Maximal information coefficient-based two-stage feature selection method for railway condition monitoring

Wen, Tao; Dong, Deyi; Chen, Qianyu; Chen, Lei; Roberts, Clive

DOI:
10.1109/TITS.2018.2881284

Maximal Information Coefficient Based Two-Stage Feature Selection Method For Railway Condition Monitoring

Tao Wen, Deyi Dong, Qianyu Chen, Lei Chen, and Clive Roberts

***Abstract*—In railway condition monitoring, feature classification is a very critical step, the extracted features are used to classify the types and levels of the faults. To achieve a better accuracy and efficiency in the classification, the extracted features must be properly selected. In this paper, Maximal Information Coefficient (MIC) is employed in two different stages to establish a new feature selection method. By using this proposed two-stage feature selection method, the strong features with low redundancy are reserved as the optimal feature subset, which makes the classification process have a moderate computational cost and also maintain a good overall performance. To evaluate this proposed two-stage selection method and prove its advantages over others, a case study focusing on the rolling bearing is carried out. The result shows that the proposed selection method can achieve a satisfied overall classification performance with a low computational cost.**

***Index Terms*—Railway condition monitoring, Maximal information coefficient, Feature selection, Bearing fault**

I. Introduction

WITH the rapid development of rail transit, increased demand in traffic capacity is obvious, which is followed by more accurate and efficient condition monitoring techniques including railway assets fault detection and diagnosis. The state-of-the-art model-free diagnosis approach is to feed the extracted features into classifiers and demonstrates the corresponding faults after evaluations and tests. In terms of feature extraction and selection techniques applied to machine status monitoring in railway engineering field, there exists some research regarding fault detection and diagnosis. [1] investigates track-circuits diagnosis by using neuro-fuzzy method, the features of which were captured through the trained networks. [2] extracts rail defects parameters from time-domain and time-frequency domain features, the optimized feature parameters are then applied to classify individual rail defects by supporting vector machine method. In railway fault diagnosis, classification is a very critical step, because it can not only enable the major faults attract enough attention, but also can make the railway service not be stopped just by the minor faults. It cannot be denied that heterogeneous feature extraction methods result in distinct features containing discriminative evidence. However, in feature classification, if there is no correlation, weak correlation or redundant correlation in the selected features, the following consequences can be resulted in: 1) When the greater number of features selected, analyse these features will be time-demanding; 2) too many features selected can cause ”dimension disaster”, as a result, the corresponding model could be very complex and not universal [3]. Therefore, feature selection process is needed to filter out the most useful information from the captured features, which is aimed to reduce the computation complexity by removing the irrelevant or redundant features and therefore enhance the diagnosis accuracy [4].

There are many existing feature selection methods, for example, in [5], Wang put forward a feature selection method based on feature clustering (FSFC) for unsupervised feature selection; in [6], Meng proposed a new selection algorithm e-GA-MTL based on the gene data; in [7], Ding proposed minimum Redundancy Feature Selection (mRMR) on microarray gene; in [8], Ge proposed the a two-step feature selection method based on maximal information coefficient (McTwo), the maximal information coefficient (MIC) is used for the first step feature selection, and the k-nearest neighbors algorithm (kNN) is utilized at the second step on the gene field. As a very important part of the mechanical components, like the wheel, bogie and pantograph in railway systems, feature selection techniques have been widely applied to bearings. [9] forms a pattern space by using selected statistical parameters, followed by nonlinear transformation and linear discriminant functions, the severity and location of bearing defects are proved to be accurately determined. [10] provides a generic methodology for machine diagnosis by introducing feature extraction and selection. The classification accuracy largely depends on the selection of right frequency band in terms of spectrum comparison, this proposed method enjoys the advantage of dealing with rather complicated rolling bearings signals with various locations, shaft speeds and bearing housing structures. [11] focuses on the feature selections of bearing vibration acceleration time signal, by taking time-domain vibration peaks and bandpass filter cut-off frequencies into considerations, the corresponding results are promising as a result of zero classification error. Finally, [12] provides a comprehensive comparison of different features from time, frequency and time-frequency domains for detecting rolling bearings defects, the measurement quality is defined as the mutual information between the bearing defects and feature

This work was supported in part by the National Natural Science Foundation of China under Grant (61806064), in part with the National Key Research and Development Program under Grant 2016YFE0200900.

Tao Wen, Qianyu Chen, Lei Chen and Clive Roberts are with the Birmingham Centre for Railway Research and Education at University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom.

Deyi Dong is with the School of Automation, Hangzhou Dianzi University, Hangzhou, 310018, China.

The corresponding author: Tao Wen(t.wen.uk@outlook.com)

parameters, the performance of the selection method is assessed using neural network with a rather high classification accuracy.

However, there are some weaknesses in the existing feature selection methods. Although the machine learning based and evolutionary based methods can improve the accuracy and performance of the classification in fault diagnosis, they are very dependent on a large number of data samples with a high complexity, an accurate but inefficient result searching could be caused, such as Xgboost [13] and e-GA-MTL. For the general correlation methods, they have a rapid classification capability, which can help to identify the function of the relationship between the signals in a very quick way, but, the downside is that the result accuracy could be compromised, only strong features can be selected, moreover, due to the presence of redundant features, a reduced overall efficiency of the classification process could be caused. Therefore, in this paper, we have a strong motivation to solve the aforementioned problems by proposing a two-stage feature selection method. In this proposed method, the linear and nonlinear relationships between the signals can be identified, the strong correlation features can be extracted without involving the redundant features, which improves both the accuracy and computational efficiency of the classification process.

The following sections are arranged: Section II gives the background knowledge about the existing feature extraction, selection and classification methods; in Section III, the methodology of the proposed MIC-based two-stage feature selection method is presented, the detailed algorithm and flowchart are illustrated; Section IV provides the performance evaluation of the feature selection result generated by the proposed approach; finally, the conclusion is drawn in Section V.

II. Background Knowledge

Three main steps are involved in fault diagnosis including feature extraction, feature selection and classification. In this section, some of the existing methods of the three steps are briefly introduced.

