFE and SVM without feature selection have been elaborated in Section 2. Relief is a Filter-method approach which calculates a feature score for each feature to rank and select top scoring feature. In Relief, the feature scoring is based on the identification of feature value differences between nearest neighbor instance pairs. FSV is a widely used feature selection method which verifies the correlation of data to measure feature importance. We think these four methods can provide a comprehensive comparison.

Moreover, to evaluate the effect of the proposed method on deep learning techniques, we also compare the proposed method with two deep learning algorithms in fault diagnosis experiment. One is named DLSVM [34] which adds a linear SVM in SoftMax layer on deep neural network and minimizes a margin-based loss instead of the cross-entropy loss. We run the proposed method to further choose the most discriminant features from DLSVM, named SFS-DLSVM. The other one is stacked denoising auto-encoder (SDAE) [35] which feeds the frequency spectrum of vibration signal into SDAE model and adopts the hidden neurons as the extracted features. Finally, SDAE uses SoftMax layer to conduct classification. As SFS-DLSVM and SFS-SVM both adopt the same feature selection strategy, the results of these two methods can both demonstrate the comparative advantage of the proposed structural feature selection method. It is worth noting that the methodology of the proposed method SFS-SVM on UCI datasets and fault diagnosis datasets are the same, both including learning the inner structure of features, choosing typical features by means of information gain and building SVM model. The only difference is the original features.

4.1. UCI Data Sets

In this section, we select four datasets (*wdbc, ionosphere, breast cancer and SPECTF heart*) from UCI Machine Learning Repository [31] for test. Here we provide a brief introduction of these four datasets. The *wdbc* and *breast cancer* datasets are both the record data of breast cancer case. For the dataset of *breast cancer*, the first 30 features are collected from a digitized image of a fine needle aspirate of a breast mass and used to describe characteristics such as radius, texture and perimeter of the cell nuclei present in the image. For the dataset of *wdbc*, the features are computed from an image of cell nucleus with most of attributes being same to the breast cancer dataset. The dataset of *ionosphere* is collected to record the radar data of a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. The dataset of *SPECTF heart* describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature patterns were created for each patient. Some statistics of these four data sets are listed in Table 1. These datasets are randomly divided into training and test set. The partition ratio is as listed in Table 1. All the experimental results are the mean value of 30 random partitions of data set.

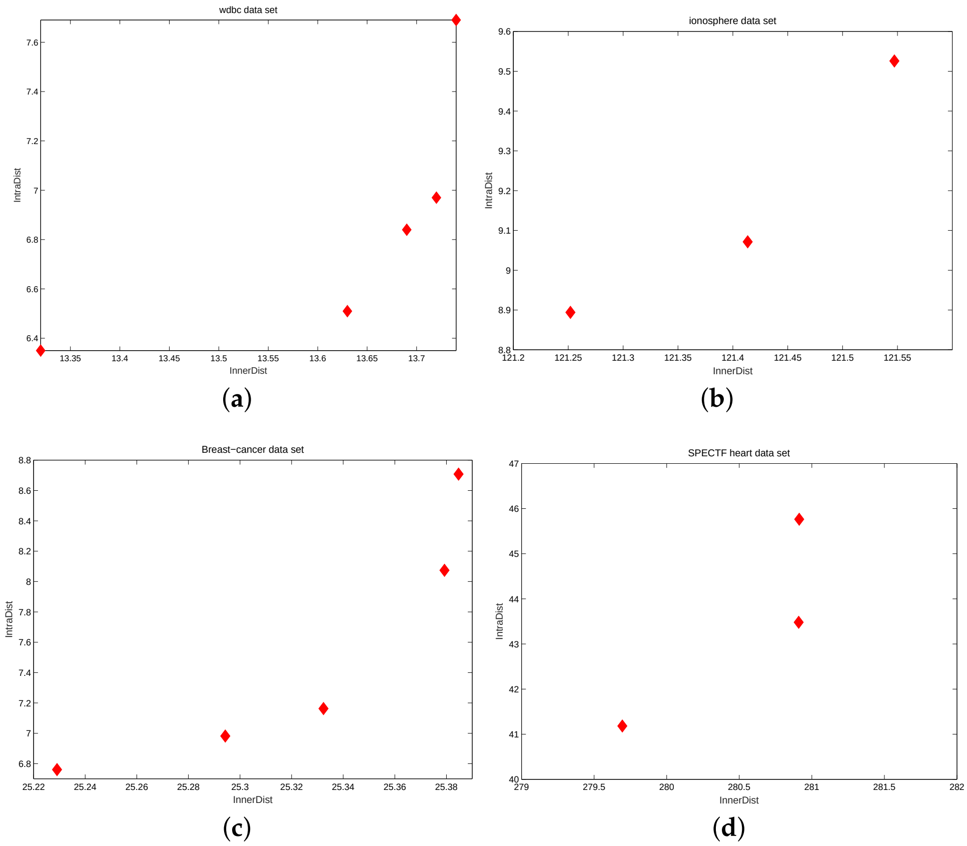
Table 1. Description of the used UCI datasets.

Name	Training	Test	Attribute	Class
wdbc	431	138	30	2
ionosphere	250	101	34	2
breast cancer	155	39	33	2
SPECTF heart	187	80	44	2

For the proposed method, the group number G can be directly set according to domain knowledge, or determined by cross validation. In this experiment, we set the group number G for data sets *wdbc* as 7, *ionosphere* as 8, *breast cancer* as 7, and *SPECTF heart* as 9, via cross validation. Larger swarm size will lead to a better searching effect of MOPSO, but the cost time will increase as well. In our experiments, we find different datasets need different swarm size to get a satisfactory result. To make a comprehensive comparison, we set the swarm size for datasets *wdbc* and *ionosphere* both as 50, *breast cancer* as 100, and *SPECTF heart* as 200. Also, to guarantee the optimization effect, the iteration number for the four data sets are all set 200 that is large enough in our experiment. Correspondingly, the value of parameter k of SVM-RFE and Relief is the same to G on each data set.

First, we check the performance of multi-objective PSO used in Section 3.2. Figure 4 provides the best individuals found in four datasets. It is clear that all the individuals have distribution approximated to the Pareto-optimal front, which indicates the intra-group weight distance sum and inter-group weight distance sum are approximately contradict. It is worth noting that Figure 4 is not the real Pareto-optimal front distribution, unless flipping the y-axis vertically. Compared with the single-objective optimization, the proposed method can get a set of Pareto-optimal solutions, which could provide the adaptable grouping structure.

