Research on intelligent diagnostic techniques for rolling bearings based on unbalanced data sets

Zhikai Xing, *Yongbao* Liu*, *Qiang* Wang, and *Jun* Li

College of Power Engineering, Naval University of Engineering, Wuhan 430032, China

Abstract. In this paper, based on the combination of comprehensive sampling and one-dimensional convolutional neural network, a bearing fault intelligent diagnosis technique is proposed for the classification of rolling bearing vibration data. At first, the fault data set is expanded by ADASYN method. Then, the data is cleaned up by Tomek link under sampling technique, the risk of overfitting caused by overlap of different classes is reduced and the data of different categories is more apparent, and finally the normal data set and fault data set after comprehensive sampling are classified by one-dimensional convolutional neural network algorithm. Compared with random forests and support vector machines, the results show that the method has a high accuracy in identifying classifications and can effectively solve the classification problem of unbalanced bearing data.

Keywords, Unbalanced data set, Comprehensive sampling, Convolutional neural network, Rolling bearings, Fault diagnosis.

1 Introduction

Rolling bearings are often exposed to high temperature, high pressure and high speed in harsh working environments, which are prone to failure and have an impact on the reliability and economy of machine life maintenance [1]. In the actual bearing operation, the fault data accounts for a small proportion in the sample set. If there is a serious imbalance in the data, the prediction conclusions tend to be biased towards the classes with more data. Most standard algorithms have the same error cost, so the classification results of complex and uneven data sets are not ideal, so it is necessary to study small sample and unbalanced data sets [2].

A lot of research has been done on the problem of data imbalance. The feature extraction of unbalanced data has become an important research field of data-driven fault diagnosis. The methods commonly used to extract features from unbalanced data can be divided into two categories [3] :(1) To make the network structure more sensitive to a few categories to improve the fault diagnosis accuracy of the small sample category [4]. (2) To use some data pre-processing techniques, such as oversampling or under sampling to reduce the imbalance between classes [5].

To solve the problem of data imbalance, in 2002 Chawla proposed the Synthetic Minority Oversampling (SMOTE)algorithm, which is an improvement based on the random

oversampling algorithm [6]. HA [7] proposed an under-extraction technique based on genetic algorithms to make sampling more reasonable and stable by minimizing losses to select the optimal subset of most classes. RAYHAN [8] proposed an under-sampling technique based on the clustering method, which was incorporated into the random algorithm, thus reducing the blindness of sampling and improving the accuracy of algorithm recognition. Cui Xin [9] uses the uneven data integration classification algorithm (IDESF) based on sampling and feature selection to first put back samples on the data set and then sample it using the SMOTE method, which increases the difference between the data sets on the basis of ensuring the reasonableness of the samples in the resulting data set in order to improve the classification performance of the algorithm.

In view of the imbalance of rolling bearing fault data, this paper puts forward the intelligent diagnostic technology combining comprehensive sampling with one-dimensional convolutional neural network. Comprehensive sampling is the expansion of samples with a small amount of data in the original signal at different scales by the Adaptive Synthetic Sampling Algorithm (Adaptive Synthetic Sampling, ADASYN) to obtain sample data equal to the normal data set sample length, and then using Tomek Link Under sampling technology processes expanded data to reduce overfitting between data classes.

2 Basic principles

2.1 Principles of the ADASYN algorithm

The adaptive synthetic sampling of ADASYN [10] (Adaptive Synthetic Sampling) is an improved oversampling algorithm based on SMOTE, which strengthens the learning ability of the classification model by giving different weights to different minority samples by not considering the distribution of adjacent samples, resulting in overlapping of a small number of samples [11]. The process is as follows:

Step 1: Calculate the number of samples that need to be synthesized, as follows:

$$G = (m_l - m_s) * \beta$$

In the above formula, the sample size m_l is the majority class, the sample size m_s is a minority class, $\beta \in$the random number of [0,1], if β=1, the positive and negative ratios after sampling are 1:1.

Step 2: Calculate the proportion of most classes in K-means, and the formula is as follows:

$$r_i = \Delta i / K$$

In the above formula, Δi is the number of samples of most classes in K-means, i is 1,2,3, ...

Step 3: Let's normalize r_i , the formula is as follows:

$$\hat{r}_i = r_i / \sum_{1}^{m_s} r_i$$

Step 4: According to the sample weight, the number of new samples *g* to be generated for each minority class sample is calculated, as follows:

$$g = \hat{r}_i * G$$

Step 5: According to the number of samples to be added for each minority class and SMOTE algorithm, new samples are generated.