A. Feature Extraction

There are different types of methods used to obtain the time-domain, frequency-domain and time-frequency analysis features from the raw data captured samples respectively [14]. All these methods can be grouped into either the stationary signal analysis methods or the non-stationary signal analysis methods [15]. Time-domain method is a statistical analysis of the time-domain signal, which are known as the time-domain features, such as Mean, Peak or Kurtosis in the field of rolling bearing [16], this type of features are useful in fault detecting, but cannot separate the fault types. Frequency-domain signal is the Fourier transform of the time-domain signal, which is used to describe the global characteristics of the signal, such as gravity frequency. Time-frequency features deal with the local characteristics of the signal, such as wavelet transform. Compared to the standard Fourier transform, wavelets are well localized in both time and frequency [17] [18]. Making use of the advantages of wavelet analysis to decompose signal into multiple layers, then the signal is reconstructed and its energy can be calculated. In the following part, the feature extraction methods are presented in detail.

1) Time and Frequency-domain Features Extraction: The formulas of the time-domain and frequency-domain features are shown in Table I.

TABLE I: Feature Extraction Formulas

Feature	Abbre.	Formula
Mean	$\bar{x}$	$\frac{1}{2}\sum_{i=1}^{n}\lvert x_i\rvert$
Root Mean-square	x_{rms}	$\sqrt{\frac{\sum_{i=1}^{n}{x_i}^2}{n}}$
Peak	x_p	$\frac{\max\lvert x_i\rvert}{x_{rms}}$
Standard Dev.	x_{std}	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
Kurtosis	x_{kur}	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{(n-1){x_{std}}^4}$
Gravity Freq.	f_g	$\frac{\sum_{f=f_1}^{f_2}[P(f)\times f]}{\sum_{f=f_1}^{f_2}P(f)}$
Mean-square Freq.	f_H	$\frac{\sum_{f=f_1}^{f_2}[P(f)\times f^2]}{\sum_{f=f_1}^{f_2}P(f)}$

[1] x_i is the amplitude signal in time-domain;
[2] n is the number of sampling points in the time-domain;
[3] f is the frequency;
[4] $P(f)$ is the amplitude after the Fourier transform

2) Wavelet and Wavelet Packet Analysis: Wavelet analysis is to decompose a signal into two parts, namely the low-frequency part and the high-frequency part, layer by layer, only the low-frequency parts (A1-A3) of the upper layer are decomposed, the high-frequency parts (D1-D3) are kept. During the decomposition, the information contained by the low-frequency parts could be captured by the high-frequency parts. The energy of the high-frequency parts will be the eigenvalues of the feature. A typical 3 layer wavelet analysis is illustrated in Fig. 1.

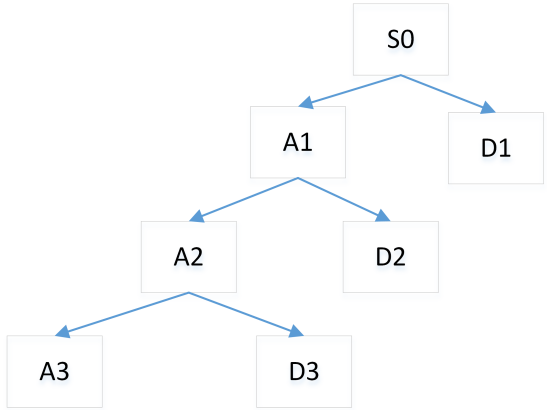


Fig. 1: The wavelet analysis process

The wavelet packet analysis is a more complex, which is widely used in signal processing. Different with the wavelet analysis, both the high-frequency part and the low-frequency part are decomposed. A typical wavelet packet analysis process is shown in Fig. (2). In this wavelet packet analysis, three layer signals (AAA3, AAD3, ADA3, DDD3, AAD3, DAD3, ADD3, and DDD3) are used to obtain 8 frequency bands ($S_3^0, S_3^1, S_3^2, S_3^3, S_3^4, S_3^5, S_3^6, S_3^7$), the energy of each frequency band will be the feature eigenvalues.

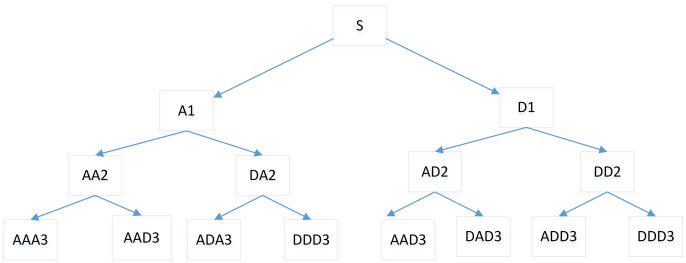


Fig. 2: The wavelet packet analysis process

B. Classification and Feature Selection Methods

In this section, some existing popular feature methods based on different criteria will be briefly reviewed.

1) Classification Methods: Support Vector Machine (SVM) is a statistical model supervised by the associated statistical learning theory and the structural risk minimization principle, which is used to analyze data used for classification and regression analysis. In [19][20][21], SVM has shown a good performance in mechanical fault classification and diagnosis.

As a widely used non-parametric method, the k-Nearest Neighbors (kNN) algorithm was firstly proposed in 1968 by Cover and Hart [22][23][24]. kNN is one of the simplest machine learning methods, which has been very matured in classification and regression. By calculating the distance of the measured test samples and the known samples to determine the class of test samples. The quality of algorithm depends on the selection of k value, improper setting of the k value can result in a high-demand in the amount of the samples and a high-complexity processing.

In this paper, to evaluate the proposed two-stage feature selection method, these two aforementioned classification methods, SVM and kNN, will be employed to test the performance of the feature selection result.

2) Feature Selection Methods: Pearson correlation analysis is a measure of the linear correlation between two features, which extracts the pole-strong features from the primary feature. However, Pearson correlation analysis is only useful in linear functions. Moreover, for the Pearson method, it is difficult to filter the redundant features.

To avoid the redundance, mRMR (Min-Redundancy and Max-Relevance) is applied in feature selection, which is based on MI (mutual information)[25]. There are two steps in implementing mRMR: get max-relevance and get min-redundancy respectively. Firstly, select the features with the strong correlation based on the maximal relevance criterion [26] [27] [28]. When two features are highly dependent on each other, there will be not big difference if any of them is removed, therefore, the min-redundancy criterion is added to select the exclusive features. However, there are some defects in this MI-based correlation analysis. For example, it is difficult to recognize the characteristics of continuous data or the data with small dispersion, some prior process need to be carried out, such as data discretion, relevance and redundancy measurement among variables, which could result in incorrect selection result. To process continuous data and identify the nonlinear function, a two-step feature selection algorithm based on maximal information coefficient (MIC) [20] has been proposed, which is known as McTwo [8]. In the first step, McTwo measures all the features for the MIC associations with the class labels, and only those with strong correlations will be reserved for further screening; in the next step, kNN is employed to further reduced the number of features. By using McTwo, a small number features with a good classification performance are selected, which is very suitable for processing the high-dimensional biomedical datasets.