From the grouping structure shown in Figure 4, a new representative feature set can be built by means of information gain. As mentioned before, we run two classical feature selection methods, SVM-RFE and Relief, and SVM with no feature selection for comparison. Considering that this comparative experiment is not for model selection but to evaluate the effect of structural feature selection, we directly set the regularization parameters of SFS-SVM, SVM-RFE and SVM as 50. The comparative results in terms of classification accuracy are shown in Table 2.



(**a**) (**b**)

(**c**) (**d**)

Figure 4. Best individuals found on (**a**) wdbc, (**b**) ionosphere, (**c**) breast cancer and (**d**) SPECTF heart data sets. It is clear that all the individuals have distribution approximated to the Pareto-optimal front.

Table 2. Comparative results on four UCI data sets in terms of classification accuracy.

Name	SFS-SVM	SVM-RFE	Relief	SVM
wdbc	**98.55%**	95.65%	92.03%	92.03%
ionosphere	**92.03%**	86.14%	87.13%	91.09%
breast cancer	**94.87%**	92.31%	79.79%	82.50%
SPECTF heart	**89.50%**	85%	80%	72.50%

From Table 2, we observe that almost all feature selection methods perform better than the SVM without feature selection except Relief, but the proposed method gets the highest classification accuracy on all four data sets. Relief is a typical Filter-kind method which does not rely on the classifier, therefore on a small-scale dataset, it may generate classification bias. The comparative results in Table 2 indicate that the proposed method can get benefit from the structural analysis for features and then find the most representative feature set for classification.

We also check the effect of the group number G. From Section 3.3, different grouping structure can generate different representative feature set. Therefore, we set different numbers of G from 2 to 16, and examine the classification accuracy of the proposed method on four UCI data sets, as shown in Figure 5. Obviously, the accuracy on *wdbc* and *SPECTF heart* data sets change less with the group number G increasing, while the accuracy on *breast cancer* fluctuates drastically. We also observe the accuracy on *ionosphere* rising heavily at an earlier stage and then remains almost constant with G increasing. These results demonstrate that the group number plays a key role in the proposed method. With a different number of G, the representative features vary and the redundant features may not be eliminated well. If the group number is less than the optimal group number, it will result in fewer representative features which could provide insufficient domain knowledge as well as reduce the classification accuracy. Conversely, with excessive group number, the representative features will increase, which will keep some redundant features. In practical applications, the group number G can be chosen via cross validation.

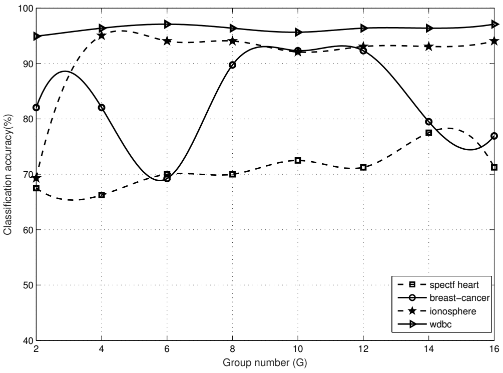


Figure 5. Classification accuracy of the proposed method on four UCI data sets.

4.2. Bearing Fault Data Sets

To further evaluate the performance of the proposed method on bearing fault diagnosis, two bearing fault data sets, CWRU and IMS, are introduced to run computer experiments. For comparison, we first extract heterogeneous fault features with a high dimension by referencing the feature extraction methods in [7]. Second, we run the proposed method SFS-SVM and FSV, Relief and SVM-RFE on the extracted heterogeneous features to compare the diagnosis accuracy. We use LibSVM toolbox [26] for SVM-based methods with linear kernel selected. The proposed method with the linear kernel can directly be applied to DLSVM for conducting structural feature selection on deep neural network, and the linear kernel has no more kernel parameter for selection. To test the effectiveness of the grouping strategy used in this work, all SVM-based methods use same regularization parameter (i.e., the parameter C in Equation (1) and the same parameter in the other SVM-based models) which is set 512. Moreover, the proposed method is initially set 7 groups in the experiments for both two datasets. Correspondingly, all other methods for comparison are all set to select 7 representative features. For SDAE, the network architecture is set [512, 50, 30, 7]. For DLSVM, two convolutional layers with 5×5 filter are adopted. To keep in line, DLSVM also generate 7-dimensional features. After feature extraction, the sample set is randomly partitioned into a training set and a test set with the ratio 7:3. To make an impartial comparison, we get the mean value of 100 repeated trials as the final result.

4.2.1. Data Description

The CWRU bearing dataset came from Case Western Reserve University (CWRU) ElectrotechnicsLab [36]. In this dataset, all kinds of bearing fault were generated by using electro-discharge machining at inner race, outer race and ball with crack size of 0.007 inch, 0.014 inch, 0.021 inch and 0.028 inch, respectively. Then the vibration signals with different fault location and crack size are recorded under motor loads of 0, 1, 2 and 3hp. Besides normal condition, this dataset adopts the fault data with sampling frequency of 12 kHz at Fan End (FE) and Drive End (DE) as well as the fault data with sampling rates of 48 kHz at drive end. Therefore, this dataset contains four kinds of health condition: normal condition, inner race fault, outer race fault and ball fault.

The IMS bearing dataset was generated by the NSF I/UCR Center for Intelligent Maintenance Systems(IMS) with support from Rexnord Corp. in Milwaukee, WI [37]. This dataset provides two different test-to-failure experiments including outer race fault and ball fault. The recording duration with outer race fault is from 22 October 2003, 12:06:24 to 25 November 2003, 23:39:56. The recording duration with ball fault is from 12 February 2004 10:32:39 to 19 February 2004, 06:22:39. The sampling rate is 20 kHz. At the end of experiment, bearing 3 and 4 come out outer race fault and ball fault, respectively. As IMS bearing data are run-to-failure data with whole degradation process, we choose the data which were collected at 164 h as the outer race fault signal and the data at 827 h as the ball fault signal.

To provide an intuitive understanding of the problem of bearing fault diagnosis, We take CWRU dataset as an example to show the raw vibration signals and their Fast Fourier Transform (FFT) for all four health conditions, as shown in Figure 6. Here the signals with 1024 time points are collected at Fun End with sampling rate 12 kHz. The load is 3 hp, and the crack size is 0.007 inch. Besides CWRU dataset, the raw signals and their FFT of IMS dataset are also shown in Figure 7.

