Check for updates

Original Article

Institution of MECHANICAL ENGINEERS

Journal of RAIL AND RAPID TRANSIT

Proc IMechE Part F:
J Rail and Rapid Transit
0(0) 1–14
© IMechE 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0954409718795089
journals.sagepub.com/home/pif
SAGE

Integrating synthetic minority oversampling and gradient boosting decision tree for bogie fault diagnosis in rail vehicles

Linlin Kou[1], Yong Qin[1,2], Xunjun Zhao[1] and Yong Fu[1]

Abstract

Bogies are critical components of a rail vehicle, which are important for the safe operation of rail transit. In this study, the authors analyzed the real vibration data of the bogies of a railway vehicle obtained from a Chinese subway company under four different operating conditions. The authors selected 15 feature indexes – that ranged from time-domain, energy, and entropy – as well as their correlations. The adaptive synthetic sampling approach–gradient boosting decision tree (ADASYN–GBDT) method is proposed for the bogie fault diagnosis. A comparison between ADASYN–GBDT and the three commonly used classifiers (K-nearest neighbor, support vector machine, and Gaussian naïve Bayes), combined with random forest as the feature selection, was done under different test data sizes. A confusion matrix was used to evaluate those classifiers. In K-nearest neighbor, support vector machine, and Gaussian naïve Bayes, the optimal features should be selected first, while the proposed method of this study does not need to select the optimal features. K-nearest neighbor, support vector machine, and Gaussian naïve Bayes produced inaccurate results in multi-class identification. It can be seen that the lowest false detection rates of the proposed ADASYN–GBDT model are 92.95% and 87.81% when proportion of the test dataset is 0.4 and 0.9, respectively. In addition, the ADASYN–GBDT model has the ability to correctly identify a fault, which makes it more practical and suitable for use in railway operations. The entire process (training and testing) was finished in 2.4231 s and the detection procedure took 0.0027 s on average. The results show that the proposed ADASYN–GBDT method satisfied the requirements of real-time performance and accuracy for online fault detection. It might therefore aid in the fault detection of bogies.

Keywords

Bogie, fault diagnosis, gradient boosting decision tree, adaptive synthetic sampling approach, imbalance dataset, multi-class

Date received: 19 February 2018; accepted: 15 July 2018

Introduction

Bogies are the main supportive parts in a rail vehicle, which are easily prone to failure. Bogie failures account for a substantial 21.1% based on the accumulation of failure data in a couple of years.[1] Thus, it is important to perform an efficient fault diagnosis on bogies used in operational rail vehicles.

There are mainly two approaches for bogie fault diagnosis, i.e. model-driven approach and data-driven approach. The model-driven approach builds physical models to simulate the mechanical degradation and failure of bogies through changing model parameters.[2] A rail vehicle bogie basically consists of a bogie frame, a wheel set, an axle box (with bearing), a driving device (only on EMU, contains motor, shaft coupling, gear box, etc.), a primary suspension mechanism (air spring, etc.), and so on. As illustrated in Figure 1, the complex structure and nonlinear coupled components make it quite difficult to develop such models. Therefore, the data-driven approach becomes more and more popular.[3]

The data-driven approach aims to derive fault information from condition monitoring data by machine learning methods. In this process, it is usually assumed that the instance number of all classes in

[1] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
[2] Beijing Key Laboratory of Performance Guarantee on Urban Rail Transit Vehicles, Beijing, China

Corresponding author:
Yong Qin, Beijing Jiaotong University, No. 3, Shang Yuan Cun, Hai Dian District, Beijing 100044, China.
Email: yqin@bjtu.edu.cn