In this paper, there three aforementioned feature selection methods, Pearson, mRMR and McTwo, will be used as the control groups of the proposed two-sage feature selection method to compare the classification performance.

III. Methodology

In this section, a new maximal information coefficient (MIC) based two-stage feature selection method is introduced. The relevant parameters are listed in Table II.

TABLE II: List of The Parameters

Parameter	Definition
D	The dataset contains two-variable samples
C	The class labels
σ_1	The first threshold of MIC to select strong relation feature
σ_2	The second threshold of MIC to eliminate redundancy
f_i	The $i-th$ feature in the first-stage selection
m_j	The $j-th$ feature in the second-stage selection
x_i	The vector about the samples with feature i
y_j	The vector about the samples with feature j
$\mathrm{MIC}(f_i, f_j)$	Calculate the MIC of the coupled feature i and j
S	The set of features after first selection
F	The optimal subset of the features after second selection

A. Maximal Information Coefficient

MIC tests the dependence between two variables. In this paper, to achieve a better feature selection result, a new two-stage feature selection method based on two-time MIC calculations is proposed. MIC was introduced by Reshef *et al.*

in 2011, which was designed to discover and classify the data with complicated associations, no matter they have a linear or other functional relationships [20]. The measurement value of the MIC is symmetric and normalized into a range [0, 1]. A higher MIC value indicates a more dependent relationship between the investigated variables, on the contrary, a lower MIC value means a less dependent relationship [8]. MIC can handle both numeric and category data, which makes it can be adapted to a wide range applications.

The calculation of MIC is on the basis of MI. To solve the MI, the variable x, y is divided into k-by-l grids in a X-Y coordinates. In practice, Y axis is divided into a number of equal parts, the X axis is divided dynamically. Therefore, the formulation of the MI is [25],

$$I(x,y) = H(x) - H(x,y) = \sum_{i=1}^{n_k} p(x_i)\log_2 \frac{1}{p(x_i)} + \sum_{j=1}^{n_l} p(y_j)\log_2 \frac{1}{p(y_j)} - \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} p(x_i,y_j)\log_2 \frac{1}{p(x_i,y_j)} = \sum_{x_i\ y_j} p(x_i,y_j)\log \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \quad (1)$$

where $p(x_i, y_j)$ is the joint probability density, $p(x_i)$ and $p(y_j)$ are the marginal density. However, as the MI value is not normalized, which makes the MI values are difficult to compare.

With the motivation to overcome the problem of MI, MIC was introduced. To calculate the MIC value of a two-variable dataset D= $(x_1, y_1), ..., (x_n, y_n)$, the maximal achievable MI in the k-by-l grid should be computed in advance, the integers (k,l) could be any pair. The calculation of the MIC(D) is shown in eq.(2) and eq.(3):

$$\mathrm{MIC(D)} = \frac{\max I(\mathrm{D}, k, l)}{\log \min(n_k, n_l)} \quad (2)$$

$$I(\mathrm{D}, k, l) = \max I(\mathrm{D}\,|G) \quad (3)$$

where G presents the different partitions of D, $k \times l \leq \mathrm{B}(n)$, B is a function of the sample size n expressed as $\mathrm{B}(n) = n^{0.6}$. As a result, the MI values can be normalized to [0-1], which enables a fair comparison between the grids are with different dimensions. When k and l change, the MIC will be the largest normalized MI value. By calculating the MIC value achieved by any grid in the X-Y coordinates, the characteristic matrix $M = (m_{x,y})$ is drawn, the maximal element of M is defined as the statistic MIC, the calculation process of the characteristic matrix M is illustrated in **Algorithm 1**

By introducing the $(\mathrm{D}\,|G)$, continuous data is able to be well processed, which is difficult for MI.

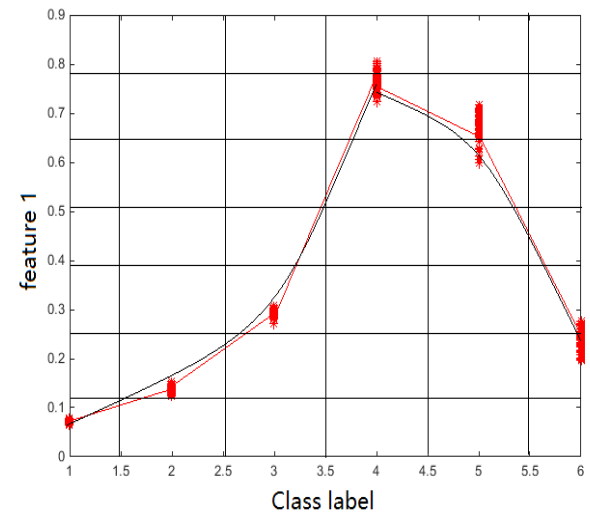
Algorithm 1: MIC Calculation

Require: $D = (x_1, y_1), ..., (x_n, y_n)$ is a set of pair about feature and category; σ_1 is the threshold, which is less than 0.9; $B \geq 3$;
for $(x, y) \leq B$ **do**
  Calculate P_x, P_y, $P_{x,y}$;
  $I_{x,y} \leftarrow \max I(D, x, y))$;
  $M_{x,y} \leftarrow I_{x,y} / \min(\log x, \log y)$;
  end
Return $M_{x,y}$;

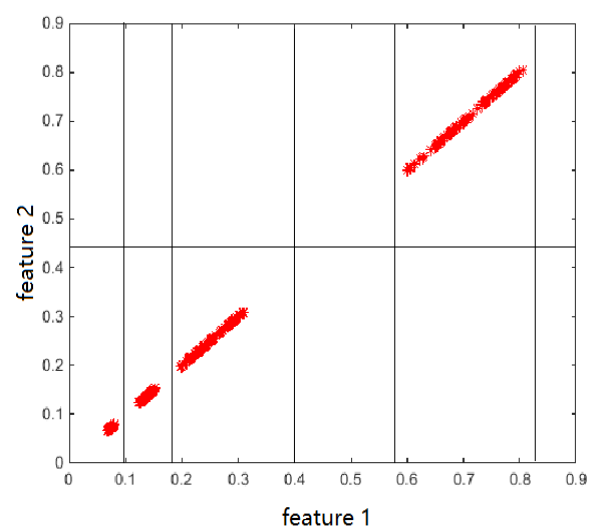
B. Two-stage Feature Selection Method

The proposed MIC-based feature selection method is implemented in two stages: 1) Select the features that have strong correlations with the class labels; 2) select the features that have low redundancy with each other [29].