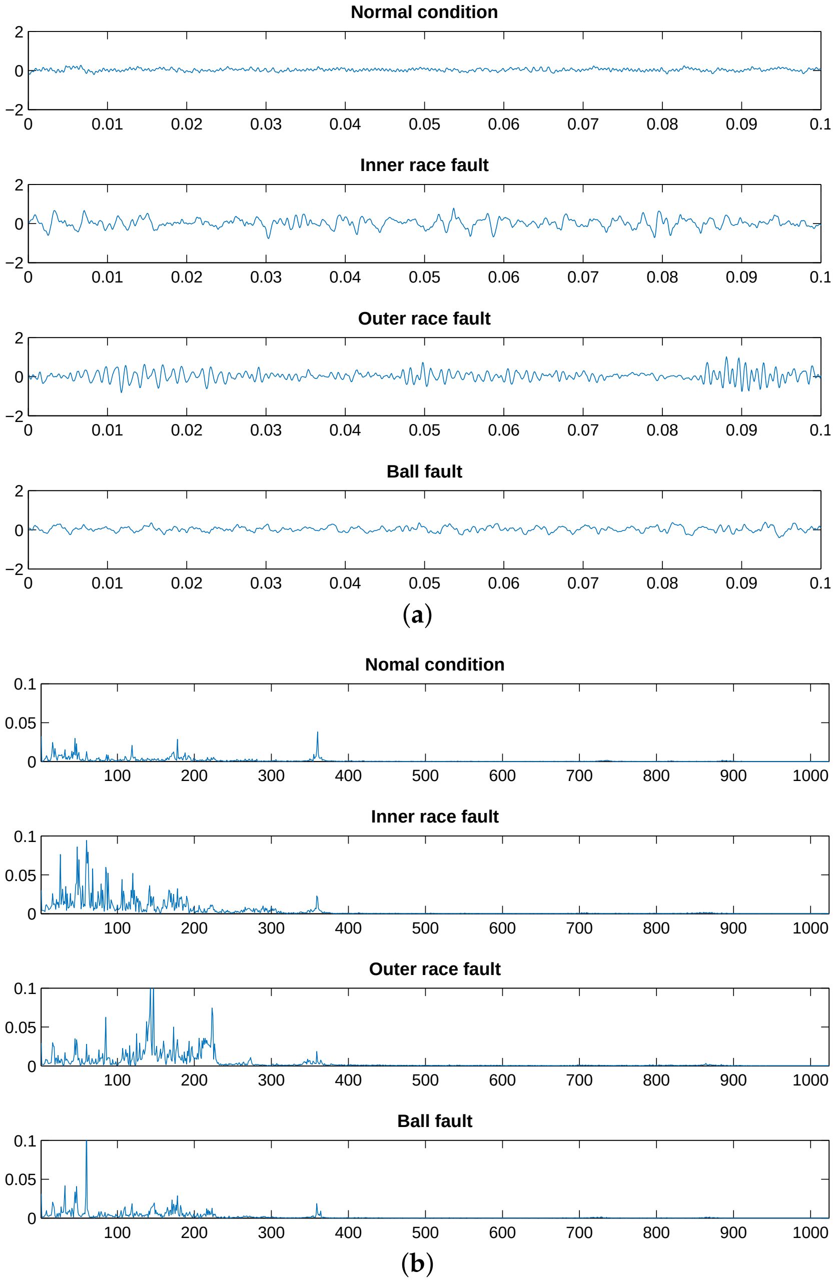
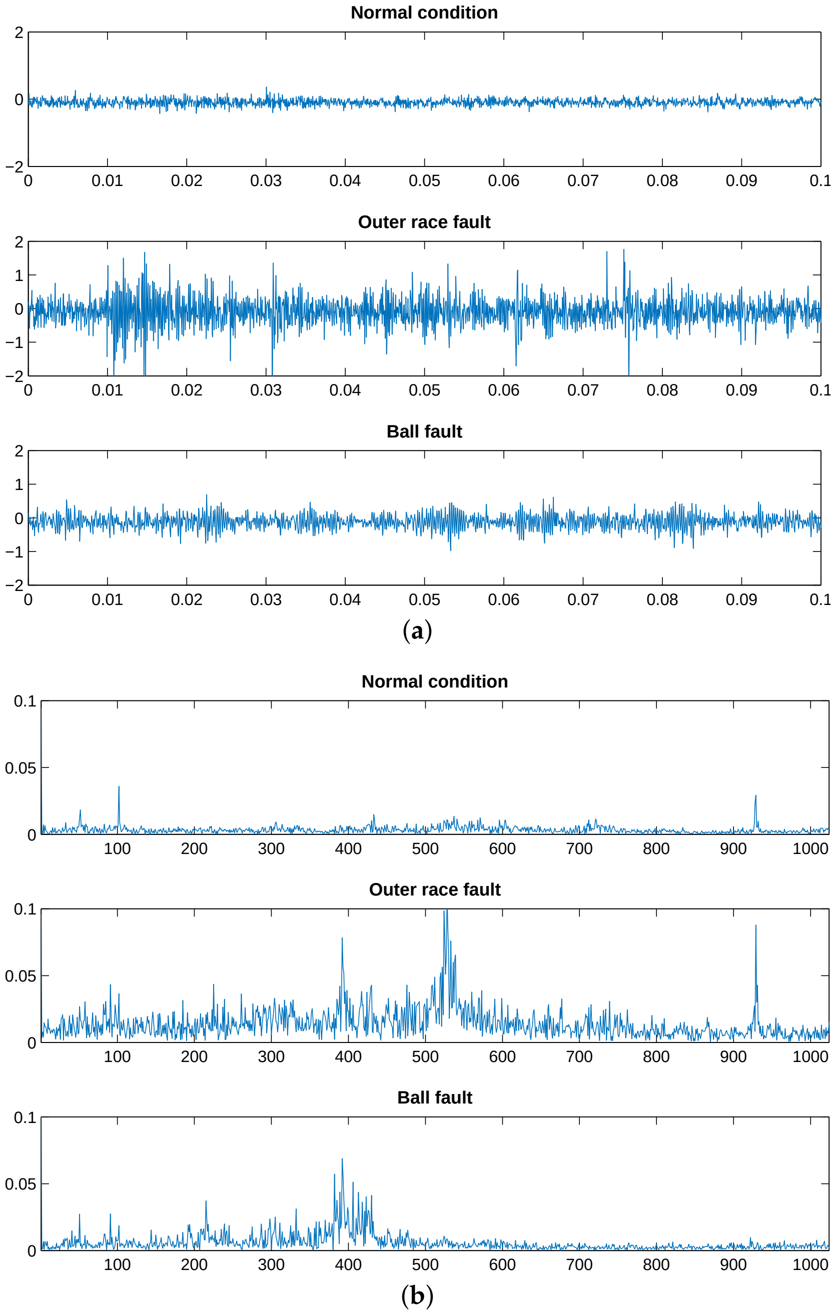


Figure 6. Four health conditions of CWRU dataset with (**a**) raw time signals and (**b**) their FFT.



