Hindawi
Journal of Sensors
Volume 2017, Article ID 8092691, 15 pages
https://doi.org/10.1155/2017/8092691

Research Article

Rotating Machinery Fault Diagnosis for Imbalanced Data Based on Fast Clustering Algorithm and Support Vector Machine

Xiaochen Zhang, Dongxiang Jiang, Te Han, Nanfei Wang, Wenguang Yang, and Yizhou Yang

State Key Lab of Power Systems, Department of Thermal Engineering, Tsinghua University, Beijing 100084, China

Correspondence should be addressed to Xiaochen Zhang; zhangxch2008@hotmail.com

Received 17 January 2017; Revised 27 June 2017; Accepted 20 September 2017; Published 22 October 2017

Academic Editor: Pietro Siciliano

To diagnose rotating machinery fault for imbalanced data, a method based on fast clustering algorithm (FCA) and support vector machine (SVM) was proposed. Combined with variational mode decomposition (VMD) and principal component analysis (PCA), sensitive features of the rotating machinery fault were obtained and constituted the imbalanced fault sample set. Next, a fast clustering algorithm was adopted to reduce the number of the majority data from the imbalanced fault sample set. Consequently, the balanced fault sample set consisted of the clustered data and the minority data from the imbalanced fault sample set. After that, SVM was trained with the balanced fault sample set and tested with the imbalanced fault sample set so the fault diagnosis model of the rotating machinery could be obtained. Finally, the gearbox fault data set and the rolling bearing fault data set were adopted to test the fault diagnosis model. The experimental results showed that the fault diagnosis model could effectively diagnose the rotating machinery fault for imbalanced data.

1. Introduction

With the development of modern large-scale production and the progress of science and technology, the structure of mechanical equipment has become more complex. During equipment operation, sudden failure of the equipment would lead to the loss of service ability or may even cause a serious disastrous accident [1, 2]. To ensure the reliability of the equipment to obtain greater economic and social benefits, the timely and accurate diagnosis of the equipment's failure mode is particularly significant to guarantee the normal operation of the equipment. Rotating machinery, such as bearings and gears, has been widely used in numerical control machine tools, aeroengine, electric power system, agricultural machinery, transport machinery, metallurgical machinery, and other modern industrial equipment [3–5]. In recent years, new technologies and theories such as artificial neural networks have been widely applied in mechanical equipment fault diagnosis, which greatly improves the accuracy of fault diagnosis. For rotating machinery, there are various kinds of faults; however, samples of some typical faults are difficult to obtain [6, 7]. Therefore, it is necessary to study rotating machinery fault diagnosis technology for the condition of imbalanced data.

At present, SVM-based fault diagnosis is one of the most widely used fault diagnosis methods for mechanical equipment [8, 9]. This method learns the process data of different operating states of the equipment and then classifies the data into different faults by constructing classification hyperplanes in high-dimensional space. Imbalanced data means that out of the data used in training classifiers, the number of some fault data is larger than other types. Adopting the imbalanced data as the training set, the classification hyperplane would be offset from the real one, thus reducing the validity of fault diagnosis. In SVM, the penalty factor indicates the error sensitivity of the classifier. Currently, one of the methods

used to solve the imbalanced data problems defines different penalty factors as positive and negative samples to increase the penalty factor of the disadvantage samples so the classifier is sensitive to them [10, 11]. However, during the process of setting penalty factors, it is difficult to choose suitable penalty factors for different faults. Different values will directly affect the performance of the classifier, and small penalty factors often result in no obvious suppression effect while larger penalty factors weaken the generalization ability of the classifier. Another method to solve the imbalanced data problems is to conduct preprocessing for data [12, 13] by reducing the number of the majority data to balance the data. Therefore, the selection of core data is the key in determining the performance of the SVM classifier for imbalanced data.

The purpose of the cluster algorithm is to classify the data according to their similarity. Therefore, we proposed an approach based on a fast clustering algorithm to reduce the number of the majority data from the imbalanced data. This fast clustering algorithm was proposed by Rodriguez and Laio in 2014 based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities [14–17]. Based on these two assumptions, the fast clustering algorithm can be used to dispose of different clusters.

To diagnose rotating machinery fault for imbalanced data, a kind of method based on fast clustering algorithm and SVM was proposed. According to the proposed method, original features of different faults are constructed by VMD. Next, PCA is applied to reduce the dimension of the original features so that sensitive features can be obtained. After that, the fast clustering algorithm is adopted to reduce the number of the majority data from the imbalanced sensitive features. Finally, SVM is trained with the data clustered by the fast clustering algorithm, so that the fault diagnosis model for imbalanced data can be obtained.

2. SVM and Imbalanced Data Classification

2.1. SVM. As a machine learning algorithm developed from statistical learning theory, SVM maps inseparable learning samples from low-dimensional space into high-dimensional space through a kernel function to obtain an optimal hyperplane [18]. If training set $\{(x_i, y_i),\ i = 1, 2, \ldots, l\}$ consists of two categories, then the computational goal can be expressed as

$$\begin{aligned} \min \quad & \frac{\|w\|^2}{2} + C\sum_{i=1}^{l}\varepsilon_i \\ \text{s.t.} \quad & y_i\left(wx_i + b\right) \geq 1 - \varepsilon_i \\ & \varepsilon_i > 0, \\ & i = 1, 2, \ldots, l, \end{aligned} \tag{1}$$

where C is the penalty factor and ε_i is the slack variable.

The constraint conditions can be defined as

$$\sum_{i=1}^{l}\alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, l. \tag{2}$$

Then, the Lagrange function is constructed as

$$\begin{aligned} \Phi\left(w, b, \alpha_i\right) = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\varepsilon_i \\ & - \sum_{i=1}^{l}\alpha_i\left[y_i\left(wx_i + b\right) - 1 + \varepsilon_i\right] - \sum_{i=1}^{l}\beta_i\varepsilon_i, \end{aligned} \tag{3}$$

where α_i and β_i are the Lagrange operators.

Then, the classification function can be formed as

$$f\left(x\right) = \operatorname{sgn}\left[\sum_{i=1}^{l}\alpha_i^* y_i K\left(x_i, x\right) + b^*\right], \tag{4}$$

where $K(x_i, x)$ is the kernel function.

2.2. Classification Boundary Migration of SVM. SVM classification algorithm assumes that the number of each class is approximately equal. In fact, for rotating machinery, the acquisition of fault samples is full of randomness, so it is difficult to guarantee the balance among different fault samples. Figure 1 shows the skewing of hyperplane.

From Figure 1, it can be seen that the hyperplane could easily distinguish two types of classes from the balanced data set. However, the hyperplane obviously shifted towards the minority class if the data set was imbalanced. Since the number of class 2 was small and two classes adopted the same penalty factor, the overall error caused by class 2 was also small. The result was that the hyperplane was easily affected by the outlier and moved to the direction of class 2, which caused a large classification error of the minority class. Therefore, to improve the classification performance of the SVM classifier for imbalanced data, a fast clustering algorithm was adopted to balance the data set.

3. Imbalanced Data Classification Based on Fast Clustering Algorithm and SVM

3.1. Fast Clustering Algorithm. In this paper, a type of fast clustering algorithm was used as the theoretical basis of balancing the original data set as the basic idea of this clustering algorithm is novel and simple and is very suitable for searching the core samples from the imbalanced data set. This fast clustering algorithm assumes that cluster centers are surrounded by neighbors with lower local densities; meanwhile, they are at a relatively large distance from the points with a higher local density [16, 17]. There are two ways to calculate local density, including cut-off kernel and Gaussian kernel. With cut-off kernel, the local density ρ_i of data point i can be calculated as

$$\begin{aligned} \rho_i &= \sum_j \chi\left(d_{ij} - d_c\right) \\ \chi\left(x\right) &= \begin{cases} 0, & x \geq 0 \\ 1, & x < 0, \end{cases} \end{aligned} \tag{5}$$