In the first stage, to extract the strong features, the correlations between the class labels of every feature and the samples



(a)



(b)

Fig. 3: (a) The process of strong feature selection; (b) The process of low-redundancy feature selection

need to be investigated. Using the class labels as a horizontal coordinate and the corresponding eigenvalues of the samples as the ordinate, an indicative example result is drawn in Fig. 3. For the same type of fault, if the feature is strong, the corresponding eigenvalues will fluctuate just within a certain small range; for the different types of fault, the corresponding eigenvalues of the samples will be fluctuating in separate ranges. More formally, if the samples satisfy the condition of eq.(4), the feature i will be assumed as a strong feature.

$$\mathrm{MIC_i}(f_i, C) \leq \sigma_1 \tag{4}$$

where x_i represents the samples of the feature i, C is the class labels containing the information of the different types of fault, σ_1 is the threshold, which is expressed as:

$$\sigma_1 = \max(\frac{1}{N}\sum_{i=1}^{N}\mathrm{MIC}_i, \mathrm{MIC}_i(i = \frac{2}{3}N)) \tag{5}$$

If the feature is not strong feature, it will be removed the from the primary feature set, the left features will form the strong features set S following a descending order.

In the second stage, it is assumed that there are k features left in the set S, namely $m_1, \ldots, m_i, \ldots, m_j, \ldots, m_k$ with a descending order, where $1 \leq i \leq j \leq k$. The MIC values between any two features in S are required to be measured, one of these two features is set as the abscissa, and the other as the ordinate. Then, a grid on the scatterplot can be drawn, as Fig. 3(b) shows. The ratio between the number of the scattered points fall into the grid and the total number of the samples is taken as the probability,

$$p(x, y) = n_{x,y}/n \tag{6}$$

where $n_{x,y}$ is the number of scatters in the grid, n is the total number of the scatters. By taking the probability $p(x, y)$ into eq. (1), the MI can be obtained, and therefore, the corresponding MIC of the two features can be deduced by using eq. (2) and (3). After measuring the MIC values between every pair of the features in S, a similarity matrix is established:

$$H = \begin{bmatrix} 1(m_{11}) & \ldots & m_{1j} & \ldots & m_{1k} \\ \vdots & \ddots & \vdots & & \vdots \\ m_{j1} & & 1(m_{jj}) & & m_{ik} \\ \vdots & & \vdots & \ddots & \vdots \\ m_{k1} & \ldots & m_{kj} & \ldots & 1(m_{kk}) \end{bmatrix} \tag{7}$$

The similarity matrix is required to be sorted as the same order as the features in set S. However, if m_i and m_j are equal, the following calculation will be carried out,

$$M_j = \sum_{p=1}^{k} m_{pj}, \quad M_i = \sum_{p=1}^{k} m_{pi} \tag{8}$$

where M_j and M_i represents the contribution of the feature j and i to the whole system respectively, the feature with a larger M will has a higher rank in the similarity matrix. Due to the symmetry of the MIC measurement, only the upper triangular of the similarity matrix need to be considered,

$$H' = \begin{bmatrix} 1(m_{11}) & \ldots & m_{1j} & \ldots & m_{1k} \\ & \ddots & \vdots & & \vdots \\ & & 1(m_{jj}) & & m_{ik} \\ & & & \ddots & \vdots \\ & & & & 1(m_{kk}) \end{bmatrix} \tag{9}$$

To eliminate the redundant features, a self-customised threshold σ_2 is applied. For the off-diagonal elements, if $m_{ij} \geq \sigma_2, (i \leq j)$, remove the feature j (set the $m_{ij} = 0$), then the updated matrix H'' will be,

$$H'' = \begin{bmatrix} 1(m_{11}) & \ldots & 0 & \ldots & m_{1k} \\ & \ddots & \vdots & & \vdots \\ & & 0 & & m_{ik} \\ & & & \ddots & \vdots \\ & & & & 1(m_{kk}) \end{bmatrix} \tag{10}$$

From eq. (10) we can see that the corresponding diagonal value becomes 0 and all retained features have a diagonal value of 1. Finally, in eq. (10) all diagonal features with the value of 1 are selected as the optimal subset of the features, which is marked as F. The algorithm of the MIC-based two-stage feature selection method is described in **Algorithm 2**. The flowchart of this algorithm is illustrated in Fig.5.

Algorithm 2: Two-Stage Feature Selection

```
get 18 features from sample data as primary set
for j ← 1 to 18 do
    m_j ← MIC(X, C)
    if m_j ≥ σ_1 then
        add j to S;
    end
sort the value of m_j in descending way
for i ← 1 to k do
    for j ← i to k do
        m_ij ← MIC(X_i, X_j);
        end
    end
for i ← 1 to k do
    for j ← i to k do
        if m_ij ≥ σ_2 then
            set m_ij = 0, i = 1, 2, ..., j and
              m_ji = 0, i = j, j + 1, ...,k
        end
    end
F=j, if m_jj = 1
```

IV. Numeric Evaluation

In this section, to evaluate the proposed two-stage MIC-based feature selection method, two widely used classifiers, kNN and SVM, are employed to test the classification performance. Both the kNN and SVM are implemented by using the built-in knnclassify and svmtrain functions in MATLAB R2014b with the default setting. All the raw data samples are

