Design of classification model on high-dimensional imbalance data of motor bearing fault

Xiaocui Zhu*, Li Hui, Qian Sai
Institute of Energy Research, Jiangxi Academy of Sciences, Nanchang Jiangxi, 330000, China
*Corresponding author e-mail: zhuxiaocui@jxas.ac.cn

Abstract. According to the characteristics of high-dimensional imbalance distribution of motor bearing fault data, a design scheme of classification model is proposed for the high-dimensional data reduction problem in the classification algorithm. For details: Combining standard particle swarm optimization algorithm and random forest algorithm, a new high-dimensional data reduction algorithm is proposed. Aiming at the imbalance problem of data categories in the classification algorithm, we propose to use machine learning under the sum of squares of dynamic deviations criterion to divide the minority sample data set into mixed regions, high-purity minority sample regions and outlier regions, and then use smote algorithm to complete the data equalization processing, so as to make the sample data equalization processing more reasonable, Focusing on the task of motor bearing fault classification, a design scheme of using standard particle swarm optimization algorithm to improve the least squares support vector machine model is proposed.

Key words: Motor, Imbalance data, Bearing fault classification.

1. Introduction
Nowadays, the classification algorithm of sample data has been the researchers committed to overcome the scientific problem [1], this project for asynchronous motor fault imbalance data [2], proposed an imbalance method to alleviate sample data, which has good scientific reference for industrial industry data imbalance classification algorithm research value. At the same time, the feature selection of high-dimensional data is a hot spot and difficulty in data mining. Combined with the demand of asynchronous motor fault diagnosis, this paper proposes the scientific and effective method of reducing the minimum subset of features from high-dimensional motor fault samples, in order to improve the operation speed of the algorithm and the accuracy of classification results, which has some significance for data reduction of high-dimensional industry.

2. Relation work
Nowadays, the processing of high-dimensional imbalance data mainly adopts two principles, starting with high-dimensional imbalance respectively. There are also two solutions for the mining of imbalance data, respectively, from the data level and at the algorithmic level:
(1) Data-level solutions belong to data pre-processing techniques, which have allowed the datasets to reach equilibrium by some specific methods before the datasets enter the mining phase. The algorithmic level solution is to optimize the data mining algorithms or models to obtain new algorithms or models that can be targeted for imbalance data mining. There are many methods for data-level processing, and stochastic undersampling and random oversampling techniques in balanced data processing techniques were first proposed in academia. For example, two proposed undersampling algorithms for Liu classification of imbalance dataset: Easy Ensemble and Balance Cascade algorithm [3]; Del mentions a KNN-Near Miss method that is based on the undersampling method of the k-nearest neighbor [4]. However, during decades of exploratory research, a considerable number of scholars believe that simple random sampling methods do not well improve the performance of classifiers on imbalance data classification, because random under-sampling methods may remove some samples with important information. For the imbalance problem classical processing algorithm for SMOTE oversampling technology, but the SMOTE algorithm also has obvious shortcomings, is to the imbalance boundary mixed data cannot be effectively solved, so the improvement of the SMOTE algorithm also caused a lot of scholars, such as Zhong et al proposed to K-means clustering algorithm to improve the SMOTE algorithm to solve the imbalance data problem [5]. However, this algorithm still does not escape the disadvantage that K-means clustering relies heavily on the initial clustering points; Feng et al. proposed an imbalance data processing method (BMS) based on boundary mixed sampling, which first introduces the "coefficient of variation" to find the boundary domain and non-boundary domain, then oversampling a few sample in the boundary region and random undersampling of most samples in non-boundary regions, but neglects to consider the disadvantages of outlier points [6]; Zhao et al. proposed a traditional SMOTE algorithm with the centroid method for the imbalance data, which improves the difficulty of boundary hybrid interpolation. Although this algorithm reflects some advantages in solving the boundary mixing, the computational steps and operation amount increase the algorithm accuracy is too small[7]; The dilemma of too low accuracy in classification mining of imbalanced datasets, Wang et al, The SMOTE algorithm was used to improve the synthetic minority class oversampling technique using the Ada Boost algorithm, The KSMOTE-Ada Boost algorithm for nonequilibrium data classification is proposed, First of all, According to the idea of the K Nearest Neighbor (KNN), The noise sample identification algorithm is proposed, Through the number of heterologous samples included in the K Nearest neighbors of the sample, Noise samples in the sample set were accurately identified and filtered, Second, the idea of clustering based divides the sample set into different subclusters during oversampling, Based on the cluster heart of the subcluster and the number of samples included, Sysynthesis of new samples was performed between the sample and the cluster heart, But the degree of discrimination of the noise samples is debatable [8].