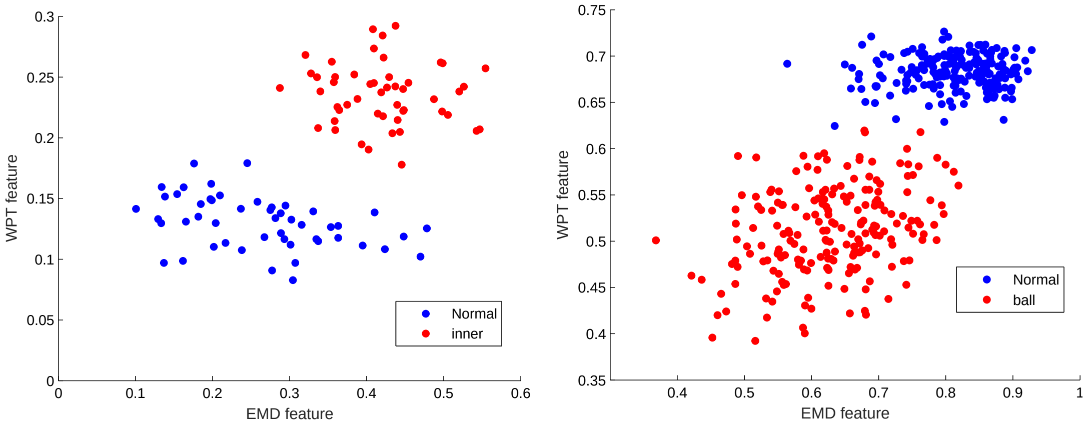
(**a**)

(**b**)

Figure 7. Three health conditions of IMS dataset with (**a**) raw time signals and (**b**) their FFT.

It is clear that different health conditions have different patterns both in time domain and frequency domain, which provides discriminant mode for classification. Especially in Figure 6a, the raw signal of outer race fault has more obvious fluctuation in time domain, but the ball fault changes less than the other fault classes, which indicates the ball fault is more difficult to be classified. The same comparative phenomenon happens in Figure 6b, as the frequency spectrum of outer race fault changes much more than the other two faults. In Figure 7, the raw signal of outer race fault still fluctuates dramatically more than the ball fault, while the frequency spectrum of outer race fault appears more changing modes, which also means the applicability of machine learning for fault diagnosis problem. To further testify the separability, we show the feature distribution of different health conditions, as shown in Figure 8. For the sake of illustration, in Figure 8a we randomly choose the first IMF component of EMD and the 4-th sub-frequency coefficient of WPT as two coordinate axes, and plot the distribution of 50 samples of normal condition and 50 samples of inner race fault.

Similarly, in Figure 8b, we plot 200 samples of normal condition and 200 samples of ball fault with the second IMF component of EMD and the first sub-frequency coefficient of WPT selected. Obviously, no matter CWRU or IMS dataset, the extracted features are all able to provide a separable distribution of different health conditions, which is the base of our following experiments.



(**a**) (**b**)

Figure 8. Data distribution using EMD and WPT feature on (**a**) CWRU and (**b**) IMS datasets.

4.2.2. Results on CWRU Dataset

In this experiment, we choose the vibration signals with crack size 0.007 inch, 0.014 inch and 0.021 inch under motor load 1, 2, and 3 as ball fault data, and the normal signals under load 1, 2, 3 as normal condition data. The sampling rate here is 12 kHz. After feature extraction, we get 50 samples for each ball fault normal condition respectively, i.e., 450 ball fault samples and 150 normal condition samples in total, as shown in Table 3.

Table 3. Sample size of ball fault and normal condition.

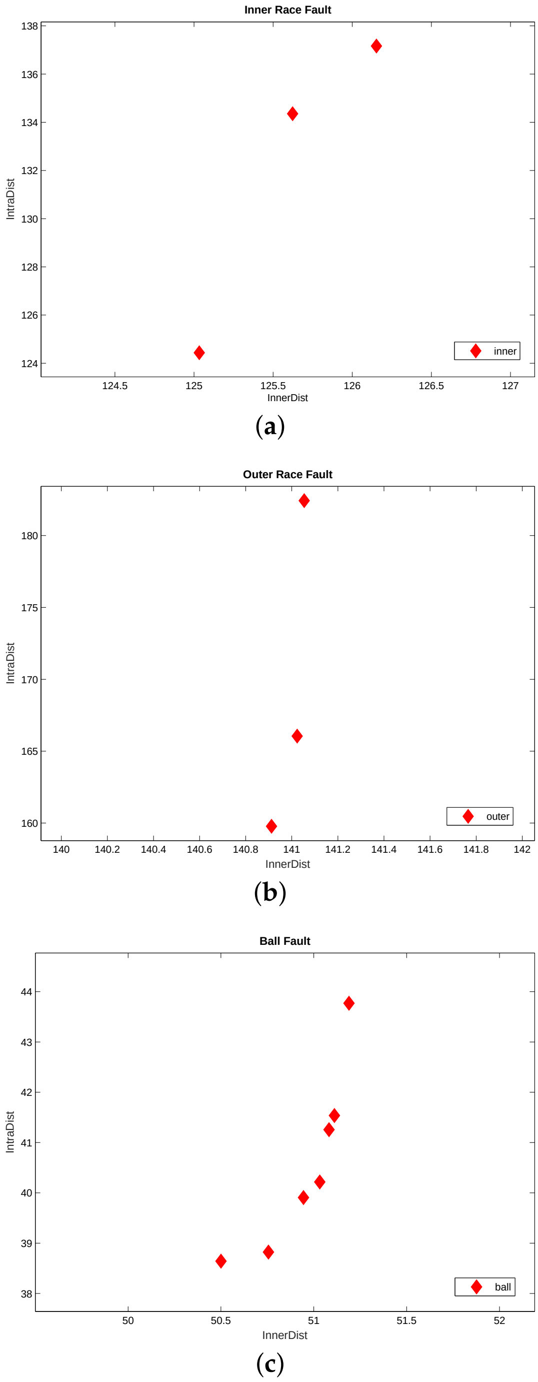
Ball Fault				**Normal Condition**		**Total Sample Size**
Load\Crack	0.007 inch	0.014 inch	0.021 inch	Load	Size	
1	50	50	50	1	50	600
2	50	50	50	2	50	
3	50	50	50	3	50	

Each sample includes 1024 signal points. For each sample, we extract heterogeneous features by using the methods in [5]. These methods contain not only the time domain/frequency domain statistical features, but also some time-frequency domain features such as EMD/WPT as well as temporal features, as shown in Table 4. Please note that to generate heterogeneity of features, we both collect the signals from Fan End (FE) and Drive End (DE) of the motor housing of the CWRU testbed. The sensors at DE, although with less confidence, can detect the faults at FE, and vise verse [5]. This phenomenon is called *cross detection* which duplicates the number of features and makes the features contain redundancy and noise. Hence, although Table 4 provides 71-dimension features, we finally obtain a feature pool whose cardinality is 142.