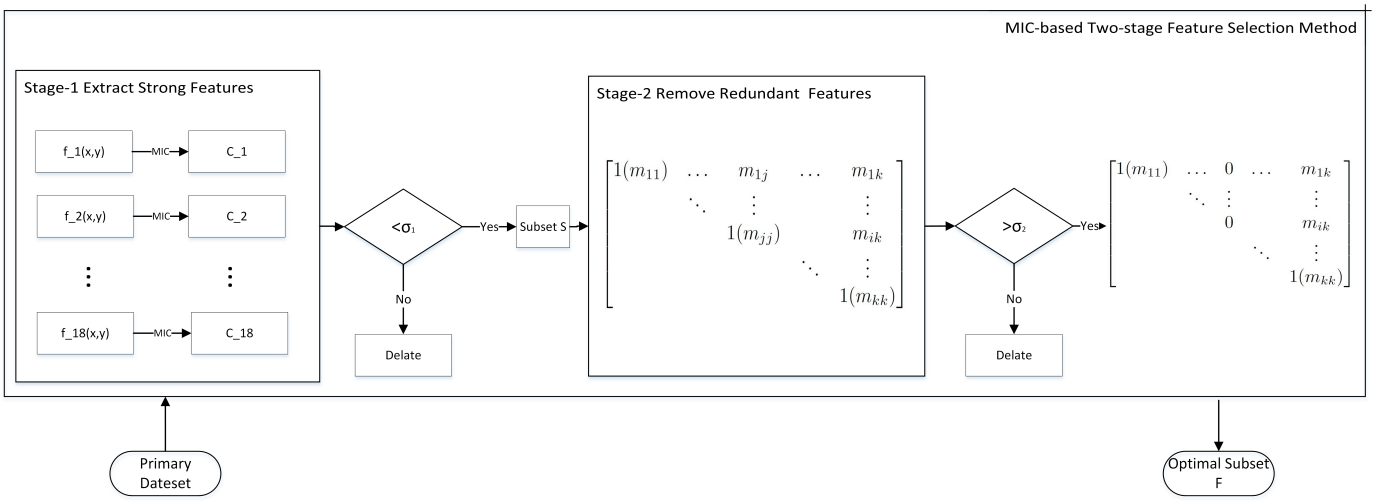


Fig. 4: The flowchart of the MIC-based two-stage feature selection method

captured from a rolling bearing, which is provided by Case Western Reserve University Bearing Data Center in USA.

A. Classification Performance Evaluation of the Two-stage Feature Selection Method

The data collection specification used in this evaluation is shown in Table III. There are 840 samples that form the primary dataset, half of samples are used as the training set, and the others form the testing set. For the samples, 18 features: Wavelet Packet Decomposition Energy (8 frequency bands), Wavelet Decomposition Energy (3 layers), Mean, Standard Deviation, RMS, Peak, Kurtosis, Gravity Frequency, Sample Entropy, are extracted, which are labeled as f_1-f_{18} respectively.

TABLE III: Data Collection Specification

Bearing states	Category	Samples
Normal	1	240
Ball fault	2	120
Inner Race fault	3	120
Outer Race fault-3 oclock	4	120
Outer Race fault-6 oclock	5	120
Outer Race fault-12 oclock	6	120

At the first-stage, selection process is implemented to find the strong features. According to the steps described in **Algorithm 2**, the MIC values between the samples in the primary dataset and the class labels of the features are calculated, the result is shown in Table IV. In this selection, in order to eliminate 1/3 of the features in the primary dataset, the threshold σ_1 is set as 0.9378. Therefore, 12 features are reserved in the subset S, 6 features are eliminated. The selection result is shown in Table V

To further remove the redundant features from the subset S, in the second-stage selection, a similarly matrix is established. In this stage, the MIC value between any two features in the subset S is calculated, and by using the self-customised threshold σ_2, the redundancy in the features can be eliminated, the left features form the set F, which is the optimal subset of the 18 features. The selection result shows that only three features are in the optimal feature subset, which are f_1, f_{13} and f_{17} respectively.

TABLE IV: The MIC Value For Each Feature

Class Labels	MIC value	Class Labels	MIC value
1	0.9989	10	0.8631
2	0.9989	11	0.9985
3	0.7922	12	0.7054
4	0.9378	13	0.9984
5	0.8573	14	0.9984
6	0.8783	15	0.8376
7	0.9378	15	0.9990
8	0.9632	17	0.9990
9	0.9378	18	0.9990

TABLE V: Strong Feature Selection Result

	Amount	Feature
Strong	12	f_1,f_2,f_4,f_7,f_8,f_9,f_{11},f_{13},f_{14},f_{16},f_{17},f_{18}
Weak	6	f_3,f_5,f_6,f_{10},f_{12},f_{15}

To verify the accuracy of the strong feature selection process in the first-stage. The correlation relations between the class labels, three selected features, f_1, f_{13}, f_{17}, and one of eliminated weak features, f_3, are plotted in Fig. 5, where shows that $f1, f13$ and f_{17} have smaller fluctuations with the class label (in red color), therefore, these three features are classified as the strong features; while features f_3 has a much bigger fluctuation with the class label (in green color), which makes it concluded as a weak feature.

To test the classification performance with using the features

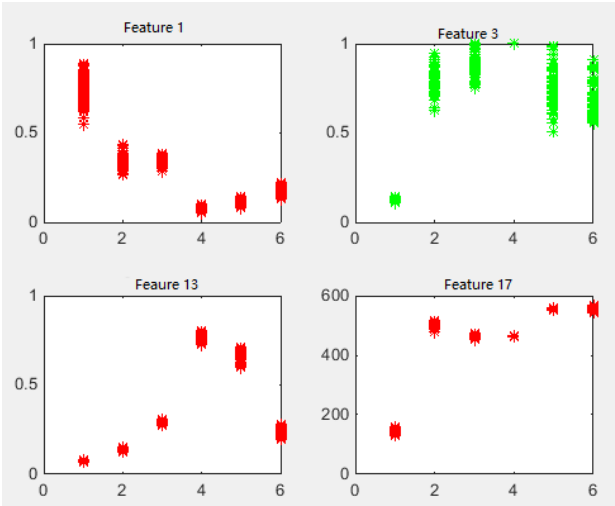


Fig. 5: Strong correlation features f_1, f_{13}, f_{17} are in red and weak correlation feature f_3 is in green.

selected by the proposed two-stage feature selection method, kNN and SVM are employed to do the classification. The classification accuracy result is illustrated in Table VII, it can be concluded that the feature combination of f_1, f_{13}, f_{17} or f_1, f_{14}, f_{17} are optimal with a lower feature redundancy and the highest classification accuracy. Moreover, by comparing the second column and other columns, we can know that feature f_{13} or f_{14} is essential for the classification; compare the first column, third column and forth column, we know that features f_{13} or f_{14} is redundant features. As a result, we can choose f_1,f_{13},f_{17} or f_1,f_{14},f_{17} as the optimal feature subset. Therefore, the proposed feature selection method reduces the 18 features to 3.