2.2 Tomek Link under sampling technology

Tomek [12] improved CNN in 1976 and proposed a new framework, that is, under sampling the samples of most categories in the boundary and detecting and deleting noise data in most categories without destroying potential information in the data space. If sample x and sample y come from different categories and meet the following criteria, they are called Tomek Links. There is no other sample z that makes $d(x,z) < d(x,y)$ or $d(y,z) < d(x,y)$ true. Where represents the distance between two samples, that is, the two samples are close neighbours to each other. At this point, sample x or sample y is most likely noise data, or two samples are near the boundary.

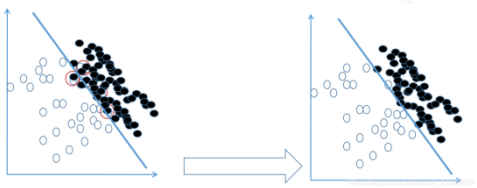

Fig. 1. How the Tomek Link algorithm works.

The image above shows the core of Tomek Link's under sampling method [13]. This method can be handled differently, such as deleting most of the class samples in the expanded samples or deleting both of the two Tomek Links samples. In this article, if the nearest neighbour of the sample belongs to a different category, the two samples will be deleted. After the Tomek Link method, a dividing line is constructed to reduce the overfitting risk caused by overlapping of different classes and make different types of data more obvious.

2.3 One-dimensional convolutional neural network

Convolutional neural network (CNN), as an integral part of deep learning, has become one of the most influential part of computer vision research. In 2012, Alwx Krizhevsky shocked the world by using convolutional neural networks to reduce the previously high classification error record from 26% to 15%. Convolutional neural networks are applied to image recognition and classification, mainly using their two-dimensional characteristics. The classic CNN network includes a series of convolutional layers, nonlinear layers, pooling layers, and fully connected layers. Take image classification as an example, take an image as input, output is the classification of image content or the probability of a group of classification.

Convolutional neural network is applied to image recognition and classification, mainly using its two-dimensional characteristics. Since the sensor data processed is time series, one-dimensional convolutional neural network (1D-CNN) is chosen as the basic framework of the nervous system in this chapter. Similar to the CNN structure introduced above, the difference is that the input of 1D-CNN is no longer an image, but a one-dimensional time series, which makes the whole network more effective in processing vibration signals.

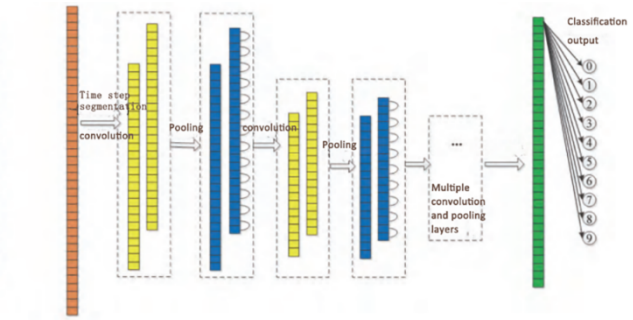

Fig. 2. The one-dimensional convolutional neural network structure diagram.

3 Engineering experiments

3.1 Collation and grouping of data results

The signal acquisition used in the verification was from SpectraQuest's comprehensive mechanical fault simulation experimental bench, which was used to conduct the ball bearing mechanical fault diagnosis experiment. The schematic diagram of the experimental platform is shown below. The test bench is mainly composed of ball bearing, removable bearing seat, motor and speed control transpose.

The experiment was carried out on the ER-16K series ball bearings supported by the test bench, with a ball count of 9. The bearing damage in this fault experiment is a single point of damage caused by the processing of electric sparks, which is prefabricated by electric sparks on the inner ring, outer ring and rolling body respectively. The acceleration sensor used in the experiment is vertically mounted on the bearing seat of the bearing to be tested, and vibration acceleration signals of the bearing under different working conditions and different loads are collected at a sampling frequency of 12kHz. The speed of the ball bearing is adjusted by the motor in the experiment site according to the need. In the three working conditions of the speed of 1800r/min, 2400r/min and 3000r/min, the signal is collected by the data acquisition system after amplification and filtering.

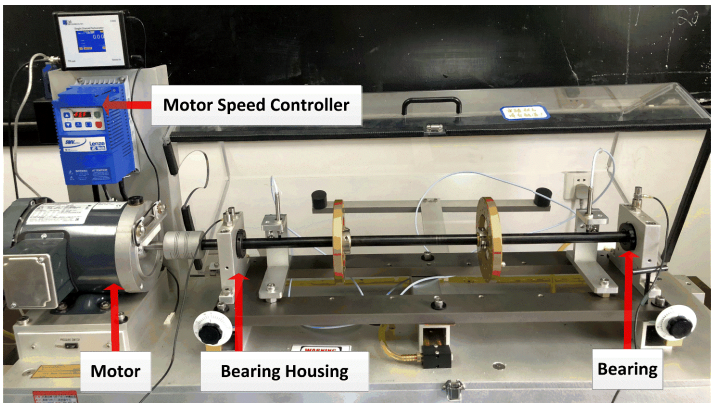

Fig. 3. Bearing data test bench diagram.