Table 4. Definition of 71-dimension fault features.

Method	Formula	Dimension
Bispectrum Analysis	$B_x(w_1,w_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} c_{3x}(\tau_1,\tau_2) e^{-j(w_1\tau_1+w_2\tau_2)}$	10
GARCH Model	$r_t = c_1 + \sum_{i=1}^{R} \phi_i r_{t-i} + \sum_{j=1}^{M} \phi_j \varepsilon_{t-j}\ {}_{t} + \varepsilon_t$ $\varepsilon_t = \mu_t \sqrt{h_t}$ $h_t = k + \sum_{i=1}^{q} G_i h_{t-i} + \sum_{i=1}^{p} A_i \varepsilon_{t-i}^2$	4
EMD	$x(t) = r_n + \sum_{i=1}^{n} IMF_i$	10
WPT	$E_j(n) = \sum_{s=0}^{S/2^j-1} \left[c_{j,n}^s\right]^2$ $x_n = \frac{E_j(n)}{\sum_{m=0}^{2^j} E_j(m)}$	16
Complex Envelope Analysis	$\widetilde{h}(t) : H\{h(t)\} := h(t) * \frac{1}{\pi t} = \frac{1}{\pi} \int_{-\infty}^{\infty} h(t) \frac{d\tau}{t-\tau}$	18
Impulse Factor	$X_{if} = \frac{\max(\lvert x_i \rvert)}{\frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert}$	1
Margin Factor	$X_{mf} = \frac{\max(\lvert x_i \rvert)}{\left(\frac{1}{N}\sum_{i=1}^{N} \sqrt{\lvert x_i \rvert}\right)^2}$	1
Shape Factor	$X_{sf} = \frac{\max(\lvert x_i \rvert)}{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{1/2}}$	1
Kurtosis Factor	$X_{kf} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^2}$	1
Crest Factor	$X_{cf} = \frac{\max(\lvert x_i \rvert)}{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{1/2}}$	1
Peak-to-Peak Value	$X_{ppv} = \max(x_i) - \min(x_i)$	1
Skewness Value	$X_{sv} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3$	1
Kurtosis Value	$X_{kv} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4$	1
Square Root of the Amplitude	$X_{sra} = \left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i \rvert}\right)^2$	1
Root Mean Square	$X_{rms} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{1/2}$	1
Frequency Center	$X_{fc} = \frac{1}{N}\sum_{i=1}^{N} f_i$	1
RMS Frequency	$X_{rmsf} = \left(\frac{1}{N}\sum_{i=1}^{N} f_i^2\right)^{1/2}$	1
Root Variance Frequency	$X_{rvf} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(f_i - X_{fc}\right)^2\right)^{1/2}$	1

For comparison, the whole 142-dimension feature pool without feature selection and the sub-feature set from FSV, Relief and SVM-RFE are input into linear LibSVM respectively to conduct diagnosis on CWRU fault data. First, we check the best solutions to the proposed method for three fault types, as shown in Figure 9. Please note that one solution in Figure 8 means a grouping structure Q_g in Section 3. Similar to Figure 4, it is also clear that the distribution of these solutions all approach the Pareto-optimal front, which indicates the strategy of multi-objective optimization suits the CWRU fault data.



(**a**)

(**b**)

(**c**)

Figure 9. Distribution of best solutions to the proposed method on CWRU dataset for (**a**) inner race fault, (**b**) outer race fault and (**c**) ball fault.

We randomly select one of the best solutions shown in Figure 9 as the optimal group structure, and run the proposed method SFS-SVM for comparison with the other seven methods. Please note that for SFS-DLSVM, we run the proposed structural feature selection strategy on the features extracted by DLSVM. The diagnosis results of all eight methods for inner race fault, outer race fault and ball fault are listed in Tables 5–7, respectively. As the collected data is sort of class imbalanced, here we provide the classification accuracy on the test set for each bi-classification problem. Moreover, we also provide G-mean value which is suitable for data imbalance problem. G-mean value is the square root of the product of two classes' accuracy; therefore, the G-mean value will improve only when the accuracy of two classes both improve.

Table 5. Diagnosis accuracy for inner race fault and normal condition on CWRU dataset.

	SVM	**SVM-RFE**	**Relief**	**FSV**	**SFS-SVM**	**DLSVM**	**SFS-DLSVM**	**SDAE**
Accuracy of normal condition(%)	99.21	98.98	31.3	99.46	100	100	**100**	100
Accuracy of inner race fault(%)	99.47	99.65	99.91	99.29	99.65	99.53	**99.96**	99.61
Whole test accuracy(%)	99.40	99.47	82.33	99.33	99.73	99.78	**99.98**	99.80
G-mean value	0.9934	0.9931	0.496	0.9937	0.9982	0.9976	**0.9998**	0.9980

Table 6. Diagnosis accuracy for outer race fault and normal condition on CWRU dataset.

	SVM	**SVM-RFE**	**Relief**	**FSV**	**SFS-SVM**	**DLSVM**	**SFS-DLSVM**	**SDAE**
Accuracy of normal condition(%)	98.73	87.34	36.71	96.20	**100**	100	**100**	100
Accuracy of outer race fault(%)	99.55	98.19	99.10	100	**100**	99.92	**100**	99.56
Whole test accuracy(%)	99.33	95.33	82.67	99.0	**100**	99.96	**100**	99.78
G-mean value	0.9914	0.9261	0.6031	0.9808	**1**	0.9996	**1**	0.9978

Table 7. Diagnosis accuracy for ball fault and normal condition on CWRU dataset.

	SVM	**SVM-RFE**	**Relief**	**FSV**	**SFS-SVM**	**DLSVM**	**SFS-DLSVM**	**SDAE**
Accuracy of normal condition (%)	100	95.92	42.60	100	100	100	100	100
Accuracy of ball fault (%)	100	98.58	93.26	100	100	100	100	100
Whole test accuracy (%)	100	97.87	80.07	100	100	100	100	100
G-mean value	1	0.9722	0.6054	1	1	1	1	1

From Tables 5–7, it is clear that SFS-GLSVM outperforms the other seven methods for all three fault types, while the propose method SFS-SVM performs the second-best. We observe that the accuracy using the whole feature pool may have good diagnosis accuracy, but with proper feature selection method (just like Relief in Table 5, FSV in Table 6), the diagnosis accuracy will be further improved. This is because some redundant information exists in the feature pool. However, by introducing grouping information among features, SFS-SVM and SFS-DLSVM both get the most representative features and then obtain the best performance in all methods, which exactly demonstrates the effectiveness of the structural feature selection strategy, no matter on traditional shallow model or deep model. We also observe that SDAE cannot get the best results, which means on small-scale data, deep learning methods would not guarantee the generalization performance.

