applied sciences

MDPI

Article

Improving Bearing Fault Diagnosis Using Maximum Information Coefficient Based Feature Selection

Xianghong Tang [1,2,4], Jiachen Wang [1,*], Jianguang Lu [1,2,4,*], Guokai Liu [3] and Jiadui Chen [1,4]

1 Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang 550025, China; xhtang@gzu.edu.cn (X.T.); chjd97@163.com (J.C.)
2 State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
3 State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China; liuguokai@hust.edu.cn
4 School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
* Correspondence: jessicaxiaowang@163.com (J.W.); jglu@gzu.edu.cn (J.L.)

Received: 8 October 2018; Accepted: 30 October 2018; Published: 2 November 2018

Abstract: Effective feature selection can help improve the classification performance in bearing fault diagnosis. This paper proposes a novel feature selection method based on bearing fault diagnosis called Feature-to-Feature and Feature-to-Category- Maximum Information Coefficient (FF-FC-MIC), which considers the relevance among features and relevance between features and fault categories by exploiting the nonlinearity capturing capability of maximum information coefficient. In this method, a weak correlation feature subset obtained by a Feature-to-Feature-Maximum Information Coefficient (FF-MIC) matrix and a strong correlation feature subset obtained by a Feature-to-Category-Maximum Information Coefficient (FC-MIC) matrix are merged into a final diagnostic feature set by an intersection operation. To evaluate the proposed FF-FC-MIC method, vibration data collected from two bearing fault experiment platforms (CWRU dataset and CUT-2 dataset) were employed. Experimental results showed that accuracy of FF-FC-MIC can achieve 97.50%, and 98.75% on the CWRU dataset at the motor speeds of 1750 rpm, and 1772 rpm, respectively, and reach 91.75%, 94.69%, and 99.07% on CUT-2 dataset at the motor speeds of 2000 rpm, 2500 rpm, 3000 rpm, respectively. A significant improvement of FF-FC-MIC has been confirmed, since the p-values between FF-FC-MIC and the other methods are 1.166×10^{-3}, 2.509×10^{-5}, and 3.576×10^{-2}, respectively. Through comparison with other methods, FF-FC-MIC not only exceeds each of the baseline feature selection method in diagnosis accuracy, but also reduces the number of features.

Keywords: feature selection; Maximum Information Coefficient (MIC); FF-MIC; FC-MIC; bearing fault diagnosis

1. Introduction

As one of the most significant parts of rotating machines, rolling bearing has a great influence on operating status of mechanical equipment. According to the statistics of mechanical faults, more than 40% of the faults are caused by rolling bearings [1,2]. Therefore, it is important and meaningful for researchers to do research on fault diagnosis of rolling bearings. In bearing fault diagnosis, an essential role to improve the diagnosis accuracy of bearings is played by feature selection, which aims at selecting a subset from the original set of features according to discrimination capability [3].

Traditional feature selection methods can be divided into three types: filters, wrappers and embedded methods [4]. Filter methods evaluate the quality of features through some feature evaluation criteria and select the top high-ranked features. Based on the advantages of high generality and low computational cost, filter-based feature selection methods are suitable for high-dimensional

datasets [5,6]. Although filter-based methods are computationally fast, they usually do not take feature relevance into considerations. Wrapper methods utilize the performance of machine learning model to judge the quality of the feature subset. They usually repeat the two steps, including searching for an optimal feature subset and evaluating the selected features, until a stopping measure is met [5,6]. The computational speed of wrapper-based methods is slow and the computational complexity is high, because of search strategies which involve sequential selection algorithms like sequential forward selection (SFS), sequential backward selection (SBS) and heuristic search algorithms like genetic algorithms (GAs), particle swarm optimization (PSO), gravitational search algorithm (GSA), and ant colony optimization (ACO). The complementary strengths of both filter and wrapper are exploited by embedded methods. Embedded methods define the feature that has the best ability to differentiate among classes in each stage and take feature selection as part of the training process. Thus, embedded methods are more effective than filter approaches since they involve interaction with the learning algorithm, and they are superior to wrapper methods as they do not need to evaluate the feature subsets iteratively [5]. However, embedded methods have the shortcomings of low computational speed and high computational complexity as well.

As mentioned above, feature selection in bearing fault diagnosis is helpful to improve the accuracy of diagnosis and reduce the dimensions of features. Fu [7] combined filter and wrapper methods to select features in bearing fault diagnosis. Ou [8,9] proposed a supervised Laplacian score (SLS)-based feature selection method. Hui [10] applied an improved wrapper-based feature selection method for bearing fault diagnosis. Liu [11] selected features through an evolutionary Monte Carlo method, which is a wrapper-based method. Islam [12] proposed a hybrid feature selection method which employed a genetic algorithm (GA)-based filter analysis to select optimal features. Luo [13] used the real-valued gravitational search algorithm (RGSA) to optimize the input weights and bias of extreme learning machine (ELM), and the binary GSA (BGSA) was used to select important features from a compound feature set. Yu [14] combined the K-means method with standard deviation to select the most sensitive characteristics. Liu [15] and Yang [16] utilized distance evaluation technique (DET) to select sensitive features. Vepa [17] presented a feature selection method involved in feature weight, monotonicity, correlation, and robustness. However, the above methods have some problems that lack considerations for relevance; for example, the computational speed is slow, and the computational complexity is high.

Some researchers recently took signal relevance into consideration. Cui [18] selected intrinsic mode functions (IMFs) as features through correlation. A correlation coefficient of two simplified neutrosophic sets (SNSs) was proposed to diagnose the bearing fault types by Shi [19]. In [20], Laplacian Score (LS) for feature selection was utilized to refine the feature vector by sorting the features according to their importance and correlations. Besides, Zhang [21] employed the Decision Tree algorithm to select the important features of the time-domain signal, and the low correlation features was selected. Jiang [22] computed the mutual information (MI) of decomposed components and the original signal, and extracted the noiseless component, in order to obtain the reconstructed signal. However, in terms of relevance, researchers did not take relevance between features and fault categories into account in feature selection, only considering the correlation between features.

The purpose of feature selection is not just simply reducing dimensions for the data. It is more about eliminating redundant and irrelevant features. Redundant features can be eliminated by the measurement of relevance among features. The stronger the relevance between two features, the stronger the redundancy and replaceability between them. Moreover, irrelevant features can be eliminated through the measurement of relevance between features and categories. The stronger the relevance between features and categories, the stronger the distinguishability of features from categories. Irrelevant and redundant features can be eliminated from a raw feature set according to relevance. Therefore, consideration of relevance plays an essential role in reducing data dimensions in feature selection.

On the other hand, some work [8–13] selected features based on wrapper methods and increased the computational complexity. To avoid the cost of searching, the idea of intersection or union operation was put forward to merge feature subsets [23,24].

When it comes to the problem of coping with redundant information, principal component analysis (PCA) has been extensively studied. PCA is an algorithm that can effectively extract subsets from data by reducing the dimensions of the data set [25]. Since the principal component includes most information of the raw features which are irrelevant and do not contain redundant information, the principal component can be used to replace raw features. Based on this, PCA can realize data reduction and reduce the complexity of data processing, preventing dimensional disaster [26]. Typical PCA algorithms like those in [27–30] are used to extract features and to reduce data dimensions.