In the experiment, the fault of the ball bearing on the motor end is studied. Three common fault types are set under each operating condition, outer ring fault (OR), inner ring fault (IR),

and rolling body fault(B), and a total of 10 classes (see Table 1). Vibration acceleration signals for the three associated positions of the motor end housing (DE), end housing (FE)and base (BA)were collected.

Table 1. Description of ten rolling bearing working conditions.

conditions	Speed (r/min)	Fault location	Labels
Condition 1	1800	Outer race fault(OR)	0
		Inner race fault(IR)	1
		Ball fault(B)	2
Condition 2	2400	Outer race fault(OR)	3
		Inner race fault(IR)	4
		Ball fault(B)	5
Condition 3	3000	Outer race fault(OR)	6
		Inner race fault(IR)	7
		Ball fault(B)	8
		Normal(N)	9

3.2 Data pre-processing is carried out using different enrichment methods

In this paper, different unbalanced proportions were selected for sample expansion to study the accuracy of the proposed method under unbalanced proportions. The imbalance ratio between faulty data and normal data is set to 2:1, 5:1, 20:1, and 50:1 respectively. The number of normal data (label 9) is set to 60,000. Therefore, the fault types of labels 0-8 are 30,000, 1200, 3000, and 1200 respectively. The original data of different proportions are respectively expanded by comprehensive sampling method combining ADASYN and Tomek Link, and the extended signal length is as follows:

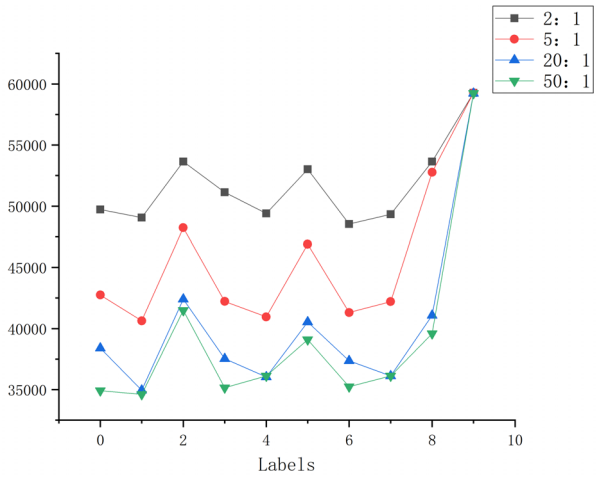

Fig. 4. Expanded signal length line chart.

By observing the data length after the expansion method, it can be found that the minimum number of label 0, label 3 and label 6 corresponds to 1,800 r/min, 2,400 r/min, and 3,000 r/min outer ring faults, respectively. It can be analyzed that there are many Tomek Links data pairs of different categories, indicating that the characteristic information of outer ring faults at different speeds is similar. With the increase of the imbalance ratio, the extended

signal length decreases continuously. As the proportion increases, when a few classes are expanded, there are more extended data and it is easier to over-fit, so there are more Tomek Links data pairs generated, and the signal length after cleaning the oversampling data of ADASYN by Tomek Link method decreases continuously.

3.3 Parameter determination

There are many parameters in the neural network, which have different influences on the final results. Some parameters may affect the accuracy of identification, and some parameters may affect the speed of network training. The influence of a single parameter on neural network is not single, but shows a primary influence aspect comparatively. In order to select the optimal neural network parameter configuration and simplify the experimental steps, this chapter compares each parameter through control variables, and then selects the optimal parameter.

Table 2. Final network model parameters.

parameter	value
The number of convolution layers	3
The number of pooled layers	3
Convolutional neuron configuration	(16, 64, 64, 128)
Optimizer	Adam
Learning rate	0.0001
The batch size	16
The maximum number of iterations	100

3.4 Experimental results

The vibration data after the unbalanced proportion expansion is input to the one-dimensional convolutional neural network, and the powerful feature extraction function of the convolutional neural network is used for fault diagnosis. The following figure is the confusion matrix of recognition results of rolling bearing fault 10 classification with an unbalanced ratio of 2:1. It can be seen from the figure that the prediction errors are mainly in labels 4 and 7, which are inner ring faults at 2400r/min and 3000r/min respectively, indicating that it is not easy to identify and classify faults with similar fault characteristics when the speed difference is small.