We also notice that the ball fault is mostly easy to be classified. As in Table 7, no matter SVM, FSV, SFS-SVM and three deep models, the accuracy of these methods are all 100%. However, in this scenario, Relief still has low accuracy for normal condition but high value for ball fault as well, which indicates FSV suffers from data imbalance problem. This comparison means improper feature selection may instead reduce the diagnosis performance. Getting help from the structural feature selection strategy, SFS-SVM and SFS-DLSVM always get the highest accuracy and G-mean value for three fault types.

Moreover, we introduce receiver operating characteristic (ROC) curve to evaluate the diagnostic performance of these methods. ROC curve is a commonly used error index that comprehensively reflects sensitivity and specificity. Setting sensitivity and specificity as vertical and horizontal coordinates, the total area under ROC curve (AUC), determines the performance of classification. For this experiment, the larger the AUC is, the better the method performs. Considering the illustration effect, we take the outer race fault diagnosis as an example, and give the ROC curve of five methods in Figure 10. As the effect of three deep learning methods are too dense to be distinguished, here we only choose five methods for analysis. Other fault types have a similar comparative effect. It is clear that the proposed method SFS-SVM gets a larger AUC area (0.9996) than SVM (0.9926), FSV (0.9889), Relief (0.6783) and SVM-RFE (0.9457) for outer race fault.

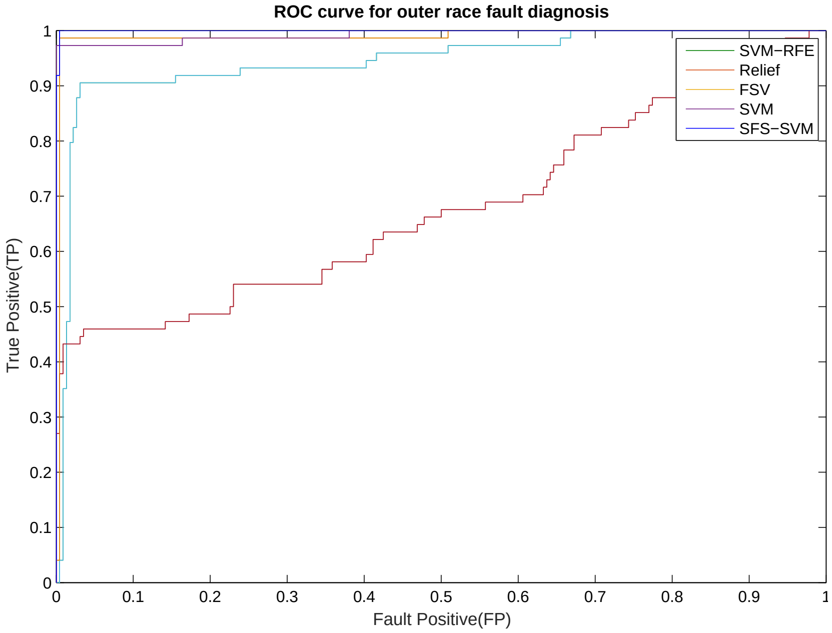


Figure 10. ROC curve of five methods for outer race fault diagnosis.

4.2.3. Results on IMS Dataset

Because IMS dataset comes from two test-to-failure experiments, we choose the vibration signals in the first hour of bearing 3 and bearing 4 as the data of normal condition. Correspondingly, we choose the signals in the last hour of bearings 3 & 4 as the data of outer race fault and ball fault, respectively. As a result, this experiment constructs 300 samples for outer race fault and 300 samples for ball fault, with 300 samples for normal condition, respectively. Each sample has 1024 signal points. We also use the feature methods in Table 4 to extract fault features, and finally obtain 53-dimension features, including 10 bispectrum features, 4 GARCH model features, 10 EMD features, 13 time domain statistical features and 16 WPT features. The network architecture and parameters of three deep learning methods are same to the ones on CWRU dataset.

First, we check the best solutions to the proposed method for two fault types, as shown in Figure 11. Here the group number is set 7. Similar to Figure 9, it is also clear that the distribution of these solutions approach the Pareto-optimal front, which indicates the optimization strategy used in the proposed method suits to the IMS fault data.

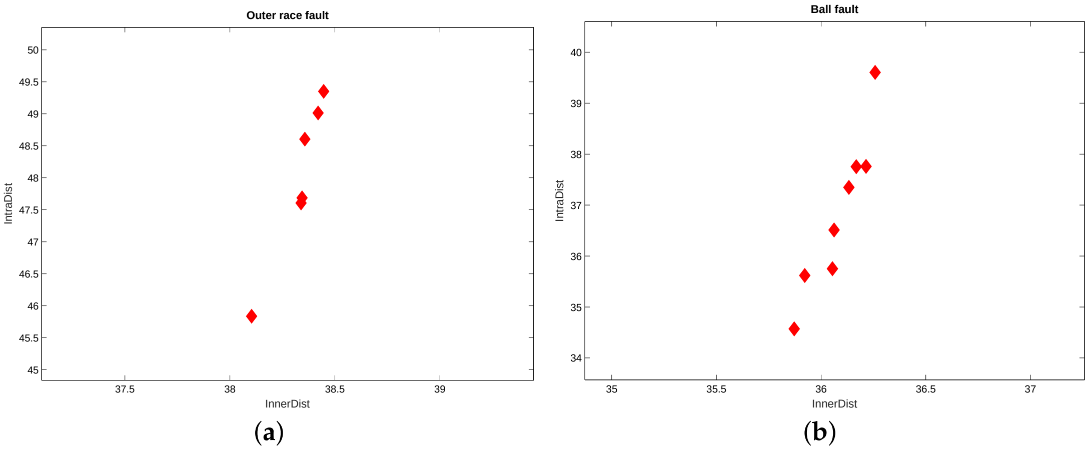


Figure 11. Distribution of best solutions to the proposed method on IMS dataset for (**a**) outer race fault and (**b**) ball fault.

Randomly select one solution in Figure 11, and then get an optimal grouping structure. Same with Table 5, we run the proposed methods for comparison. The comparative results are listed in Tables 8 and 9.

Table 8. Diagnosis accuracy for outer race fault and normal condition on IMS dataset.