TABLE VI: Classification Accuracy Performance For Different Feature Selection

Features	1, 13, 17	1, 17	1, 14, 17	1,13,14,17
kNN	96.15%	90.5%	96.15%	96.15%
SVM	100%	68.3%	100%	100%

B. Comparison with Other Feature Selection Methods

To prove that the proposed two-stage feature selection method has advantages over other selection methods, by using kNN as the classifier, the classification performances of the features selected by the proposed feature selection method and other three feature selection methods, namely mRMR, Pearson and McTwo, are evaluated and compared.

In binary classification problems, we can define that there are two samples sets, namely the Positive (P) set and Negative (N) set respectively. The number of P and N are set as $P = n$ and $N = m$ respectively. The total number of samples is $s = n + m$. Each sample contains p features. A binary classifier assigns each sample to a feature by either P or N. Precision (R), Recall (P), Harmonic (F_1) and accuracy (Acc) are widely used to measure how well a binary classification model performs [30], which define that TP and FN are the total number of the positive samples that are predicted by the model as positive and negative respectively, TN and FP are the total number of the negative samples that are predicted by the model as negative and positive respectively. Therefore, there is

$$\begin{cases} R = \frac{TP}{TP+FN} \\ P = \frac{TP}{TP+FP} \\ F = \frac{2\times P\times R}{P+R} \\ Acc = \frac{TP+TN}{TP+FN+TN+FP} = \frac{TP+TN}{P+N} \end{cases} \tag{11}$$

where precision (R) is the ratio of positive samples that are correctly predicted, and Recall (P) is the ratio of positive samples are correctly predicated, harmonic (F) is the indicator of the balance between R and P, the models overall accuracy is defined as Acc [31].

In the evaluation, 580 samples are used as the training set and 260 samples are used as the testing set. To assess the complexity of each feature selection method, EI is introduced,

$$\text{EI} = \frac{(Acc - p)}{100} \tag{12}$$

The Acc and EI of each feature selection method are shown in Fig. 6. From the figure we can see that MIC, McTwo and Pearson all show a good performance in terms of classification accuracy; however, in terms of EI, the proposed two-stage MIC-based feature selection method shows the best performance. As a result, it can be concluded that the MIC has the best overall performance.

In Table VII, the selected feature number of each methods are illustrated. For Pearson, due to there is no consideration for the redundancy elimination, the selected features are always f_1, f_2, f_{11}, f_{16} and f_{17}, which is more than other methods and makes the model efficiency is limited; mRMR can select the fewest feature, only one, and the selected feature will be f_{13} or f_{17} alternatively; for the McTwo and the proposed two-stage MIC, the selected features are a subset of f_{11}, f_{14}, f_{16} and f_{17}. Compared to the Pearson, MIC and McTwo can achieve a high accuracy with fewer features; compared to mRMR, MIC and McTwo can achieve a much higher classification accuracy, which shows MIC and McTwo are more cost-effective.

To further evaluate the performance of the different feature selection methods in processing multi-class classification problems, instead of using P, R and F, $macro - P$, $macro - R$ and $macro - F$ are considered [32],

$$\begin{cases} macro - P = \frac{1}{n}\sum P_{ii} \\ macro - R = \frac{1}{n}\sum R_{ii} \\ macro - F_1 = \frac{2\times macro-P\times macro-R}{macro-P+macro-R} \end{cases} \tag{13}$$

where ii is the binary classification times. The SVM is employed as the classifier, the test data is divided into 6 categories, 15 times binary classification are implemented. The evaluation result of $macro - P$, $macro - R$ and $macro - F$ achieved by SVM is shown in Fig.7. The overall performance of SVM is good, the main error of classification is

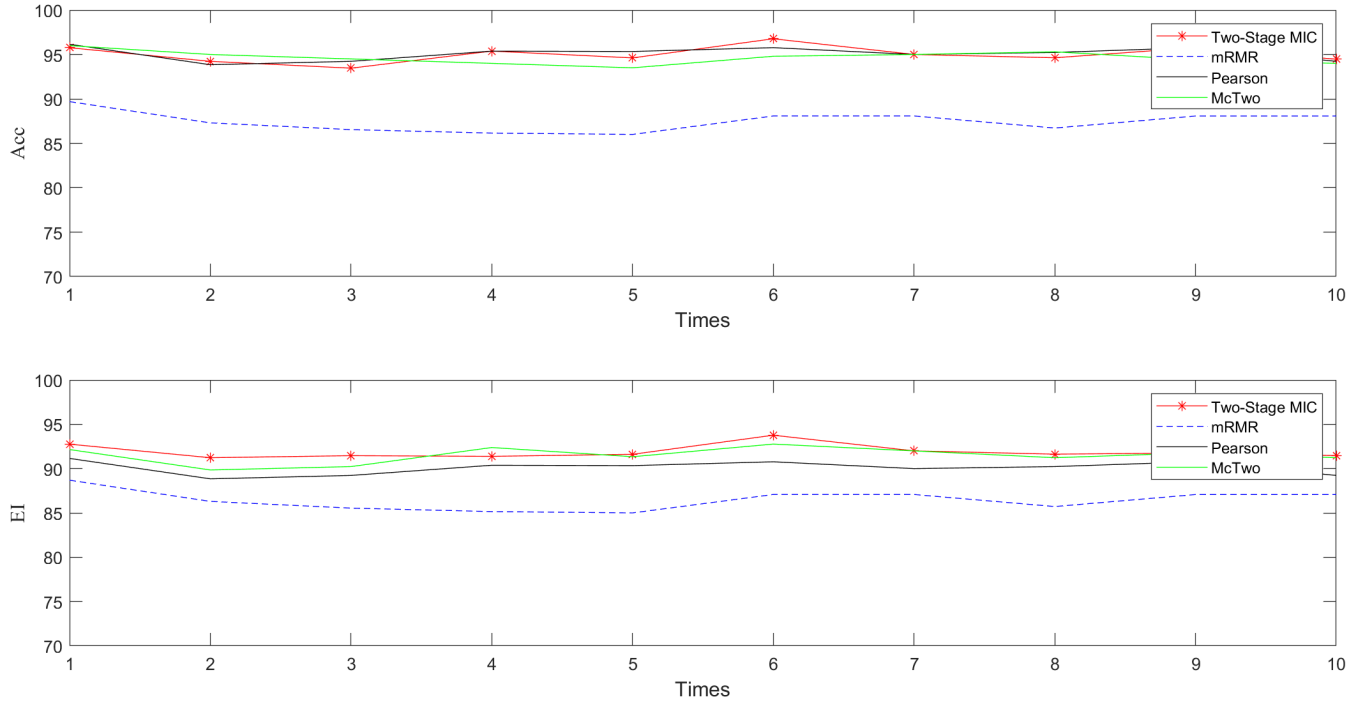


Fig. 6: The Acc and EI of kNN classification with four different select methods

TABLE VII: The Number of Feature Selected of Different Selection Methods

Times	1	2	3	4	5	6	7	8	9	10
MIC	3	3	2	4	3	3	3	3	4	3
mRMR	1	1	1	1	1	1	1	1	1	1
Pearson	5	5	5	5	5	5	5	5	5	5
McTwo	4	4	4	3	4	3	3	4	4	3

concentrated in label 3 and 4, label 5 and 6, especially the error between the fault 5 and 6. For the proposed MIC-based two-stage feature selection method, it achieves the best performance for all the metrics, including $macro - P$, $macro - R$ and $macro - F_1$.

