A Fault Diagnosis Method based on Improved Synthetic Minority Oversampling Technique and SVM for Unbalanced Data

MingHong Han, Yaman Wu, Yunfeng Huang and Yumin Wang

1School of Reliability and Systems Engineering, Beihang University, Beijing, China

E-mail: hanminghong@buaa.edu.cn

Abstract. Equipment usually breaks down suddenly and irregularly, so most of the data sets obtained for fault diagnosis have unbalanced characteristics, and the amount of data varies greatly from different fault types. In this paper, three problems in the application of synthetic minority oversampling technique (SMOTE) are studied, and the improved SMOTE algorithm combined with support vector machine (SVM) is proposed. The validity of the model is verified by CWRU bearing data compared with SVM and SMOTE+SVM methods, and the result of fault diagnosis is satisfactory.

1. Introduction

With the development of computers, the number and types of data continue to increase, the classification of unbalanced data has received extensive attention in various fields of society. When facing unbalanced data sets, the classification surface of traditional classifiers will be biased towards the side of minority samples, so the classifier is likely to incorrectly classify minority samples into majority samples. This will result in lower classification accuracy for minority samples. In the field of fault diagnosis, it is very difficult to collect some fault signals, but the minority sample data often implies key information. Misclassification of minority samples may lead to catastrophic consequences.

In order to increase the accuracy of the classifier when facing unbalanced data, Chawla et al.[1] proposed the synthetic minority oversampling technique (SMOTE) method. The basic idea of this method is to artificially increase the number of minority samples according to certain rules, so that the ratio of the two types of samples reaches a balance. Compared with the traditional oversampling algorithm, the SMOTE algorithm can effectively alleviate the overfitting problem in the classification process and improve the classification performance of the classifier to a certain extent. H He et al.[2] innovatively proposed an adaptive synthetic sampling method. H Han et al.[3] proposed the Borderline-SMOTE sampling method. The core idea of this method is to oversample the minority samples near the classification boundary, so that the classification boundary between the majority class samples and the minority class samples is clearer. J Mathew et al. [4] proposed a weighted kernel-based SMOTE, which uses oversampling in the feature space of the SVM classifier to overcome the limitation of SMOTE on the nonlinear problem. Bunkhumpornpat C et al.[5] proposed the Safe-Level-SMOTE algorithm. By
setting a safety factor for each minority sample, the newly synthesized sample is located in a safe area as much as possible.

Support vector machine (SVM) is usually considered as a classification method, but it can be used for both classification and regression problems. SVM generates the best hyperplane in an iterative manner to separate different classes. The core idea of the SVM algorithm is to increase the dimension of the linear inseparable feature space to make it linearly separable. In this way, an optimal interval classifier can be constructed in a high-dimensional space to achieve the classification goal. Siamese Support Vector Machine is a binary data classifier proposed by Jayadeva et al. [6]. Its basic idea is to generate a classification plane for the majority class sample and the minority class sample respectively, and then optimize on the basis of these two classification planes to construct a classification hyperplane. This method has a more efficient training speed than SVM, but lacks sufficient theoretical support, and its application in practice needs further research. The ranking support vector machine proposed by Herbrich et al. [7] is widely used in the field of information retrieval, and its core is how to establish a corresponding damage function for specific problems to improve the accuracy of the ranking. In the field of fault diagnosis, Han et al. [8] proposed to combine the LMD algorithm with SVM to improve the accuracy of diagnosis, and Liu et al. [9] applied a multi-scale entropy algorithm on this basis. Lin et al. [10] proposed the joint application of fuzzy mathematics and support vector machines. The basic idea is to add membership to sample data, and to reduce the influence of noise samples on the optimum hyperplane by giving support vectors and non-support vectors different membership degrees.

This paper investigates the current mainstream algorithms for processing unbalanced data, selects artificial synthesis of minority oversampling technology (SMOTE) and support vector machine (SVM), builds the SMOTE-SVM classifier model. The corresponding optimization and improvement are carried out for the main defects of the SMOTE algorithm. Finally, use the improved SMOTE+SVM classifier to train the fault sample data sets, and then realize the detection and recognition of the fault.