	SVM	SVM-RFE	Relief	FSV	SFS-SVM	DLSVM	SFS-DLSVM	SDAE
Accuracy of normal condition (%)	95.53	96.22	96.88	96.92	97.42	96.59	**98.95**	98.11
Accuracy of outer race fault (%)	89.66	98.07	91.55	94.95	98.60	97.85	**99.33**	98.87
Whole test accuracy (%)	92.57	97.13	94.27	95.90	98.0	97.72	**99.20**	98.48
G-mean value	0.9252	0.9713	0.9411	0.9591	0.98	0.9722	**0.9914**	0.9849

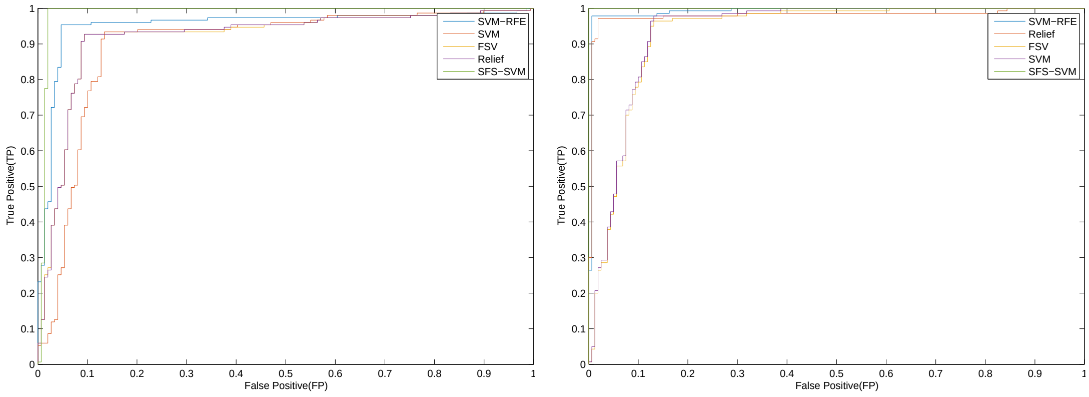
Table 9. Diagnosis accuracy for ball fault and normal condition on IMS dataset.

	SVM	SVM-RFE	Relief	FSV	SFS-SVM	DLSVM	SFS-DLSVM	SDAE
Accuracy of normal condition (%)	95.96	96.67	96.78	98.03	97.46	97.56	**98.87**	98.13
Accuracy of ball fault (%)	87.57	97.36	88.97	96.17	98.28	98.16	**99.13**	97.45
Whole test accuracy (%)	91.83	96.97	92.93	97.13	97.83	97.86	**98.99**	97.76
G-mean value	0.9166	0.9701	0.9269	0.9708	0.9786	0.9786	**0.9901**	0.9779

In this experiment, we choose the signals in the last hour of total 827 h as fault data. At that time, the fault is fully formed, therefore, the fault state is easily to be classified with high diagnosis accuracy of all five methods. However, in these methods, SFS-DLSVM still gets the best results and SFS-SVM get almost equal results with the other two deep learning methods. Just a little different from Tables 5–7, here SVM with no feature selection all gets the lowest accuracy in all methods, which indicates the effect of feature selection. For outer race fault diagnosis, SVM-RFE gets a higher accuracy than FSV, but for ball fault, this comparison is just the opposite, which means different fault type needs proper selection method. By getting help from structural information among heterogeneous features, SFS-SVM always gets the best performance in all shallow models in terms of whole accuracy and G-mean value. Same comparative effect also exists in deep learning methods where SFS-DLSVM gets the best results. Thus, we can conclude that the structural feature selection strategy works well no matter on heterogeneous features or deep features.

Actually, we test different parameters and sample split for the proposed method, and it always gets the best or the second-best diagnosis performance, which gives a firm demonstration about the effect of the structural information for fault diagnosis. Please note that in this experiment, two classes have almost same samples for training, so G-mean value is just close to the whole test accuracy.

Figure 12 provides the ROC curve of five shallow methods for outer race fault and ball fault diagnosis. It is clear that the proposed method SFS-SVM has the largest AUC area in all methods, even though the other four methods already have large area. This comparison also keeps line with Figure 10.



(**a**) (**b**)

Figure 12. ROC curve of five methods on IMS dataset for (**a**) outer race fault and (**b**) ball fault diagnosis.

5. Conclusions

In this paper, a new bearing fault diagnosis method is proposed based on structural feature selection. In contrast to most of common fault diagnosis methods, the proposed method transfers the problem of fault diagnosis into a new multi-objective 0–1 programming problem, and then uses the MOPSO algorithm to exploit the structural relationship among heterogeneous fault features. This operation is helpful to choose the most representative fault features for constructing diagnosis model. Then the proposed method is universally applicable to various rotating machineries. From the experimental results, we have the following conclusions:

(1) Heterogeneous features of bearing may have intrinsic structures such as group or others, which will provide useful information for fault diagnosis;

(2) The proposed method does not guarantee the highest diagnosis accuracy each time, but it can almost get the second-best performance at least, which shows the effect of structural information.

(3) Feature selection does not always work well for fault diagnosis. Too few selected features may perform worse than no feature selection. To guarantee the diagnosis performance, structural information should be used to select more representative features in the case of limited feature number.

(4) The proposed structural feature selection strategy works well on not only traditional heterogeneous features but also deep features. Even extracted by deep neural networks, the deep features still have some inner structure which is also helpful for fault diagnosis.

In our next work, we plan to study theoretically the generalization analysis of the proposed method. Ideally speaking, if we can estimate an upper error bound, the inner structure of heterogeneous features can be evaluated more accurately. Moreover, incipient fault diagnosis generally has inconspicuous fault features. How to exploit the inner structure of such kind of features is still challenging.

Author Contributions: Conceptualization, W.M. and L.W.; Methodology, W.M.; Software, L.W.; Validation, W.M., and N.F.; Formal analysis, W.M.; Investigation, W.M.; Resources, W.M.; Data curation, W.M.; Writing—original draft preparation, W.M.; Writing—review and editing, N.F.; Visualization, W.M.; Supervision, N.F.; Funding acquisition, W.M.