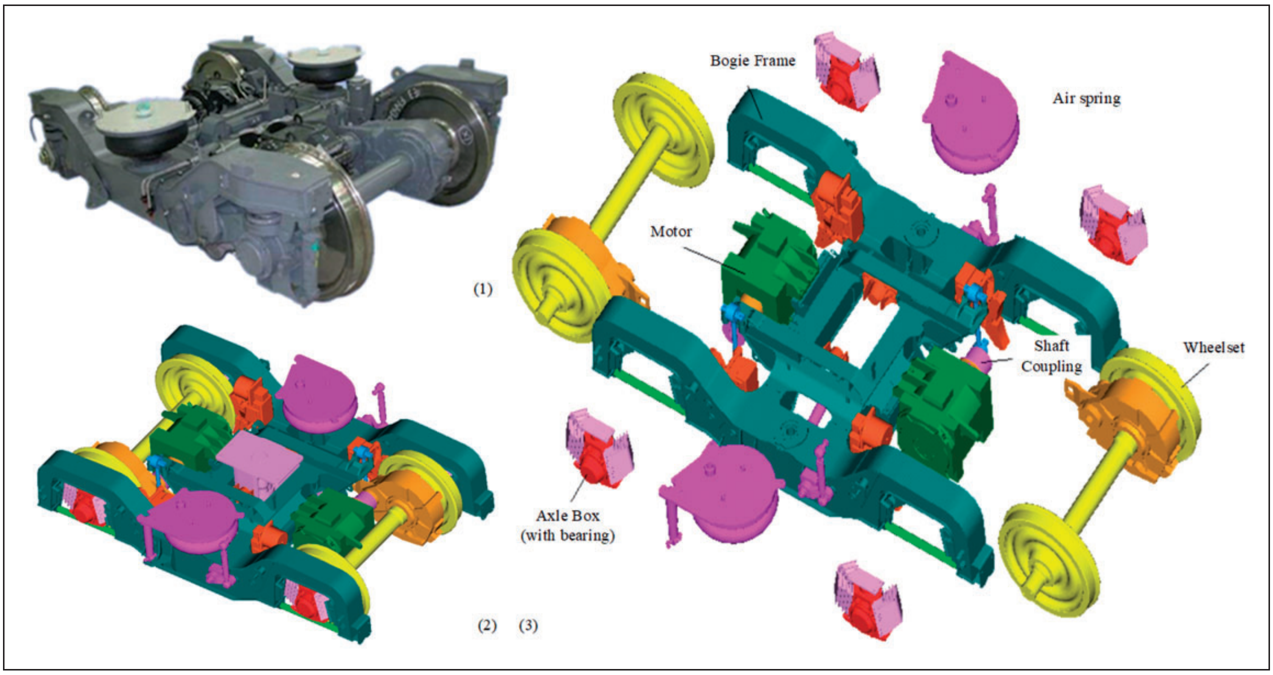

Figure 1. System composition of a bogie.

a dataset is equal to each other. However, bogie field data show a typical class imbalance characteristic, i.e. compared with data in normal condition, fault data are much smaller in size. Furthermore, in railway transit systems, the main attention is paid on fault conditions, which are the minority classes in the dataset, and misclassification of those classes comes at a high price. Therefore, the class imbalance problem has a great impact on the accuracy of data-driven bogie fault diagnosis. Another essential issue is the algorithm efficiency problem. Actual data are more complicated than the simulated ones. In order to obtain good results, high-dimensional feature space is commonly designed in bogie fault diagnosis. Dimensionality reduction method like principal component analysis (PCA) is necessary for rapid calculation. However, such procedure is also time consuming, which is contrary to the computational efficiency requirement. Finally, strong controllability is also an important factor in choosing the classifier for practical application. The working mechanism of the method and principle should be clear for controllable results. Therefore, a controllable algorithm that can deal with class imbalance problem and handling high-dimensional features efficiently is urgent in need.

