Research of System Fault Diagnosis Method Based on Imbalanced Data

QingYu Zhu[1], *Hengyu Liu*[2], *Junling Wang*[3], *Shaowei Chen*[4], *Pengfei Wen*[5], *Shengyue Wang*[6]
1. China Aero-polytechnology establishment, Beijing, PR China (zqy985@126.com)
2. Northwestern Polytechnical University, Xi'an, PR China (liuhengyu@mail.nwpu.edu.cn)

***Abstract*—For the training of multiple models, the performance of single algorithm will be unstable, and an adaptive imbalance classification algorithm is proposed. The algorithm combines the area under curve (AUC) value to optimize the support vector description algorithm, the random forest algorithm and the gradient boosting decision tree algorithm respectively to generate the sub-model for classification. This paper uses the model fusion method to generate the sub-model for fusion. Finally, this paper selects the optimal sub-model to get stable classification results based on the AUC value. The algorithm in this paper is verified on the equipment running data set published by the American Prognostic and Health Management (PHM) Society in 2015. The result shows that the fault recognition rate is higher than that of the single imbalance classification algorithm. The classification effect is superior and the performance is stable.**

Keywords-imbalanced classification; AUC; adaptable model fusion

I. INTRODUCTION

In recent years, with the development of science and technology and the advance of industrialization, modern industrial equipment, such as rail vehicles, aerospace objects, is gradually becoming larger, more intelligent and more complex. At the same time, the fault diagnosis for equipment is becoming more important. Once the complex equipment breaks down, it will cause a shutdown in the production, thereby causing financial loss to the enterprise, or even major accidents, casualties and the destruction of the ecological environment [1]. For example, on October 29, 2018, Indonesian Lion Air JT610 flight crashed into the outer seas of northern Java, killing 189 people. The official accident report showed that six faults had been found before the accident. In 2017, there were 219 coal mine accidents in China, with 375 deaths. Fault diagnosis is a vital method to ensure safe, reliable and stable operation of equipment.

At present, most fault diagnosis systems need abundant and complete historical data and establish diagnosis models by machine learning. However, in actual industrial processes, the acquisition of such data often takes a long time, especially for complex industrial equipment. It is very difficult to obtain the data in different states of the equipment due to the shortage of fault data. The problem of data shortage in fault diagnosis belongs to imbalanced data diagnosis.

Currently, the researches on the diagnosis of imbalanced data mainly focus on two aspects [2]: i) At the data level, the data sampling method is adopted to adjust the distribution of original data and to reduce the imbalanced ratio of imbalanced data sets; ii)At the algorithm level, the problem of fault diagnosis of imbalanced data can be solved by some existed algorithms, such as one class learning, cost-sensitive learning and improving the traditional diagnosis algorithm, etc.

As can be seen from previous researches, the most common approaches at the data level include the artificial small sampling technique--synthetic minority oversampling technic (SMOTE) proposed by Gong Chunlin and Gu Liangxian [3]. A guided up-sampling method proposed by Shengguo Hu, Yanfeng Liang, Lintao Ma and Ying He [4]. Xuan Tho Dang, Dang Hung Tran, Osamu Hirose and Kenji Satou proposed an up-sampling method [5] that integrates data information (such as data set distribution, imbalance factor, inter-class distance, etc.), one-sided selection (OSS) algorithm [6] and neighborhood cleaning rule (NCL) [7]. The proposed methods at the algorithm level include the cost-sensitive support vector machine (SVM) algorithm proposed by Zhang Pengxiang, Liu Limin and Ma Zhiqiang [8], AdaCost algorithm proposed by Hamed Masnadi Shirazi and Nuno Vasconcelos [9], the support vector data description (SVDD) algorithm proposed by Raskutti and Kowalczyk [10] and isolation forest [11], etc.

The current researches have solved the imbalanced fault diagnosis problem from different aspects, with respective advantages and disadvantages. There are still many problems in practical engineering applications. For example, with data-level methods, it is prone to introduce noise and delete useful information, cost sensitive learning can be less effective because of unreasonable costs and so on. This paper proposes an adaptive imbalance classification (AIC) algorithm based on model fusion, which fuses the advantages of each method and automatically adapts to data sets with different imbalance levels.