(2)Improved research of imbalance data from the algorithmic level has also been explored in academia. Park et al. introduced the heterogeneous cost penalty function to adjust the misclassification so that the decision tree can split with different evaluation criteria and named this algorithm a cost-sensitive support vector machine algorithm [9]; Xin et al. found that the improved algorithm KSVM algorithm to solve the imbalance data is defective, because the valve value is fixed and cannot be dynamically adjusted with the data. Therefore, it is proposed to first use the genetic algorithm to optimize the valve value $\epsilon$, and then conduct data mining, and improve the disadvantages of the fixed valve value [10]; Qiu et al [11] proposed an improved algorithm for random forest, which adopts hierarchical self-sampling method to over-sample a few classes, dynamically calculate the cost-sensitive matrix of each partition, uses the cost-sensitive matrix to introduce the sensitive matrix into the key steps of building the base classifier, and weaken the effect of data bias on the classifier. When mining high-dimensional data, the method commonly adopted in the academic circle is to first make dimensionality reduction and then use the classifier to train and learn the dimension reduction data set. Two common dimension reduction methods include feature selection and feature extraction [12-13]. For example, a robust geometric mean-based subspace discriminant analysis feature extraction method for image set classification is proposed by Gao et al [14]. Li et al. introduce the framework of feature selection methods in detail, focusing on the two processes of generating feature subsets and evaluation
criteria, then classify feature selection algorithms according to the different combination methods of feature selection and learning algorithms, analyze the advantages and disadvantages of various methods, discuss the problems of existing feature selection algorithms, and put forward research difficulties and research directions [15].

3. Main researched issues and key problems needed to solved

(1) Facing the fault sample data of electric vehicle bearings and focusing on the dimensional reduction method of high-dimensional data, we will solve the problem of how to transform the high-dimensional fault sample data into the low-dimensional data with strong correlation and good representative;

(2) In view of the characteristics of the imbalance distribution of classified sample data, focus on studying the oversampling technology of imbalance data, and reasonably enhance the distribution number of samples of a few important categories;

(3) Establishing a classification model based on the fault diagnosis of motor bearings, which greatly improves the fault diagnosis accuracy rate and enhances the robustness, efficiency and application of the classification model.

In the scheme proposed in this paper, the main study content and their interrelations are shown in Figure 1.

Centering on the research content, the research task is to focus on breaking through the impact problem of high dimension and imbalance of classification samples on the classification results, and explore the applicable method for motor bearing fault classification. Therefore, the Key problems to be solved are:

(1) How to screen the best combination of dimensional features contributing most to the classification algorithm from high-dimensional features scientifically and efficiently.

(2) How to effectively alleviate the imbalance problem in sample data, that is, how to maximize to avoid the overfitting or underfitting problem of classification algorithm when taking oversampled, undersampling or SMOTE methods.