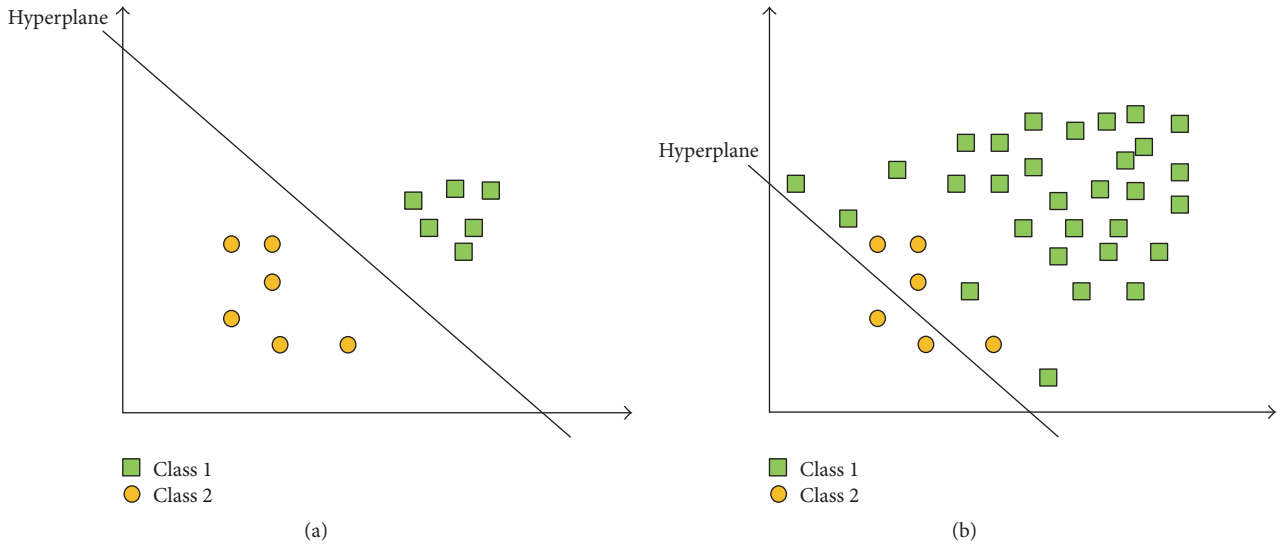

Figure 1: Hyperplane of SVM. (a) Balance data set and (b) imbalance data set.

where d_{ij} is the distance between data point i and data point j and d_c is the cut-off distance.

With Gaussian kernel, the local density ρ_i of data point i can be calculated as

$$\rho_i = \sum_j e^{-(d_{ij}/d_c)^2}. \tag{6}$$

From (5) to (6), the local density ρ_i means the number of the data points that are closer to data point i compared with d_c.

Distance δ_i is defined as

$$\delta_i = \begin{cases} \min\limits_{j \in I} \left(d_{ij}\right), & I \neq \varnothing \\ \max\limits_{j \in I} \left(d_{ij}\right), & I = \varnothing, \end{cases} \tag{7}$$

where set $I = \{\rho_j > \rho_i\}$.

From (7), we know that distance δ_i is the minimum distance between point i and the point with higher density, except that point i has the highest density.

For each data point, we can calculate its local density ρ_i and distance δ_i. Then, the weight of clustering center γ_i is constructed as

$$\gamma_i = \rho_i \delta_i. \tag{8}$$

Obviously, points with larger weights are clustering centers. The sequence n_i is constructed as

$$n_{q_i} = \arg\min d_{q_i q_j}\left(q_j\right), \quad i \geq 2,\ i > j, \tag{9}$$

where sequence q_i is the index number of local density ρ_i sorted in descending order. The sequence n_i represents the index number of the point closest to point i, while the local density of this point is larger than point i.

Then, the nonclustering center points can be categorized as

$$c_{q_i} = c_{n_{q_i}}, \tag{10}$$

where c is the label of the clustering centers.

For each cluster, the mean local density of this cluster is calculated. By comparing the mean local density, the points of this cluster can be divided into core points or halo points.

The synthetic point distributions data set [16] was adopted to test the effectiveness of the algorithm. Figure 2(a) shows the distribution before clustering, while Figure 2(b) shows the distribution after clustering. It is clear that core points of five class data were correctly chosen from the raw synthetic point distributions data set and showed that the fast clustering algorithm could be well applied to eliminate the halo points of the raw data.

3.2. Imbalanced Data Classification. With the fast clustering algorithm, the imbalanced data set was preprocessed and the number of the majority classes reduced. Therefore, the raw data set was reassembled into a balanced data set. Then, the SVM classification algorithm was adopted to learn the balanced data set. The movement of the SVM hyperplane during the process of clustering is shown in Figure 3.

As shown in Figure 3, affected by the number of the data sets, the hyperplane was obviously biased to the minority class. The purpose of the fast clustering algorithm was to search the core points of the majority class and reconstruct a balanced data set so that the hyperplane could return to the side of the majority class. Therefore, the classification accuracy of the SVM classifier could be improved.

3.3. Evaluation of Imbalanced Data Classification. For imbalanced data, the proportion of minority samples was not large, so the classification results of the minority samples had little effect on overall accuracy of classification. Therefore, there were some unique classification evaluation indexes for imbalanced data [19, 20]. Based on the confusion matrix, we defined the positive class (minority class) as P and the negative class (majority class) as N in the imbalanced data. As shown in Figure 4, TP and TN denote the correctly identified

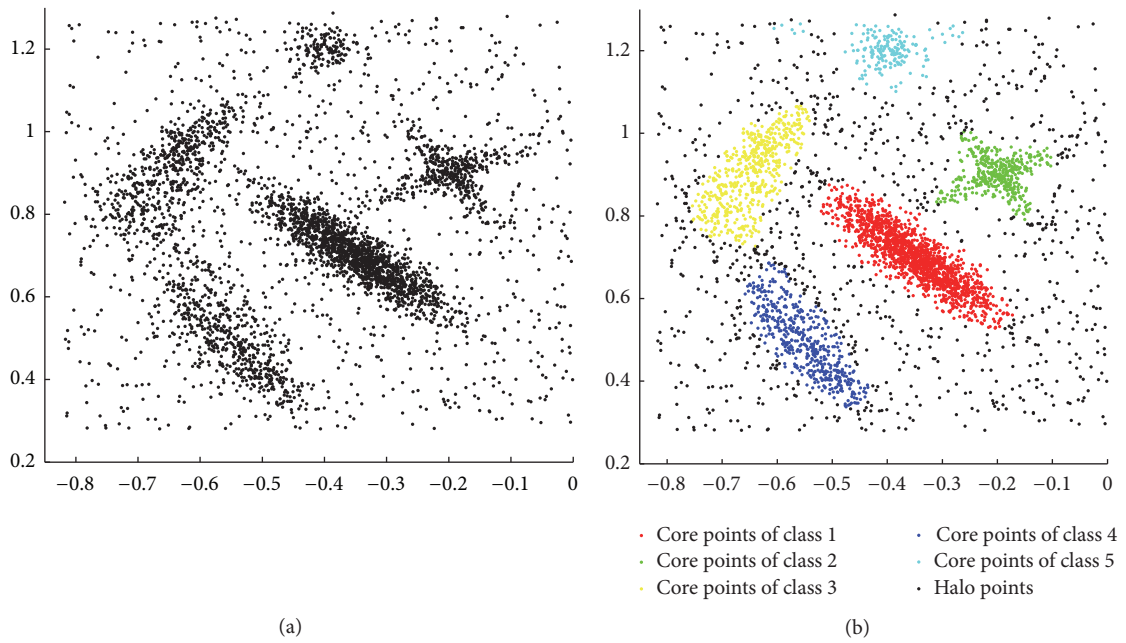

Figure 2: Synthetic point distributions data set: (a) before clustering and (b) after clustering.

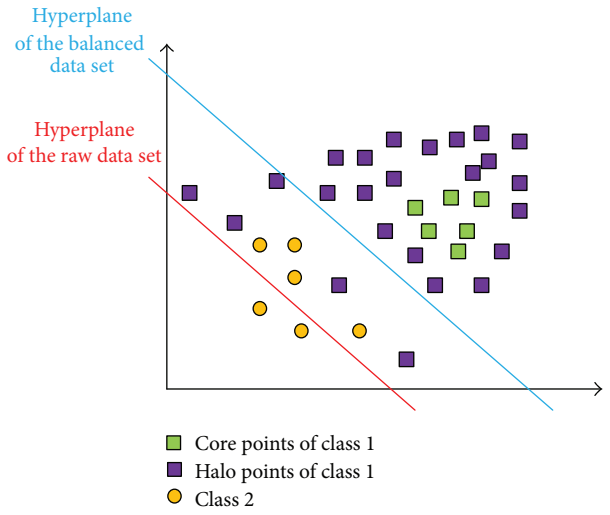

Figure 3: Schematic diagram of SVM hyperplane movement.

positive and negative samples, respectively. FP indicates that the negative samples are misclassified into positive class, while FN indicates that the positive samples are misclassified into negative class.