II. ALGORITHMIC ANALYSIS

A. Support Vector Description Algorithm

SVDD is a one-class classification algorithm. Its basic idea is to transform the original sample space into a new high-dimensional feature space by non-linear mapping and to find an optimal hypersphere in the high-dimensional space. As a result, it can contain all the samples as far as possible and weigh the maximum number of samples as well as the minimum radius of the sphere.

Consider the training data set $D = \{x_i\}, x_i \in R^n, i = 1, 2, .., n$ and that there is a non-linear mapping ϕ that transforms data sample x_i from the original space R^n into the new high-dimensional space H, which means $\phi(x_i) \in H$。In SVDD, it aims to find a hypersphere with

978-1-7281-0860-5/19/$31.00 ©2019 IEEE

a center and a radius (marked by a and r respectively). Besides the concept of slack variables $\xi_i, \xi_i > 0$ is introduced. Therefore, SVDD solves the problem mathematically as shown in (1), where $\|.\|$ represents Euclidean Distance and C is the penalty parameter。

$$\begin{gathered} \min_{R,a,\xi} \varepsilon(R,a,\xi) = R^2 + C\sum_i \xi_i \\ \text{s.t.} \left\|\phi(x_i) - a\right\|^2 \le R^2 + \xi_i, i = 1,2,...,n; \\ \text{and } \xi_i \ge 0, i = 1,2,...,n \end{gathered} \tag{1}$$

Construct the Lagrange function, introduce the kernel function $K(x_i, x_j)$, and solve the dual problem. After solving the dual problem, the problem can be expressed as (2).

$$\begin{gathered} \max_{\alpha_i} L_D = \sum_i \alpha_i K(x_i, x_j) - \sum_i \sum_j \alpha_i \alpha_j K(x_i, x_j) \\ s.t.\ 0 \le \alpha_i \le C, \sum_i \alpha_i = 1 \end{gathered} \tag{2}$$

In SVDD, assuming that the distance function from any sample point to the center of the sphere is recorded as $d^2(x)$, the classification is defined by whether the test sample is in the hypersphere or not, as shown in (3).

$$f(x) = \text{sgn}(R^2 - d^2(x)) \tag{3}$$

B. Random Forest Algorithm

Random Forest [12] (RF) is an extension of the Bagging method. The main idea of the Bagging method [13] is: use Bootstrap sampling on the training data set D_{train}; the training subset D_{train}^i is composed of the random sampling and times with put-back; the sub-classifier model h is then trained on the training subset D_{train}^i, repeat sampling and training k times to get k sub-classifier models $h_i, i = \{1,2,...,k\}$. Finally, the k sub-classifier models are merged by voting method. The final classifier model is shown in (4).

$$H(x) = sign(\sum_{i=1}^{k} h_i(x)) \tag{4}$$

The extension is mainly embodied in two aspects. In the first part, it is assumed in RF method that the sub-classifier model of Bagging method is classification and regression trees （CART）. In the second part, it is assumed the random attribute selection is introduced in RF method. The RF algorithm forms the training subset D_{train}^i by Bagging method, randomly selects k non-repeating attributes from the d attribute features to form a new training subset $D_{train}'^i$.

C. Gradient Boosting Decision Tree

Gradient boosting decision tree [14] (GBDT) is an extension of Boosting method. It uses decision tree as a sub-classifier for model training and belongs to Boosting Tree model.