4. Research on technical routes

The project will be carried out in two aspects, firstly integrating standard particle group autoalgorithm and random forest algorithm to establish high-dimensional data dimension reduction algorithm, combining with dynamic deviation sum of square criterion theory and SMOTE oversampling technology; then building a new fault classification model with SPA, and finally through the electric vehicle training sample data. The overall technical route of the project study is shown in Figure 2.
(1) High-dimensional and imbalance data processing optimization algorithm research technical route

The proposed integration of the Gini coefficient node purity and the out-of-pocket data accuracy estimation to jointly determine the characteristic severity estimation value, and the calculation formula can be designed as:

\[ \text{Infactor} \ x_i = \alpha_i \text{Gini}_i + \beta_i \text{OOB}_i \ i = 1,2, \ldots, k \ , \quad (1) \]

The \( i \) represents the \( i \)th feature of the data, the \( \text{Gini}_i \) coefficient of the \( i \)th feature, and the \( \text{OOB}_i \) (off-bag data) accuracy estimation of the \( i \)th feature, where the coefficient \( \alpha_i \) and \( \beta_i \) meet the \( \alpha_i + \beta_i = 1 \), and the \( \alpha_i \) and \( \beta_i \) are estimated by the particle optimization group algorithm, and then the feature importance of \( \text{Infactor} \ x_i \) is rearranged according to the value.

SMOTE algorithm: interpolation between center of mass and sample of sample set, the formula is:

\[ X_{i-new} = x_{(i)}^{(2)} + \text{RAND}(0,1) \times (x_{(i)}^{(2)} - d) , \ i = 1,2,\ldots,n_2^{(2)} \]  \quad (2)

\( x_{(i)}^{(2)} \) is the \( i \)th sample in the sample set of \( C \), \( d \) is the centroid of mass in the sample set, \( n_2^{(2)} \) is the number of samples in a few category sample sets, and the newly synthesized samples \( X_{i-new} \).

The technical route that can be taken is as follows: 1. First, based on the high-dimensional characteristic dataset of motor bearing failure, Establish a random forest model, Several taxonomic trees were obtained; 2. tested the taxonomic tree acquired in the first step, And its accuracy is calculated; 3. constructs the feature importance formula, Calcuite the weights of each feature, That is, the importance of the valuation; 4, utilizes the formula (1), Parparameters were calculated using the particle optimization group algorithm, And to optimize it; 5, screens out the best set of features required for the fault classification of motor bearings; The 6, cleaves the dataset by region; 7, then performs oversampling of regional samples using dynamic sum of squared criteria and SMOTE. Finally, the dataset was qualified, and if not met, a continuous iterative optimization algorithm was performed. The intuitive description of the research technical route is shown in Figure 3.

**Figure 2.** Overall Roadmap for project research
5. Experimental scheme and model evaluation method

(1) Experimental data: Two public datasets will be held at the Western Reserve University and Germany-Paderburn University.

(2) Algorithm or model development tools: part of the experiments of the project, while the main model is completed using JetBrains PyCharm 2019.1.3 x64 development platform integrated with python.

(3) Model evaluation method: The classification model of this project adopts the mainstream evaluation method. Table 1 is the confusion matrix (Confusion Matrix), which defines the basic indicators of the model evaluation.

Table 1. Confusion matrix

|                  | Ground truth |
|------------------|--------------|
|                  | Positive     | Negative    |
| Prediction values| TP           | FN           |
|                  | FP           | TN           |

Figure 3. Study of the technical route for the high-dimensional and nonequilibrium data processing optimization algorithm

(2) Research on technical route of fault classification model of motor bearings

The radial basis and regularization parameters of the least squares vector machine will be optimized using the standard particle optimization group algorithm, with the radial basis function kernel being:

$$K(x, z) = \exp\left(\frac{|x-z|^2}{\sigma^2}\right)$$  \hspace{1cm} (3)

The classification problem of least-squares SVM can establish the following classification formula:

$$\min_{w, b, e} F(w, b, e) = \frac{1}{2}w^T\omega + \frac{1}{2}\sum_{k=1}^{N}e_k^2$$  \hspace{1cm} (4)