The recall of the positive class can be defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}. \tag{11}$$

The recall of the negative class is

$$\mathrm{TNR} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})}. \tag{12}$$

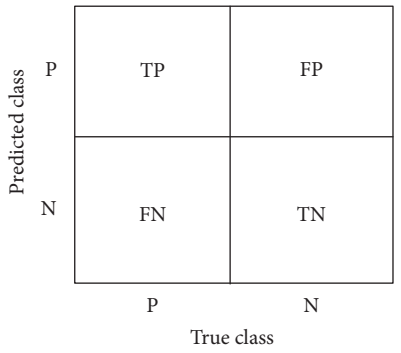

Figure 4: Confusion matrix of the imbalanced data.

The precision of the positive class can be formed as

$$\mathrm{precision} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}. \tag{13}$$

Then G-mean can be constructed as

$$G = \sqrt{\mathrm{TPR} * \mathrm{TNR}}. \tag{14}$$

F-mean can be constructed as

$$F = \frac{2 * \mathrm{TPR} * \mathrm{precision}}{\mathrm{TPR} + \mathrm{precision}}. \tag{15}$$

As the evaluation index, G-mean takes into account the classification performance of both positive and negative class. If the classification of the classifier is biased towards one class, it will directly affect the classification accuracy of another class where the G value will be very small. From (15), we can see that F-mean considers the recall and precision of

the positive class. Therefore, F-mean can comprehensively show the classification effect of the classifier on positive class (minority class).

4. Rotating Machinery Fault Diagnosis for Imbalanced Data

4.1. Feature Extraction. For rotating machinery, the vibration signal is composed of multiple components. VMD, a novel adaptive signal decomposition method, was adopted to construct the original features in this study. The target of the VMD was to decompose the original signal f into a number of IMFs (Intrinsic Mode Functions), u_k, that had specific sparse properties while reproducing the original signal f [21, 22]. Since the decomposition of VMD was sparse, it could be considered as a constrained variational problem as follows:

$$\begin{aligned} \min_{\{u_k\},\{w_k\}} \quad & \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right\} \\ \text{s.t.} \quad & \sum_k u_k = f, \end{aligned} \tag{16}$$

where $\{u_k\}$ are shorthand notations of the modes and $\{w_k\}$ are center frequencies of the modes.

To solve this constrained variational problem, the augmented Lagrange function is introduced as

$$\begin{aligned} & L(\{u_k\}, \{w_k\}, \lambda) \\ & := \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \\ & \quad + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 \\ & \quad + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle, \end{aligned} \tag{17}$$

where α is the penalty factor and λ is the Lagrange multiplier.

The process of decomposing was as follows. First, $\{u_k\}$, $\{w_k\}$, λ, and n were all initialized as 0. Then, u_k, w_k, and λ were updated through the circulative iteration. u_k was updated as

$$\begin{aligned} & u_k^{n+1} \\ & = \arg\min_{u_k \in X} \left\{ \alpha \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right. \\ & \quad \left. + \left\| f(t) - \sum_i u_i(t) + \frac{\lambda(t)}{2} \right\|_2^2 \right\}. \end{aligned} \tag{18}$$

The center frequencies w_k can be calculated as

$$\begin{aligned} & w_k^{n+1} \\ & = \arg\min_{w_k} \left\{ \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right\}. \end{aligned} \tag{19}$$

The condition for convergence is the following:

$$\sum_k \frac{\left\| u_k^{n+1} - u_k^n \right\|_2^2}{\left\| u_k^n \right\|_2^2} < \varepsilon, \tag{20}$$

where ε is the discriminant accuracy.

Finally, the original signal f was decomposed into a number of IMFs, u_k. Then, the energy of each IMF was calculated to constitute original feature vector, which was used to distinguish the original signal.

To test the validity of VMD, a pure harmonic signal affected by noise was adopted. Furthermore, we also conducted a comparison with empirical mode decomposition (EMD) based on the exact same testing signal. Here, the pure harmonic signal was the following:

$$f_1 = 2 \sin\left(58.2\pi t + \frac{\pi}{7} \right). \tag{21}$$

The noisy input signal was the pure harmonic signal affected by noise with the expression as follows:

$$f = f_1 + 0.3 \cos(1400\pi t) + 0.36 \cos(576\pi t) + 0.7\eta, \tag{22}$$

where $\eta \sim N(0, \sigma)$ represents the Gaussian additive noise.

The signal waveforms of the pure harmonic signal and the noisy input signal are shown in Figure 5.

Figure 6 shows the decomposition of the noisy input signal. It was clear that the VMD algorithm correctly extracted the pure harmonic signal from the noisy input signal and that the EMD algorithm extracted seven IMFs from the noisy input signal. There was no pure harmonic signal in the seven IMFs.

With VMD, the vibration signal of the rotating machinery was decomposed into a number of IMFs. Then, the energy of each IMF was calculated to constitute original feature vector. Since these original feature vectors were high-dimensional features, dimensionality reduction algorithm was applied to reduce the computational complexity.

4.2. Feature Dimension Reduction. A kind of traditional dimensionality reduction algorithm, PCA, was adopted to reduce the dimension of the original feature vectors. PCA is a statistical method which adopts orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ expresses the n-dimensional original features, while $\mathbf{Y} = [y_1, y_2, \ldots, y_n]$ is used to express the linearly uncorrelated sensitive features. With PCA, the contribution of the ith component η_i can be defined as follows:

$$\eta_i = \frac{\lambda_i}{\sum_{k=1}^n \lambda_k}, \tag{23}$$

where λ_i means the variance of the y_i.

Then, the contributions of the first m principal component η_m' can be calculated as follows:

$$\eta_m' = \frac{\sum_{i=1}^m \lambda_i}{\sum_{k=1}^n \lambda_k}. \tag{24}$$

Finally, the principal components with high contributions can be chosen as the sensitive features.

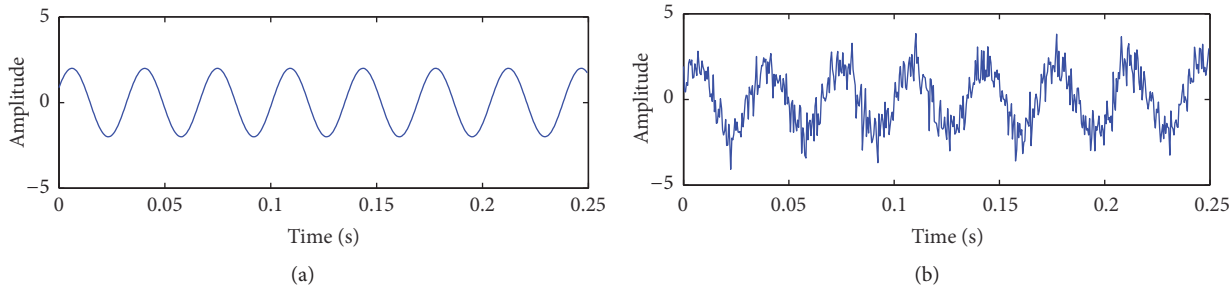

Figure 5: Signal waveforms. (a) Pure harmonic signal and (b) the noisy input signal.