To the authors' best knowledge, research on using class imbalance data in rail vehicle fault diagnosis has not been fully documented to date. Many approaches have been developed to solve the class imbalance problem in other fields. Under-sampling techniques[4] decrease the frequency of the majority class to balance data. However, it may be at the expense of losing important information of majority.[5] Cost-sensitive methods use different cost matrices that describe the costs for misclassifying any particular sample, instead of creating a balanced dataset. Some studies have shown that cost-sensitive methods perform better than sampling method,[6,7] but they show a high degree of variance in the performance measures compared to the least expected cost over the evaluated datasets.[5] Oversampling is the third approach to solve the class imbalance problem. The basic idea of oversampling methods is to increase the minority class samples by altering the samples in the dataset distribution. The synthetic minority oversampling technique (SMOTE) is a popular method proposed by Chawla et al.[8] It constructed the synthetic minority samples through the interpolation between minority training data and its K-nearest neighborhoods. However, SMOTE would be problematic of categories overlap on newly generated samples, as it synthesized new sample of each existing data point without considering that the selected K-nearest neighborhoods were not in the same class of the current sample points. Adaptive synthetic (ADASYN) sampling approach overcame the previous problem of SMOTE, and was also more efficient[9] than those methods mentioned above. He et al.[9] applied ADASYN to deal with the class imbalance problem to recognize abnormal heart sounds and got a sensitivity improvement of 58.6–84.4%. It can not only reduce the learning bias introduced by the original imbalance data distribution, but can also adaptively shift the decision boundary to focus on those difficult-to-learn samples.[9] However, bogie fault diagnosis is a multi-class classification problem in which imbalance exists among more than two classes; no previous work has been done to address multi-class oversampling problem for ADASYN.

Classification modeling plays an important role in bogie fault diagnosis. Many researches had been conducted on different classifiers for fault diagnosis of the rail vehicle. Hu et al.[10] got a high accuracy of over 98% by the deep neural network with the cost of 6.2 s on a high-performance computer, and the accuracy rate depends much on the sample size. A support vector machine (SVM) method got a relatively high recall value, with the calculation of 43 variables and feature selection procedure for braking system in high-speed trains.[11] The permutation entropy of IMFs after EEMD and initial signals are calculated as multi-scale complexity measure feature vectors together with SVM to classify and identify the operating conditions of high-speed train bogies.[12] But it worked on a binary system. To the best of our knowledge, most of them are expensive, time consuming or do not produce good results for a binary system. No previous research has studied the requirements of both less time computation and multi-class identification with consideration of imbalanced dataset in the fault diagnosis of bogies.

Gradient boosting decision tree (GBDT) method is a powerful boosting method based on the decision tree model, originally derived by Friedman.[13] It sequentially generates base models from a weighted version of the training data and searches the optimal combination of trees. At each step, a new base model is incorporated to correct the mistakes made by previous base models. Therefore, the gradient boosting method has the potential to provide a more accurate classifier. Moreover, GBDT derives feature contributing information from the fitted regression trees, which intrinsically perform feature selection by choosing appropriate split points, and it is able to handle different types of predictor variables with fit complex nonlinear relationship. The method shows a great performance in classification accuracy and timeliness on EEG,[14] online visual tracking,[15] prediction of hospital transfer and mortality,[16] as well as short-term subway ridership.[17] There are limited studies on the application of tree-based ensemble methods in rail vehicle fault diagnosis, and GBDT gets into trouble when it comes to class imbalance dataset.

In this paper, an integrated approach of ADASYN–GBDT is proposed to build a rapid and high-performance multi-class classifier for the online detection of bogie fault based on the real-running data, especially with the treatment of imbalanced dataset. In the proposed method, the improved ADASYN is used to deal with the class imbalance problem; and GBDT works for achieving high classification accuracy and reduced time consumption. More importantly, the real-running data obtained from a Chinese subway company is used in the validation. A plain probe into the real-running data is also made to give a preliminary complex vibration characteristic of the bogie, in this paper.

The rest of this paper is organized as follows. Section "Methodology" provides the proposed approach of bogie fault diagnosis. In section "Background and data analysis", description of bogie system composition, and common failure mode, and signal acquisition system are introduced. The comparison between the proposed integrated approach of ADASYN–GBDT and three commonly used classifiers (K-nearest neighbor, KNN; SVM; and Gaussian naïve Bayes, GaussianNB) in bogie fault detection was done in section "Experiments". The final section concludes the whole work.

Methodology

A multi-class classifier is proposed in this section. The proposed synthetic approach integrates the improved ADASYN and GBDT for more timely and precise diagnostic performance of a bogie fault, with consideration of the class imbalance dataset. GBDT works pretty well with respect to efficiency and accuracy as a classifier with the balanced dataset of high-dimension features.