The loss function is assumed as $L(y, f(x))$, the optimization problem of boosting tree model can be described in (5), where n represents the number of samples for data set D.

$$\min_{\beta,\Theta} \sum_{i=1}^{n} L(y_i, \sum_{m=1}^{M} \beta_m h(x, \Theta_m)) \tag{5}$$

The boosting tree uses a forward-step algorithm to solve the model. Forward-step algorithm is executed from front to back, which learns only one tree model h and its coefficients β_m per step. The greedy algorithm is used to approximate the optimal solution step by step. Let the decision tree model parameter that k step needs to solve is Θ_k and the model coefficient is β_k, the model of k th step can be expressed as (6).

$$f_k(x) = f_{k-1}(x) + \beta_k h(x, \Theta_k) \tag{6}$$

The optimization problem of k th step can be described as the (7)：

$$\min_{\beta_k, \Theta_k} \sum_{i=1}^{n} L(y_i, f_{k-1}(x) + \beta_k h(x, \Theta_k)) \tag{7}$$

The k th step of gradient boosting tree is used to fit the negative gradient of the loss function of the $k-1$ th step model, which is shown in (8). The negative gradient of the loss function of the $k-1$ th step model can be approximately equivalent to the residual between the predicted result and the true value of the $k-1$ th step. In each iteration of the gradient boosting tree, the residual error is reduced.

$$y_i^k = -\left.\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right|_{f(x)=f_{k-1}(x)} \tag{8}$$

III. ADAPTIVE IMBALANCE CLASSIFICATION ALGORITHM

Adaptive imbalanced classification algorithm uses different strategies in multiple child training model y_i from the perspective of integration model, select the optimal sub-model according to the area under curve (AUC) value and get the stable classification results aiming to actively adapt to data sets with different imbalance degrees. The adaptive imbalance classification algorithm adopts four strategies.

Strategy 1: Optimize the support vector description algorithm and generate sub-model y_1 from the new algorithm idea.

Strategy 2: Optimize stochastic forest algorithm and generate sub-model y_2 from the perspective of improving the

existing algorithms

Strategy 3: From the perspective of data, through data sampling method, reduce the number of health samples to balance different types of samples, and combine the gradient lifting tree algorithm to generate sub-model y_3.

Strategy 4: From the perspective of model fusion, the weighted average method is used to fuse the above three sub-models to generate sub-model y_4.

A. Optimized SVDD

In this paper, the decision boundary of SVDD is optimized based on the category label information and the AUC value of the performance evaluation index.

It needs to calculate the probability output value for AUC. However, SVDD outputs discrete quantities $\{+1,-1\}$. As shown in (3), the AUC value cannot be calculated directly. Therefore, this paper uses Sigmoid function to modify the output of SVDD, as is shown in (9). The modified output $g(x)$ can be thought of as the sample in the state of normal forecast probability $p(y=0\mid x)$.

$$g(x)=S(R^2-d^2(x)) \tag{9}$$

The key steps of the optimized SVDD algorithm are as follows:

Step 1: Initialize the parameters of SVDD algorithm;

Step 2: Train the hypersphere model and combine the Sigmoid function to obtain the probability output value of the SVDD algorithm;

Step 3: Calculate the AUC value of the model on the data set;

Step 4: Use the differential evolution algorithm to iteratively optimize the algorithm parameters and to maximize the AUC value of the model on the data set.

B. Optimized RF

This paper focuses on improving the accuracy of the RF sub-classifier and indirectly improving the accuracy of the RF algorithm.

First of all, based on the idea of "hierarchical sampling", this paper divides the original training set D_{train} into majority class data set $D_{train,1}$ and minority class data set D_{train2} according to the sample category. Resample several times at the same time with playback for the majority class data set $D_{train,1}$ and minority class data set D_{train2} so as to make the "majority class subset $D^i_{train,1}$" and the "minority class subset $D^i_{train,2}$". The optimized training subset D^i_{train} consists of "majority class subset $D^i_{train,1}$" and "minority class subset $D^i_{train,2}$". In this paper, the proportion of the training subset D^i_{train} samples is balanced by means of "hierarchical sampling", which improves the probability of participation of minority samples in the sub-classifier training, and reduces the influence on the "accuracy" of the sub-classifier due to the size of minority samples.

Secondly, in this paper, the overall accuracy of the sub-classifier h_i is optimized. The accuracy of the sub-classifier is improved by deleting the sub-classifier with "low accuracy" h_i. The AUC value is selected to evaluate the "accuracy" of the sub-classifier h_i. By setting the AUC value threshold ε of the sub-classifier h_i, the sub-classifier h_i below this threshold is deleted.

