Page 1

1、Mr Chairman，honorable guest, good morning！My name is Liu Hengyu. I am from Northwestern polytechnical university. I'm pleased to be able to attend this meeting. Today , I'd like to present my paper "Research of System Fault Diagnosis Method Based on Imbalanced Data".

Page 2

My presentation will cover four aspects. Introduction、proposed framework、case study、and the last part is about the summary and future research.

Page 3

1、First, some background information about this research is introduced.

2、This page illustrates the importance of the classification techniques in diagnosis .

3、The first part is the information acquisition and processing system. Generally, the different attribute data is collected with multi-sensors and is supposed to be preprocessed. Then it will be sent to the classification system.

Using the classification model to get the fault classes which can achieve the fault diagnosis.

4、possessing the anomaly information, fault information and the predict information

Page 4

1、As for the fault diagnosis, this paper uses a novel classification model to obtain the fault classes providing useful information for the maintenance strategies.
2、The classification model is trained by the historical data. Then, when the new monitoring data is arrived, the trained classification model will give the fault classes to finish the fault diagnosis.
3、The identification of the fault classes can implement diagnosis accurately and carry out suitable maintenance in complex systems, which is of great significance in the safe and reliable operation of the practical application.

Page 5

The proposed framework in this research includes four parts.

Page 6

Support Vector Domain Description Model 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.

Page 7

In our method, we uses Sigmoid function to modify the output of SVDD, as is shown in formula

$f(x) = sigmoid(R^2 - d^2(x))$

And this classification is a recursive process

1: First, we initialize the parameters of SVDD algorithm;

2: Second, we need train the hypersphere model and combine the Sigmoid function to obtain the probability output value of the SVDD algorithm;

3: Then, we calculate the AUC value of the model on the dataset;

4: Finally, we use the differential evolution algorithm to iteratively optimize the algorithm parameters and to maximize the AUC value of the model on the data set.

Page 9

Also, Random Forest Model is a one-class classification algorithm. Random Forest is an extension of the Bagging method. The main idea of the Bagging method is: use Bootstrap sampling on the training data set ; the training subset   is composed of the random sampling and times with put-back; the sub-classifier model   is then trained on the training subset   , repeat sampling and training   times to get   sub-classifier models

Page 10

Optimized Random Forest Model

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

1: First of all, we initialize RF algorithm parameters , such as the number of trees, the depth of the tree and the maximum number of tree features, etc.

2: Secondly, we randomly generate 100 CART subtrees in combination with the stratified sampling method;

3: Besides, we calculate the AUC value of the CART subtree, and use the 30% quantile to determine the threshold parameter   of the AUC value;

4: Then, we 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   ;

5: And then, we repeat the generation of CART subtrees until the termination condition is satisfied;

6: Also we choose voting method to fuse multiple subtree models;

7: Finally, we 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.

Page 12

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

Page 13

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.

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

1: First of all, we run NCL algorithm several times until the imbalance degree of the data set meets the requirements;

2: Then, we find the optimal model of gradient lifting tree by iterative optimization method such as the differential evolution algorithm.

Page 15

Adaptive Imbalance Classification

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

2: Next, we train the sub-model with the optimized SVDD algorithm on the data set .

3: Then, we train the sub-model with an optimized RF algorithm on the data set .

4: After that, we combine the NCL algorithm with the GBDT algorithm on the data set to train the sub-model .

5: Next, we use the weighted average method to fuse the sub-models , and . Then output the fusion model.

6: Then, we use the cross-validation set to determine the final model of the state.

7: At last, we traverse the data set of the device states, and repeat steps 2~6 to obtain optimal models .

Page 17

The Adaptive Imbalance Classification method tends to acquire more accurate classification result.

After introducing the proposed method, now a case study is used to verify the effectiveness of the proposed method.

Next section is the case study.

We choose 2015 PHM Data Challenge dataset as the case study. This dataset is a typical imbalanced classification problem.

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 and the end time and the fault type .

The dataset aims to detect plant failure events in advance. Given the labeled fault type of the past time, our aim is to predict future failure events of types and the time of their occurrence from past data.

Page 18

In this paper, 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.

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

Page 19

Next section is the Performance illustration and discussion of the case study.

Page 20

These three common imbalance diagnosis methods, voting, averaging and learning to combine the classification results of sub-classifiers. So, in this paper, we compare modified GBDT, modified RF, modified SVDD with proposed method.

It is also worth being mentioned that the parallel structure helps to reduce the variance of the model and makes the final result more stable. To verify this theory, the standard deviation of training accuracy and testing accuracy has been recorded as above.

These results prove that the accuracy mean performance of AIC algorithm is better . As a whole, AIC algorithm basically maintains the minimum accuracy standard deviation , that is to say, it can effectively improve classification accuracy.

We can see that the performance of AIC algorithm. From the Table , the conclusion can be drawn that AIC algorithm has the best accuracy and stability compared with the other four algorithms.

Next part is the summary and future research.

1、This paper combine modified GBDT, modified RF with modified SVDD and propose a novel ensemble method for imbalanced classification application.
2、In this model, from the perspective of data sampling and algorithm, we optimize the traditional SVDD, RF and GBDT.
3、modified GBDT, modified RF with modified SVDD were involved to establish classification models in the data sets.
4、 Experiments indicate that our method has good generalization performance and stability in the PHM 2015 data sets.
5、The future research can be conducted from the aspect of the ensemble learning and practical industry.

The last part is about some references.

Okay, it has been my pleasure to give you my viewpoints on the classification problems. That’s all. Thanks for your kind attention!