GBDT got good results in online visual tracking[15] and short-term subway ridership,[17] etc. in both accuracy and operating implicity. By combining weak classification models, typically decision tree, with 'poor' performance, it usually produces high classification accuracy, and in contrast to other machine learning methods that have been treated as black-boxes, gradient boosting method provides interpretable results, while requiring little data preprocessing, and is able to handle different types of predictor variables with fit complex nonlinear relationship. It works pretty well in prediction and system identification with balance data. However, there is little research shown that it can get a good result on imbalanced dataset of multi-class system. An improved ADASYN method is proposed to overcome such substance mentioned before.

The improved ADASYN will make GBDT more suitable in our bogie multi-fault detection, where the characteristic of imbalance with multi-class is extremely prominent.

The improved adaptive synthetic sampling approach

ADASYN is mainly applied in medical and in binary systems. It increases the size of the minority class adaptively according to their density distributions,[9] and makes the classifiers more sensitive to minority class, especially the difficult-to-learn samples. However, to our best knowledge, ADASYN has only been used in binary systems, which is the biggest limitation in many cases with multi-class in real world. The ignorance of outliers also makes it a little inferior.

Some improvements have been made to generate new datasets to deal with multi-class data generation in this subsection.

Step 1. Sort the classes according to sample number in each class, and find out the majority class.

$$\begin{cases} C_{\max} = \max\{C_0, C_1, \ldots, C_N\} \\ sort\{C_0, C_1, \ldots, C_N\} = \{SC_0, SC_1, \ldots, SC_N\} \end{cases} \tag{1}$$

where $C_{\max} = SC_N$.

Step 2

1. Calculate the ratio of class imbalance of second largest class, that is

$$r_{N-1} = SC_{N-1}/C_{\max} \tag{2}$$

where $r_{N-1} \in (0,\ 1]$. The classes that need to be resampled are arranged in order of their sample size to reduce the possibility of generating more samples on the decision boundary.

2. If $r_{N-1} > r_{th}$, there is no need to generate new sample for SC_{N-1}; If $r_{N-1} < r_{th}$ (r_{th} is a present threshold for the maximum tolerated degree of class imbalance ratio), then:

① Calculate the number of synthetic data examples that need to be generated for the minority class:

$$G_{N-1} = (C_{\max} - SC_{N-1}) \times \beta \tag{3}$$

where $\beta \in [0,\ 1]$ specifies the desired balance level after generation of the synthetic data. It represents the percentage of sample in minority class after generation to that in majority class. A fully balanced dataset is created when $\beta = 1$, which means the number of samples in minority class after generating is equal to that in the majority class. When $\beta = 0$, no new sample will be generated.

② For $x_i \in SC_{N-1}$, find K-nearest neighbors based on the Euclidean distance, and calculate the ratio E_n

$$E_{n,N-1} = \Delta_{n,N-1}/K \tag{4}$$

where $\Delta_{n,\ N-1}$ is the example number in K-nearest neighbors of x_i that belong to $C_{\max}$, therefore $E_{n,N-1} \in [0,\ 1]$;

③ To ensure $\hat{E}_{n,N-1}$ is a density distribution ($\sum_{n=1} \hat{E}_{n,N-1} = 1$), we normalized $E_{n,N-1}$ according to

$$\hat{E}_{n,N-1} = E_{n,N-1} \Big/ \sum_{n=1}^{SC_{N-1}} E_{n,N-1}$$

④ Calculate the number of synthetic data examples that need to be generated. In this regard

$$g_{N-1} = \hat{E}_{n,N-1} \times G_{N-1} \tag{5}$$

where G_{N-1} is the total number of synthetic data examples that need to be generated as defined in equation (3).

⑤ For each example x_i in minority classes, generate g_{N-1} synthetic data examples. There are $newSC_{N-1}$ samples in second largest class,

$$newSC_{N-1} = SC_{N-1} + g_{N-1} \tag{6}$$

Step 3. It is almost the same as step 2; however, we have to calculate the ratio with the largest and the second largest class which has added new samples.

$$\begin{aligned} r_{N-2,1} &= SC_{N-2}/C_{\max} \\ r_{N-2,2} &= SC_{N-2}/newSC_{N-1} \end{aligned} \tag{7}$$

and compare them with r_{th}, calculate the number of synthetic data examples that need to be generated

$$\begin{aligned} G_{N-2,1} &= (C_{\max} - SC_{N-2}) \times \beta \times \left(\frac{r_{N-2,1}}{r_{N-2,1} + r_{N-2,2}} \right) \\ G_{N-2,2} &= (newSC_{N-1} - SC_{N-2}) \times \beta \\ &\quad \times \left(\frac{r_{N-2,2}}{r_{N-2,1} + r_{N-2,2}} \right) \end{aligned} \tag{8}$$