Based on the problems mentioned above, a novel feature selection method called FF-FC-MIC based on the Feature-to-Category-Maximum Information Coefficient is proposed. The main contributions of this paper are as follows:

(1) A new frame of feature selection is proposed for bearing fault diagnosis, which aims at considering the relevance among features and the relevance between features and fault categories. In the new feature selection frame, strong relevance features are eliminated by an FF-MIC matrix based on the relevance between features and features. On the contrary, strong relevance features are selected by the FC-MIC matrix based on the relevance between features and categories. The proposed frame can eliminate not only redundant features but also irrelevant features from the feature set.

(2) In order to avoid computational complexity brought by wrapper methods, an intersection operation is applied to merge the obtained feature subsets. The intersection operation has advantages of forming a final subset with a lower dimension and saving time, instead of subset searching.

(3) This paper applied two datasets to validate the effectiveness and adaptability of proposed method. It turned out the proposed method has a good applicability and stability.

The remainder of this paper is organized as follows. Section 2 introduces the basic theory and algorithm applied in this paper. Section 3 gives the details of the proposed method. Section 4 shows the experimental results. Section 5 concludes the findings shown in this paper.

2. Related Basic Theory and Algorithm

Based on the two main problems mentioned before, relevance and computational complexity, the related theory for solving the problems are listed in this section. First, the maximum information coefficient (MIC) is used to measure the relevance between features and features, and the relevance between features and categories. Second, an intersection operation is employed so that it can avoid high computational complexity. Third, in order to reduce data dimensions for further and to shorten the training time of classification model, the PCA algorithm is utilized.

2.1. Maximum Information Coefficient

The MIC cannot only measure linear relationships and nonlinear relationships, but can extensively excavate non-functional dependencies between variables. MIC mainly works with MI and meshing methods. MI is an indicator of the degree of correlation between variables. Given a variable $A = \{a_i, i = 1, 2, \ldots, n\}$ and $B = \{b_i, i = 1, 2, \ldots, n\}$, where n is the number of samples, the MI is defined as follow:

$$\mathrm{MI}(A, B) = \sum_{a \in A} \sum_{b \in B} p(A, B) log_2 \frac{p(A, B)}{p(A)p(B)} \quad (1)$$

where $p(A, B)$ is the joint probability density of A and B, and $p(A)$ and $p(B)$ are the boundary probability densities of A and B, respectively. Suppose $D = \{(a_i, b_i), i = 1, 2, \ldots, n\}$ is a set of finite ordered pairs. It defines a division G, dividing the value range of variable A into x segments and also dividing the value range of variable B into y segments. Therefore, the G is a grid with a size of $x \times y$. Meanwhile, $\mathrm{MI}(A, B)$ in each grid is calculated. Since the same grid with a size of $x \times y$ has several

ways of dividing, the maximum value of $\mathrm{MI}(A, B)$ in different ways of dividing is chosen as the MI value of a division G. Additionally, a definition of maximum MI under a division G is as follow:

$$\mathrm{MI}^*(D, x, y) = max\mathrm{MI}(D|G) \tag{2}$$

where $D|G$ denotes data D are divided by G. Although it utilizes MI to indicate the quality of the grid, MIC is not just simply estimating the MI. A characteristic matrix is formed by maximum normalized MI values under different divisions. The characteristic matrix is defined as follow:

$$\mathrm{MI}(D)_{x,y} = \frac{\mathrm{MI}^*(D, x, y)}{logmin\{x, y\}} \tag{3}$$

Moreover, the MIC is defined as:

$$\mathrm{MIC}(D) = \max_{xy<B(n)} \left\{ \mathrm{MI}(D)_{x,y} \right\} \tag{4}$$

where $B(n)$ is the upper limit of the $x \times y$ grid. Generally, $\omega(1) < B(n) < O\left(n^{1-\varepsilon}\right)$, where $0 < \varepsilon < 1$.

2.2. Merging Feature Subsets

Merging feature subsets usually has two ways. One is the intersection approach and the other is the union approach.

Let S be the set of samples and $F = \{f_1, f_2, \ldots, f_i\}$ be the feature set after preprocessing. $FS_1 = \{f_1, f_2, \ldots, f_m\}$ is a subset of features selected with filter (M_1), where m denotes the number of features which are selected by M_1 and $m < i$. $FS_2 = \{f_1, f_2, \ldots, f_n\}$ is a subset of features selected with filter (M_2), where n denotes the number of features which are selected by M_2 and $n < i$.

The union approach is to create a feature subset FS_3, which has the number of features p ($p \geq \{m, n\}$), by merging all features in FS_1 and in FS_2:

$$FS_3 = FS_1 \cup FS_2 \tag{5}$$

The intersection approach is to create a feature subset FS_4, which has the number of features $q (q \leq \{m, n\})$, including these features that are present in both feature subsets FS_1 and FS_2.

$$FS_4 = FS_1 \cap FS_2 \tag{6}$$

Usually, the two approaches mentioned above are utilized to merge feature subsets selected by different filter-based feature selection methods. The union approach selects all features in both subsets. At the same time, it increases the number of features and does not achieve the goal of reducing data dimensions. On the contrast, the intersection approach selects only common features. It reduces the total number of features. However, it is possible to lose some features which are proficient [31].

3. Proposed Method

A new feature selection method was proposed to address two issues in feature selection: relevance and computational complexity.

The MIC cannot only measure the relevance between features and features, but also measure the relevance between features and categories. The method is divided into two parallel steps to select features from both aspects mentioned above. First, strong irrelevance features can be selected based on the MIC between features and features. Second, strong relevance features can be selected based on the MIC between features and fault categories. This frame has the advantage of a comprehensive consideration of relevance to eliminate redundant and irrelevant features from a feature set.

In term of computational complexity, an intersection operation is employed to merge feature subsets which are selected by the relevance between features and features, and by the relevance between features

and fault categories. The intersection operation has the advantage of obtaining a lower-dimension final subset with a lower computational complexity, containing as much as information.

Based on the above consideration, MIC is applied to measure the relevance among features and the relevance between features and fault categories. In addition, to avoid computational complexity caused by wrapper methods, an intersection operation is applied, instead of wrapper methods, to merge subsets.

The detailed implementation steps are as follows. First, MIC among features and MIC between features and fault categories are calculated to obtain two MIC matrixes. MIC among features is called FF-MIC and MIC between features and categories is named FC-MIC. Second, strong relevance values in FF-MIC and in FC-MIC are calculated to distinguish strong and weak relevance features. Third, strong irrelevance features selected by FF-MIC and strong relevance features selected by FC-MIC are merged through an intersection operation to form a final feature subset. Finally, PCA is applied to further reduce dimensions of the final feature set. The whole process of feature selection is shown in Figure 1.