Where, $\sigma^2$ and $\gamma$ from the above formula (3) and formula (4), it is known that the classification model of this project has two important parameters, in which the parameters are the radial basis parameters. In the expression of the least squares SVM, the changes of the two parameters have a great impact on the mining effect of the least squares SVM model, which directly determines the classification accuracy and the degree of fitting. The kernel function $\sigma^2$ is called the kernel width, which reflects the distribution characteristics of the data. $\gamma$ known as the regularization parameter, it has a regulator on the ratio between confidence intervals and empirical risk of LSSVM. The intuitive description of the motor bearing classification model is shown in Figure 4.
6. Conclusion

This paper proposes a design scheme of classification model for fault data of motor bearings using the standard particle group algorithm and random forest algorithm, and then adjusts the dynamic sum of square criterion and the SMOTE method to form the low-dimensional equilibrium dataset. Finally, the standard particle group optimization algorithm is used to iteratively optimize the least squares support vector machine classification model in the low-dimensional dataset. The classification model is subsequently tested and classified, and the classification result is output.
vector machine parameters, and finally form the optimization classification model, in order to improve the fault classification accuracy of motor bearings. The classification model design scheme proposed in this paper is instructive for the practical development of motor bearing fault classification project.

References

[1] Randall R B, Antoni J. Rolling element bearing diagnostics: A tutorial[J]. Mechanical Systems and Signal Processing, 2018, 25(2): 485-520.

[2] Sawalhi N, Randall R B, Endo H. The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis [J]. Mechanical Systems and Signal Processing, 2017, 21(6): 2616-2633.

[3] Liu X Y, Wu J, Zhou Z H. Exploratory under sampling for class-imbalance learning [J]. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009, 39(2): 539-550.

[4] Del Ro S, Lpez V, Bentez J M, et al. On the use of Map Reduce for imbalanced big data using Random Forest [J]. Information Sciences, 2014, 285: 112-137.

[5] Zhong Longshen, Gao Xuejun, Wang Zhenyu... Application of a new K-means-based improved SMOTE algorithm in the classification of imbalance datasets [Practice and knowledge of [J]. mathematics, 2015, 45 (19): 198-205.

[6] Feng Hongwei, Yao Bo, Plateau, Wang Huiyan, Feng Jun. - -An imbalance Data Processing Algorithm based on boundary mixed sampling [J]. Control and Decision Making, 2017, 32 (10): 1831-1836.

[7] Zhao Qinghua, Zhang Yihao, Ma Jianfei, Duan Qianqian. Classification Algorithm for Improving SMOTE Non-balanced Dataset [J]. Computer Engineering and Application, 2018, 54 (18): 168-173.

[8] Wang Zhongjun, Huang Bo, Fang Zhijun, Gao Yongbin, Zhang Juan....... Improving the SMOTE imbalance Data Integrated Classification Algorithm [J]. Computer Application, 2019, 39 (09): 2591-2596.

[9] Peng L, Zhang H, Yang B, et al. A new approach for imbalance data classification based on data gravitation [J]. Information Sciences, 2014, 288: 347-373.

[10] Ge G, Jin X, Lu X, et al. Threshold optimization algorithm-ksvm for imbalance data classification prediction[C]// 2019 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017). 2019.

[11] Qin Jing, Qian Xuezong, Wang Weitao, Xie Guowei, et al. A parallel random forest algorithm for processing imbalance big data [J]. Microelectronics and Computer, 2020, 34 (04): 22-27.

[12] Zhou ZH. Ensemble methods: foundations and algorithms [M]. CRC press, 2012.

[13] Chandrashekar G, Sahin F. A survey on feature selection methods [J]. Computers Electrical Engineering, 2014, 40(1): 16-28.

[14] Jianqiang Gao, Li Li. A robust geometric mean-based subspace discriminant analysis feature extraction approach for image set classification [J]. Optik, 2019, 199: 2116-2124.

[15] Li Zhiqin, Du Jianqiang, Nie Bin, Xiong Wanping, Huang Canyi, Li Huan. Review of Characteristic Selection Methods [J]. Computer Engineering and Application, 2020, 55 (24): 10-19.