IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 2, 2019, accepted April 17, 2019, date of publication June 21, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919406

Diagnosis of Rolling Bearing Based on Classification for High Dimensional Unbalanced Data

QI HANG[1], JINGHUI YANG[1], AND LINING XING[2]
[1]School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai 200120, China
[2]College of Information Systems and Management, National University of Defense Technology, Changsha 410073, China

Corresponding authors: Jinghui Yang (Jhyang@sspu.edu.cn) and Lining Xing (13874845346@qq.com)

This work was supported by the Discipline Construction of Mechanical Engineering, Shanghai Polytechnic University, under Grant XXKZD1603.

ABSTRACT Motor systems are becoming more and more vital in modern manufacturing and bearings play an important role in the performance of a motor system. Many problems that arise in motor operation are related to bearing faults. In many cases, the accuracy of the devices for monitoring or controlling a motor system highly depends on the dynamic properties of motor bearings. Thus, fault diagnosis of a motor system is inseparably related to the diagnosis of the bearing assembly. The fault diagnosis of rolling bearings is substantially a classification problem. The traditional application of random forest (RF) to fault diagnosis methods is based on balanced data. However, in a practical situation, it is difficult to collect the fault data that are usually unbalanced. In order to solve this problem, in the first step, we propose a two-step (TS) clustering algorithm to enhance the original synthetic minority oversampling technique (SMOTE) algorithm for the unbalanced data classification. Then, based on the improvement of the SMOTE algorithm, we propose the principal component analysis (PCA) and apply it in the field of high-dimensional unbalanced fault diagnosis data. In this paper, we apply this new method to the fault diagnosis of rolling bearings, and the experiments conducted in the end show that the improved algorithm has a better classification performance.

INDEX TERMS Fault diagnosis of rolling bearing, high dimensional unbalanced data, random forests.

I. INTRODUCTION

Nowadays, it generates a large amount of data in the field of finance, Internet and intelligent manufacturing. By studying the world's authoritative information, consulting and analysis company IDC proposes that the data would grow 50 times by 2020 [1]. In the era of big data, various decisions are inseparable from data mining and analysis. Moreover, the massive data generated by these practical applications often have features like imbalance and high dimensionality. How to store and extract important information and classify data has become a hot topic. Especially the solution of classification has been the highlight of data digging and has been used in broad fields such as medical imaging [2], fault detection [3], text categorization [4] and gene selection [5], [6]. Although traditional classification algorithms can achieve good results in low-dimensional data, it has a bad performance under the high-dimensional data. For example in text classification, data usually can hold thousands or even millions of dimensions [7]. If we deal with the original data directly, it usually comes up a model which is so complex that will easily end up in overfitting. Furthermore, the redundancy and noise interference that high-dimension data can't get rid of increases computation complexity and lengthen the training period [8]. Therefore, it is necessary to reduce the dimensionality of high-dimensional data which will improve the performance of the classification algorithm.

The unbalance data sample can be found in every corner of an industry. For example, in the telecommunications industry, the number of regular calls is far much bigger than fraud calls [9]. And in medical issue, the consequences of misdiagnosing a healthy man as having a cancer is not even close to misdiagnosing a cancer man as healthy [10]. And in case of identifying the nature of users, the number of regular users is larger than fraud users, and it is banks' job to find out the fraud users to avoid potential loss [11]. Therefore, the

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

identification of minority is important when handling the unbalance data. The traditional classification algorithms acquiescence that the numbers of samples of each category are the same and they often take the improvement of classification accuracy as priority. But the essence of handling the unbalance data focuses on the minority data which means that the traditional algorithms are not suitable for handling the unbalance data. It has been an emergency problem to improve the accuracy of identifying the minority data of the unbalance data.

However, the traditional classification algorithm is based on the balance data and generally aiming at improving the classification accuracy which often results in poor performance of minor classes. But in the unbalanced data, minor classes are what we should pay more attention to which means the traditional classification algorithm based on classification accuracy is not quite applicable in the field of dealing with unbalanced data. How to improve the recognition rate of minor classes in unbalanced data has become an urgent problem to be solved in data mining.

The random forest (RF) is integrated by the decision trees. It uses the bagging sampling method which randomly extracts the samples from the training samples. So when it meets the unbalanced data, due to the random extraction of data, the problem of imbalance gets more serious. It affects the performance of the decision tree in the random forest algorithm. When we train the high-dimensional data, it will contain a large number of redundant attributes. Furthermore, high-dimensional data sets often contain nonlinear characteristics but the decision tree can only be used to segment the attribute space by linear. Due to these two problems, random forest algorithm still has rooms for improvement.

The fault diagnosis of rolling bearing [12] is essentially a process of pattern recognition which means categorizing the data into normal or failure operation. However, the traditional fault diagnosis method based on random forest is under the condition of sample equalization. Although the random forest classifier shows good results for balanced data sets, in the condition of a high-dimensional unbalanced data set, the accuracy would drop especially in the field of fault diagnosis. Due to the difficulty in collecting and sorting, the number of fault samples is far less than normal samples which results in forming an unbalanced data set [13]. There are multiple sets of physical quantities in the fault diagnosis analysis [14], and for each time series data set, many frequency and time domain feature quantities can be extracted. If the classification is performed before the redundant and interfering data is removed, it will lead to heavy calculation and low accuracy. Above all, how to improve the classification performance of random forest algorithm in machine fault diagnosis has always been a main issue in research [15], [16]. In this paper, we use the high-dimensional unbalanced data of rolling bearing as the original data of research and propose a new fault diagnosis method. The main work is divided into three parts: firstly, we propose a two-step clustering algorithm (TS) to enhances the original synthetic minority oversampling technique (SMOTE) algorithm for the unbalanced data classification which can solve the shortcomings of using SMOTE algorithm alone, and we call this combined algorithm as TS-SMOTE algorithm. Secondly, we combine the principle component analysis (PCA) algorithm with the TS-SMOTE algorithm which we call it PCA-TS-SMOTE algorithm. We use the PCA algorithm to dimensionality reduction before the data is interpolated. Finally, this paper proposes PCA-TS-SMOTE-RF fault diagnosis method. That is to combine PCA-TS-SMOTE with random forest (RF) to diagnose faults under unbalance and high-dimensional data. The rolling bearing is used as the fault diagnosis object for experimental verification. The study indicates that this new fault diagnosis method, PCA-TS-SMOTE-RF, has better performance in setting evaluating indicator like recall, specificity, accuracy, AUC and G-mean than directly categorizing the original data by random forest algorithm or classifying after applying TS-SMOTE or PCA.

II. REVIEW

This section gives a review of classification algorithms and applies Random Forest algorithm to high dimensional unbalanced data as rolling bearing fault diagnosis method and analyses the conclusion..

A. CLASSIFICATION ALGORITHMS

Machine learning includes supervised learning [17], [18] and unsupervised learning [19], [20]. Classification problem belongs to supervised learning. Specifically, given a training sample, each sample X is used as an input, corresponding to an explicit Y as an output. At this time, a specific model is trained (mapping f: X –> Y), and then given an unknown sample X', a prediction of the result Y' is made. For example, the demands to tell apart whether the mail is spam or whether the user will purchase the product or whether the tumor is malignant or benign are basically classification problems, but when the response is a continuous variable, these demands turn into a regression problem. The classification algorithm is a method to solve classification problems, and is used to assign specific categories to data objects with unknown category. It includes training processes and testing processes:

Training Process: Training Set - Feature Selection - Training - Classifier

Test Process: Classifier - Test Set - Test - Classification Results - Evaluation

There are various traditional data classification methods and commonly used methods are Logistic Regression (LR) [21], Artificial Neural Network (ANN) [22], Support Vector Machine (SVM) [23], decision tree [24] and Ensemble Learning [25].

The LR algorithm uses sigmoid function and minimizes the loss function for classification by nonlinear mapping. The performance is equivalent to the decision tree and neural network. The LR algorithm runs fast and has high accuracy, suitable for large data sets.

ANN is a structure that mimics the synapses of the brain, composed of different neurons, and the network needs to be trained. Currently there are BP Propagation, Radial Basis Function (RBF), Generalized Regression Neural Network (GRNN), Probabilistic Neural Network (PNN), etc.