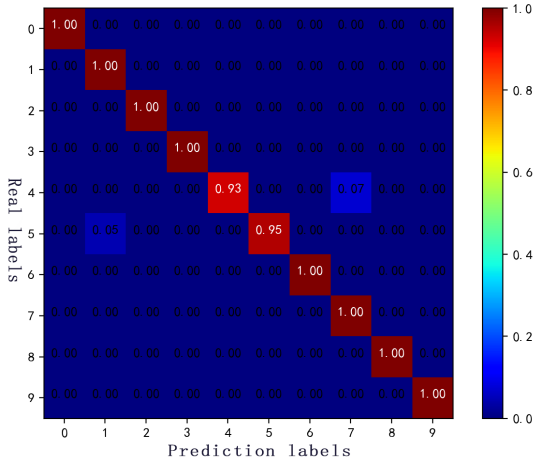

Fig. 6. Imbalance Ratio 2:1 Predictive results confuse the matrix.

In order to compare the classification results of the model classified in this paper with the current mainstream intelligent troubleshooting algorithm, select a representative BP neural network and random forest integrated learning to identify the samples. Among them, BP neural networks use time-frequency domain characteristic parameters including peak factor, pulse factor, margin factor, waveform factor, cliff factor and bias as inputs. The Random Forest method is an integrated learning algorithm based on multiple decision trees, using the above time domain feature parameters as input, combined cross-evidence method, random selection of 80% of the data in the sample as training data, the remaining 20% of the sample data as a test sample, the comparison results as shown in Table 4. The recognition accuracy of BP neural network and random forest algorithm is lower than that of convolutional neural network model classification mentioned in this paper.

Table 3. Different scales under algorithms identify accuracy.

	2:1	5:1	20:1	50:1
CNN	0.988	0.920	0.750	0.630
Random forest	0.880	0.630	0.560	0.496
BPNN	0.690	0.532	0.396	0.354

4 Conclusion

In order to solve the problem of multi-classification of unbalanced data at different scales, t a new technique combining ADASYN+ Tomek Link sampling and one-dimensional convolutional neural network is proposed.

Using comprehensive sampling to process uneven data, the data volume of a few class samples is expanded to different quantities by oversampling method, and the under-sampling method is applied to reduce the occurrence of overfitting phenomenon after data expansion. The results of classification show that using this method to classify is better than the common base classifier. In data at different scales, as the scale increases, the fewer fault characteristics are included in the fault dataset, the more obvious the overlap in the different categories after expansion, resulting in a decrease in recognition accuracy. The accuracy of classification at different speeds at different speeds in the same position needs to be improved when the method is classified 10.

References

1. Zhou Xuqiang. Discussion on the fault diagnosis technology of gas turbine[J].Chemical Enterprise Management, 2019(16):180-181.
2. Yi Wei, Mao Li,Sun Jun, Wu Linhai. Research on Classification of Improved Smote Algorithm on Imbalanced Datasets[J]. Computer and Modernization,2018(03):83-88.
3. Li Zhongzhi, Yin Hang, Zuo Jiankai, Liu Hedan. Bearing Fault Diagnosis Based on Generative Adversarial Network on Imbalanced Data[J].Journal of Chinese Computer Systems, 2021, 42(01):46-51.
4. Tapkan P, zbakir L, Kulluk S, et al. A cost-sensitive classification algorithm: BEE-Miner［J］. Knowledge-Based Systems, 2016, 95: 99-113.
5. Zughrat A, Mahfouf M, Yang Y Y, et al. Support vector machines for class imbalance rail data classification with bootstrapping-based over-sampling and under-sampling ［J］. IFAC Proceedings Vol-umes, 2014, 47(3):8756-8761.
6. BLAGUS R, LUSAL.SMOTE for high-dimensional class-imbalanced data[J]. BMC Bioinformatics, 2013, 14(1):106.
7. HA J, LEE J S. A new under-sampling method using genetic algorithm for imbalanced data classification[C]. IMCOM '16: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. NY, United States: Association for Computing Machinery, 2016: 1-6.
8. RAYHAN F, AHMED S, MAHBUB A, et al. CUSBoost: cluster-based under-sampling with boosting for imbalanced classification[C].2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). Bangalore, India: IEEE, 2017:1-5.
9. Cui Xin, Xu Hua, Zhu Liang. Multi-classification ensemble algorithm for imbalanced data [J/OL]. Computer Engineering and Applications:1-10[2021-0413]
10. He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C].Proceedings of the International Joint Conference on Neural Networks,2008,3:1322-1328.
11. Lin Yu, Huang Xun, Chun Weide, Huang Dengshi. An early warning study of extreme financial risk based on ODR-ADASYN-SVM[J]. Journal of Management Sciences in China,2016,19(05):87-101.
12. Duan Huajuan, Wei Yongqing, Liu Peiyu. An improved multi-decision tree algorithm for imbalanced classification[J]. Journal of Guangxi Normal University(Natural Science Edition),2020,38(2):72-80.
13. Tomek I. Two Modifications of CNN.IEEE Transactions on Systems Man and Communications SMC-6(11), 1976:769-772.