The ratio $E_{n,N-2,1}$, $E_{n,N-2,2}$

$$\begin{aligned} E_{n,N-2,1} &= \Delta_{n,N-2,1}/K \\ E_{n,N-2,2} &= \Delta_{n,N-2,2}/K \end{aligned} \tag{9}$$

where $\Delta_{n,N-2,1}$ is the example number in K-nearest neighbors of x_i that belong to $C_{\max}$, and $\Delta_{n,N-2,2}$ is the examples number in K-nearest neighbors of x_i that belong to $newSC_{N-1}$. They are also changed into density distribution $\hat{E}_{n,N-2,1}$, $\hat{E}_{n,N-2,2}$.

The number of synthetic data examples that needs to be generated is

$$\begin{aligned} g_{N-2,1} &= \hat{E}_{n,N-2,1} \times G_{N-2,1} \\ g_{N-2,2} &= \hat{E}_{n,N-2,2} \times G_{N-2,2} \end{aligned} \tag{10}$$

And the total number of SC_{N-2} is

$$newSC_{N-2} = SC_{N-2} + g_{N-2,1} + g_{N-2,2} \tag{11}$$

The third step will be repeated for SC_{N-3}, SC_{N-4} to SC_1, and each new step will compare the target generating class with all the previous generated classes.

In order to eliminate the outliers, we make a little bit difference. If $E_n = 1$, $n = 1, \ldots, SC_j$ in each step, x_i is surrounded by samples belonging to other classes, then x_i is regarded as an outlier and removed from the training dataset.

Overall ADASYN–GBDT fault diagnosis architecture

ADASYN is a data per-processing method which balances the dataset for classification in next procedure where GBDT is the classifier to detect the bogie fault. It adaptively shifts the decision boundary to focus on those difficult-to-learn samples, which are the fault samples in our case, and more efficient than other resample methods.[9]

Fault diagnosis algorithm based on time-domain parameters is characterized as simple and rapid direction, which makes it more capable in online project application. Related indexes are root mean square (RMS), peak, skewness, kurtosis, skewness factor, kurtosis factor, shape factor, crest factor, impulse factor, etc. They show good performance in fault diagnosis individually or in combination.[18–22] The energy and entropy parameters, like energy, torque, Shannon entropy, energy entropy, which contain more information, are also added to this paper to show the characteristic of bogie vibration and get a better diagnosis result in different situations.

Too many features lead to high computational complexity and weak timeliness. These irrelevant and redundant features, which are detrimental for a classifier, should be avoided.[23,24] Feature dimension reduction can improve the model generalization ability, and avoid over-fitting. It is a necessary step when feature is in high dimension.

GBDT[25] is an ensembled machine-learning technique of decision trees as the weak prediction model for regression and classification. It builds the model in a stage-wise fashion. An arbitrary differentiable loss function is used in the sequential error-correcting process to converge to an accurate model. GBDT classifier learns from part of the features instead of learning from all of them recursively and repeatedly, and it can also do interaction between multiple sets of features. These properties make GBDT easier to use, more accurate and timesaving.

The procedure for the ADASYN–GBDT based-bogie fault detection method is as follows.

1. Feature extraction. We calculate 12 time-domain parameters, RMS, peak, skewness, kurtosis, skewness factor, kurtosis factor, shape factor, crest factor, impulse factor, together with energy, Shannon entropy and energy entropy to represent the bogie vibration dataset.
2. Identify whether the training dataset is imbalanced or not. The standard ratio of abnormal samples to normal ones is set to be 0.5 intuitively, if it is smaller than 0.5, go to (3), otherwise, go to (4).
3. They are sent to ADASYN method to resample more abnormal datasets. In this procedure, we try to get the similarly sized resampled datasets with normal samples. Resample abnormal datasets and original normal datasets are mixed as the input of GBDT for training.
4. Train the GBDT model. Cross-validation is conducted here in order to limit problems like overfitting. The processed dataset is divided into two segments: one used to train the model and the other used to validate the model.
5. New sample identification. The new sample is sent to the trained ADASYN–GBDT model.