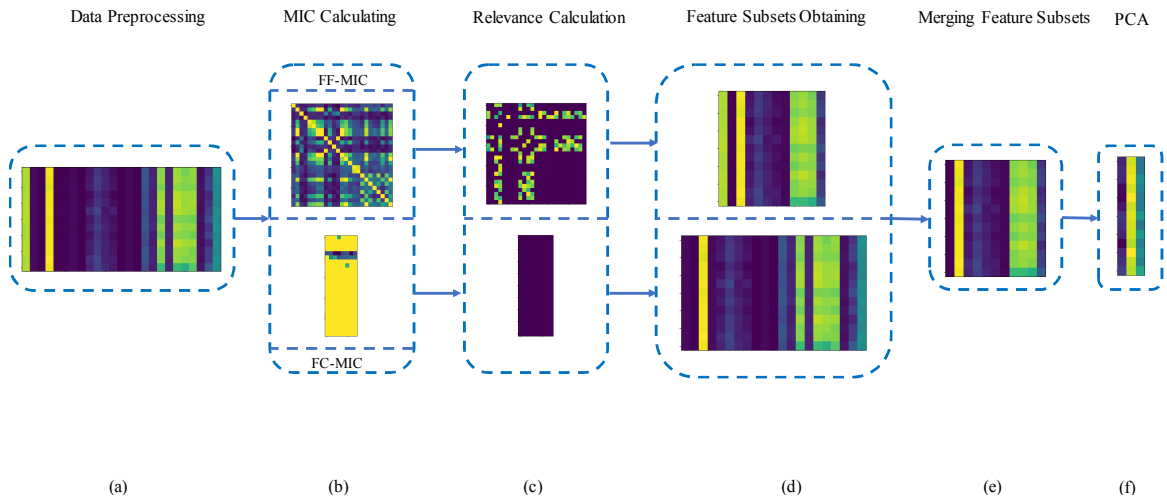


Figure 1. Flowchart of the proposed method. (**a**) Data preprocessing: Features are obtained through the time domain and frequency domain, forming a feature set (FS) matrix with a size of $n \times m$; (**b**) MIC (Maximum Information Coefficient) calculation: relevance between features and features are calculated to obtain an FF-MIC matrix with a size of $m \times m$ by MIC, and relevance between features and fault categories are calculated to obtain the FC-MIC matrix with a size of $m \times p$ by MIC; (**c**) relevance calculation: a threshold $FF\_threshold$ is set to distinguish strong irrelevance features from FS, and a threshold $FC\_threshold$ is set to distinguish strong relevance features from FS; (**d**) obtaining of feature subsets: according to the thresholds, Subset1 and Subset2 are obtained; (**e**) merging of feature subsets: An intersection operation is applied to merge Subset1 and Subset2, obtaining a final subset $F - Subset$; (**f**) PCA: PCA is employed to further reduce the dimension of $F - Subset$.

3.1. Data Preprocessing

Since the vibration signal measured by sensors carries vital information, it should be transformed by the appropriate action and sensitive features, which can efficiently reflect the bearing working condition will be obtained. To avoid missing sensitive features, almost all features in the time and frequency domains are acquired. In order to improve the convergence speed of the model and the accuracy of the model, all features will be performed by feature scaling after the features are obtained. The details are given in Algorithm 1.

Algorithm 1. Data preprocessing.

Input: samples: $S = \{s_1, s_2, \ldots s_n\}$, where n is the number of samples.
Output: the preprocessed feature set FS.

1. **for** each s_i in S
2. calculate each feature value from both the time domain and frequency domain and scale each feature value to [0,1], forming a feature set: $FS = \{f_1, f_2, \ldots f_m\}$, where m is the number of features
3. **end for**

3.2. MIC Calculation

MIC is used to define the relevance between features and features, and the relevance between features and categories. A matrix called FF-MIC is formed by the MIC between features and features, and a matrix called FC-MIC is formed by the MIC between features and fault categories. The construction details of FF-MIC and FC-MIC are described in Algorithm 2.

Algorithm 2. MIC calculation.

Input: $FS = \{f_1, f_2, \ldots f_m\}$, where $C = \{c_1, c_2, \ldots c_p\}$ are fault categories, and p is the number of fault categories.
Output: FF-MIC and FC-MIC.

1. **for** each $f_i, f_j (1 \leq i, j \leq m)$ in FS
2. calculate MIC between f_i and f_j, obtaining the FF-MIC matrix with a size of $m \times m$; the row and column denote the serial number of features.
3. **for** each f_i in FS
4. **for** each c_j in C
5. calculate MIC between f_i and c_j, obtaining the FC-MIC matrix with a size of $m \times p$; the row denotes the serial number of features and the column denotes the serial number of fault categories.
6. **end for**

3.3. Relevance Calculation

As mentioned before, the value of MIC ranges from 0 to 1. FF-MIC denotes a matrix which can measure the relevance between features and features. The closer to 0 each element in the FF-MIC matrix approaches, the stronger the irrelevance between the features corresponding to the row and column of the element is. On the contrary, the closer to 1 each element in the FC-MIC matrix approaches, the stronger the relevance between the feature and category is. On this basis, a feature can be judged whether it is a weak relevance or strong relevance through the value in the FF-MIC matrix.

In each column of the FF-MIC matrix, each minimum value is found to form a set $FFmin = \{min_1, min_2, \ldots, min_m\}$, and then the maximum value $FF\_threshold$ is found in the set $FFmin$. Therefore, the features, of which corresponding FF-MIC values are smaller than $FF\_threshold$ are, are called strong irrelevance features. The reason of setting up $FFmin$ is that the condition of the strong irrelevance between a certain feature and each one in the rest can be observed. Through the set $FFmin$, the maximum of $FFmin$ is utilized to judge whether a feature is strong irrelevant with other features.

At the same time, in each row of FC-MIC, each maximum value is found to form a set $FCmax = \{max_1, max_2, \ldots, max_m\}$, and then the minimum value $FC\_threshold$ is found in the set $FFmax$. The features, of which corresponding FC-MIC values are bigger than the $FC\_threshold$, are called strong relevance features, which are being selected soon. The reason of setting up $FCmax$ is that the condition of the strongest relevance between a certain feature and each category can be observed. Through the set $FCmax$, the minimum of $FCmax$ is utilized to judge whether a feature is strong relevant with categories.

The details of relevance calculation are shown in Algorithm 3.

Algorithm 3. Relevance calculation

Input: $FS = \{f_1, f_2, \ldots f_m\}$, FF-MIC, and FC-MIC.
Output: *FF_threshold*, and *FC_threshold*.

1. **for** each column in the FF-MIC matrix
2. **for** elements in each column
3. find each minimum value to form a set $FFmin = \{min_1, min_2, \ldots, min_m\}$
4. **end for**
5. **for** each min_i in *FFmin*
6. find the maximum *FF_threshold* in *FFmin*
7. **end for**
8. **for** each row in FC-MIC
9. **for** elements in each row
10. find each maximum value to form a set $FCmax = \{max_1, max_2, \ldots, max_m\}$
11. **end for**
12. **for** each max_i in *FCmin*
13. find the minimum *FC_threshold* in *FCmax*
14. **end for**

3.4. Obtaining Feature Subsets

After relevance calculation, two feature sets are formed called *Subset*1 and *Subset*2, respectively, according to the *FF_threshold* and *FC_threshold*. The features, of which corresponding FF-MIC values are smaller than the *FF_threshold*, will be the elements of *Subset*1, because the closer to 0 the FF-MIC value, the stronger the irrelevance between features. Additionally, the features, of which corresponding FC-MIC values are bigger than the *FC_threshold*, will become the members of *Subset*2, because the closer to 1 the FC-MIC value, the stronger the relevance between features and categories. The details of obtaining feature subsets are shown in Algorithm 4.

Algorithm 4. Obtaining feature subsets

Input: $FS = \{f_1, f_2, \ldots f_m\}$, *FF_threshold*, and *FC_threshold*.
Output: *Subset*1 and *Subset*2.