2. Algorithm principle

2.1. SMOTE

The core idea of the SMOTE algorithm is to add a certain number to the minority samples, to balance the minority samples, and to synthesize minority samples, to balance. The basic principle of SMOTE is shown in Figure 1:

![Figure 1. SMOTE algorithm interpolation principle diagram.](image)

The specific steps for implementing the SMOTE algorithm are as follows: suppose there are A minority samples in the data set and the up-sampling rate is B (integer), then the SMOTE algorithm will synthesize A*B new minority samples. First, randomly select a sample i from the minority samples, use its feature vector \( x_i \) as the root sample of the new sample, and the Euclidean distance method is used to find the k nearest neighboring samples of the same category of \( x_i \) from all A samples of the minority class, randomly select a sample from the k neighboring samples as an auxiliary sample for the new
sample, and then use the equation (1) to perform linear interpolation between the sample \( x_i \) and the auxiliary sample, repeat the above steps \( B \) times to obtain \( B \) newly synthesized minority samples.

\[
x_{\text{new}} = x_i + (x_{ij} - x_i) \times \gamma
\]

\( \gamma \) is a random number between \([0,1]\); \( x_{ij} \) is the \( j \)-th neighboring sample of sample \( x_i \); \( j=1, 2...k \); \( x_{\text{new}} \) is a sample obtained by interpolation between the samples \( x_{ij} \) and \( x_i \).

Although the SMOTE algorithm effectively alleviates the over-fitting problem caused by random copying samples by artificially synthesizing minority samples, it also has some shortcomings. First, when the root sample or auxiliary sample is a noise sample, the newly synthesized sample will also become a noise sample with a high probability. Secondly, if both the root sample and the auxiliary sample are located at the boundary of the two classes, the new sample synthesized by linear interpolation will also fall in the overlapping part of the two classes, making the boundary between classes more difficult to distinguish. Finally, the areas with dense distribution of minority samples are more dense after being processed by SMOTE algorithm, resulting in the reduction of classification accuracy of minority samples.

2.2. ISMOTE

In order to improve the classification effect of SVM and make the classification boundary clearer, this paper makes certain improvement to the SMOTE algorithm. In this paper, the optimized and improved SMOTE algorithm is called ISMOTE. The improvement requires the following operations for each minority class sample:

1. Arbitrarily select a minority sample \( x_i \), calculate its \( K \) nearest neighboring samples, and assume that there are \( A \) majority samples among the \( K \) nearest neighboring samples;

2. When \( K=A \), that is, when the \( K \) nearest neighboring samples of the minority sample \( x_i \) are all majority samples, divide the minority sample \( x_i \) into noise samples and eliminate them;

3. If \( 0 \leq A < K/2 \), that is, when the majority of the \( K \) neighboring samples of the minority sample \( x_i \) account for less than half, we consider the sample to be a safe sample and store it in the Safe set;

4. If \( K/2 \leq A < K \), that is, when the majority of the \( K \) neighboring samples of the minority sample \( x_i \) account for more than half of the majority, we consider the sample to be a boundary sample and store it in the Border set;

5. For each sample \( y_i \) in the Border set, calculate the \( K \) nearest neighboring samples of \( y_i \);

6. Two samples \( y_{i1} \) and \( y_{i2} \) are randomly selected from the \( K \) neighboring samples of \( y_i \), and then a point \( p \) is randomly determined between the two sample points using equation (2), and \( \gamma \) is a random number between 0 and 1;

7. Finally, use equation (3) to synthesize new sample points between \( p \) point and \( y_i \);

8. Repeat this process for each sample point in the Border set until the number of newly synthesized minority samples reaches the required number.

\[
y_{p} = y_i + (y_{i2} - y_{i1}) \times \gamma
\]

\[
y_{p} = y_i + (y_{i2} - y_{i1}) \times \gamma
\]

\[
y_{\text{new}} = y_i + (y_{p} - y_i) \times \gamma
\]
2.3. SVM
Support vector machines (SVM) are mainly used for binary classification data. It is a kind of linear classifier, and its function is to find a classification surface with the largest interval in a given feature space to distinguish two types of samples.