The key steps of the optimized RF algorithm are as follows:

Step 1: Initialize RF algorithm parameters Θ, such as the number of trees, the depth of the tree and the maximum number of tree features, etc.

Step 2: Randomly generate 100 CART subtrees in combination with the stratified sampling method;

Step 3: Calculate the AUC value of the CART subtree, and use the 30% quantile to determine the threshold parameter ε of the AUC value;

Step 4: Generate a training CART subtree by using a hierarchical sampling method, calculate AUC values of the CART subtree, and delete the CART subtree if the AUC values are less than the threshold value ε;

Step 5: Repeat the generation of CART subtrees until the termination condition is satisfied;

Step 6: Use voting method to fuse multiple subtree models;

Step 7: Use the iterative optimization method, differential evolution algorithm to find the optimal algorithm parameters, making the model in the data set on the AUC value is the largest.

C. Optimized GBDT

The optimization of the GBDT algorithm mainly focuses on the data level. In this paper, using the data sampling method to reduce the number of majority samples, the number of samples in different categories of the training data set is balanced, and then the model training is completed by using the gradient lifting tree.

In this paper, the neighborhood cleaning rule (NCL) is used to undersample most samples. NCL algorithm combines the distribution and structure characteristics of samples and effectively retains the information of most samples, which is more accurate than the traditional undersample method.

In this paper, the imbalance degree of data set is defined as the ratio of the number of samples in most classes to the number of samples in a few classes. The key steps of the optimized GBDT algorithm are as follows:

Step 1: Run NCL algorithm several times until the imbalance degree of the data set meets the requirements;

Step 2: Find the optimal model of gradient lifting tree by iterative optimization method such as the differential evolution algorithm.

D. Flow Chart Of The Algorithm

Combined with the above analysis, the block diagram of the adaptive imbalanced classification algorithm is shown in Figure 1. The execution process of the entire algorithm is as follows:

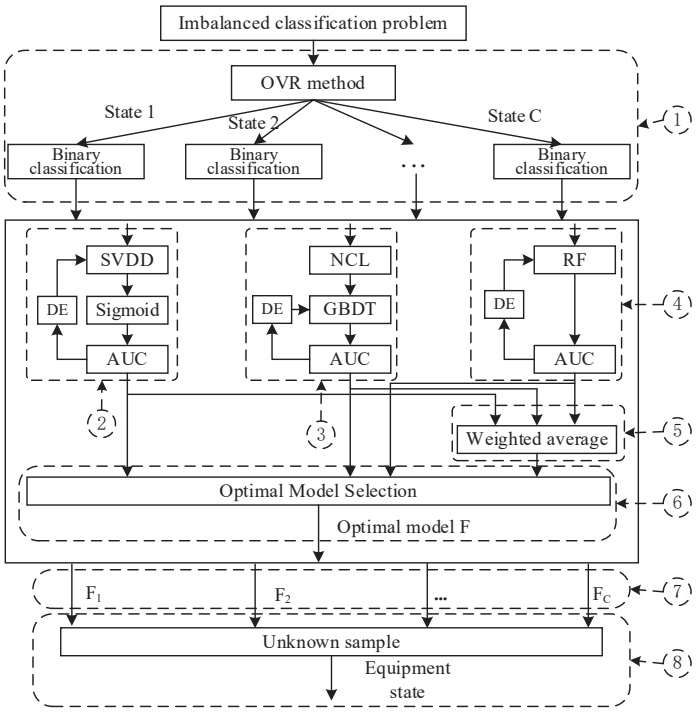

Figure 1. Adaptive imbalanced classification algorithm block diagram

Step 1: The algorithm combines the one vs rest (OVR) method to decompose the imbalanced classification problem into C binary classification problems. The binary classification data set of state k is represented as D^k.

Step 2: Train the sub-model f_k^1 with the optimized SVDD algorithm on the data set D^k.

Step 3: Train the sub-model f_k^2 with an optimized RF algorithm on the data set D^k.

Step 4: Combine the NCL algorithm with the GBDT algorithm on the data set D^k to train the sub-model f_k^3.