The SVM considers the support vector, implicitly maps the features to higher dimensional feature space by a kernel trick, and uses all the mapped features for classification. It constructed the hyperplane to minimize the structural risk and maximize the classification interval. So this algorithm is suitable for nonlinear, high-dimensional and local optimum problems.

The decision tree is a tree structure, constructed recursively from top root node to bottom leaves. Non-leaf nodes represent attribute features, and leaf nodes represent categories. Several algorithms generating such optimal trees have been proposed, such as ID3/4/5, and CART. The algorithms are simple to understand and interpret, and can be combined with other decision techniques [26].

In statistics and machine learning, ensemble learning methods by using multiple learning algorithms to obtain better predictive performance than what could be obtained from any of the constituent learning algorithms alone [27]. Bagging [28] boosting [29] and random forest algorithms are the only two ensemble learning algorithms which can reduce the error of a single classifier and have higher classification accuracy. Studies have shown that compared with ANN, regression tree and SVM, RF algorithm has higher stability and robustness, and proper training parameters can obtain better classification accuracy [30].

B. APPLY RANDOM FOREST ALGORITHM TO HIGH DIMENSIONAL UNBALANCED DATA

The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, and in Ho's formulation, it is a way to implement the ''stochastic discrimination'' approach to classification proposed by Kleinberg [31]. An extension of the algorithm was developed by Leo Breiman. He clearly defined the concept of random forest and also proved that random forests algorithm is very good at avoiding overfitting [32]. Selecting some of the decision trees to make up an ensemble algorithm, may be not only smaller in the size but also stronger in the generalization than ensembles generated by non-selective algorithms. At present, most ensemble algorithms utilize all the trained learners to make up an ensemble. Zhou and Tang proposes GASEN-b algorithm to show that when the learners are decision trees, it is better to build selective ensembles [33]. It is proposed in their paper that the combination of the KM-SMOTE algorithm and the RF algorithm to process the unbalanced data set [34]. Similarly, an algorithm combining RF and cure-smote presents a better performance than other traditional algorithms [35]. Zhou *et al.* [36] propose a feature selection algorithm based on random forest that incorporates the feature cost into the base decision tree construction process to produce low-cost features. For multi-class unbalanced data. It is shown in literature [37] that a combination of SMOTE and Bagging with Random Forest produced the best overall accuracy of minority class.

C. ROLLING BEARING FAULT DIAGNOSIS METHOD

1) OVERVIEW OF DATA-DRIVEN TROUBLESHOOTING

In general, fault diagnosis methods can be divided into three types: experience-based monitoring and detection, analysis based on model and data-driven fault diagnosis techniques.

To diagnose the malfunction of a large complicated system, it's not an easy task for company just relying on their experience but requires a long term program to build a sophisticated and professional knowledge base. And to diagnose the fault by modeling and analysis, we have to obtain machine's operation model precisely so that we can go through the whole system to find the disability. Therefore, fault diagnosis based on data reveals its advantages which makes up the shortage of modeling-based and experience-based diagnosis methods. Data-based diagnosis method only needs to process the data under the normal and abnormal working conditions of the machine followed by general pattern which is collecting data first and then diagnosing the causes of fault and finally classifying the original data. So, it draws broad attention academically and industrially [38]. Data-based diagnosis can be viewed as three categories: methods based on statistics, signal features processing and artificial intelligence [39].

(1) Method based on multivariate analysis. Multivariate analysis (MVA) is based on the statistical principle of multivariate statistics which is to consider the intrinsic relationship among all variables. By using the existed ways to extract the eigenvalue of the original data, such as mean value or variance, and setting threshold, we can monitor data fluctuation in real-time and tell the fault of machine instantly.

(2) Principle of characteristic signal processing: Firstly, collect the variable signal values which contain rich information in the production, and then extract and utilize the characteristics of these signal values, and finally the related processing techniques are applied to diagnose the faults in the frequency domain and the time domain. Since different fault signals can produce different spectral characteristics, common methods of signal processing include wavelet transform [40], spectral analysis [41].

(3) Machine learning algorithm. This method conducts artificial intelligence diagnosis by simulating the human decision-making progress which means using the computer to accomplish the decision-making task to the fault diagnosis without using a certain mathematical model. The most widely used machine learning classification algorithms are SVM, artificial neural networks, and decision trees.

2) MACHINE TOOL FAULT DIAGNOSIS BASED ON RANDOM FOREST ALGORITHM UNDER MACHINE LEARNING

Although support vector machine has better generalization ability in small sample, nonlinear dataspace, etc., its classification performance is poor under multi-dimensional and large sample setting. The neural network can handle large samples,

but it is easy to fall into local optimum and long learning time. The random forest algorithm is an integrated algorithm using unbiased estimation for generalization error, and because its feature subset is randomly selected, it can handle high-dimensional data for fault detection purposes. Based on the above advantages of random forests, its application in fault diagnosis is one of the current research hotspots.

In modern manufacturing, bearings, as necessary rolling elements, are important parts of machinery. It is also a frequent reason of equipment failures. The operation status of the machine directly affects the overall performance of the mechanical system, and troubleshooting can effectively prevent major accidents. Therefore, the state monitoring and fault diagnosis of bearings has extremely important practical significance [42]. It is proposed in their paper [43], [44] that most methods of fault diagnosis of bearings apply the traditional random forest algorithm, but the fault diagnosis method proposed by the original algorithm is carried out under the condition of high-dimensional data set without dimension reduction processing. Meanwhile, the original random forest algorithm is under the condition of balanced data set, but in actual production process the fault samples are unbalanced. Due to these reasons, when using the original random forest algorithm for rolling bearing fault diagnosis, the recognition rate of fault samples will drop which directly leads to poor performance of the classification. Therefore, it is meaningful to find a suitable classification method of fault diagnosis for high-dimensional unbalanced data.

III. BASICS CONCEPT OF RANDOM FOREST

This section describes the basics concept of random forest Algorithm in detail.

A. CLASSIFICATION ALGORITHMS

Random forest is an Ensemble Learning algorithm in the field of machine learning. It uses Bagging synthesis technology [28] to select m batches of samples with certain size from original dataset and generates m decision trees to form a random forest. The final decision is made by majority voting to aggregate the predictions of all the decision trees. The flow chart of random forests algorithm is shown as Figure 1.

There are two random procedures in RF. The first one is for each tree, it will randomly and reversibly extract N training samples from the training set. The training set for each tree is different and contains duplicate training samples. The second one is the method to inject randomness into the trees, so that features to be chosen for splitting the tree node can be random. Features are selected with non-replacement from the total features when the nodes of the trees are split. The size of the feature subset is usually far less than the size of the total features. Random procedures can help to reduce the correlation between tree classifiers in a random forests algorithm.

B. THE PROGRESS OF ALGORITHM

The best node for splitting can be computed by three methods: information gain, information gain rate and Gini coefficient,

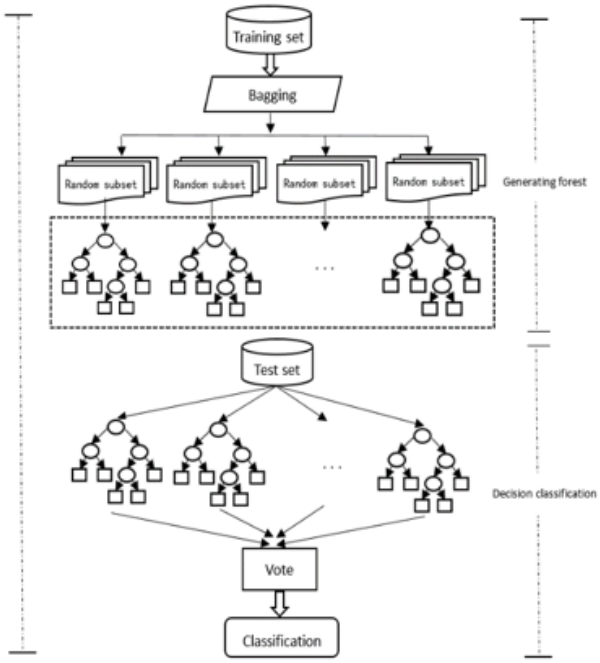

FIGURE 1. Flow chart of random forests algorithm.