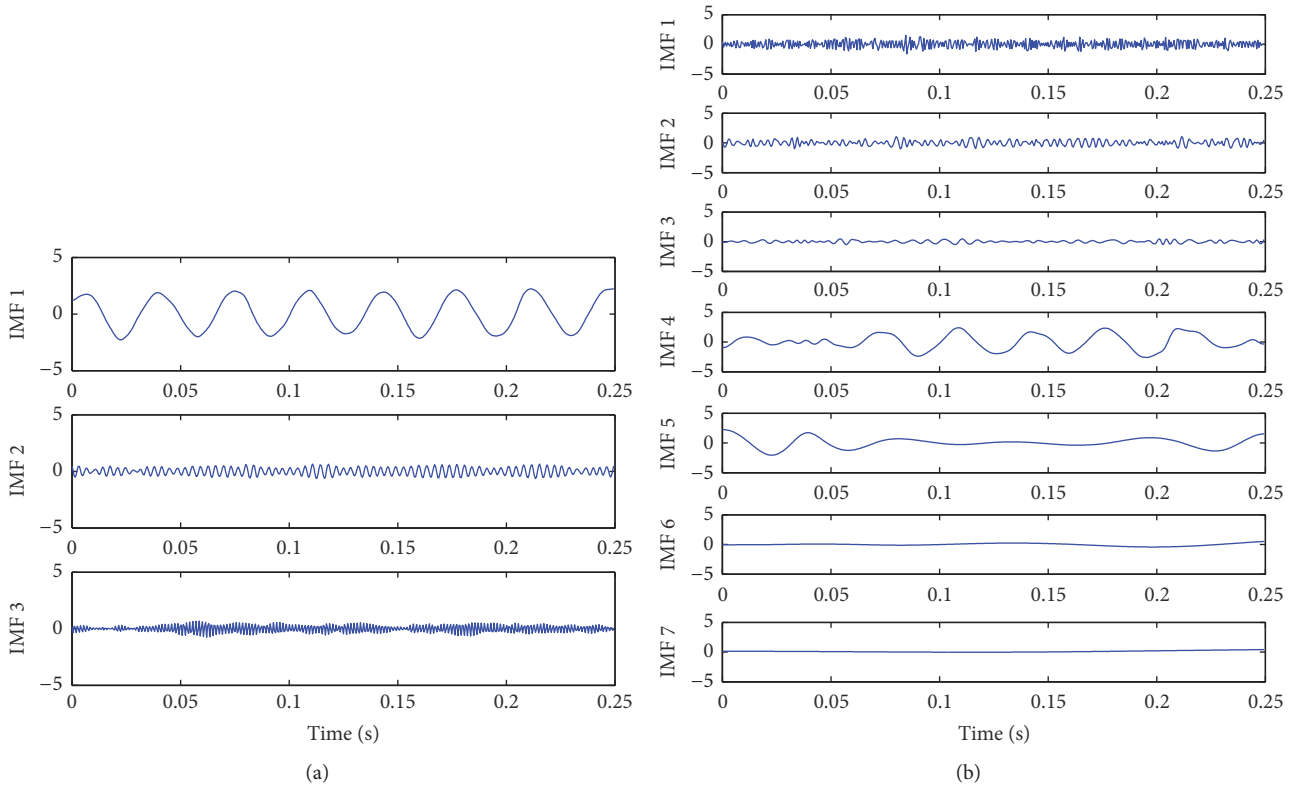

Figure 6: Decomposition of the noisy input signal: (a) IMFs extracted by VMD and (b) IMFs extracted by EMD.

4.3. Sample Selection and Fault Diagnosis Model. The rotating machinery fault sample set (an imbalanced data set) is made up of several kinds of faults. Some faults are majority class while others are minority class. Each fault sample contains a number of sensitive features. The distance between the ith fault sample and the jth fault sample can be calculated as

$$d_{ij} = \sqrt{\sum_{k=1}^{K} \left(t_{ik} - t_{jk}\right)^2}, \tag{25}$$

where t_{ik} are the sensitive features of the ith fault sample and t_{jk} are the sensitive features of the jth fault sample. K is the number of sensitive features for each fault sample.

According to (5)–(7), the local density ρ_i and the distance δ_i were obtained. Then, based on (8), the weight γ_i of each fault sample was calculated. With reference to the number of samples of the minority class, the same number of fault samples with higher weight γ_i were selected from the majority class. Whole samples of the minority class and selected samples of the majority class constructed balanced fault sample sets. Finally, the SVM classification algorithm was adopted to learn the balanced fault sample set. The flowchart of building fault diagnosis model is shown in Figure 7.

From Figure 7, it can be seen that the training samples chosen from the balanced fault sample set were used to train SVM, while the imitative testing samples chosen from the imbalanced fault sample set were applied to test the identification accuracy of the trained SVM. The SVM would be retrained until the classification accuracy of the trained SVM was acceptable; then this trained SVM could be adopted as the fault diagnosis model.

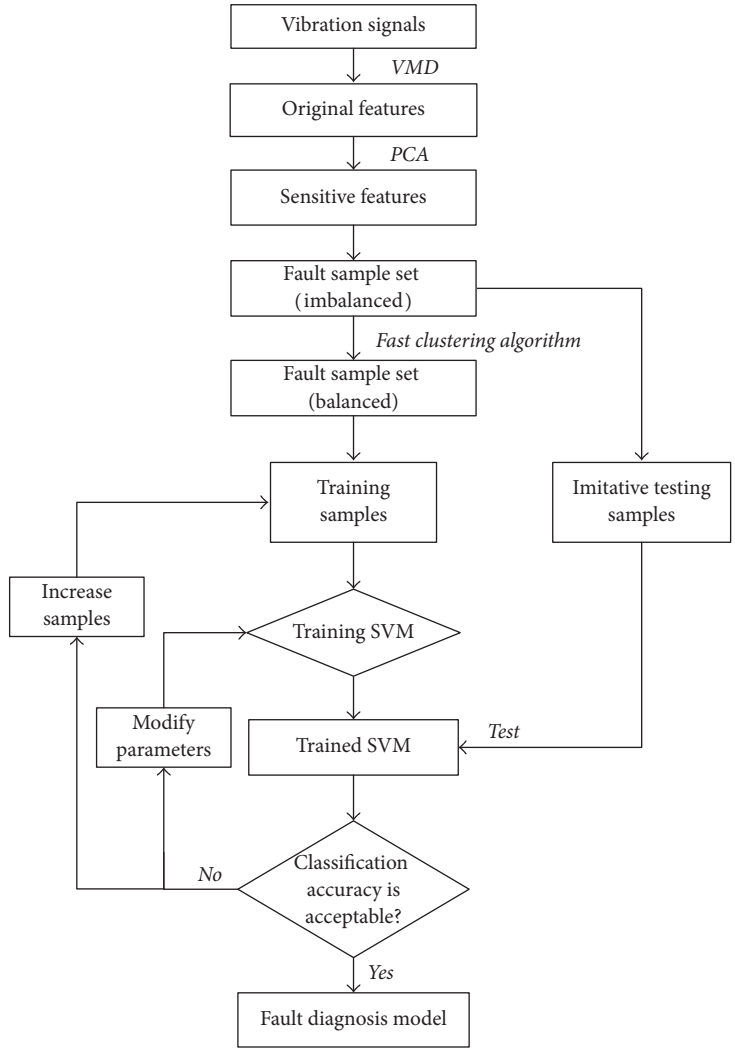

Figure 7: Flowchart of building the fault diagnosis model.

5. The Experimental Results

To verify the viability and effectiveness of the proposed algorithm, the gearbox fault data set and the rolling bearing fault data set were adopted to test the proposed fault diagnosis model.

5.1. Gearbox Fault Diagnosis. A wind turbine transmission chain fault simulation test bed is shown in Figure 8. The test bed mainly consisted of a motor driver, motor, gearbox, wind wheel, sensors, and data acquisition system. The wind wheel was driven by the motor through the gearbox and the motor speed was controlled by the motor driver. An acceleration sensor was installed on the top of the gearbox while the signal was acquired by the data acquisition system. The tested gearbox was a kind of single-stage planetary transmission, while the number of the planetary gear teeth was 20. In this test, two faults of gearbox were simulated: half fracture and full fracture for planetary gear. To simulate the real working condition of the wind turbine transmission chain, different wind wheel speeds were also considered. For each fault, three kinds of working conditions (wind wheel speed: 197 r/min, 237 r/min, and 277 r/min) were simulated. Therefore, as is shown in Table 1, the gearbox fault data set consisted of three condition modes. Figure 9 shows the pictures of planetary gears.

Table 1: Three kinds of condition modes for gearbox.

	Planetary gear	Wind wheel speed (r/min)
Condition mode 1	Normal	197/237/277
Condition mode 2	Half fracture	197/237/277
Condition mode 3	Full fracture	197/237/277

As is shown in Figure 10, the vibration signal of the gearbox can be decomposed into a number of IMFs by VMD. Then, the original features can be obtained by calculating the energy of each IMF.

With PCA, the original features are mapped to another plane and replaced with the sensitive features. In the sensitive

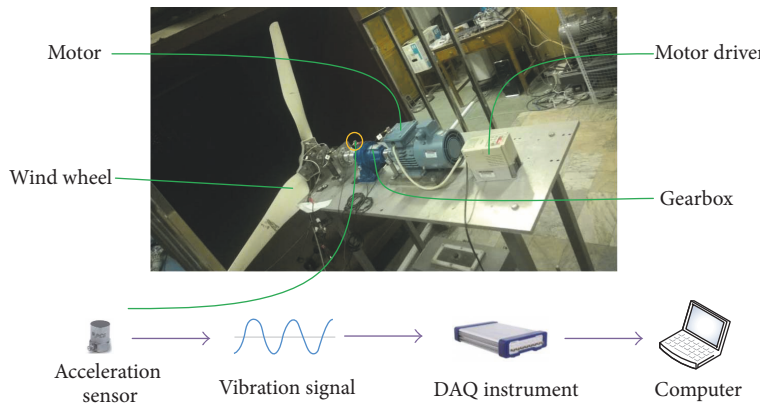

Figure 8: Wind turbine transmission chain fault simulation test bed.