1. **for** each f_i in *FS*
2. select the features corresponding FF-MIC values smaller than the *FF_threshold* to form a *Subset*1
3. **end for**
4. **for** each f_i in *FS*
5. select the features corresponding FC-MIC values bigger than the *FF_threshold* to form a *Subset*2
6. **end for**

3.5. Merging Feature Subsets and Reducing Dimensions with PCA

In above steps, two feature subsets are obtained. *Subset*1 is a strong-irrelevance subset which contains features with strong irrelevance among them. *Subset*2 is a strong-relevance subset which consists of features which have strong correlation with fault categories. First, an intersection operation is carried between *Subset*1 and *Subset*2 to obtain a final subset $F - Subset$. $F - Subset$ contains the common elements of *Subset*1 and *Subset*2. Next, the variance of each feature is calculated and then the five features with the largest variance are selected. The reason of selecting the features with the five largest variances is that the smaller a variance of the feature, the less information this feature contains. Therefore, the five features with the largest variance means that almost all information is retained. Finally, PCA is employed to reduce dimensions of the $F - Subset$. The details are given in Algorithm 5.

Algorithm 5. Merging feature subsets.

Input: *Subset*1 and *Subset*2.
Output: $F - Subset$.

1. Merge *Subset*1 and *Subset*2 with an intersection operation to obtain $F - Subset$.
2. **for** each f_i in $F - Subset$
3. employ PCA to select five features with the largest variance
4. **end for**

4. Experimental Results and Analysis

In this paper, the proposed method was applied to two datasets. One was the vibration dataset provided by the Bearing Data Center of Case Western Reserve University (CWRU) [32], and the other was the vibration dataset which was obtained by the CUT-2 platform. Based on both datasets, the proposed method was compared with 3 traditional feature selection methods, which are filter based on variance, wrapper based on recursive feature elimination (RFE), and embedded based on gradient boosting decision tree (GBDT). Performances of these methods were evaluated by the diagnosis accuracy on two popular fault classification models, which are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

The features were calculated both in the time domain and frequency domain. All the features are shown in Table 1.

Table 1. Features in the time domain and frequency domain [14,33,34].

Features in the Time Domain		Features in the Frequency Domain	
$f_0 = \frac{\sum_{n=1}^{N} x(n)}{N}$	$f_6 = \frac{\sum_{n=1}^{N} (x(n)-f_0)^4}{(N-1)f_2^4}$	$f_{12} = \frac{\sum_{k=1}^{K} s(k)}{K}$	$f_{19} = \sqrt{\frac{\sum_{k=1}^{K} f_k^4 s(k)}{\sum_{k=1}^{K} f_k^2 s(k)}}$
$f_1 = \sqrt{\frac{\sum_{n=1}^{N} (x(n))^2}{N-1}}$	$f_7 = \frac{f_4}{f_3}$	$f_{13} = \frac{\sum_{k=1}^{K} (s(k)-f_{12})^2}{K-1}$	$f_{20} = \sqrt{\frac{\sum_{k=1}^{K} f_k^2 s(k)}{\sum_{k=1}^{K} s(k) \sum_{k=1}^{K} f_k^4 s(k)}}$
$f_2 = (\frac{\sum_{n=1}^{N} \sqrt{\lvert x(n) \rvert}}{N})^2$	$f_8 = \frac{f_4}{f_2}$	$f_{14} = \frac{\sum_{k=1}^{K} (s(k)-f_{12})^3}{K(\sqrt{f_{13}})^3}$	$f_{21} = \frac{f_{17}}{f_{16}}$
$f_3 = \sqrt{\frac{\sum_{n=1}^{N} (x(n))^2}{N}}$	$f_9 = \frac{f_3}{\frac{1}{N}\sum_{n=1}^{N} \lvert x(n) \rvert}$	$f_{15} = \frac{\sum_{k=1}^{K} (s(k)-f_{12})^4}{K(f_{13})^2}$	$f_{22} = \frac{\sum_{k=1}^{K} (f_k-f_{16})^3 s(k)}{K{f_{17}}^3}$
$f_4 = \max \lvert x(n) \rvert$	$f_{10} = \frac{f_4}{\frac{1}{N}\sum_{n=1}^{N} \lvert x(n) \rvert}$	$f_{16} = \frac{\sum_{k=1}^{K} f_k s(k)}{\sum_{k=1}^{K} s(k)}$	$f_{23} = \frac{\sum_{k=1}^{K} (f_k-f_{16})^4 s(k)}{K{f_{17}}^4}$
$f_5 = \frac{\sum_{n=1}^{N} (x(n)-f_1)^3}{(N-1)f_2^3}$	$f_{11} = \sum_{n=1}^{N} \lvert x(n) \rvert^2$	$f_{17} = \sqrt{\frac{\sum_{k=1}^{K} (f_k-f_{16})^2 s(k)}{K}}$	$f_{24} = \frac{\sum_{k=1}^{K} (f_k-f_{16})^{1/2} s(k)}{K\sqrt{f_{17}}}$
		$f_{18} = \sqrt{\frac{\sum_{k=1}^{K} f_k^2 s(k)}{\sum_{k=1}^{K} s(k)}}$	
$x(n)$ is the time-domain signal sequence, $n = 1, 2, \ldots, N$, N is the number of each sample points.		$s(k)$ is the frequency-domain signal sequence, $k = 1, 2, \ldots, K$, K is the number of spectral lines.	

4.1. Experiments on CWRU Datasets

4.1.1. Experimental Setup and Datasets

Our datasets were constructed on CWRU bearing datasets [32] to apply the proposed method. As shown in Figure 2, the experimental test rig included an electric motor (left), a torque transducer/encoder (center), a dynamometer (right), and a control circuitry (not shown). The deep groove ball bearings of the type 6205-2RS JEM SKF at the drive end (DE) was used for vibration signal collection. The vibration signals were collected by accelerometers at 4 different motor speeds of 1797 rpm, 1772 rpm, 1750 rpm and 1730 rpm, where the sampling frequencies were 12 kHz and 48 kHz. The faults were set by an electric discharge machine with diameters of 0.007 inches, 0.014 inches, 0.021 inches and 0.028 inches.

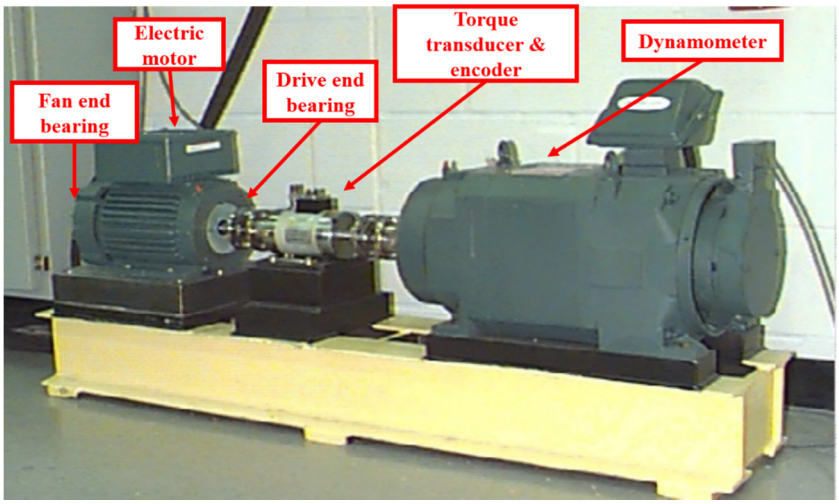


Figure 2. CWRU bearing experimental platform.

There were 4 bearing categories, including 3 fault categories and 1 normal category. One hundred fifty samples were acquired in each category and each sample contains 1024 continuous data points. Two same datasets were constructed at the motor speeds of 1750 rpm and 1772 rpm. Table 2 shows the details of the datasets. The training process and the test process were carried out at the same speed. The training set included 420 samples and the testing set included 180 samples.