Assuming that there is a hyperplane $\omega \cdot x + b = 0$ that can accurately divide the sample set $\{(x_i, y_i) \mid i = 1, 2, ..., z\}$, where $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $z$ is the number of training sample points, and $\mathbb{R}^n$ is the $n$-dimensional real number vector space. The optimum hyperplane is the classification surface with the largest distance from the closest sample point of the two types of sample points, and the sample points closest to the hyperplane are called support vectors. The schematic diagram of the support vector machine is shown in Figure 3. As shown, the optimum hyperplane is only determined by the support vector, so the maximum segmentation hyperplane problem of the support vector machine model can also be expressed as an optimization problem with the following constraints:

$$\min \frac{1}{2} \|w\|^2$$

s.t. $y_i(w^Tx + b) \geq 1, i = 1, 2, ..., n$
(3) Use ISMOTE to sample minority samples to reduce the unbalanced rate of data;
(4) Use support vector machine for training and complete classification.

![Flowchart](image)

**Figure 4.** ISMOTE+SVM fault diagnosis flowchart.

4. ISMOTE+SVM method application

4.1. Evaluation index
The former classification algorithms generally use the classification accuracy of the overall sample data to measure the performance of the algorithm, but for unbalanced data, the classification accuracy is no longer applicable to evaluate the performance of the algorithm. When facing the sample data set with high imbalance ratio, if all the samples of minority class are divided into majority class by the classifier, the classification accuracy of majority class samples is still very high, so this paper uses F-value and G-mean values to verify the performance of the model. The calculation formulas for these values are obtained according to the confusion matrix shown in Table 1.

| Predicted minority | Predicted majority |
|--------------------|--------------------|
| Actual minority    | TP                 | FN                |
| Actual majority    | FP                 | TN                |

TP is the number of samples which are minority samples and are divided into minority classes. TN is the number of samples that are majority samples and are divided into majority classes. FN is the number of samples which are minority samples but are divided into majority classes. FP is the number of samples which are majority samples but are divided into minority classes. The definition of the equations are as follows:

\[ A = \frac{TP}{TP + FN} \]  
\[ B = \frac{TN}{FP + TN} \]  
\[ C = \frac{TP}{TP + FP} \]  
\[ G - mean = \sqrt{A \times B} \]
\[
F-value = \frac{(2 \times A \times C)}{(A + C)}
\]  

(8)

Among them, A stands for the accuracy of the classifier to recognize the minority samples, B stands for the accuracy of the classifier to recognize the majority class samples, and C represents the proportion of the samples divided into minority groups that are actually minority class samples. Suppose A is a variable value, no matter the value of B (belonging to 0 to 1), the final value of G-mean cannot be affected. Similarly, the value of A also cannot determine the final size of the G-mean value. Formula (8) shows that F-value is more focused on the evaluation of the classification accuracy of minority samples than G-mean value. When the prediction accuracy and precision of minority samples are relatively high, the value of F-value will be higher.

4.2. Application

This paper uses the bearing data set of Case Western Reserve University (CWRU) to conduct application research on the optimized and improved fault diagnosis method. The experimental device is shown in Figure 5.

![Figure 5. Physical image of rolling bearing fault simulation test bench.](image)

In order to verify the performance of ISMOTE+SVM in practical applications, this paper selects 500 normal samples from the CWRU data set, 500 fault samples of inner ring, outer ring and rolling element with diameter 7, and 500 fault samples of inner ring, outer ring and rolling element with diameter 14. Then mix the normal samples with the outer ring fault samples, the inner ring fault samples, and the rolling element fault samples at a ratio of 5:1, and use the ISMOTE+SVM classifier for training. This paper uses a cross-validation method, and split the training set and the test set at a ratio of 4:1. During the experiment, the SVM algorithm selects a Gaussian kernel, the gamma value (width) is 1, the penalty factor C is 1000, and the number of neighbour of the ISMOTE algorithm is 5.