The framework can be represented as shown in Figure 2.

Background and data analysis

Bogie description

A bogie does a critical role in railway vehicle operation. It supports the car body, transfers load, traction and braking force between wheel–rail and car body, guarantees safety when railway cars running on curve, relaxes impact and attenuates vibration. The interested reader can refer to Dong[26] for details.

The common failure modes related to the bogie system[1] are crack, wear, abrasion and warping on bogie frame; tread peeling and scrape, defect and molten slag on the wheel set (collectively called wheel flat); axle box problems; defect of air spring; leakage and damage of gearbox; defect of anti-snake movement damper, etc. We would like to give some examples of bogie failure modes which are not acceptable even in safe operation.

Peeling, scrape, crack and sloughing on the surface of wheel tread are collectively called wheel flats.[27,28] Wheel tread peeling is a phenomenon that happens

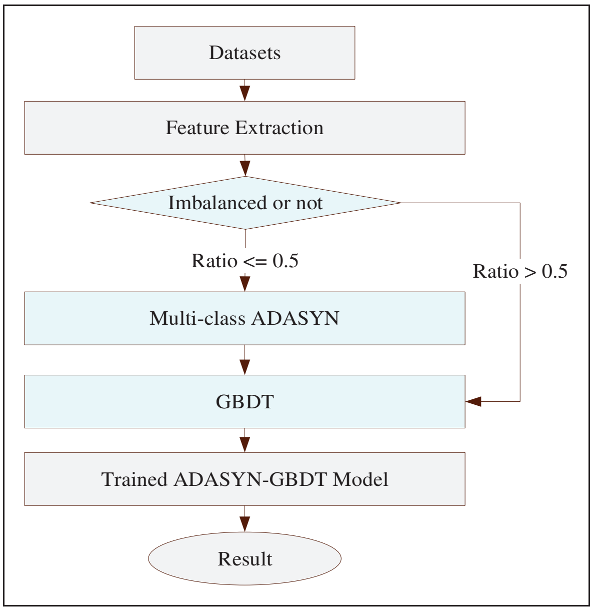

Figure 2. ADASYN–GBDT-based method for bogie fault diagnosis.
ADASYN–GBDT: adaptive synthetic sampling approach–gradient boosting decision tree.

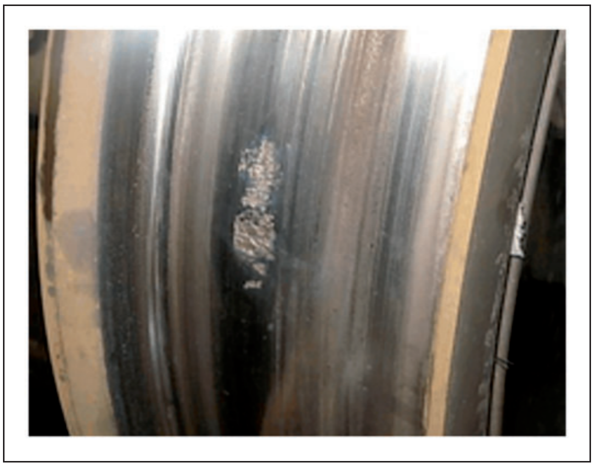

Figure 3. Peeling of the wheel tread.

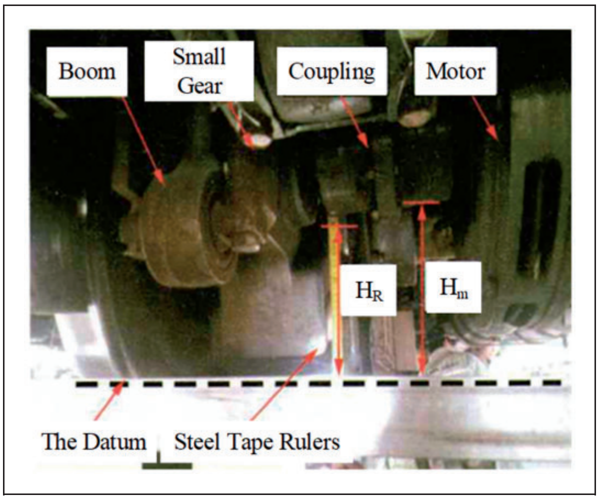

Figure 4. Misalignment of the shaft coupling.[29]