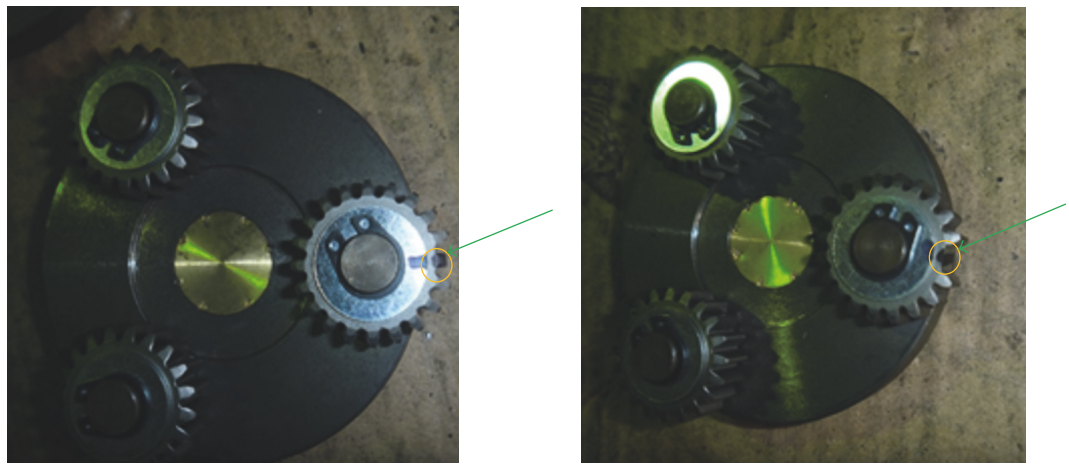

Figure 9: Pictures of planetary gears: (a) half fracture and (b) full fracture.

features, feature is sorted according to its contribution degree. Figure 11 shows the first five principal component contributions of PCA. It is clear that the accumulated contribution of the first three sensitive features was 89.28%; thus, sensitive feature 1, sensitive feature 2, and sensitive feature 3 were selected as the sensitive features.

Figure 12 was obtained by drawing three kinds of condition modes in a space formed by sensitive feature 1, sensitive feature 2, and sensitive feature 3. From Figure 12, it was clear that the distribution area of the normal planetary gear had been distinguished from half fracture and full fracture, but for half fracture and full fracture, an aliasing region exists in the distribution areas where it is difficult to make a distinction. The aliasing region can easily lead to the miscarriage of different failures, especially for the imbalanced failure data set. Thus, half fracture data and full fracture data were used to construct an imbalanced data set to test the classification of the proposed fault diagnosis model.

The imbalanced data sets under different proportions were constructed, while the distributions of imbalanced data sets are shown in Figure 13. Full failure was defined as the positive class (minority class) while half failure was the negative class (majority class). The number of the positive classes varied from 10 to 100. In the meantime, the number of the negative classes was 150. G-mean and F-mean were adopted as the evaluation indexes.

The proposed fault diagnosis model was adopted to classify the imbalanced data sets under different proportions. To test the validity of the proposed fault diagnosis model, the random undersampling (RU) algorithm, the synthetic minority oversampling technique (SMOTE) algorithm, the backpropagation (BP) neural network, and the radial basis function (RBF) neural network were introduced simultaneously. Table 2 and Figure 14 show a comparison of the evaluation indexes of the fault diagnosis model and other models. It is clear that the fault diagnosis model obtained good classification performances in different data sets. It is particularly worth mentioning that the fault diagnosis model was less affected by the proportion of the data set. A good classification effect could still be obtained with even less samples of the positive class (small class).

The testing samples consisted of 300 samples of which 150 samples were from half failure data while another 150 samples were from the full failure data. The classification accuracy

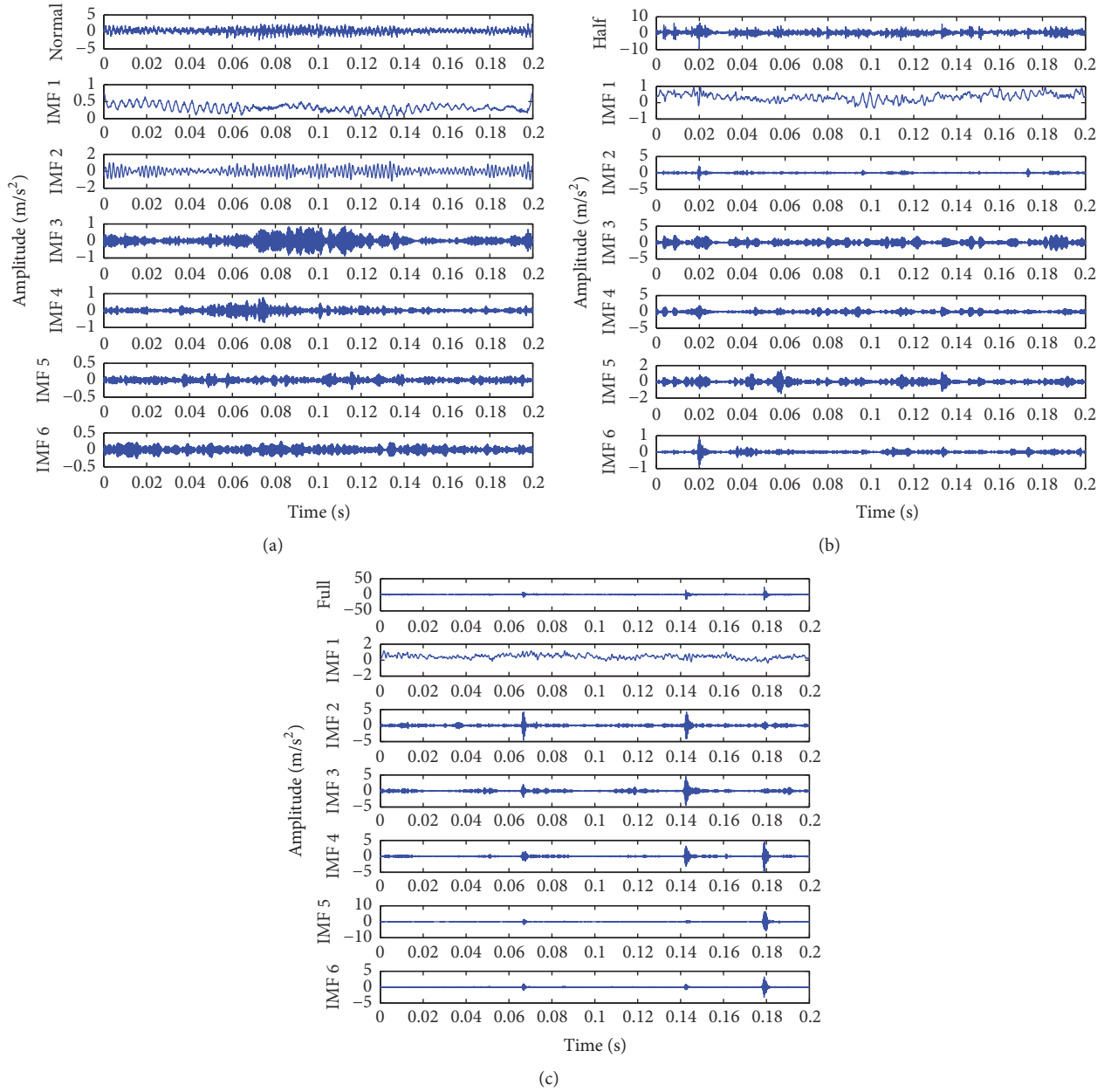

Figure 10: VMD decomposition of vibration signal (wind wheel speed: 237 r/min): (a) normal; (b) half fracture; and (c) full fracture.

comparisons of the fault diagnosis model and other models are shown in Table 3 and Figure 15, where it was obvious that the fault diagnosis model achieved good classification results. The classification accuracies in different data sets were all more than 80%.

5.2. Rolling Bearing Fault Diagnosis Based on Casing Vibration. In the case of gearbox fault diagnosis, the gearbox fault data set was used to test the performance of the fault diagnosis model when the model was applied to distinguish two failure modes. In this case, the fault diagnosis model was applied to distinguish multiple failure modes in the imbalanced data set.