From Table 2 to Table 7, we can clearly see that the values of F-value and G-mean of bearings with different diameters have been greatly improved by using SMOTE algorithm to pre-process data than without using SMOTE algorithm in six cases of normal and inner ring fault, outer ring fault and rolling element fault. This is because the traditional SVM algorithm in the face of unbalanced data, due to the large difference between the number of samples, the classification hyperplane of SVM algorithm will tend to the side of less samples, which may lead to some minority samples being misjudged as majority samples, which greatly reduces the accuracy of the algorithm. The SMOTE algorithm artificially synthesizes minority samples to make the number of minority samples and majority samples close to balance, thereby improving the classification performance of the SVM algorithm. In addition, we can also see that compared with the SMOTE+SVM algorithm, the F-value and G-mean values obtained after training with the ISMOTE+SVM algorithm have been significantly improved, which means the ISMOTE+SVM algorithm has a great advantage over the SMOTE+SVM algorithm in the field of fault detection. This is because the ISMOTE algorithm recognizes the samples located in the overlap area of the majority class and the minority class samples, and deletes the minority class samples (noise samples) that completely fall in the majority class region, making the class boundaries of the two types of samples clearer. Moreover, the ISMOTE algorithm also effectively alleviates the over-generalization problem of
the original SMOTE algorithm. By over-sampling the fault samples at the boundary, the number of fault samples at the boundary is greatly increased, which means that when the SVM algorithm is used for training, the number of support vectors required to construct the classification hyperplane is more and the possibility of hyperplane offset is reduced. This effectively improves the algorithm's recognition rate of normal samples and fault samples. As mentioned above, using the ISMOTE+SVM algorithm for fault diagnosis of unbalanced data sets can indeed improve the accuracy of fault diagnosis to a certain extent than the traditional SVM algorithm and SMOTE+SVM algorithm. For a bearing with a diameter of 7, when the normal sample and the inner ring fault sample are mixed, comparing the SVM algorithm and the SMOTE algorithm, the F-value of the ISMOTE algorithm has increased by 27.1% and 6.3%, and the G-mean value has increased by 21.2% and 10.3%; when the normal sample and the outer ring fault sample are mixed, the F-value of the ISMOTE algorithm increases by 27.1% and 8.4%, and the G-mean value increases by 21.4% and 10.1% respectively; when the normal sample and the rolling element fault sample are mixed, the F-value of the ISMOTE algorithm increases by 24.3% and 9.1%, and the G-mean value increases by 18.1% and 11.5% respectively. For a bearing with a diameter of 14, when the normal sample and the inner ring fault sample are mixed, comparing the SVM algorithm and the SMOTE algorithm, the F-value of the ISMOTE algorithm has increased by 27.7% and 13.6% respectively, and the G-mean value has increased by 21.5% and 9.7%; when the normal sample and the outer ring fault sample are mixed, the F-value of the ISMOTE algorithm increases by 27.0% and 12.4%, and the G-mean value increases by 21.1% and 10.9% respectively; when the normal sample and the rolling element fault sample are mixed, the F-value of the ISMOTE algorithm increases by 24.1% and 10.5%, and the G-mean value increases by 22.8% and 9.2%.