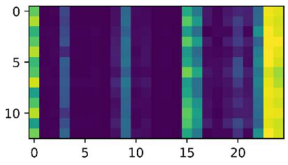
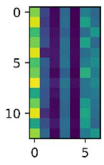
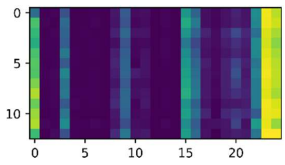
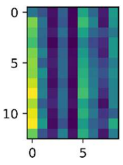
Table 2. Datasets obtained at the motor speeds of 1750 rpm and 1772 rpm on the CWRU bearing experimental platform.

Conditions of the Bearings	Fault Size (inch)	Number of Samples	Class
Normal		150	0
Inner race fault	0.007	50	1
	0.014	50	
	0.021	50	
Outer race fault	0.007	50	2
	0.014	50	
	0.021	50	
Baller fault	0.007	50	3
	0.014	50	
	0.021	50	

4.1.2. Result Analysis of CWRU Datasets

According to the proposed feature selection method, features selected by the proposed method are shown in Table 3.

Table 3. CWRU dataset features selected by the proposed method. (The pictures are the visualization results for the corresponding speed samples. The horizontal axis in the figure represents the number of features, and the vertical axis represents the first 13 samples. For the sake of beauty, the first 13 samples are chosen).

Motor Speed (rpm)	All Features	Selected Features	Selected Number
1750			0, 3, 8, 9, 11, 16, 22
1772			0, 3, 8, 9, 11, 15, 16, 20, 22

As is shown in Table 3, at the motor speed of 1750 rpm, the proposed method selected 7 features from 25 features in total, and selected 9 features at the speed of 1772 rpm.

KNN and SVM classifiers and 3 traditional feature selection methods were applied to validate the proposed method. As mentioned before, 2 datasets were constructed at 1750 rpm and 1772 rpm. Both of them were used in the training and testing processes. The results are shown in Table 4.

Table 4. Comparison of classification accuracy of SVM and KNN at motor speeds of 1750 rpm and 1772 rpm using different feature selection methods.

Motor Speed (rpm)	Feature Selection Methods	Classification Models	
		SVM	KNN
1750	Var_FS	0.8750	0.8750
	RFE_FS	0.8500	0.8333
	GBDT_FS	0.9417	0.9600
	FF_FC_MIC	0.9583	0.9750
1772	Var_FS	0.7750	0.7417
	RFE_FS	0.8000	0.7666
	GBDT_FS	0.9833	0.9917
	FF_FC_MIC	0.9917	1.0000

As shown in Table 4, the proposed method has a higher accuracy than the other 3 feature selection methods. Especially, the best classification accuracy has achieved 100% at the motor speed of 1772 rpm with KNN. Comparing with the other 3 methods, the proposed method can improve the accuracy by about an average of 1.21% at least, and by about an average of 15.25% at most. Meanwhile, the classification accuracies of the proposed method at different motor speeds with 2 classifiers reached are all above 95%. Above all, it is obvious that the proposed method has a pretty performance on the aspect of diagnosis accuracy.

In addition, in order to validate the propose method's effectiveness, the number of features selected is compared between the proposed method and the other 3 traditional methods. All the numbers of features selected by these 4 methods are shown in Figure 3.

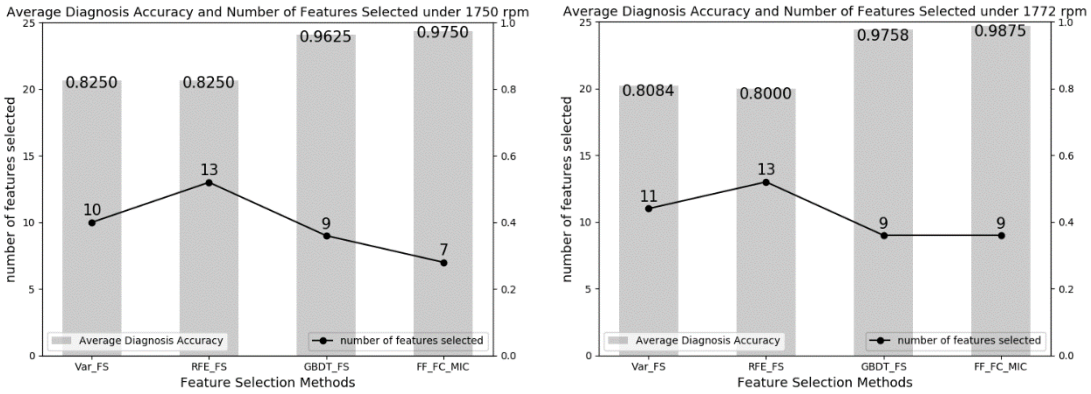


Figure 3. Average diagnosis accuracy of single fault and number of features selected at two motor speeds. (**a**) Average diagnosis accuracy and number of features selected at 1750 rpm; (**b**) average diagnosis accuracy and number of features selected at 1772 rpm.

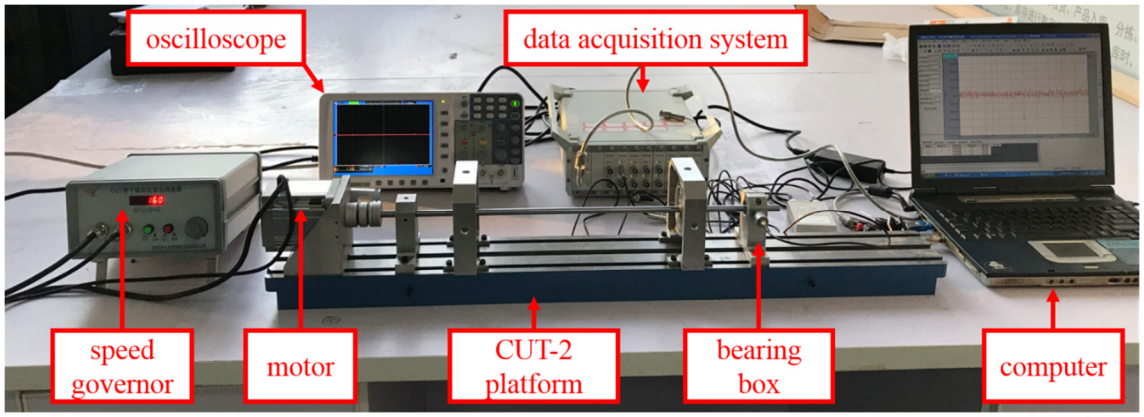
In Figure 3, the number of features selected by the proposed method under both motor speeds is the minimum, comparing to those in the other 3 methods. In other words, the proposed method achieves a higher diagnosis accuracy based on fewer features. Fewer features indicate lower data dimension and lower computation complexity for classifiers, which means the proposed method can reach a better performance with respect to feature dimension.

The reason why the proposed method can reach a higher diagnosis with less number of features can be explained as follow. First, in terms of relevance, not only the relevance among features but also relevance among features and fault categories are taken into account. It can eliminate redundant features based on relevance among features and fault categories, and eliminate irrelevant features based on relevance among features. Second, an intersection operation is used, which deletes redundant and weak relevance features in two subsets, avoiding complex computation.