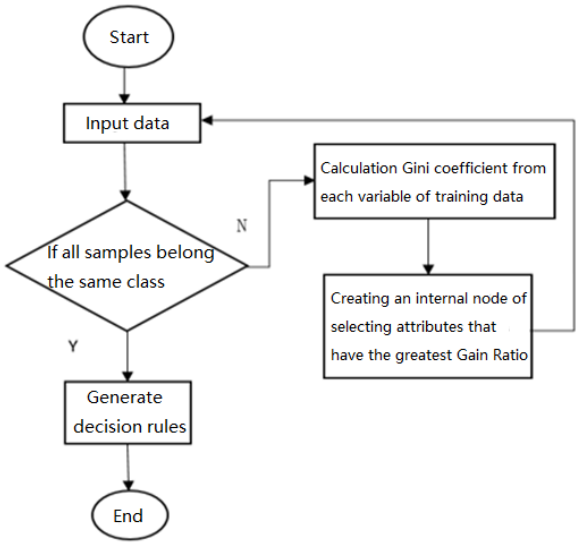

FIGURE 2. Flowchart of decision tree algorithm.

which correspond to ID3, C4.5 [45] and CART [46]. In this paper we use the CART method, which a smaller Gini coefficient indicates a better classification result. Assuming that there are n categories in the sample T, the formula to calculate the Gini index of the sample T is as follows:

$$Gini(\mathrm{T}) = \sum_{\mathrm{k}} = 1^{\mathrm{n}} p_k(1\text{-}p_k) = 1 - \sum_{\mathrm{k=1}}^{\mathrm{n}} p_k^2 \qquad (1)$$

P_k is defined as the proportion of samples in Kth category. We calculate the Gini coefficient of each feature and select

TABLE 1. Confusion matrix of dichotomous data.

Confusion Matrix	Classified positive	Classified negative
Positive	TP	FN
Negative	FP	TN

the one with the smallest Gini coefficient as the segmentation threshold point of decision tree. T represents the number of all samples which will be divided into m parts by the smallest Gini coefficient of feature A.

$$Gini(\mathrm{T,A}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \frac{|T_{\mathrm{i}}|}{|T|} Gini(T_i) \tag{2}$$

The CART decision tree is an unstable algorithm. The random forest algorithm uses the Bagging algorithm to form a random forest by generating different training sets to form mutually independent decision trees. Flowchart of Decision Tree algorithm is as follows:

C. PERFORMANCE EVALUATION CRITERIA

The measures of the quality of binary classification are built using a confusion matrix. According to the literature [35], it uses confusion matrix of dichotomous data which is showed in table 1.

In table 1, TP represents the number of positive samples that is classified as true by the model; TN represents the number of negative samples that is classified as true by the model; FP represents the number of positive samples that is classified as false by the model; FN represents the number of negative samples that is classified as false by the model.

The rate of recall (sensitivity) shows the classification accuracy of model to the positive samples. The formula to calculate the rate of recall is showed as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Specificity indicates the classification accuracy of model to negative samples. Its calculation formula is showed as follows:

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

Precision shows the proportion of the actual true positive samples to all the samples that model classifies as positive. Its formula is showed as follows:

$$\mathrm{Pr}\,ecesion = \frac{TP}{TP + FP} \tag{5}$$

Accuracy indicates the general classification accuracy of model. Its calculation formula is showed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

OBB error or out-of-bag error implies the classification performance of random forest. The lower average OBB error of n decision trees gets, means the better performance of

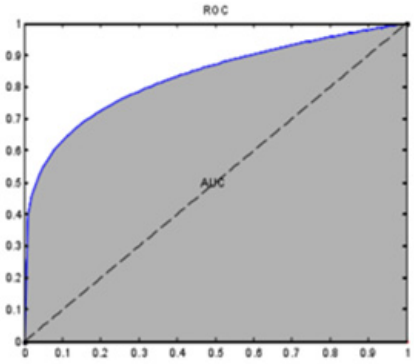

FIGURE 3. ROC curve and AUC.

RF algorithm. The following formula shows how OBB error is calculated:

$$OBB \quad \text{error} = \frac{\sum_{i}^{nT_{ree}} OBB\ error_i}{nTree} \tag{7}$$

Due to the imbalance of OBB error of each decision tree, the OBB error of samples with huge amount of data can take higher weight in the average OBB error which lower the reliability of the result. Literature [47] proposes an AUC-based permutation variable importance measure for random forests. To determine the value of AUC, we have to draw a ROC curve (Receiver Operating Characteristic curve) which goes through the original point and point (1,1), and the value of the area enclosed by ROC curve and axes is the value of AUC. Figure 3 shows the relationship between ROC curve and AUC value which the horizontal coordinate means False Positive Rate (FPR) and the vertical coordinate means True Positive Rate (TPR).

G-mean is another general comprehensive performance indicator that can efficiently evaluate the imbalanced data set. The value of G-mean depends on two factors: the rate of recall and specificity. Only when the value of recall and specificity both get bigger, the value of G-mean can get bigger which means a better performance of the classifier. The following formula shows how G-mean is calculated.

$$G - mean = \sqrt{Recall * Specificity} \tag{8}$$

IV. EXPERIMENT AND ANALYSIS

The purpose of the simulation experiment in this section is to compare the algorithm performance between balanced data and unbalanced data for different classifiers. The test data is selected from the UCI data sets iris and Breast-Cancer-Wisconsin. The iris data set contains 50 positive samples and 50 negative samples, which belong to the balanced data set. The Breast-Cancer-Wisconsin data set contains 243 positive samples and 459 negative samples, which belongs to the unbalanced data set. The details of data are shown in Table 2.

The two sets of data sets are used in the algorithm of random forest algorithm, support vector machine and artificial

TABLE 2. UCI data set.

Data	Number of dimension	Number of positive samples	Number of positive samples
iris	4	50	50
breast-cancer-wisconsin	10	243	459

TABLE 3. The result comparison of each algorithm applied to iris data set.

	Recall	Specificity	Accuracy	G-Mean	AUC
SVM	0.960	0.920	0.940	0.930	0.912
ANN	0.880	0.920	0.900	0.899	0.878
RF	0.960	0.880	0.920	0.910	0.901

TABLE 4. The result comparison of each algorithm applied to Breast-Cancer-Wisconsin data set.

	Recall	Specificity	Accuracy	G-Mean	AUC
SVM	0.82	0.90	0.91	0.85	0.84
ANN	0.94	0.82	0.84	0.80	0.81
RF	0.89	0.92	0.93	0.87	0.87

neural network respectively. The accuracy of each algorithm are show in table 3 and table 4.

In summary, when the data set is balanced, the SVM, ANN and RF algorithms performs better, and the classification advantage of the RF algorithm is not very obvious. However, when the data set is unbalance and the number of dimensions is too large, the performance of the classification algorithm will be affected, but the effect of the RF algorithm is still generally higher than other algorithms. It shows that the RF algorithm is RF algorithm is more adaptable in high dimensional unbalanced data.

V. CLASSIFICATION METHODS BASED ON UNBALANCED DATA

A. DEFINITION AND IMPACT OF UNBALANCED DATA

The unbalanced classification problem begins with the skewed distribution of data in different categories [48]. Imbalanced data sets generally refer to data that is distributed unevenly among different categories where the data in the smaller category is far less prevalent than the data in the larger category. The Imbalance Ratio (IR) is defined as the ratio of the number of minor class samples to the number of major class samples.

Unbalanced data is ubiquitous in many applications [49], [50]. For example, in the medical records used for disease diagnosis prediction, the number of rare but very important disease samples is much smaller than the number of common disease samples. The data used for Internet intrusion detection has more normal samples than the invasive samples. If traditional classifiers are applied to these scenarios without any pre-treatment of unbalanced data, the data in the categories of larger samples will overwhelm the data in smaller samples categories and will not achieve good classification results. Due to the imbalanced data, the training set for each decision tree will be imbalanced during the first ''random'' procedure, and the random forest algorithm will not become a good ''expert'' of the small samples. This leads to classification with high accuracy on large samples, in contrast to the low accuracy on small samples [51]. Therefore, the unbalanced data is a very important problem in data classification.