The rolling bearing fault simulation test bed is shown in Figure 16. The motor was connected to the axis by a coupling, while the other end of the axis fits together with blades and the testing rolling bearing. The motor was responsible for driving blades and a casing was installed around the blades. Two one-way accelerometers were installed on the surface of the casing at a 90-degree angle. A data acquisition system was used to acquire the accelerometers' signals. The rotating speed was 1800 rpm and the sampling frequency was 16 kHz. The rolling bearing data set consisted of four modes such as normal, rolling element failure, inner race failure, and outer race failure.

Table 4 shows the composition of the rolling bearing data set where it was clear that the inner race failure and outer race failure were positive classes (small classes) in this imbalanced rolling bearing data set.

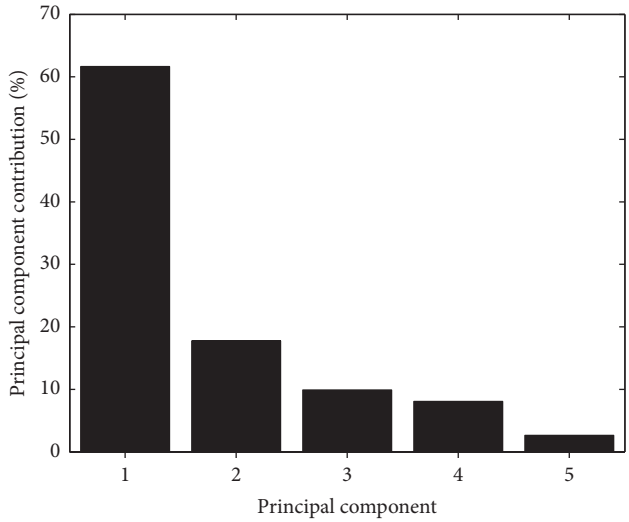

Figure 11: The first five principal component contributions.

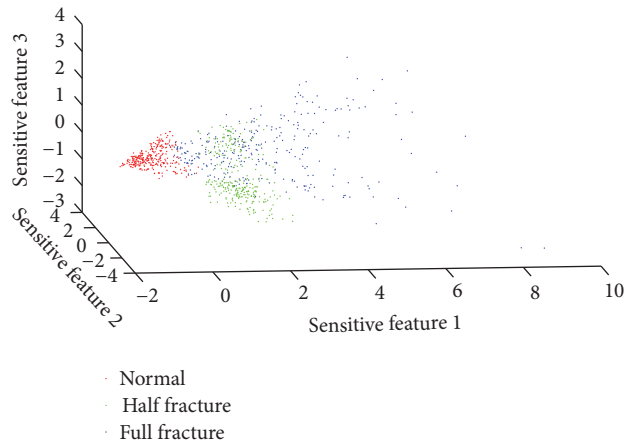

Figure 12: Distribution of three condition modes after VMD and PCA.

Table 2: Evaluation indexes comparisons of fault diagnosis model and other models.

Proportion of the data set	10 : 150	15 : 150	40 : 150	50 : 150	80 : 150	100 : 150
G-mean of fault diagnosis model	0.94	0.95	0.96	0.94	0.92	0.88
F-mean of fault diagnosis model	0.54	0.65	0.86	0.85	0.88	0.86
G-mean of FCA + BP	0.61	0.69	0.68	0.87	0.89	0.87
F-mean of FCA + BP	0.16	0.28	0.48	0.77	0.86	0.84
G-mean of FCA + RBF	0.92	0.93	0.92	0.90	0.91	0.89
F-mean of FCA + RBF	0.47	0.59	0.78	0.79	0.87	0.86
G-mean of RU + SVM	0.93	0.91	0.90	0.91	0.91	0.90
F-mean of RU + SVM	0.50	0.54	0.75	0.81	0.86	0.88
G-mean of SMOTE + SVM	0.85	0.91	0.92	0.92	0.92	0.91
F-mean of SMOTE + SVM	0.36	0.55	0.78	0.83	0.88	0.89
G-mean of SVM	0.60	0.68	0.85	0.87	0.92	0.89
F-mean of SVM	0.16	0.28	0.69	0.77	0.88	0.87

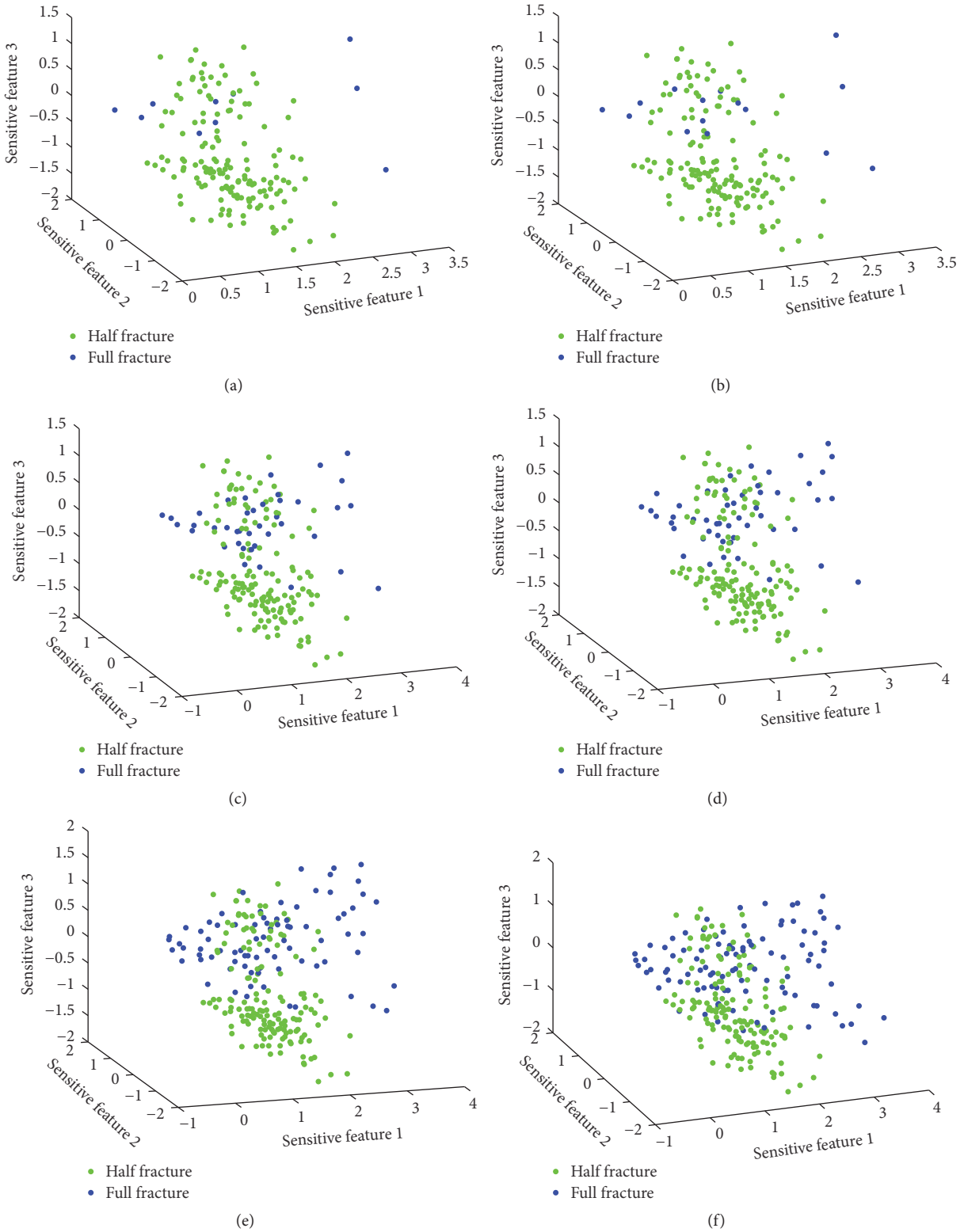

FIGURE 13: Distributions of the imbalanced data set under different proportions: (a) 10 : 150; (b) 15 : 150; (c) 40 : 150; (d) 50 : 150; (e) 80 : 150; and (f) 100 : 150.