Table 2. Comparison of G-mean value and F-value value of each algorithm (normal sample and inner ring faulty sample, diameter 7).

| Type          | F-value | G-mean |
|---------------|---------|--------|
| SVM           | 0.7647  | 0.7738 |
| SMOTE+SVM     | 0.9142  | 0.8506 |
| ISMOTE+SVM    | 0.9722  | 0.9380 |

Table 3. Comparison of G-mean value and F-value value of each algorithm (normal sample and outer ring faulty sample, diameter 7).

| Type          | F-value | G-mean |
|---------------|---------|--------|
| SVM           | 0.7493  | 0.7580 |
| SMOTE+SVM     | 0.8791  | 0.8390 |
| ISMOTE+SVM    | 0.9527  | 0.9241 |

Table 4. Comparison of G-mean value and F-value value of each algorithm (normal sample and rolling element failure sample, diameter 7).

| Type          | F-value | G-mean |
|---------------|---------|--------|
| SVM           | 0.7619  | 0.7901 |
| SMOTE+SVM     | 0.8682  | 0.8369 |
| ISMOTE+SVM    | 0.9472  | 0.9330 |

Table 5. Comparison of G-mean value and F-value value of each algorithm (normal sample and inner ring faulty sample, diameter 14).
| Type          | F-value | G-mean |
|--------------|---------|--------|
| SVM          | 0.7258  | 0.7460 |
| SMOTE+SVM    | 0.8162  | 0.8260 |
| ISMOTE+SVM   | 0.9271  | 0.9063 |

Table 6. Comparison of G-mean value and F-value value of each algorithm (normal sample and outer ring faulty sample, diameter 14).

| Type          | F-value | G-mean |
|--------------|---------|--------|
| SVM          | 0.7429  | 0.7601 |
| SMOTE+SVM    | 0.8397  | 0.8298 |
| ISMOTE+SVM   | 0.9438  | 0.9205 |

Table 7. Comparison of G-mean value and F-value value of each algorithm (normal sample and rolling element failure sample, diameter 14).

5. Conclusion
In the field of fault detection, due to the randomness and irregularity of mechanical equipment failures, the data sets obtained through fault diagnosis often have unbalanced characteristics. The classification performance of traditional fault detection algorithms in the face of unbalanced data sets is greatly reduced, and it cannot meet our requirements for detection accuracy. In view of this, this paper studies the bearing fault diagnosis technology for unbalanced data, tries to optimize and improve SMOTE algorithm, and combines it with SVM algorithm. Compared with the SVM algorithm and the SMOTE+SVM algorithm, the diagnostic accuracy of the ISMOTE+SVM algorithm is approximately increased by 24% and 10% respectively.

References
[1] Chawla N V, Bowyer K W, Hall L O, and et.al. 2002 SMOTE: synthetic minority over-sampling technique J. Journal of artificial intelligence research, 16 321-57.
[2] He H, Bai Y, Garcia E A, and et.al. 2008 Adaptive synthetic sampling approach for unbalanced learning. J. IEEE International Joint Conference on Neural Networks.
[3] Han H, Wang W Y and Mao B H. 2005 Borderline-SMOTE: a new over-sampling method in unbalanced data sets learning International conference on intelligent computing. Springer, Berlin, Heidelberg 878-87.
[4] Mathew J, Pang C K, Luo M and et.al. 2017 Classification of unbalanced data by oversampling in kernel space of support vector machines J. IEEE transactions on neural networks and learning systems, 29(9) 4065-76.
[5] Bunkhumpornpat C, Sinapiromsaran K and Lursinsap C. 2009 Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class unbalanced problem Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg 475-82.
[6] Khemchandani R and Chandra S. 2007 Twin support vector machines for pattern classification J. IEEE Transactions on pattern analysis and machine intelligence, 29(5) 905-10.
[7] Herbrich R. 2000 Large margin rank boundaries for ordinal regression J. Advances in large margin classifiers, 115-32.

[8] Han M and Pan J. 2015 A fault diagnosis method combined with LMD, sample entropy and energy ratio for roller bearings J. Measurement, 76 7-19.

[9] Liu H and Han M. 2014 A fault diagnosis method based on local mean decomposition and multi-scale entropy for roller bearings J. Mechanism and Machine Theory, 75 67-78.

[10] Lin C F and Wang S D. 2002 Fuzzy support vector machines J. IEEE transactions on neural networks, 13(2) 464-71.