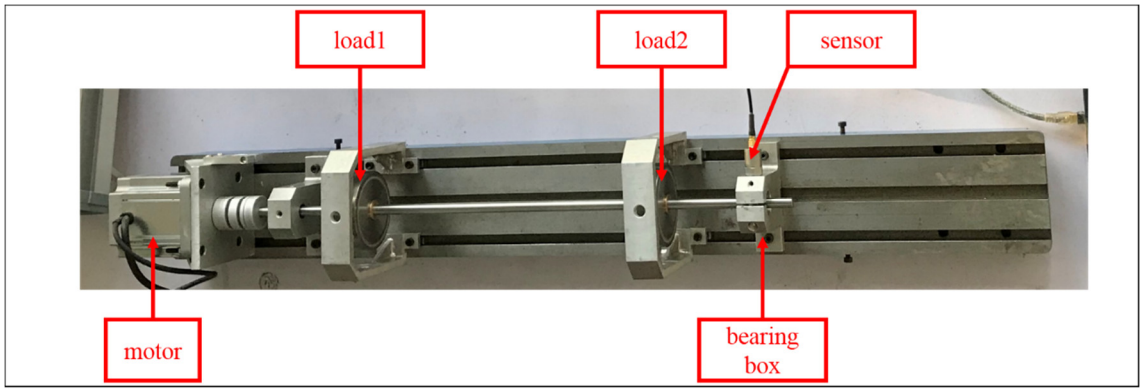
In summary, a conclusion can be drawn that the proposed feature selection method performs better in both reducing feature dimensions and achieving higher diagnosis accuracy, compared to the other 3 feature selection methods on the datasets of CWRU.

4.2. Experiments on CUT-2 Datasets

In order to validate the adaptability of the proposed feature selection method, vibration signals were collected and experiments were conducted on the CUT-2 platform. The whole experimental test rig is shown in Figure 4. The test rig was composed of an oscilloscope, a data acquisition system, a speed governor, a CUT-2 platform and a computer. All the measurements in the experimental system were carried out without load. In terms of vibration signal collection, the type of sensor was CT1010L, which is a kind of piezoelectric sensor, and the sensor was located in the horizontal direction of the bearing box. As shown in Figure 4b, the bearing tested in the experiment was placed at the far end of the motor. In addition, in order to validate the proposed method on the CUT-2 platform, an electric discharge machine was utilized to set three different kinds of faults. Figure 5 shows the bearing, of which the type was 6900ZZ in the experiments. The fault diameters were 0.2 mm and 0.3 mm. The vibration signals were collected by accelerometers at 3 different motor speeds of 2000 rpm, 2500 rpm and 3000 rpm, where the sampling frequency was 2 kHz.



(**a**)



(**b**)

Figure 4. Experimental test rig: (**a**) CUT-2 data acquisition system; (**b**) CUT-2 platform.

(**a**) (**b**) (**c**) (**d**)

Figure 5. Locations of bearing faults: (**a**) outer race fault; (**b**) ball fault; (**c**) inner race fault; and (**d**) combination of parts.

4.2.1. Experimental Setup and CUT-2 Datasets

In this section, the proposed method was applied on fault datasets from CUT-2. The details of the datasets are shown in Table 5. There were 4 kinds of bearing condition, including 1 normal condition and 3 fault conditions. The 3 fault conditions were inner race fault, outer race fault and baller fault. Two hundred samples were collected for each condition. Each sample contained 1024 continuous data points. Each dataset included 800 samples in total. Through data seeded, 3 kinds of datasets were acquired at 2000 rpm, 2500 rpm, and 3000 rpm. The training process and the testing process were carried out at the same speed. The training set included 560 samples and the testing set included 240 samples.

Table 5. Datasets obtained at motor speeds of 2000 rpm, 2500 rpm and 3000 rpm on the CUT-2 platform.

Conditions of the Bearings	Fault Size (mm)	Number of Samples	Class
Normal condition		200	0
Inner race fault	0.2 0.3	100 100	1
Outer race fault	0.2 0.3	100 100	2
Baller fault	0.2 0.3	100 100	3

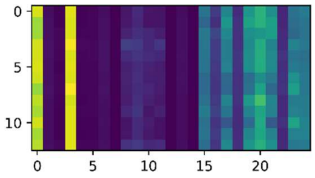
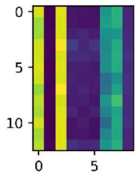
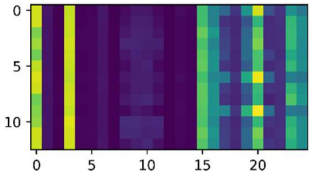
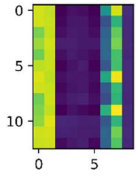
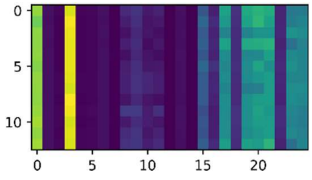
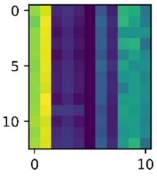
4.2.2. Result Analysis of CUT-2 Datasets

According to the proposed feature selection method, the features selected on 3 datasets are shown in Table 6. From Table 6, it can be seen that 9, 9 and 11 features were selected at 3 speeds.

Under most circumstances of 3 motor speeds from Table 7, the proposed feature selection method can reach the highest diagnosis accuracy, comparing to the other 3 methods. In addition, it performs a certain adaptability and stability on SVM and KNN.

At the speeds of 2000 rpm, 2500 rpm, and 3000 rpm, the average diagnosis accuracies are 91.75%, 94.69% and 99.07%, respectively. From this phenomenon, it was found that the diagnosis accuracy increases while the motor speed increases. There are 2 factors to cause this phenomenon. First, our bearings are small in both size and fault size, which means the vibration signal is less obvious in a lower speed than in a higher speed, leading to the phenomenon of a lower diagnosis accuracy in low speed and a higher diagnosis accuracy in high speed. Second, a large difference in motor speed, ranging from 2000 rpm to 3000 rpm, causes such a diversity of diagnosis accuracy increases while the motor speed increases.

Table 6. Dataset features selected by the proposed method. (The pictures are the visualization results for the corresponding speed samples. The horizontal axis in the figure represents the number of features, and the vertical axis represents the first 13 samples. For the sake of beauty, the first 13 samples are chosen).

Motor Speed (rpm)	All Features	Selected Features	Selected Number
2000			0, 2, 3, 8, 10, 11, 19, 20, 22
2500			0, 3, 8, 9, 10, 11, 19, 20, 22
3000			0, 3, 8, 9, 11, 14, 15, 16, 20, 21, 23

For further analysis, compared with the other 3 methods, the proposed method can improve the accuracy by about an average of 1.32% at least, and by about an average of 6.94% at most. In term of diagnosis accuracy, the proposed method has a better performance than the other 3 feature selection methods.

Table 7. Comparison of classification accuracy of SVM and KNN at motor speeds of 2000 rpm, 2500 rpm and 3000 rpm using different feature selection methods.

Motor Speed (rpm)	Feature Selection Methods	Classification Models	
		SVM	KNN
2000	Var_FS	0.8688	0.8750
	RFE_FS	0.8938	0.8875
	GBDT_FS	0.9313	0.9012
	FF_FC_MIC	0.9225	0.9125
2500	Var_FS	0.8562	0.8250
	RFE_FS	0.8688	0.8812
	GBDT_FS	0.9187	0.9175
	FF_FC_MIC	0.9625	0.9313
3000	Var_FS	0.9437	0.9250
	RFE_FS	0.9187	0.9062
	GBDT_FS	0.9750	0.9875
	FF_FC_MIC	0.9938	0.9875

In terms of number of features selected by different methods, the numbers of features selected among the 4 methods were compared. All the numbers of features selected by 4 methods are shown in Figure 6.