frequently where metal on the surface exfoliates or flakily lifts (shown in Figure 3). Friction heat makes the local tissue on the wheel tread to change and then the tread falls off when wheels slide on the rail. The other reason is the poor quality of wheel materials. Fatigue failure occurs after the wheel tread was extruded by the rail continuously. The shaft coupling is used to pass motor torque to the gear box, so as to drive the wheels. Shaft coupling misalignment (shown in Figure 4) may happen as results of shaft vibration and beat, bearing abrasion, bad stress, variable load, etc. during train operation.

Signal acquisition

We present a simplified structure of the signal acquisition network on the railway vehicle (shown in Figure 5). A brief introduction of the system is given as below.

1. Vibration and temperature compound sensor is the general purpose accelerometer with internal temperature sensor of type 787T, produced by Wilcoxon company.
2. Access point transfers data from the sensors to fusion point. Signal isolation processing, analog-to-digital conversion, and digital signal filtering would be conducted by a signal conditioner in access point.
3. Fusion point and network point are used to do data concentration, data transmission and priority assignment (because this system is not only for running part. It works for real-time monitoring and per-warning of the whole vehicle), dynamic network organizing and so on.
4. Service host receives data from the network point, and divides them into packets, then does data direction controlling and storage based on different subsystems of the railway vehicle. Actually, service host can also conduct safety assessment, fault diagnosis and per-warning, and support terminal display. Our study is to find a more efficient method for this part.

More details of this system are beyond our competence, and cannot be provided in this paper.

Data preparation and analysis

This section describes data preparation, summary of the datasets and the mechanism.

Data preparation. Real-running vibration data of railway vehicles provided by a Chinese subway company are used in our study.

Data were collected from a bullet-train carriage of type-A vehicle. The bullet-train carriage of the tested vehicle weighs about 38 tons, and its load is zero while testing. Each bullet-train contains two bogies, and each bogie has four wheels. Therefore, the load on each wheel is about 4.75 tons.

The rail vehicles were running on the track in a metro depot at a speed of 35 ± 5 km/h while gathering the data. The schematic of the bogie and locations of vibration sensors are shown in Figure 6, and sample frequency is set to be 10 kHz. The permissible range of wheel diameter is 770–840 mm. We chose the half wear (also the mean value) wheel diameter to compute the rotate speed, which is about 3.846 r/min/s.

The dataset consists of four types of bogie conditions, operation with no trouble (Normal), wheel out-of-roundness or flat (Fault 1), shaft misalignment (Fault 2), wheel run out (Fault 3). We got 2569–2002 samples of operation with no trouble, 96 samples of wheel out-of-roundness or flat, and 96 samples of shaft misalignment, 375 samples of wheel run out. Each sample contains 32,768 points. The number of samples is not governed by us. In most cases, rail vehicles would be stopped even there is a suspected fault or a slight fault, due to the strict actual