V. Conclusion

In this paper, a new feature selection method using two-time MIC calculations in different stages is proposed. This proposed feature selection method has a strong identification capability on a wide range of relationships for both continuous or discrete data. By using this method, strong features that contain more useful information can be identified and reserved, which makes the high classification accuracy is guaranteed; on the other hand, the features with redundant information can be eliminated, which reduces the complexity, computational cost and the storage demand of the classification process. A rolling bearing focused case study is carried on, the result shows that the proposed feature selection method has a very good overall performance for both accuracy and efficiency. It is worthy mentioning that the proposed feature selection method is generic, which can be applied to many types of feature classification problems, especially for using in railway condition monitoring, such as bogie, track, point machine, switch and wheel. Apart from the railway filed, this method can also be adapted to other areas, such as condition monitoring in intelligent transportation systems, fault detection in lager industrial equipment. In our future work, more methodology applications will be investigated.

References

[1] J. Chen, C. Roberts, and P. Weston, "Fault detection and diagnosis for railway track circuits using neuro-fuzzy systems," *Control Engineering Practice*, vol. 16, no. 5, pp. 585–596, 2008.

[2] M. Sun, Y. Wang, X. Zhang, Y. Liu, Q. Wei, Y. Shen, and N. Feng, "Feature selection and classification algorithm for non-destructive detecting of high-speed rail defects based on vibration signals," in *Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, 2014 IEEE International*. IEEE, 2014, pp. 819–823.

[3] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1044700

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[5] L. X. Wang and S. Y. Jiang, "A feature selection method based on feature clustering," *Computer Application Research of Computers*, vol. 32, no. 5, pp. 1305–1308, 2015.

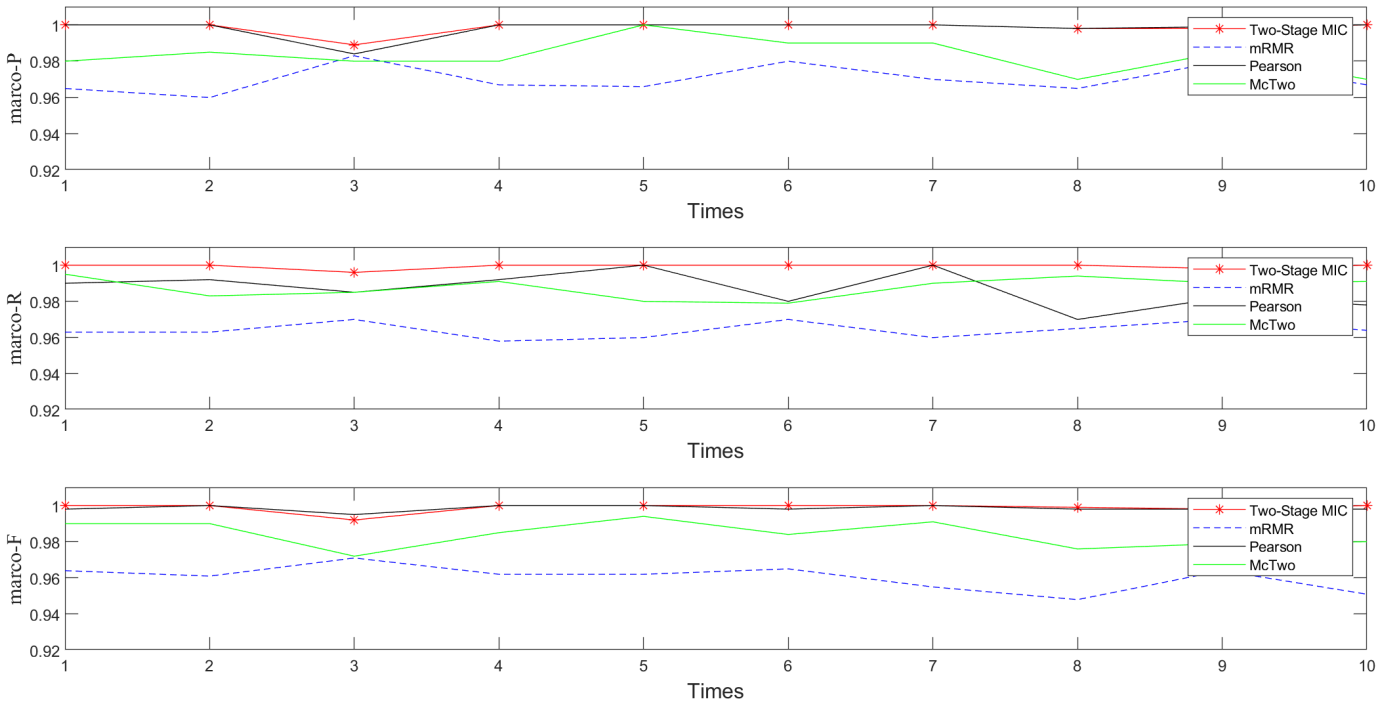


Fig. 7: The $macro-P$, $macro-R$ and $macro-F$ of SVM classification with four different feature selection methods

[6] H. H. Meng, "Study on multi-task learning based on genetic algorithm," *Computer Science*, 2008.

[7] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Bioinformatics Conference, 2003. Csb 2003. Proceedings of the*, 2003, pp. 523–528.

[8] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, and F. Zhou, "Mctwo: a two-step feature selection algorithm based on maximal information coefficient," *Bmc Bioinformatics*, vol. 17, no. 1, p. 142, 2016.