Funding: This research was funded by the National Natural Science Foundation of China grant number U1704158 and China Postdoctoral Science Foundation Special Support grant number 2016T90944.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gao, Z.; Cecati, C.; Ding, S.X. A survey of fault diagnosis and fault-tolerant techniques–Part I: fault diagnosis with model-based and signal-based approaches. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3757–3767. [CrossRef]
2. Mao, W.; Chen, J.; Liang, X. A new online detection approach for rolling bearing incipient fault via self-adaptive deep feature matching. *IEEE Trans. Instrum. Meas.* **2019**, in press. [CrossRef]
3. Caesarendra, W.; Widodo, A.; Yang, B.S. Application of relevance vector machine and logistic regression for machine degradation assessment. *Mech. Syst. Signal Process* **2010**, *24*, 1161–1171. [CrossRef]
4. Prieto, M.D.; Cirrincione, G.; Espinosa, A.G. Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks. *IEEE Trans. Ind. Electron.* **2013**, *60*, 3398–3407. [CrossRef]
5. Rauber, T.W.; Boldt, F.A.; Varejao, F.M. Heterogeneous feature models and feature selection applied to bearing fault diagnosis. *IEEE Trans. Ind. Electron.* **2015**, *62*, 637–646. [CrossRef]
6. Antoni, J.; Randall, R.B. The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mech. Syst. Signal Process.* **2006**, *20*, 308–331. [CrossRef]
7. Wang, Y.; Xu, G.; Liang, L.; Jiang, K. Detection of weak transient signals based on wavelet packet transform and manifold learning for rolling element bearing fault diagnosis. *Mech. Syst. Signal Process.* **2015**, *54–55*, 259–276. [CrossRef]

8. Lei, Y.G.; Lin, J.; He, Z.J.; Zuo, M.J. A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.* **2013**, *35*, 108–126. [CrossRef]
9. Mao, W.; He, L.; Yan, Y.; Wang, J. Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine. *Mech. Syst. Signal Process.* **2017**, *83*, 450–473. [CrossRef]
10. Qiu, H.; Lee, J.; Lin, J.; Yu, G. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *J. Sound Vib.* **2006**, *289*, 1066–1090. [CrossRef]
11. Ren, L.; Lv, W.; Jiang, S.; Xiao, Y. Fault diagnosis using a joint model based on sparse representation and SVM. *IEEE Trans. Instrum. Meas.* **2016**, *65*, 2313–2320. [CrossRef]
12. Soualhi, A.; Medjaher, K.; Zerhouni, N. Bearing health monitoring based on Hilbert-Huang transform, support vector machine, and regression. *IEEE Trans. Instrum. Meas.* **2015**, *64*, 52–62. [CrossRef]
13. Cai, J.; Xiao, Y. Time-frequency analysis method of bearing fault diagnosis based on the generalized S transformation. *J. Vibroeng.* **2017**, *19*, 4221–4230. [CrossRef]
14. Ali, J.B.; Fnaiech, N.; Saidi, L.; Chebel-Morello, B.; Fnaiech, F. Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. *Appl. Acoust.* **2015**, *89*, 16–27.
15. Jia, F.; Lei, Y.G.; Lin, J.; Zhou, X.; Lu, N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **2016**, *72–73*, 303–315. [CrossRef]
16. Mao, W.; Feng, W.; Liang, X. A novel deep output kernel learning method for bearing fault structural diagnosis. *Mech. Syst. Signal Process.* **2019**, *117*, 293–318. [CrossRef]
17. Mao, W.; He, J.; Li, Y.; Yan, Y. Bearing fault diagnosis with auto-encoder extreme learning machine: A comparative study. *Proc. Inst. Mech. Eng. Part C: J. Mech. Eng. Sci.* **2017**, *231*, 1560–1578. [CrossRef]
18. Jia, F.; Lei, Y.; Guo, L.; Lin, J.; Xing, S. A neural network constructed by normalized sparse autoencoder and its application to intelligent fault diagnosis of machines. *Neurocomputing* **2018**, *272*, 619–628. [CrossRef]
19. Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. *Measurement* **2016**, *93*, 490–502. [CrossRef]
20. Liu, H.; Yu, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 491–502.
21. Han, L.; Zhang, Y. Discriminative Feature Grouping. In Proceedings of the AAAI Conf. Artificial Intelligence, Austin, TX, USA, 25–29 January 2015; pp. 2631–2637.
22. Mao, W.; Xu, W.; Li, Y. Sparse Feature Grouping based on L1/2 Norm Regularization. In Proceedings of the American Control Conference 2018 (ACC2018), Milwaukee, WI, USA, 27–29 June 2018.
23. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proc. R. Soc. Lond. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
24. Looney, D.; Mandic, D.P. Multiscale Image Fusion Using Complex Extensions of EMD. *IEEE Trans. Signal Process.* **2009**, *57*, 1626–1630. [CrossRef]
25. Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing multiple parameters for support vector machines. *J. Mach. Learn. Res.* **2002**, *46*, 131–159. [CrossRef]
26. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
27. Duan, K.B.; Rajapakse, J.C.; Wang, H.; Azuaje, F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans. NanoBiosci.* **2005**, *4*, 228–234. [CrossRef] [PubMed]
28. Huang, V.L.; Suganthan, P.N.; Liang, J.J. Comprehensive learning particle swarm qptimizer for solving multiobjective optimization problems. *Int. J. Intell. Syst.* **2006**, *21*, 209–226. [CrossRef]
29. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995; IEEE Service Center: Piscataway, NJ, USA, 1995; Volume 4, pp. 1942–1948.
30. Mao, W.; Mu, X.; Zheng, Y.; Yan, G. Leave-one-out cross-validation-based model selection for multi-input multi-output support vector machine. *Neural Comput. Appl.* **2014**, *24*, 441–451. [CrossRef]
31. Dua, D.; Graff, C. *UCI Repository of Machine Learning Databases*; School of Information and Computer Science, University of California: Irvine, CA, USA. Available online: http://archive.ics.uci.edu/ml (accessed on 24 November 2019).

32. Sun, Y.J. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1035–1051. [CrossRef]
33. Rusiecki, A.; Aniserowicz, K.; Duffy, A.P.; Orlandi, A. The feature selective validation technique as analysis tool for shielding effectiveness of slotted enclosures. *IEEE Trans. Electromagn. Compat.* **2015**, *57*, 1472–1480. [CrossRef]
34. Tang, Y. Deep learning using linear support vector machines. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013.
35. Shao, H.; Jiang, H.; Zhao, H.; Wang, F. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. *Mech. Syst. Signal Process.* **2017**, *95*, 187–204. [CrossRef]
36. These Data Comes from Case Western Reserve University Bearing Data Center. Available online: https://csegroups.case.edu/bearingdatacenter/pages/download-data-file (accessed on 24 November 2019).
37. Lee, J.; Qiu, H.; Yu, G.; Lin, J.; Rexnord Technical Services. *IMS Bearing Data Set, NASA Ames Prognostics Data Repository*; NASA Ames Research Center, Moffett Field, CA, 2007. Available online: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/ (accessed on 24 November 2019)