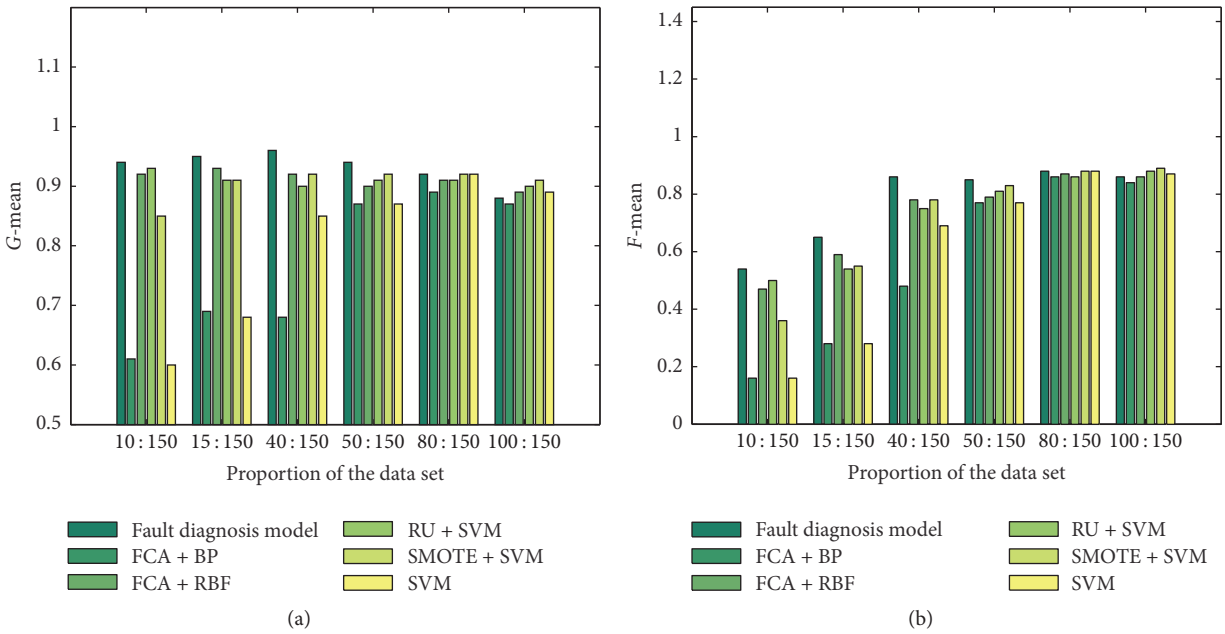

Figure 14: Evaluation indexes comparisons: (a) G-mean and (b) F-mean.

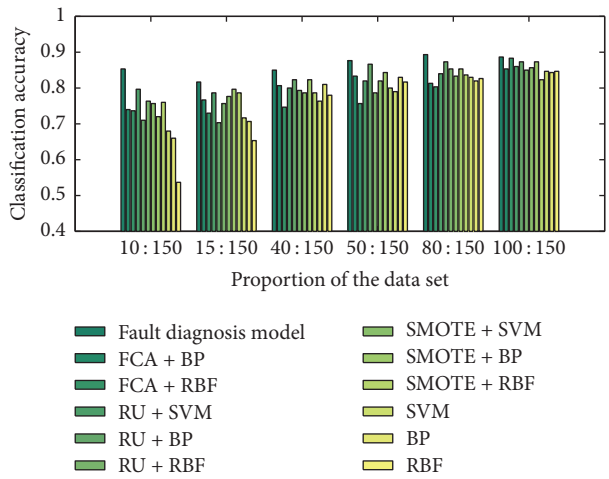

Figure 15: Classification accuracy comparisons.

With the proposed approach, the fault diagnosis model was obtained from the imbalanced rolling bearing data set. To test the classification accuracy of the model, testing samples consisting of 400 samples (100 samples from each mode) were constructed.

Figure 17 shows the confusion matrixes of the fault diagnosis model and other models. From Figure 17(a), the classification accuracy of the fault diagnosis model was 93.25%. Obviously, Figure 17(b) shows that the FCA + BP model was unable to identify the inner race and outer race failures. From Figure 17(c), the classification accuracy of the SMOTE + SVM model was 90.25%, less than the fault diagnosis model. Figure 17(d) shows the confusion matrix of the SVM model as the SVM model is confused with inner race failure and outer race failure. The reason for this situation was that the SVM model was trained by the imbalanced data set. Since the inner race and outer race failures were small classes, the hyperplane of the SVM model was biased to small classes. Therefore, the trained SVM model found it difficult to identify the inner race and outer race failures from the testing samples. In conclusion, the fault diagnosis model could distinguish the mode of the rolling bearing and obtain good classification accuracy, which proved the validity of the proposed approach.

6. Conclusions

In this paper, a kind of data-based approach was proposed. The experiment results showed that our proposed approach achieved better classification accuracy when compared to the other models. Some conclusions were obtained:

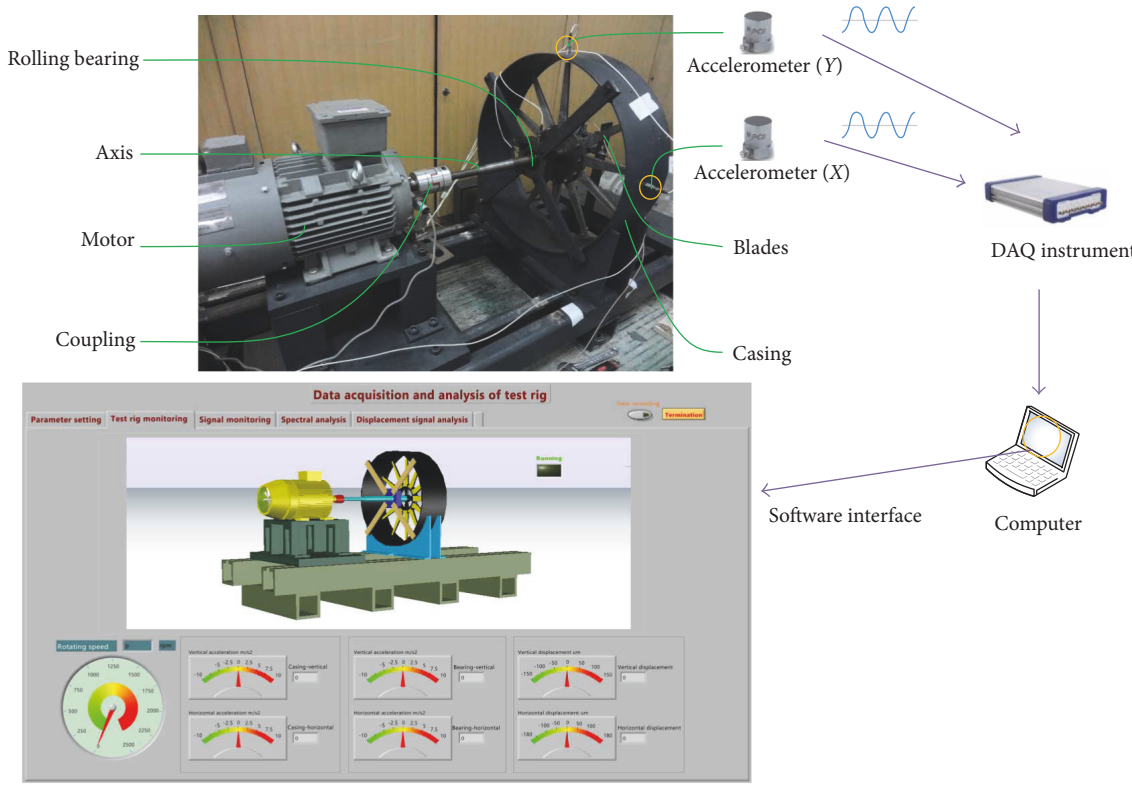

Figure 16: Rolling bearing fault simulation test bed.

Table 3: Classification accuracy comparisons of fault diagnosis model and other models.

Proportion of the data set	10 : 150	15 : 150	40 : 150	50 : 150	80 : 150	100 : 150
Fault diagnosis model	85.33%	81.67%	85.00%	87.67%	89.33%	88.67%
FCA + BP	74.00%	76.67%	80.67%	83.33%	81.33%	85.33%
FCA + RBF	73.67%	73.00%	74.67%	75.67%	80.33%	88.33%
RU + SVM	79.67%	78.67%	80.00%	82.00%	84.00%	86.00%
RU + BP	71.00%	70.33%	82.33%	86.67%	87.33%	87.33%
RU + RBF	76.33%	75.67%	79.33%	78.67%	85.33%	85.00%
SMOTE + SVM	75.67%	77.67%	78.67%	82.00%	83.33%	85.67%
SMOTE + BP	72.00%	79.67%	82.33%	84.33%	85.33%	87.33%
SMOTE + RBF	76.00%	78.67%	78.67%	80.00%	83.67%	82.33%
SVM	68.00%	71.67%	76.33%	79.00%	83.00%	84.67%
BP	66.00%	70.67%	81.00%	83.00%	82.00%	84.33%
RBF	53.67%	65.33%	78.00%	81.67%	82.67%	84.67%

Table 4: Composition of the rolling bearing data set.