B. TS-SMOTE ALGORITHM

Several methods [52]–[54] have existed for processing unbalanced data, including over-sampling and under-sampling techniques. In particular, a type of synthesis resampling technique algorithm is called the synthetic minority oversampling technique (SMOTE) [55], [56], has a positive effect on the unbalanced data problem. The SMOTE algorithm is an improved algorithm based on the random sampling method in random forests. It is artificially synthesized by producing new samples according to the characteristics of a few samples in small categories. The SMOTE algorithm proposes a hypothesis based on the idea of clustering algorithms: samples that are closer to the positive class sample are also positive class samples. Based on this assumption, for any X1 n a minor category, the algorithm obtains the k–nearest neighbors of X from the whole data set, and then selects n samples randomly with replacement from the k-nearest neighbors, Denote these n samples by Q j (j $=$ 1, . . . , n), and the original data in minor category by X1, then the new sample Xj1 is defined by interpolation as follows:

$$X_{j1} = X_1 + U * (Q_j - X_1) \quad j = 1, \ldots, n \tag{9}$$

where U is a random number uniformly distributed within the range (0,1). Finally, new samples are generated by iterating formula (3) multiple times until the data become balanced. However, some flaws exist in the SMOTE algorithm. Firstly, the selection of a value for k is an open question, and it needs multiple iterations, increasing the computation burden of algorithm. Secondly, the artificial samples generated by the minor class samples at the edges may make a fuzzy boundaries between the positive and negative classes.

TS-SMOTE algorithm is an improved algorithm of SMOTE. Before inserting the new samples, the first step is to cluster the samples of the minor class by using two-step algorithm. During the clustering process of the TS-SMOTE algorithm, noisy points must be removed because they are far away from the normal points and hinder the merge speed in the corresponding class. Then, get the Cluster cores of each cluster sample and Calculate the centroid of all cluster cores. Finally, choose the minor class samples that is farthest from the centroid (representative original sample) and generate artificial samples randomly between representative original sample and the centroid.

Two-step cluster [57] is a hierarchical algorithm. It can automatic determine the number of clusters and handle large huge amount of data [58]. It is performed in two steps.

(1) Pre-clustering stage: Using hierarchical algorithms BIRCH (Balance Iterative Reducing and Clustering using hierarchical) algorithm, one of the hierarchical algorithm which comes from dealing with samples of large size, to derived into several sub-cluster.

(2) Clustering stage: Taking sub-cluster, the result of pre-clustering, as target and apply the agglomerative hierarchical clustering method to merge the sub-cluster until we get the expected number of clusters.

The general idea of the TS-SMOTE algorithm is as follows: cluster the samples of the minor class using two-step cluster, remove the noise and outliers from the original samples, and then, generate artificial samples randomly between representative point and the centroid. The implementation steps of the TS-SMOTE algorithm are as follows:

(1) Use the Two-Step Cluster Algorithm to cluster a small number of data and record its cluster cores and calculate the centroid of all cluster cores.

(2) Go through all the original data of minor class samples and find out the set of data which is the furthest to the centroid.

(3) Generate a new sample according to the interpolation formula. a_i represents the centroid of Cluster cores after clustering by the Two-Step Cluster algorithm. X_{max} represents set of data of original data which is furthest to the centroid.

$$X^* = a_i + rand(0, 1) * (X_{max} - a_i) \quad i = 1, \ldots, N \quad (10)$$

During the clustering process of the TS-SMOTE algorithm, noise points must be removed because they are far away from the normal points. For the sample points after clustering, the interpolation can effectively improve generalization ability. In the interpolation formula, X1 is replaced by the centroid of Cluster cores after clustering by the Two-Step Cluster algorithm; Q j is replaced by the data of original which is the furthest to the cluster cores. Consequently, the samples are generated only between the representative samples and the centroid of all cluster cores, which effectively avoids the influence of fuzzy boundary between positive and negative classes. The combination of the clustering and interpolation to eliminate the noise points at the end of the process and reduce the complexity. And this interpolation method allows all new samples to be obtained at once, this can also reduce the algorithm complexity. Moreover, the two-Step Cluster Algorithm can automatic determination of number of clusters, avoid setting the k value of the original SMOTE algorithm and thus, reduce the instability of the proposed algorithm.

C. EXPERIMENT AND ANALYSIS

The purpose of the simulation experiment in this section is to compare the accuracy between the TS-SMOTE algorithm and SMOTE algorithm. Data set uses the Breast-Cancer-Wisconsin data which used in Section 2.3. The TS-SMOTE algorithm proposed in this paper can generate samples near the center point and the representative point, avoiding the introduction of noise, and the generated sample follows the original distribution. The performance evaluation criteria of the data set under different sampling methods are shown in the table 5.

TABLE 5. The accuracy of SMOTE algorithm and TS-SMOTE algorithm.

	G-mean	AUC	OBB error
Smote	0.82	0.90	0.91
TS-Smote	0.94	0.82	0.84

From the table we can tell that the G-mean, AUC and OOB errors of the TS-SMOTE algorithm perform better than the smote algorithm under different sampling methods.

VI. BASICS CONCEPT OF RANDOM FOREST

A. RESEARCH CONCERNING DIMENSIONALITY REDUCTION

In recent years, with the rapid development of information technology, data acquisition technology and data storage capacity have been improved, resulting in high-dimensional unbalanced data in fields. The classifier becomes more complicate when meeting high-dimensional data. Those data make the classifier easy to over-fit. There are also some irrelevant or redundant attributes, which easily lead to bad classifier performance. Jimenez and Landgrebe [59] conducted an in-depth analysis of the geometric properties of high-dimensional attribute spaces. They pointed out that as the number of attributes increases, the data spreads in all directions, making the central data sparse. The increase in the dimension of data also increases the difficulty of analyzing the data exponentially. This phenomenon is what the scholars often call ''curse of dimensionality''.

According to the conclusion of literature [60], most of the high-dimensional space is empty, with most of the data lying in a low-dimensional subspace, so the high-dimensional data can be mapped to a low dimension by some methods. In this way, the data still maintains the original distribution. Feature extraction is the process of deriving new features from original features to reduce the cost of feature measurement, increase the efficiency of classifiers and allow higher accuracy. Therefore, for high-dimensional unbalanced data, we employ this technique in combination with the TS-SMOTE algorithm.

B. DIMENSIONALITY REDUCTION PRINCIPLE OF PCA

Principle Component Analysis (PCA) [61], [62] is one of the most used dimensionality reduction methods [63]. The basic idea is to apply orthogonal transformation on the high-dimensional data that turns the correlated variables into a new set of linearly independent variables, with descending variance [64]. The way to reduce dimension of data is to choose new orthogonal feature vectors with largest variances.

As linear combinations of original features, the new variables contain most of the information in features, due to their large variances, and also eliminate the correlation of the original data that affects the prediction accuracy. The following paragraph shows the basic steps of PCA.

(1) Calculate covariance matrix. In this first step, we have to represent each sample by a vector, and then calculate the sample covariance matrix of these vectors. The covariance matrix is of dimension n∗n, where n is the number of features in the original data.

(2) Get eigenvalues and eigenvectors of the covariance matrix. In this step, we sort the eigenvalues from largest to the smallest, $\lambda 1 > \lambda 2 > \lambda 3$.

(3) Find the accumulative contribution of eigenvectors and select the principle components. The accumulative contribution G(r) tells how much variation the directions of first r eigenvectors contribute to the variation of the whole data set. Formula 11 shows how G(r) is calculated and s represents the number of eigenvalues. Next we select the eigenvectors $V_1,\ldots,V_r$, corresponding to the largest eigenvalues such that accumulative contribution G(r) reaches 85% or above.

$$G(r) = \sum_{i=1}^{r} \lambda_i / \sum_{j=1}^{s} \lambda_j \quad (i, j = 1, 2, 3 \ldots) \tag{11}$$

VII. EXPERIMENT ON FAULT DIAGNOSIS OF ROLLING BEARING BASED ON PCA-TS-SMOTE-RF

A. THE PRINCIPLE OF PCA-TS-SMOTE-RF

In chapter 3.2, we have applied TS-SMOTE algorithm to solve the problem of unbalanced data and fuzzy boundaries between the positive and negative classes caused by interpolation and heavy computation. However, TS-SMOTE algorithm interpolates data randomly in minor class, which may affect distribution of original data and intervene the result of prediction. Therefore, we use PCA before interpolation to reduce the dimension of features. This step is to erase the data from category with few samples that mixed or close to category with many samples and thus ensure the consistency between interpolated data and original data. Based on the features of PCA and TS-SMOTE, this chapter combines PCA-TS-SMOTE algorithm with random forest to classify the fault data of rolling bearings.