[9] F. Xi, Q. Sun, and G. Krishnappa, "Bearing diagnostics based on pattern recognition of statistical parameters," *Journal of Vibration and Control*, vol. 6, no. 3, pp. 375–392, 2000.

[10] Q. Sun, P. Chen, D. Zhang, and F. Xi, "Pattern recognition for automatic machinery fault diagnosis," *Journal of vibration and acoustics*, vol. 126, no. 2, pp. 307–316, 2004.

[11] S. Goreczka and J. Strackeljan, "Optimization of time domain features for rolling bearing fault diagnostics," in *The Sixth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, 2009, pp. 642–652.

[12] K. Kappaganthu and C. Nataraj, "Feature selection for fault detection in rolling element bearings using mutual information," *Journal of vibration and acoustics*, vol. 133, no. 6, p. 061001, 2011.

[13] X. Zhao and Y. Shi, "Xgboost application in rolling bearing fault diagnosis," *Noise and Vibration Control*, vol. 37, no. 4, pp. 166–170, 2017.

[14] T. W. Rauber, F. de Assis Boldt, and F. M. Varejão, "Heterogeneous feature models and feature selection applied to bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 637–646, 2015.

[15] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131 – 156, 1997. P. Langley, "Selection of relevant features in machine learning," *Proc.aaai Fall Symp.on Relevance*, pp. 140–144, 1994.

[16] Z. Li, "Study on fractal feature extraction and diagnosis methods," PhD Thesis, Chongqing University, Chongqing, China, 2013.

[17] D. Y. Wang, W. Z. Zhang, and J. G. Zhang, "Application of wavelet packet energy spectrum in rolling bearing fault diagnosis," *Bearing*, 2010.

[18] S. Li and Z. Li, "Fault monitoring method for rolling bearing based on wavelet packet energy features," *Journal of System Simulation*, vol. 15, no. 1, pp. 76 – 80, 2013.

[19] D. Q. Liu, G. J. Tang, and C. Zhang, "Application of improved support vector machine in fault diagnosis of rotating machinery," *Noise and Vibration Control*, 2011.

[20] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. Mcvean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, p. 1518, 2011.

[21] V. Sugumaran and K. I. Ramachandran, "Effect of number of features on classification of roller bearing faults using svm and psvm." *Expert Systems with Applications*, vol. 38, no. 4, pp. 4088–4096, 2011.

[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans.inf.theory*, vol. 13, no. 1, pp. 21–27, 1967.

[23] H. J. Jiang, "Maximum information coefficient and its application in brain network analysis," PhD Thesis, University of Chinese Academy of Sciences, Wuhan, China, 2013.

[24] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[25] P. Viola and W. M. W. Iii, *Alignment by Maximization of Mutual Information*. IEEE Computer Society, 1995.

[26] B. Q. Li, L. L. Hu, L. Chen, K. Y. Feng, Y. D. Cai, and K. C. Chou, "Prediction of protein domain with mrmr feature selection and analysis," *Plos One*, vol. 7, no. 6, pp. e39 308–e39 319, 2012.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, 2005.

[28] C. Yang, "Fault diagnosis of rolling based on vibration signal analysis," Master Thesis, Lanzhou University of Technology, Lanzhou, China, 2014.

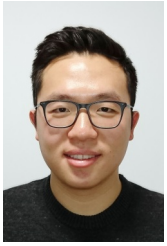
[29] D. Dong, "The diagnostic methods of bearing failure based on correlation analysis of output signals," Master Thesis, School of Automation, Hangzhou Dianzi University, Hangzhou, China, 2018.

[30] Y. Liu, B. Li, R. Tan, X. Zhu, and Y. Wang, "A gradient-boosting approach for filtering de novo mutations in parent-offspring trios." *Bioinformatics*, vol. 30, no. 13, pp. 1830–1836, 2014.

[31] C. C. Yu, J. C. Yang, Y. C. Chang, C. Jiing-Guang, C. W. Lin, M. S. Wu, and C. Lu-Ping, "Vcp phosphorylation-dependent interaction partners prevent apoptosis inhelicobacter pylori-infected gastric epithelial cells," *Plos One*, vol. 8, no. 1, p. e55724, 2013.

[32] H. E. Zhifen, M. Yang, and H. Liu, "Multi-task joint feature selection

for multi-label classification," *Chinese Journal of Electronics*, vol. 24, no. 2, pp. 281–287, 2015.



Tao Wen received his B.Eng degree from the School of Computer Science, Hangzhou Dianzi University, Hangzhou, China, and Master degree from the Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK, in 2011 and 2013, respectively. From 2013 to 2017, he was a PhD candidate at the Birmingham Centre for Railway Research and Education at the University of Birmingham, Birmingham, UK, and received his PhD degree in 2018. His research interests include CBTC system optimization, railway signalling simulation, railway condition monitoring, wireless signal processing and digital filter research.

Deyi Dong Born in 1992. She received the B.Eng. degree the Jinling Institute of Technology, Nanjing, China, in 2015. She is currently working toward the Master degree at the School of Automation, Hangzhou Dianzi University, Hangzhou, China. Her research interests include fault diagnosis, feature classification and pattern recognition.

Qianyu Chen Received the degree of Bachelor of Engineering (Electrical and Electronic) with the award of First Class Honours, from University of Leicester in 2016, and the degree of Master of Science in the College of Science and Engineering in Signal Processing and Communications with Distinction, from The University of Edinburgh in 2017. In November 2017, she started Ph.D. research in Birmingham Centre for Railway Research and Education (BCRRE), University of Birmingham. Her current research interests include railway condition monitoring and pattern recognition.

Lei Chen received the B.Eng. degree in automation engineering from Beijing Jiaotong University, Beijing, China, in 2005, and the Ph.D. degree in railway traffic management from the University of Birmingham, Birmingham, U.K., in 2012. He is currently a Birmingham Fellow for Railway Traffic Management with the Birmingham Centre for Railway Research and Education, University of Birmingham. His research interests include railway traffic management and control, railway safety critical system design, and railway simulation.

Clive Roberts is Professor of Railway Systems at the University of Birmingham. Clive is Director of the Birmingham Centre for Railway Research and Education, which is the largest railway research group in Europe with just over 100 researchers. He works extensively with the railway industry and academia in Britain and overseas. He leads a broad portfolio of research aimed at improving the performance of railway systems, including a leading a strategic partnership in the area of data integration with Network Rail. His main research interests lie in the areas of railway traffic management, condition monitoring, energy simulation and system integration.