Mode	Processing method	Fault size (width × depth) (mm)	Sample size
Normal	~	~	100
Rolling element failure	Line cutting	0.3 × 1	100
Inner race failure	Line cutting	0.3 × 0.5	5
Outer race failure	Line cutting	0.3 × 0.5	5

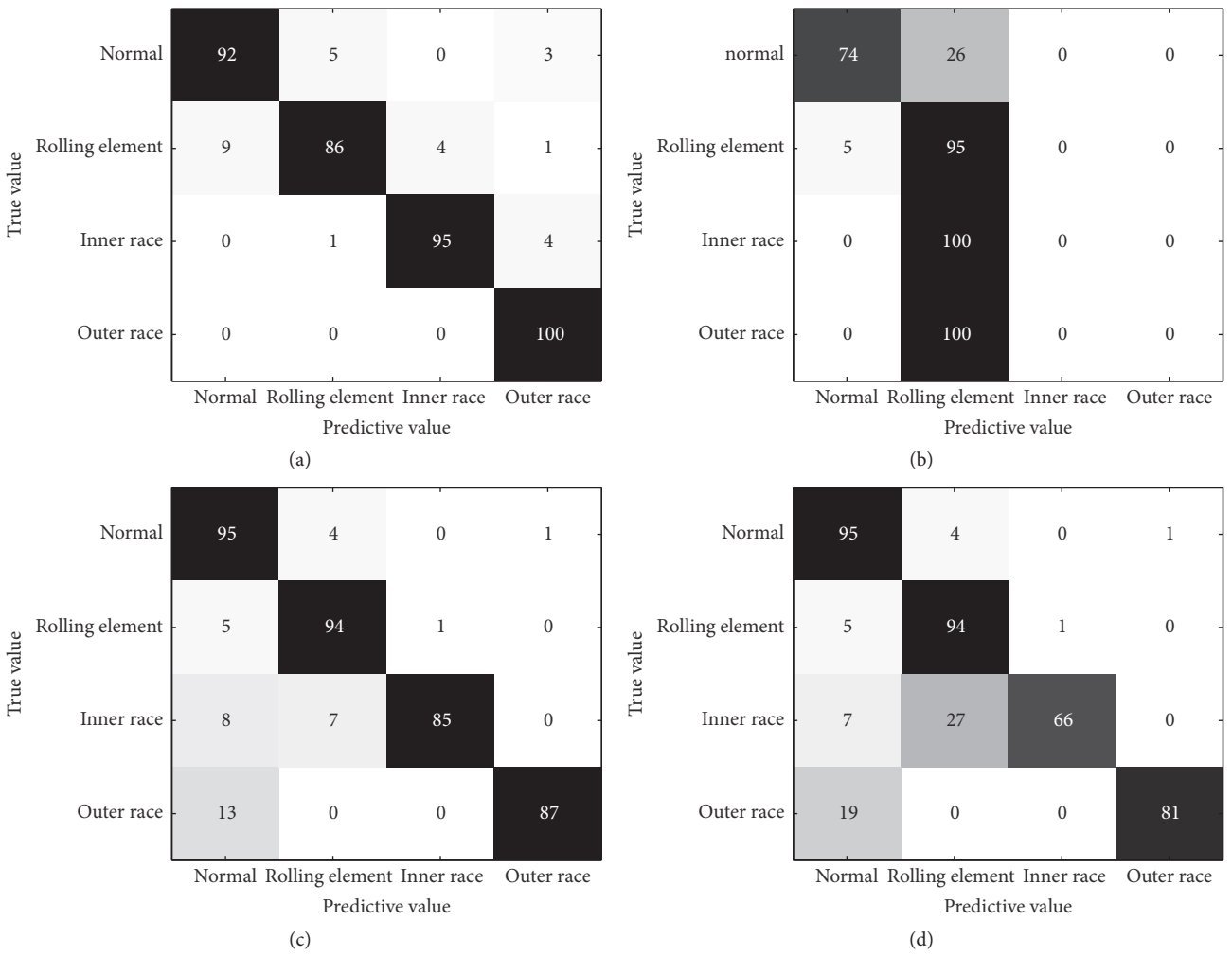

Figure 17: Confusion matrix comparisons: (a) fault diagnosis model; (b) FCA + BP; (c) SMOTE + SVM; (d) SVM.

(1) The signals of the accelerometers could be acquired to diagnose the rotating machinery fault.

(2) By combining VMD and PCA, sensitive features of the rotating machinery fault could be extracted.

(3) To diagnose the rotating machinery fault for imbalanced data, a kind of data-based approach was proposed in this paper. The fast clustering algorithm was adopted to reduce the number of the majority data from the imbalanced sensitive features. Then, the SVM was trained and tested with the data clustered by the fast clustering algorithm so the fault diagnosis model for the imbalanced data was obtained. The fault diagnosis model showed a very good classification capability in both the gearbox fault data set and rolling bearing fault data set. Therefore, our approach was suitable to the rotating machinery fault diagnosis for imbalanced data.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research is supported by National Natural Science Fund of China (11572167).

References

[1] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 108–126, 2013.

[2] S. Yang, D. Xiang, A. Bryant, P. Mawby, L. Ran, and P. Tavner, "Condition monitoring for device reliability in power electronic converters: A review," *IEEE Transactions on Power Electronics*, vol. 25, no. 11, pp. 2734–2752, 2010.

[3] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—a tutorial," *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, 2011.

[4] H. Wang, R. Li, G. Tang, H. Yuan, Q. Zhao, and X. Cao, "A Compound fault diagnosis for rolling bearings method based on blind source separation and ensemble empirical mode decomposition," *PLoS ONE*, vol. 9, no. 10, article e109166, 2014.

[5] Z. P. Feng, M. Liang, Y. Zhang, and S. M. Hou, "Fault diagnosis for wind turbine planetary gearboxes via demodulation analysis

based on ensemble empirical mode decomposition and energy separation," *Journal of Renewable Energy*, vol. 47, pp. 112–126, 2012.

[6] T. Y. Wu, J. C. Chen, and C. C. Wang, "Characterization of gear faults in variable rotating speed using Hilbert-Huang Transform and instantaneous dimensionless frequency normalization," *Mechanical Systems and Signal Processing*, vol. 30, pp. 103–122, 2012.

[7] S. Singh and N. Kumar, "Combined rotor fault diagnosis in rotating machinery using empirical mode decomposition," *Journal of Mechanical Science and Technology*, vol. 28, no. 12, pp. 4869–4876, 2014.

[8] N. Li, R. Zhou, Q. Hu, and X. Liu, "Mechanical fault diagnosis based on redundant second generation wavelet packet transform, neighborhood rough set and support vector machine," *Mechanical Systems and Signal Processing*, vol. 28, pp. 608–621, 2012.

[9] X. Zhang and J. Zhou, "Multi-fault diagnosis for rolling element bearings based on ensemble empirical mode decomposition and optimized support vector machines," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 127–140, 2013.

[10] X.-M. Tao, D.-X. Zhang, S.-Y. Hao, and D.-D. Fu, "SVM classifier for unbalanced data based on spectrum cluster-based under-sampling approaches," *Control and Decision*, vol. 27, no. 12, pp. 1761–1768, 2012.

[11] H. Yi, X. F. Song, B. Jiang, Y. F. Liu, and Z. H. Zhou, "Fault diagnosis based on self-tuning support vector machine in sample unbalance condition," *Transactions of Beijing Institute of Technology*, vol. 33, no. 4, pp. 394–398, 2013.

[12] Y. Zhang, X. Zhou, H. Shi, Z. Zheng, and S. Li, "Corrosion pitting damage detection of rolling bearings using data mining techniques," *International Journal of Modelling, Identification and Control*, vol. 24, no. 3, pp. 235–243, 2015.

[13] Y. Lan, W. Zong, X. Ding et al., "A two-step fault diagnosis framework for rolling element bearings with imbalanced data," in *Proceedings of the 13th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2016*, pp. 620–625, Xian, China, August 2016.

[14] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[15] S. M. Razavi Zadegan, M. Mirzaie, and F. Sadoughi, "Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets," *Knowledge-Based Systems*, vol. 39, pp. 133–143, 2013.

[16] A. Laio and A. Rodriguez, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[17] J. Xie, H. Gao, W. Xie, X. Liu, and P. W. Grant, "Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors," *Information Sciences*, vol. 354, pp. 19–40, 2016.

[18] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[19] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[20] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.

[21] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[22] Y. Wang and R. Markert, "Filter bank property of variational mode decomposition and its applications," *Signal Processing*, vol. 120, pp. 509–521, 2016.