The steps of applying PCA-TS-SMOTE-RF algorithm are as follows:

(1) Using PCA to reduce the dimensions of eigenvectors. PCA will select several principle components from the top contribute rates based on accumulative contribute rate. And we use the selected principle components as the input matrix for TS-SMOTE algorithm.

(2) Interpolation for the class with few samples by TS-SMOTE algorithm. Interpolation for the input matrix of TS-SMOTE algorithm which can avoid the shortage of the original SMOTE algorithm. This step aims to balance the number of samples of each category and make the ratio of the number of major class samples to the number of minor class samples.

(3) Use random forest algorithm to classify the processed data set.

B. EXPERIMENTAL DATA

This paper uses the data downloaded from the NASA website to bring up PCA-TS-SMOTE-RF algorithm for bearings fault diagnosis. This series of data comes from the whole life time experiment of rolling bearings carried out by center for intelligent maintenance systems of Univ. of Cincinnati [65]. The experiment took samples of time domain acceleration signal every ten minutes under 20KHz. We take one set of this experiment data which started at ten thirty-two and thirty-nine second a.m. on 12/2/2004 and ended at six twenty-two and thirty-nine second a.m. on 19/2/2004. This set of data recorded every acceleration Signals that revealed the very stages of bearing being failure. The capacity of this set of data is 984.

C. FEATURE EXTRACTION AND PCA DIMENSIONALITY REDUCTION

Due to the huge amount of monitoring data of rolling bearings and the noise interference, we need to extract the eigenvalue of original data. The amplitude parameters of time domain signal are often used in condition monitoring and fault diagnosis for motor systems. We establish the connection between input and output of fault diagnosis for motor spindle based on TS-SMOTE-RF algorithm. Those formula below presents the amplitude parameters this paper extracts. N represents the total number of monitoring data samples, x_i represents the value of each sample.

Kurtosis factor:

$$C_q = \frac{\frac{1}{N}\sum_{i=1}^{N}(|x_i| - \bar{x})^4}{X_{rms}^4} \tag{12}$$

Peak factor:

$$I_p = \frac{X_p}{X_{rms}} \tag{13}$$

Pulse factor:

$$C_f = \frac{X_p}{\bar{X}} \tag{14}$$

Skewness factor:

$$C_w = \frac{\frac{1}{N}\sum_{i=1}^{N}(|x_i| - \bar{x})^3}{X_{rms}^3} \tag{15}$$

Kurtosis:

$$\beta_q = \frac{1}{N}\sum_{i=1}^{N}(|x_i| - \bar{x})^4 \tag{16}$$

Skewness:

$$\beta_w = \frac{1}{N}\sum_{i=1}^{N}(|x_i| - \bar{x})^3 \tag{17}$$

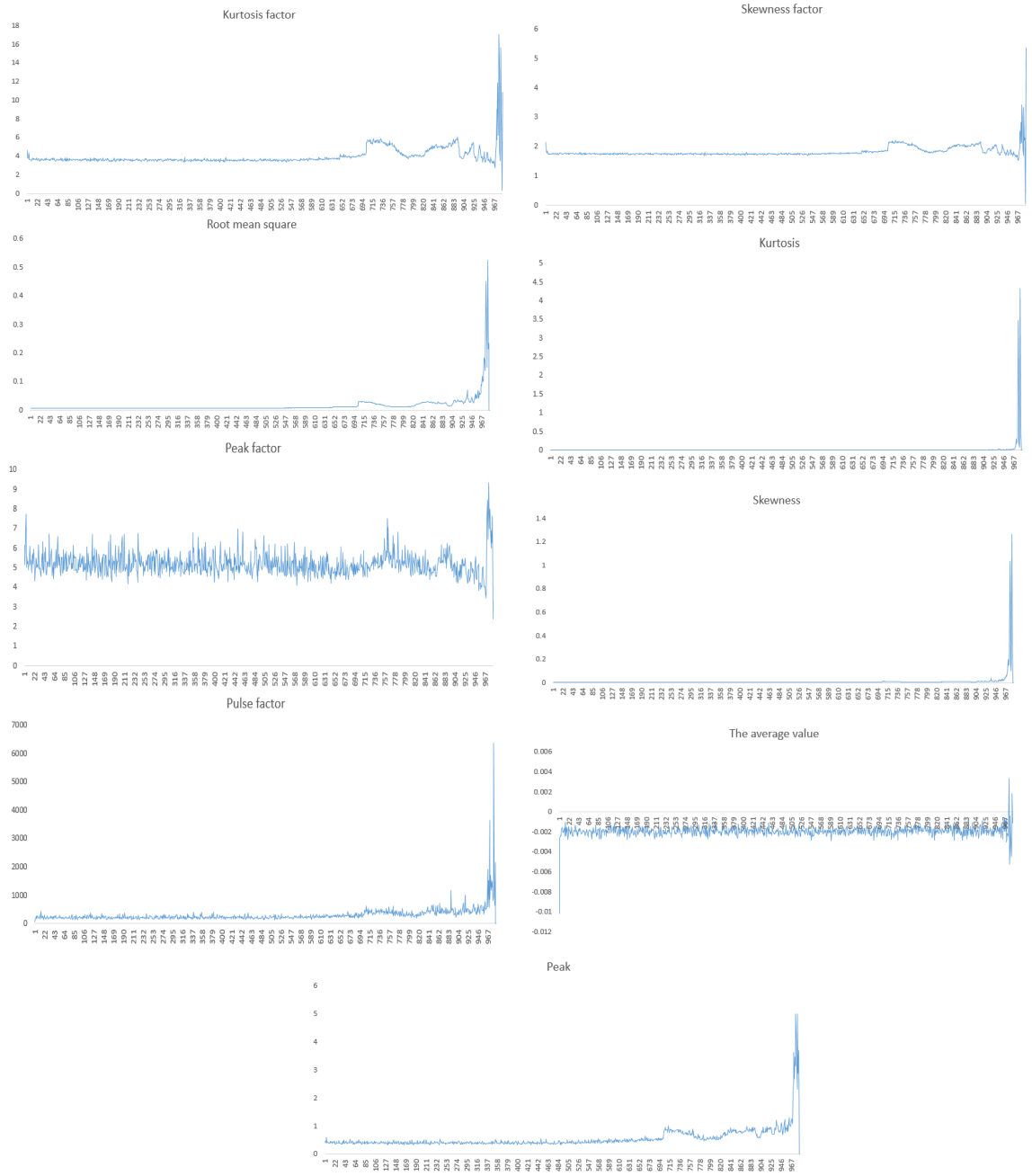

FIGURE 4. Line charts of transformation of time domain feature parameters.

The average value:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{18}$$

Peak:

$$X_p = \frac{\sum_{i=10} \max\{x_i\}}{10} \tag{19}$$

Root mean square:

$$X_{rms}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{20}$$

The original data of every set comes from samples taken by every ten minutes, thus, we can get 984 sets of data totally. After calculating the parameters, we select the value of parameter as vertical coordinate and samples' serial number as horizontal coordinate (at range from 1 to 984 in chronological order) to make line charts (Figure 4). It is not difficult to tell that some of the parameters have similar trend from the line charts. So to reduce the redundancy and improve the accuracy of prediction, we use PCA to decrease the covariates' dimensions that will generate new covariates which are linear combination of old covariates. The total contribution of variance of selected components reaches 85%.

TABLE 6. Contribution rates and eigenvalues of feature components.

	Component1	Component2	Component3	Component4
Total contribution rate (%)	56.196	13.562	10.837	9.264
the total of initial eigenvalue	5.058	1.221	0.975	0.834

TABLE 7. Component matrix.

	Component			
Bearing No.1	1	2	3	4
Kurtosis factor	.888	-.237	-.031	.313
Root mean square	.948	.098	-.074	-.119
Peak factor	.399	-.257	.668	-.394
Pulse factor	.599	.421	.491	.161
Skewness factor	.685	-.384	-.034	.571
Kurtosis	.832	.100	-.398	-.300
Skewness	.904	.089	-.302	-.275
The average value	.070	.864	.020	.221
Peak	.931	.007	.175	-.014

In this study, we choose first four parameters of which total contribution reaches 89.859%. Table 6 shows the contribution rates and eigenvalue of feature components and table 7 shows the component matrix.