Step 5: Use the weighted average method to fuse the sub-models f_k^1, f_k^2 and f_k^3. Then output the fusion model.

Step 6: Use the cross-validation set to determine the final model f_k^* of the state.

Step 7: Traverse the data set $D^i, i = \{1,2,...,C\}$ of the C device states, and repeat steps 2~6 to obtain C optimal models $f_i^*, i = \{1,2,...,C\}$.

Step 8: For the unknown sample x_{n+1}, traverse the C optimal models f_i^* and output the probability of the state of the device $i, i = \{1,2,...,C\}$. The maximum probability of the state i of the device is the corresponding state of the unknown sample x_{n+1}.

IV. Experimental Results and Analysis

The data set is based on the equipment operation data set published by the American PHM Association in 2015. The data set contains operation data of 60 pieces of equipment from 2009 to 2012. The data operation can be divided into two categories. The first category is operation information, which is the operation status information of the device. The second category is the fault information, recording fault information during the operation of the equipment, including the start time t_1 and the end time t_2 and the fault type t_3.

The problem with the data set is that the data set contains 60 sets of equipment operation data, of which equipment failure information of 20 sets is incomplete. The goal of this data set is to correctly predict the equipment failure information of the 20 sets.

A. Data Set Analysis

The goal of the data set is to correctly predict the failure information of 20 groups of equipment, including the start time, end time and the type of the fault. In this paper, the problem of predicting fault information is transformed into classification problem. The specific strategy is to determine whether the current time is the start time of the fault type or the end time of the fault. Since that the problem of predicting the start of the fault is an imbalanced classification problem, and that the problem of predicting the fault start time of different fault types of different devices is to handle the data set with different degrees of imbalance, this paper focuses on the prediction problem at the start of the fault.

The occurrence of faults may cause changes in equipment operation information. Therefore, it is reasonable to use equipment operation information training classification model to judge whether faults occur at the current time.

However, the start time of the failure reflects a change from the health state to the failure state. The operation information of the equipment at a sole time cannot adequately express the change of the operation state of the equipment. Therefore, this paper introduces the concept of time window. Time window refers to the equipment operation data within a period of time. The data at a single time cannot reflect the change of equipment state, but the data over a period of time can capture this change.

The introduction of the time window enables the classifier to capture the changes of the device state but increases the dimension of the data. High-dimensional data will not only lead to a sharp increase in the amount of computation but also lead to dimensional disasters, resulting in a significant decline in the performance of the classifier. This paper uses principal components analysis (PCA) method to reduce dimension.

Therefore, the key steps of data processing are:

Step 1: For the original equipment operation matrix $A_{n\times m}$, the time window method is used to expand the data dimension and change it into a new matrix $A_{n\times m}$.

Step 2: Use the PCA method to reduce the dimension of matrix $A_{n\times m}$ and retain 95% of the data variance to obtain the final training matrix $A_{n\times m}$.

B. Algorithm Evaluation Index

In this paper, (10) is selected as the performance evaluation index of the algorithm, which is defined as GCS. The symbol TP means that the non-fault start time is correctly identified; the symbol TN indicates that the fault start time is correctly identified; the symbol FN means that the non-fault start time is incorrectly identified as the fault start time; The symbol FP indicates that the fault start time was incorrectly identified as the non-fault start time.

$$GCS = TP\times 0 + TN\times 10 - FN\times 0.1 - FP\times 1 \qquad (10)$$

C. Results Analysis

This paper compares the improved correlation algorithms such as RF algorithm, GBDT algorithm and SVDD algorithm, as well as the difference between AIC algorithm proposed in this paper, on the issue of predicting the failure start time. Table 1 shows the performance differences of each algorithm for predicting the start time of failure on all faults of 40 sets of equipment. Table 1 compares the differences of improved GBDT algorithm, improved RF algorithm, improved SVDD algorithm and AIC algorithm from GCS mean and GCS variance dimensions.