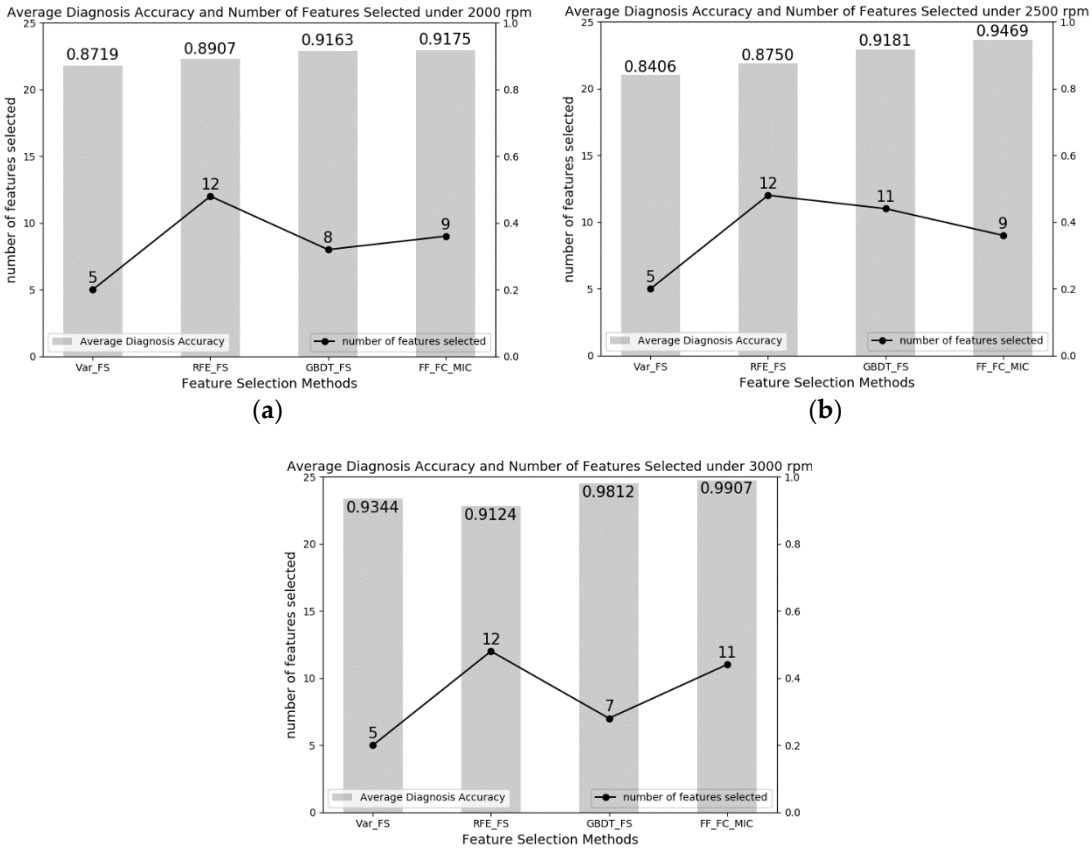


Figure 6. Average diagnosis accuracy of fault and number of features selected at three motor speeds: (**a**) average diagnosis accuracy and number of features selected at the speed of 2000 rpm; (**b**) average diagnosis accuracy and number of features selected at the speed of 2500 rpm; (**c**) average diagnosis accuracy and number of features selected at the speed of 3000 rpm.

4.3. Comparison of Significant Differences

In this section, a comparison of significant differences was made to validate the proposed method. The proposed method was compared to each of the traditional feature selection method. Because of the reason that the experimental results were obtained based on small samples, the t-test was applied to test whether the mean difference between the two methods was significant. According to the t-test, the assumption that judges whether the expectations of two compared methods were equal was made:

$$H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2.$$

If p-value < 0.05, the assumption will be accepted and it means the there is a significant difference between the two methods. If p-value > 0.05, the assumption will not be accepted and it means that there is not a significant difference between the two methods.

Table 8 shows that different p-values were calculated under the t-test. From Table 8, it is can be seen that under comparisons between each traditional method and the proposed method, the p-values of proposed method are all smaller than 0.05.

A conclusion can be drawn that the proposed method can achieve the highest diagnosis accuracy with fewer features and has a significant improvement in accuracy performance, compared with other 3 feature selection methods.

Table 8. Comparison between traditional feature selection methods and the proposed method.

Methods Compared	p-Value	Whether the Mean Difference between the Two Methods Is Significant (Y/N)
Var_FS, FF_FC_MIC	1.166×10^{-3}	Y
RFE_FS, FF_FC_MIC	2.509×10^{-5}	Y
GBDT_FS, FF_FC_MIC	3.576×10^{-2}	Y

5. Conclusions

This paper presents a bearing fault diagnosis method based on the feature selection method called FF-FC-MIC by exploiting the capability of MIC to capture nonlinear relevance. The results can be summarized as follows: First, the most intuitive indicator, diagnosis accuracy, shows that the proposed method can reach 97.50%, and 98.75% in terms of average diagnosis accuracy in the CWRU dataset, and reach 91.75%, 94.69%, and 99.07% in terms of average diagnosis accuracy in the CUT-2 dataset. All the accuracies are the highest compared with those in the other feature selection methods. Second, on the basis of relevance among features and relevance between features and categories, the proposed method can select relatively low-dimension feature subsets to approach a higher diagnosis accuracy, compared with the other feature selection methods. Third, by calculating p-values under the t-test, the proposed method has a significant performance improvement, compared with traditional feature selection methods. The reasons can be summarized as follows: First, since the proposed method utilizes the MIC to measure the nonlinear and non-functional relationships between features and features, and between features and categories, it can eliminate redundant and irrelevant features. Second, the proposed method employs the intersection operation to merge two subsets, avoiding subset searching. Third, extensive experiments on the CWRU dataset and CUT-2 dataset were conducted to validate the effectiveness and adaptability of the proposed method. It turns out that the proposed method performs better in both reducing feature dimensions and achieving higher diagnosis accuracy, compared with the other 3 feature selection methods.

The relationship between the weight of features and feature subset is as essential as the work of feature selection. Besides, feature fusion technology should also be taken into consideration for feature selection. Our next work would investigate the relationship between the weight of features and feature subset and feature fusion technology to present more effective feature selection methods to reach the higher diagnosis accuracy.

Author Contributions: Conceptualization, X.T. and J.W.; methodology, J.W. and X.T.; software, J.W.; validation, X.T. and J.W.; investigation, X.T. and J.W.; data curation, J.W.; writing of original draft preparation, X.T. and J.W.; writing of review and editing, X.T., J.W., J.L., G.L. and J.C.; visualization, J.W.; supervision, X.T., J.L., G.L., and J.C.

Funding: This research was funded by Science and Technology Major Project of Guizhou Province ([2013]6019) and Project of Guizhou High-Level Study Abroad Talents Innovation and Entrepreneurship (2018.0002), Open Fund of Guizhou Provincial Public Big Data Key Laboratory (2017BDKFJJ019), and Guizhou University Foundation for the introduction of talent ((2016) No. 13).

Acknowledgments: We gratefully acknowledge the support of Jianjun Hu with the guiding for this research.