After determining four principle components, we also need to identify the expression of every principle component to display the linear relationship between each eigenvalue and principle components. Formula 21 presents the calculate method of every element in coefficient matrix of principle components and formula 22 shows what the coefficient matrix looks like. In formula 16, C_{ij} represents the effect factor (component matrix) of eigenvalue i to the principle component j and T_j represents the total of initial eigenvalue of principle component j. Formula 23-26 indicates expressions of every four principle components.

$$X_{ij} = \frac{C_{ij}}{\sqrt{T_j}} \quad (i, j = 1, 2, 3, 4 \ldots) \tag{21}$$

$$\begin{pmatrix} X_{11} & X_{21} & X_{31} & \ldots & X_{i1} \\ X_{12} & X_{22} & X_{32} & \ldots & X_{i2} \\ X_{13} & X_{23} & X_{33} & \ldots & X_{i3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1j} & X_{2j} & X_{3j} & \ldots & X_{ij} \end{pmatrix} \tag{22}$$

Formula 23:

$$F_1 = 0.394 \cdot x_1 + 0.421 \cdot x_2 + 0.177 \cdot x_3 + 0.266 \cdot x_4 + 0.304 \cdot x_5 + 0.369 \cdot x_6 + 0.401 \cdot x_7 + 0.031 \cdot x_8 + 0.413 \cdot x_9$$

Formula 24:

$$F_2 = -0.214 \cdot x_1 + 0.088 \cdot x_2 - 0.232 \cdot x_3 + 0.380 \cdot x_4 - 0.347 \cdot x_5 + 0.090 \cdot x_6 + 0.080 \cdot x_7 + 0.781 \cdot x_8 + 0.006 \cdot x_9$$

Formula 25:

$$F_3 = -0.031 \cdot x_1 - 0.074 \cdot x_2 + 0.676 \cdot x_3 + 0.497 \cdot x_4 - 0.034 \cdot x_5 - 0.403 \cdot x_6 - 0.305 \cdot x_7 + 0.020 \cdot x_8 + 0.177 \cdot x_9$$

Formula 26:

$$F_4 = 0.342 \cdot x_1 - 0.130 \cdot x_2 - 0.431 \cdot x_3 + 0.176 \cdot x_4 + 0.625 \cdot x_5 - 0.328 \cdot x_6 - 0.301 \cdot x_7 + 0.241 \cdot x_8 - 0.015 \cdot x_9$$

D. EXPERIMENT ON FAULT DIAGNOSIS OF ROLLING BEARING BASED ON PCA-TS-SMOTE-RF

Though analyzing the line charts (Figure 4) of every time domain feature parameters, we can first define every stage of motor spindle's working condition from good to failure. The indicator value remains stable until the 694th data, where there is a sudden increase. After 694, there is some fluctuation, followed by a sharp increase and drop. Therefore, 694 is chosen as the initial failure point of rolling bearing. There are 290 fault samples in the whole original sample, which is much smaller than the normal samples.

Considering huge data set and various parameters in diagnosing fault of rolling bearings, we prefer to combine the more universal PCA algorithm with TS-SMOTE algorithm which has strong variability and then apply random forest algorithm as classifier to fault diagnosis of rolling bearings. The components matrix we get in chapter VI is used as the input of TS-SMOTE algorithm to balance the unequal quantity of each category of original data. For the categories with minor class samples after data point 649, the point when bearing begins to failure, we expand this category by interpolation so that the number of categories with little data (Tr-) can be close to those with lots of data set (Tr+). And next, we build random forest model to diagnose the fault based on new generated data and set the traditional value with fixed parameters.

E. RESULTS AND ANALYSIS

In order to verify PCA-TS-SMOTE-RF classification performance, this paper compare the performance evaluation criteria of random forest (RF), PCA-RF, TS-SMOTE-RF with PCA-TS-SMOTE-RF algorithms. The result are shown in Table 8.

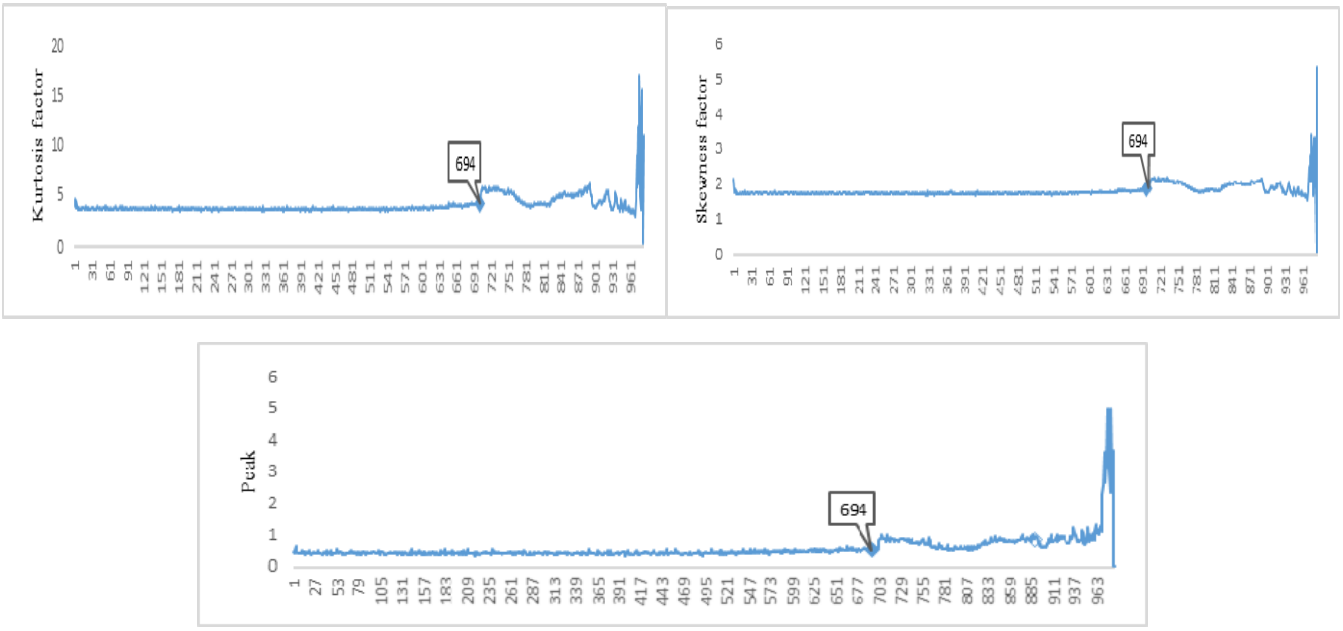

FIGURE 5. Line charts of transformation of characteristics.

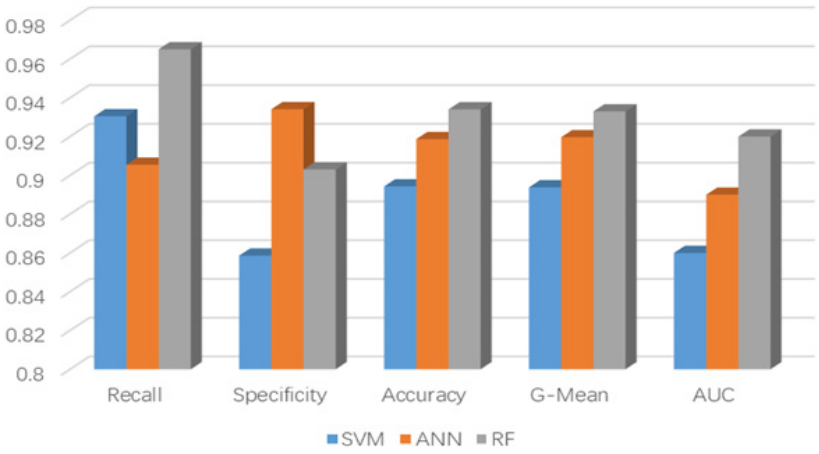

FIGURE 6. Experiment results of different algorithms based on PCA-TS-SMOTE.