TABLE I ALGORITHMS PERFORMANCE ON THE FAILURE START PROBLEM

	Modified GBDT	Modified RF	Modified SVDD	AIC
GCS mean	1436.14	2153.43	1864.93	2343.07
GCS variance	2533473	2212145	1897230	1176700
TNR mean	54.1%	64.9%	60.4%	68.3%
Average accuracy	84.1%	91.4%	86.7%	89.2%

Comparing the AIC algorithm and the improved algorithm proposed in this paper, it can be seen that the AIC algorithm has the best comprehensive performance. Its TNR mean and GCS mean are the highest, and its GCS variance is lower than the other three improved algorithms, and its accuracy mean performance is better. The reason is that the imbalance degree of 40 sets of equipment varies greatly and the performance of a single algorithm is unstable, which leads to the high variance of single algorithm GCS. However, AIC algorithm combines the advantages of the three improved algorithms and chooses the optimal sub-model according to AUC value, which stabilizes the classification results of the model. Therefore, the AIC algorithm can effectively solve the problem of imbalanced data fault classification, obtaining better fault diagnosis effect, higher fault recognition rate and more stable performance.

V. CONCLUSION

This paper proposes a classification method with better stability and adaptability to different imbalanced data sets, namely an adaptive imbalanced classification algorithm. This paper combines the data set of equipment operation published by the American PHM Association in 2015 to analyze the prediction of the failure start time. The results show that the AIC algorithm proposed in this paper is better than the single imbalanced classification algorithm, with a better fault diagnosis, a higher fault recognition rate and a more stable performance.

ACKNOWLEDGMENT

The authors are grateful for the financial support from Beijing Municipal Natural Science Foundation (Grant No. 4172067) and Aeronautical Science Foundation of China (Grant No. 2017ZD41006).

REFERENCES

[1] Nan Z. Mechanical Fault Diagnosis Method Based on Machine Learning[C]. Seventh International Conference on Measuring Technology & Mechatronics Automation. IEEE, 2015.

[2] Feng Y, Feng D. Research And Application Of Unbalanced Data Classification[J]. Computer Applications and Software, 2018.

[3] Gong C, Gu L. A Novel SMOTE-Based Classification Approach to Online Data Imbalance Problem[J]. Mathematical Problems in Engineering,2016,(2016-5-25), 2016, 2016(35):1-14.

[4] Hu S , Liang Y , Ma L , Ma L, He Y. MSMOTE: Improving Classification Performance When Training Data is Imbalanced[C] Second International Workshop on Computer Science & Engineering. IEEE, 2010.

[5] Dang X T , Tran D H. SPY : a novel resampling method for improving classification performance in imbalanced data[C] Seventh International Conference on Knowledge & Systems Engineering. IEEE, 2015.

[6] Jia C, Zuo Y. S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique[J]. Journal of Theoretical Biology, 2017, 422:84-89.

[7] Al Abdouli N O, Aung Z, Woon W L, Svetinovic D. Tackling Class Imbalance Problem in Binary Classification using Augmented Neighborhood Cleaning Algorithm[M] Information Science and Applications. 2015.

[8] Zhang P, Liu L, Ma Z. Research On Cascade-Grouping Parallel Svm Algorithm Based On Mapreduce[J]. Computer Applications & Software, 2015.

[9] Masnadishirazi H, Vasconcelos N. Cost-Sensitive Boosting[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 33(2):294-309.

[10] Jiang H, Wang H, Hu W, Kakde D, Chaudhuri A. Fast Incremental SVDD Learning Algorithm with the Gaussian Kernel[J]. 2017.

[11] Liu F T, Kai M T, Zhou Z H. Isolation Forest[C] Eighth IEEE International Conference on Data Mining. 2009.

[12] Fang W, Suxia M, He W, Yaodong L, Zhiguo Q, Junjie Z. A hybrid model integrating improved flower pollination algorithm-based feature selection and improved random forest for NO X emission estimation of coal-fired power plants[J]. Measurement, 2018, 125:303-312..

[13] Bifet A, Holmes G, Pfahringer B. Leveraging Bagging for Evolving Data Streams[M] Machine Learning and Knowledge Discovery in Databases. 2010.

[14] Son J. Tracking-by-Segmentation with Online Gradient Boosting Decision Tree[C] IEEE International Conference on Computer Vision. 2016.