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

References

1. Ciabattoni, L.; Ferracuti, F.; Freddi, A.; Monteriu, A. Statistical Spectral Analysis for Fault Diagnosis of Rotating Machines. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4301–4310. [CrossRef]
2. Tra, V.; Kim, J.; Khan, S.A.; Kim, J. Bearing Fault Diagnosis under Variable Speed Using Convolutional Neural Networks and the Stochastic Diagonal Levenberg-Marquardt Algorithm. *Sensors* **2017**, *17*, 2834. [CrossRef] [PubMed]
3. Rodriguez-Galiano, V.F.; Luque-Espinar, J.A.; Chica-Olmo, M.; Mendes, M.P. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Sci. Total Environ.* **2018**, *624*, 661–672. [CrossRef] [PubMed]
4. Li, Y.; Li, T.; Liu, H. Recent advances in feature selection and its applications. *Knowl. Inf. Syst.* **2017**, *53*, 551–577. [CrossRef]
5. Kashef, S.; Nezamabadi-Pour, H.; Nikpour, B. Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **2018**, *8*, e1240. [CrossRef]
6. Yao, X.; Wang, X.; Zhang, Y.; Quan, W. Summary of feature selection algorithms. *Control. Decis.* **2012**, *27*, 161–313.
7. Fu, X.; Zhang, Y.; Zhu, Y. Rolling bearing fault diagnosis approach based on case-based reasoning. *J. Xi'an Jiaotong Univ.* **2011**, *45*, 79–84.
8. Ou, L.; Yu, D. Rolling bearing fault diagnosis based on supervised laplaian score and principal component analysis. *J. Mech. Eng.* **2014**, 88–94. [CrossRef]
9. Ou, L.; Yu, D. Rolling bearing fault diagnosis based on laplaian score fuzzy C-means clustering. *China Mech. Eng.* **2014**, *25*, 1352–1357.
10. Hui, K.H.; Ooi, C.S.; Lim, M.H.; Leong, M.S.; Al-Obaidi, S.M. An improved wrapper-based feature selection method for machinery fault diagnosis. *PLOS ONE* **2017**, *12*, e189143. [CrossRef] [PubMed]
11. Liu, X.; Zheng, H.; Zhu, T. Feature selection in machine fault diagnosis based on evolutionary Monte Carlo method. *J. Vib. Shock* **2011**, *30*, 98–101.
12. Islam, R.; Khan, S.A.; Kim, J. Discriminant Feature Distribution Analysis-Based Hybrid Feature Selection for Online Bearing Fault Diagnosis in Induction Motors. *J. Sens.* **2016**, *2016*, 1–16. [CrossRef]
13. Luo, M.; Li, C.; Zhang, X.; Li, R.; An, X. Compound feature selection and parameter optimization of ELM for fault diagnosis of rolling element bearings. *ISA Trans.* **2016**, *65*, 556–566. [CrossRef] [PubMed]
14. Yu, X.; Dong, F.; Ding, E.; Wu, S.; Fan, C. Rolling Bearing Fault Diagnosis Using Modified LFDA and EMD with Sensitive Feature Selection. *IEEE Access* **2018**, *6*, 3715–3730. [CrossRef]
15. Liu, Y.; Hu, J.; Ren, L.; Yao, K.; Duan, J.; Chen, L. Study on the Bearing Fault Diagnosis based on Feature Selection and Probabilistic Neural Network. *J. Mech. Transm.* **2016**, *40*, 48–53.
16. Yang, Y.; Pan, H.; Wei, J. The rolling bearing fault diagnosis method based on the feature selection and RRVPMCD. *J. Vib. Eng.* **2014**, *27*, 629–636.
17. Atamuradov, V.; Medjaher, K.; Camci, F.; Dersin, P.; Zerhouni, N. Railway Point Machine Prognostics Based on Feature Fusion and Health State Assessment. *IEEE Trans. Instrum. Meas.* **2018**, 1–14. [CrossRef]
18. Cui, B.; Pan, H.; Wang, Z. Fault diagnosis of roller bearings base on the local wave and approximate entropy. *J. North University China (Nat. Sci. Ed.)* **2012**, *33*, 552–558.
19. Shi, L. Correlation Coefficient of Simplified Neutrosophic Sets for Bearing Fault Diagnosis. *Shock. Vib.* **2016**, *2016*, 1–11. [CrossRef]

20. Zheng, J.; Cheng, J.; Yang, Y.; Luo, S. A rolling bearing fault diagnosis method based on multi-scale fuzzy entropy and variable predictive model-based class discrimination. *Mech. Mach. Theory* **2014**, *78*, 187–200. [CrossRef]
21. Zhang, N.; Wu, L.; Yang, J.; Guan, Y. Naive Bayes Bearing Fault Diagnosis Based on Enhanced Independence of Data. *Sensors* **2018**, *18*, 463. [CrossRef] [PubMed]
22. Jiang, X.; Wu, L.; Ge, M. A Novel Faults Diagnosis Method for Rolling Element Bearings Based on EWT and Ambiguity Correlation Classifiers. *Entropy* **2017**, *19*, 231. [CrossRef]
23. Wang, Y.; Feng, L. Hybrid feature selection using component co-occurrence based feature relevance measurement. *Expert Syst. Appl.* **2018**, *102*, 83–99. [CrossRef]
24. Bharti, K.K.; Singh, P.K. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Syst. Appl.* **2015**, *42*, 3105–3114. [CrossRef]
25. Xu, J.; Huang, F.; Mu, H.; Wang, Y.; Xu, Z. Tumor feature gene selection method based on PCA and information gain. *J. Henan Univ. (Nat. Sci. Ed.)* **2018**, *46*, 104–110.
26. Zhu, L.; Wang, X.; Zhang, J. An engine fault diagnosis method based on ReliefF-PCA and SVM. *J. Beijing Univ. Chem. Technol. (Nat. Sci.)* **2018**, *45*, 55–59.
27. Xiao, Y.; Liu, C. Improved PCA method for SAR target recognition based on sparse solution. *J. Univ. Chin. Acad. Sci.* **2018**, *35*, 84–88.
28. Du, Z.; Xiang, C. A Wavelet Packet Decomposition and Principal Component Analysis Approach. *Control Eng. China* **2016**, *23*, 812–815.
29. Fadda, M.L.; Moussaoui, A. Hybrid SOM–PCA method for modeling bearing faults detection and diagnosis. *J. Braz. Soc. Mech. Sci. Eng.* **2018**, 40. [CrossRef]
30. Wang, T.; Xu, H.; Han, J.; Elbouchikhi, E.; Benbouzid, M.E.H. Cascaded H-Bridge Multilevel Inverter System Fault Diagnosis Using a PCA and Multiclass Relevance Vector Machine Approach. *IEEE Trans. Power Electron.* **2015**, *30*, 7006–7018. [CrossRef]
31. Sun, G.; Song, Z.; Liu, J.; Zhu, S.; He, Y. Feature Selection Method Based on Maximum Information Coefficient and Approximate Markov Blanket. *Acta Autom. Sin.* **2017**, *43*, 795–805.
32. The Case Western Reserve University Bearing Data Center Bearing Data Center Seeded Fault Test Data[EB/OL]. Available online: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file (accessed on 10 June 2016).
33. Wei, Z.; Wang, Y.; He, S.; Bao, J. A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection. *Knowl.-Based Syst.* **2017**, *116*, 1–12. [CrossRef]
34. Lei, Y.; He, Z.; Zi, Y. Fault Diagnosis Based on Novel Hybrid Intelligent Model. *Chin. J. Mech. Eng.* **2008**, *44*, 112–117. [CrossRef]