From the classification results obtained by the different sampling algorithms discussed in Table 4, the Specificity, Accuracy, G-mean and AUC achieved by PCA-TS-SMOTE-RF are superior to the other sampling algorithms, and its Recall is slightly lower. The best value of every performance evaluation criteria obtained by the algorithms are marked in boldface. At the meantime, to prove the excellence of random forest algorithm, we apply PCA-TS-SMOTE on both SVM and ANN algorithm, and Table 9 shows the value of each indicator of these three algorithms.

In order to show the results of different algorithms clearly, we use bar chart, as Figure 6, to present the value of each indicator between different algorithm.

TABLE 8. Experiment results of different algorithms.

	Recall	Specificity	Accuracy	G-Mean	AUC
RF	0.979	0.629	0.809	0.784	0.62
PCA-RF	0.981	0.701	0.841	0.829	0.74
TS-SMOTE-RF	0.973	0.735	0.837	0.845	0.81
PCA-TS-SMOTE-RF	0.965	0.903	0.934	0.933	0.92

In conclusion, the classification results of the PCA-TS-SMOTE-RF algorithm as measured by the Specificity, Accuracy, G-means and AUC are substantially enhanced, whereas

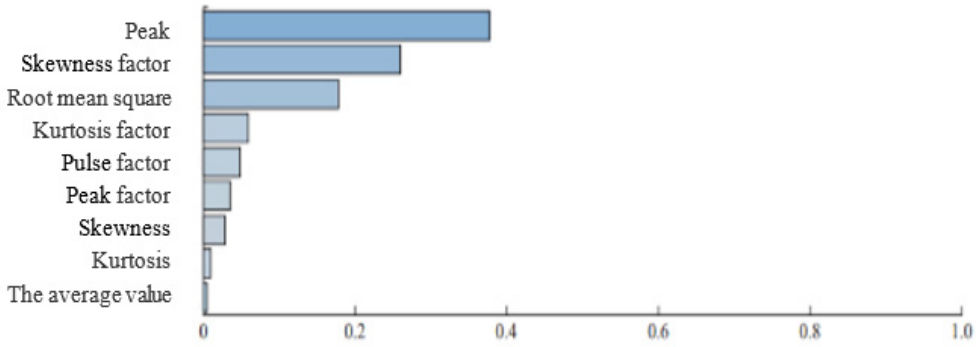

FIGURE 7. The importance of different parameters for PCA-KNN model.

TABLE 9. The results of different algorithms based on PCA-TS-SMOTE.

	Recall	Specificity	Accuracy	G-Mean	AUC
SVM	0.930	0.859	0.894	0.894	0.860
ANN	0.905	0.934	0.919	0.920	0.890
RF	0.965	0.903	0.934	0.933	0.920

the results using TS-SMOTE-RF or PCA-RF alone are not particularly stable. Thus, the PCA-TS-SMOTE-RF algorithm combined with RF has a substantial effect on classification. At meanwhile, through table 9 and Figure 6, it can tell that PCA-TS-SMOTE has better comprehensive performance than the improved SVM and ANN. In addition, Random Forrest algorithm can also calculate the prediction error from OOB (Out of Bag) data. Each tree in the Random Forrest is extracted randomly from original data by Bagging algorithm. Every extraction there will be one third of original data won't be extracted and these data are called OOB data. To each feature, apply OOB data to every tree to calculate the prediction error and then add interference noise which means to randomly alter its Eigenvalues to calculate the noise error. The average error of all kind which we calculate before and after adding the interference noise is the estimated value of the importance of this certain characteristic variable. The bigger the estimated value is, the deeper this characteristic variable can affect the evaluation process. Figure 7 shows the estimated value of the importance of characteristic variables by OOB. The horizon axis presents the average of prediction error of variables; and we can tell from the figure that three variables which has the biggest average of error are peak, skewness and root mean square.

VIII. CONCLUSION

In the era of big data, data tends to be characterized by high dimensionality and imbalance. If the traditional classification algorithm is used to classify it directly, the performance of the classifier will behave bad. And among the domestic and foreign research, these two characteristics are often studied separately, considering the imbalance in the high dimension, or directly studying the imbalance to ignore the high dimension. This paper combines the improved SMOTE algorithm and PCA dimension reduction to solve the problem of high dimension and imbalance of data in the model. Random forests are superior to other classification algorithms in processing classification performance. However, in the face of high-dimensional unbalanced big data, traditional random forest algorithms will have shortcomings such as long time to modeling and sensitive to unbalance data. Therefore, it is necessary to improve random forests to suitable for classification of high dimensional unbalanced data. Based on the rolling bearing data, this paper proposes a PCA-TS-SMOTE-RF algorithm to improve the classification accuracy of fault diagnosis. The experimental data of the rolling bearing life cycle provided by the Intelligent System Maintenance Center of the University of Cincinnati was used to verify the classification and prediction, and the superiority of the algorithm was proved. The steps of algorithms this paper proposes to solve the high dimensional unbalanced data are showed as follows:

(1) To classify the unbalanced data set, TS-SMOTE algorithm shows its excellent classification performance in balancing the data set than SMOTE algorithm. And the classification performance can be move on further when combine with PCA algorithm.

(2) PCA-TS-SMOTE algorithm efficiently avoids changing the data distribution pattern after interpolation by the path that using PCA to decrease data dimension first and then interpolation to balance the data of each category.

(3) During the classification experiment of fault diagnosis data, it is showed apparently that PCA-TS-SMOTE-RF algorithm achieves a much better result in every evaluation for fault diagnosis by comparing with using random forest directly, classifying by TS-SMOTE algorithm and only applying PCA. And it is also better than the SVM and ANN algorithm after applying PCA-TS-SMOTE.

REFERENCES

[1] J. Zhai, S. Zhang, and C. Wang, "The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 3, pp. 1009–1017, Jun. 2017.

[2] D. Seo, J. Ho, and B. C. Vemuri, "Covariant image representation with applications to classification problems in medical imaging," *Int. J. Comput. Vis.*, vol. 116, no. 2, pp. 190–209, Jan. 2016.

[3] Z. He, L. Fu, S. Lin, and Z. Bo, "Fault detection and classification in EHV transmission line based on wavelet singular entropy," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2156–2163, Oct. 2010.

[4] C. Liu, W. Wang, G. Tu, Y. Xiang, S. Wang, and F. Lv, "A new centroid-based classification model for text categorization," *Knowl. Based Syst.*, vol. 136, pp. 15–26, Nov. 2017.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[6] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, "Gene selection for tumor classification using neighborhood rough sets and entropy measures," *J. Biomed. Inform.*, vol. 67, pp. 59–68, Mar. 2017.

[7] D. Ö. Şahin, N. Ateş, and E. Kiliç, "Feature selection in text classification," in *Proc. 24th Signal Process. Commun. Appl. Conf. (SIU)*, Zonguldak, Turkey, May 2016, pp. 1777–1780.

[8] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56–62, Sep. 2017.

[9] A. Almas, M. A. H. Farquad, N. S. R. Avala, and J. Sultana, "Enhancing the performance of decision tree: A research study of dealing with unbalanced data," in *Proc. 7th Int. Conf. Digit. Inf. Manage.*, Macau, China, Aug. 2012, pp. 7–10.

[10] M. F. Ganji, M. S. Abadeh, M. Hedayati, and N. Bakhtiari, "Fuzzy classifcation of imbalanced data sets for medical diagnosis," in *Proc. 17th Iranian Conf. Biomed. Eng. (ICBME)*, Isfahan, Iran, Nov. 2010, pp. 1–5.

[11] X. W. Chen and M. Wasikowski, "Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Nevada, USA, 2008, pp. 124–132.

[12] V. Purushotham, S. Narayanan, and S. A. N. Prasad, "Multi-fault diagnosis of rolling bearing elements using wavelet analysis and hidden Markov model based fault recognition," *NDT E Int.*, vol. 38, no. 8, pp. 654–664, 2005.

[13] W.-Y. Chen, J.-X. Xu, and S. K. Panda, "A study on automatic machine condition monitoring and fault diagnosis for bearing and unbalanced rotor faults," in *Proc. IEEE Int. Symp. Ind. Electron.*, Gdansk, Poland, Jun. 2011, pp. 2105–2110.

[14] T. W. Rauber, F. De A. Boldt, and F. M. Varejão, "Heterogeneous feature models and feature selection applied to bearing fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 637–646, Jan. 2015.

[15] W. Yan, "Application of random forest to aircraft engine fault diagnosis," in *Proc. Conf. Comput. Eng. Syst. Appl.*, Beijing, China, Oct. 2006, pp. 468–475.

[16] R. Shrivastava, H. Mahalingam, and N. N. Dutta, "Application and evaluation of random forest classifier technique for fault detection in bioreactor operation," *Chem. Eng. Commun.*, vol. 204, no. 5, pp. 591–598, May 2017.

[17] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Mach. Learn.*, vol. 56, nos. 1–3, pp. 209–239, 2004.

[18] X. Zhu and A. Goldberg, "Introduction to semi-supervised learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 3, no. 1, p. 130, Jan. 2009.

[19] H. B. Barlow, "Unsupervised Learning," *Neural Comput.*, vol. 1, no. 3, pp. 295–311, 1989.

[20] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, Jan. 2001.

[21] M. Maalouf, "Logistic regression in data analysis: An overview," *Int. J. Data Anal. Techn. Strategies*, vol. 3, no. 3, pp. 281–299, Jul. 2011.

[22] S. R. Bhatikar, C. DeGroff, and R. L. Mahajan, "A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics," *Artif. Intell. Med.*, vol. 33, no. 3, pp. 251–260, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365704001204

[23] A. Rakotomamonjy, "Variable selection using SVM-based criteria," *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, Mar. 2003.

[24] E. Ekinci and H. Takçi, "Using authorship analysis techniques in forensic analysis of electronic mails," in *Proc. 20th Signal Process. Commun. Appl. Conf. (SIU)*, Mugla, Turkey, Apr. 2012, pp. 1–4.

[25] B. Verma and A. Rahman, "Cluster-oriented ensemble classifier: Impact of multicluster characterization on ensemble classifier learning," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 605–618, Apr. 2012.

[26] M. Z. F. Nasution, O. S. Sitompul, and M. Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4. 5 classification," *J. Phys., Conf. Ser.*, vol. 978, no. 1, Mar. 2018, Art. no. 012058.

[27] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmos. Environ.*, vol. 80, pp. 426–437, Dec. 2013.

[28] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[29] K. Tieu and P. Viola, "Boosting image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Hilton Head Island, SC, USA, Jun. 2000, pp. 228–235.

[30] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, Dec. 2015.

[31] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[32] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[33] Z. H. Zhou, W. Tang, and Z. Zhou, "Selective ensemble of decision trees," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (Lecture Notes in Computer Science), vol. 2639. Apr. 2003, pp. 476–483.

[34] B. Chen, Y. Su, and S. Huang, "Classification of imbalance data based on KM-SMOTE algorithm and random forest," *Comput. Technol. Develop.*, vol. 34, pp. 17–21, Sep. 2015.

[35] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *Bmc Bioinf.*, vol. 18, no. 1, p. 169, Mar. 2017.

[36] Q. Zhou, H. Zhou, and T. Li, "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features," *Knowl. Based Syst.*, vol. 95, pp. 1–11, Mar. 2016.

[37] T. J. Lakshmi and C. S. R. Prasad, "A study on classifying imbalanced datasets," in *Proc. 1st Int. Conf. Netw. Soft Comput. (ICNSC)*, Guntur, India, Aug. 2014, pp. 141–145.

[38] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process," *J. Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.

[39] Z. Ge, Z. Song, and F. Gao, "Review of recent research on data-based process monitoring," *Ind. Eng. Chem. Res.*, vol. 52, no. 10, pp. 3543–3562, 2013.

[40] J. Chen, Z. Li, J. Pan, G. Chen, Y. Zi, J. Yuan, B. Chen, and Z. He, "Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vols. 70–71, pp. 1–35, Mar. 2016.

[41] L. Ciabattoni, F. Ferracuti, A. Freddi, and A. Monteriù, "Statistical spectral analysis for fault diagnosis of rotating machines," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4301–4310, May 2018.

[42] M. He and D. He, "Deep learning based approach for bearing fault diagnosis," *IEEE Trans. Ind. Appl.*, vol. 53, no. 3, pp. 3057–3065, May/Jun. 2017.

[43] Q. Yao, J. Wang, L. Yang, H. Su, and G. Zhang, "A fault diagnosis method of engine rotor based on Random Forests," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Ottawa, ON, Canada, Jun. 2016, pp. 1–4.

[44] Z. Wang, Q. Zhang, J. Xiong, M. Xiao, G. Sun, and J. He, "Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5581–5588, Sep. 2017.

[45] J. R. Quinaln, *C4.5: Programs for Machine Learning*, vol. 16, no. 3. Burlington, MA, USA, Morgan Kaufmann, Sep. 1994, pp. 235–240.

[46] B. S. Everitt and D. Howell, *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ, USA: Wiley, 2005.

[47] S. Janitza, C. Strobl, and A.-L. Boulesteix, "An AUC-based permutation variable importance measure for random forests," *BMC Bioinf.*, vol. 14, no. 1, p. 119, Apr. 2013.

[48] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[49] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[50] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.

[51] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 14, no. 1, p. 103, Mar. 2013.

[52] 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," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
[53] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in *Proc. 19th IEEE Int. Conf. Tools Artif. Intell.*, Patras, Greece, Oct. 2007, pp. 310–317.
[54] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Kamel MS. Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, Jun. 2009.
[55] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[56] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-unbalanced data," *BMC Bioinf.*, vol. 14, no. 1, pp. 1–16 Mar. 2013.
[57] I. N. Martínez, J. M. Morán, and F. J. Peña, "Two-step cluster procedure after principal component analysis identifies sperm subpopulations in canine ejaculates and its relation to cryoresistance," *J. Androl.*, vol. 27, no. 4, pp. 596–603, Jul. 2006.
[58] V. Babic, J. Vancetovic, S. Prodanovic, V. Andjelkovic, M. Babic, and N. Kravic, "The identification of drought tolerant maize accessions by two-step cluster analysis," *Romanian Agricult. Res.*, vol. 29, pp. 53–61, Jan. 2012.
[59] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 28, no. 1, pp. 39–54, Feb. 1998.
[60] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Trans. Evol. Comput.*, vol. 4, no. 2, pp. 164–171, Jul. 2002.
[61] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Autom. Control*, vol. AC-26, no. 1, pp. 17–32, Feb. 1981.
[62] H. P. Deutsch, "Principle component analysis," in *Derivatives and Internal Models*. 2004, pp. 615–623.
[63] E. Dobriban, "Sharp detection in PCA under correlations: All eigenvalues matter," *Ann. Statist.*, vol. 45, no. 4, pp. 1810–1833, Feb. 2017.
[64] H. Hosoya and A. Hyvärinen, "Learning visual spatial pooling by strong PCA dimension reductio," *Neural Comput.*, vol. 28, no. 7, pp. 1249–1264, Jul. 2016.
[65] National Aeronautics and Space Administration. (Nov. 10, 2018). *Acoustics and Vibration Database. IEEE PHM 2012 Data Challenge Bearing Dataset*. [Online]. Available: http://data-acoustics.com/measurements/bearing-faults/bearing-6/

QI HANG graduated from the Department of Civil Engineering, Anhui Xinhua University, China, in 2016. She is currently pursuing the master's degree in environmental engineering with Shanghai Polytechnic University. Her research interests include machine learning, intelligent manufacturing, and pattern recognition.

JINGHUI YANG graduated from the Department of Electronic Science and Technology, East China Normal University. She received the M.E degree from the Department of Automation, Dalian University of Technology, and the Ph.D. Diploma degree in management science and technology from the Department of Management, Dalian University of Technology, in 2005. She is currently a Professor with the School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, China. Her research interests include enterprise system integration, intelligent manufacturing systems, scheduling, and optimization.

LINING XING graduated from Xi'an Jiaotong University. He received the M.E. and Ph.D. Diploma degrees from the Department of Control Science and Engineering, National University of Defense Technology, China. He is currently a Professor with the College of Information Systems and Management, National University of Defense Technology. His research interests include system planning, resource scheduling, and management decision systems.