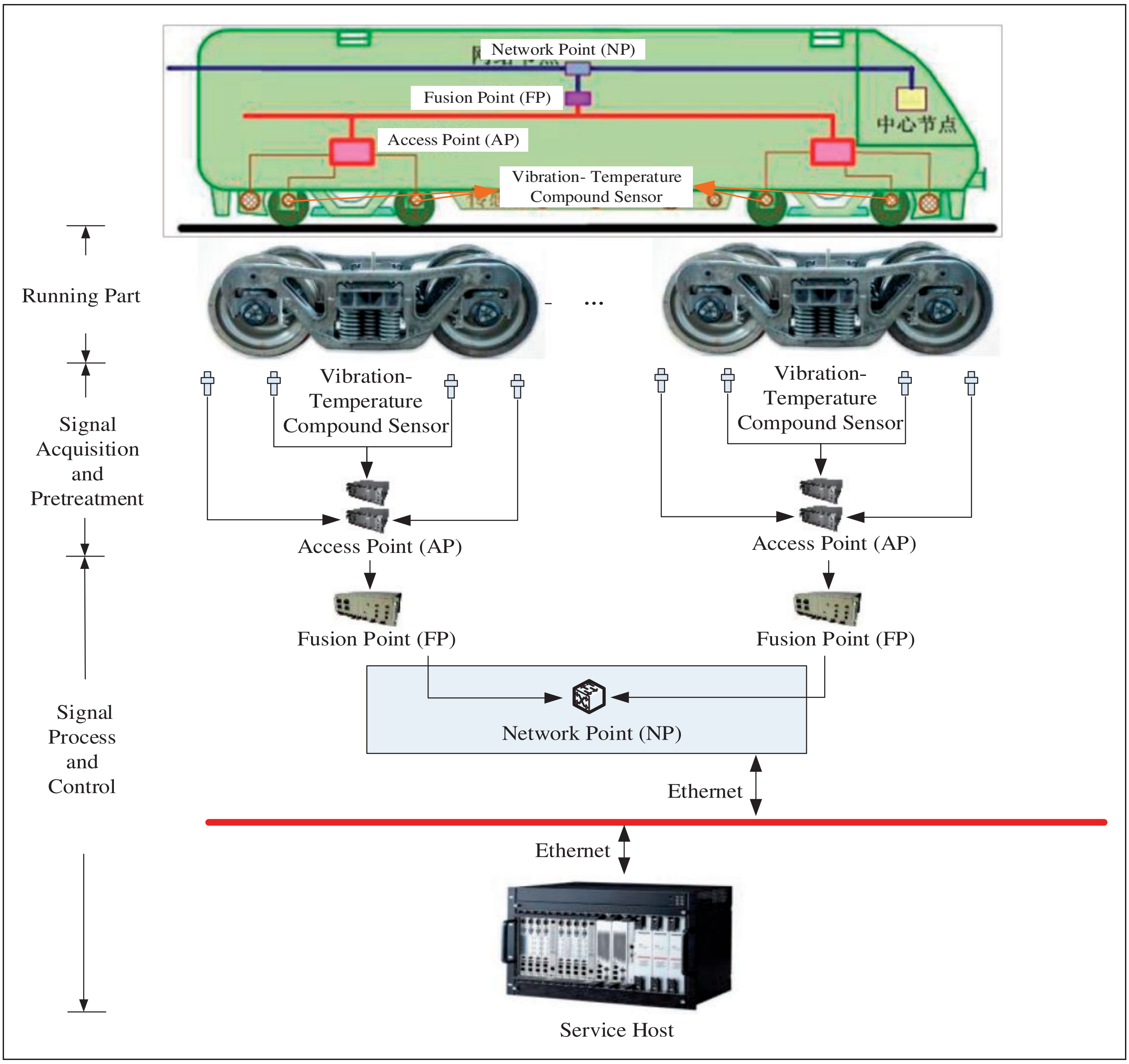

Figure 5. Part of the signal acquisition network on a railway vehicle.

application requirements. It is a common situation that the fault diagnosis of rail vehicle is desperately short of useful samples in different failure modes, especially in various fault degrees.

Data analysis. Time-domain waveform and their details under four conditions are shown in Figure 7. There are obvious periodic feature in bogie vibration data from its time-domain waveform. Amplitude of vibration in abnormal conditions was significantly larger than that in normal ones (shown as red rectangles in Figure 7). Vibration shows slight difference from the operation with no trouble to wheel run out condition. Signal possesses a clear sinusoidal property with random noise under condition of operation with no trouble, and shock characteristics in abnormal situations.

Based on the above analysis and the real situation, the difficulties of online fault diagnosis of railway vehicle can be summarized as follows

1. The velocity is hard to stay stable while testing, due to the route protection, switch protection, vehicle and its over-speed protection, as well as the driver's behavior. The running process increased the non-stationary level of the dynamic status signals and features.
2. The data were collected on the same type vehicle, but not the same one. In addition, to get different fault type dataset, we detected vehicles in a long period. The data were easily interfered by the changeable environment. The characteristic of multi-formity and complexity affects method's forecast accuracy and suitability of application.

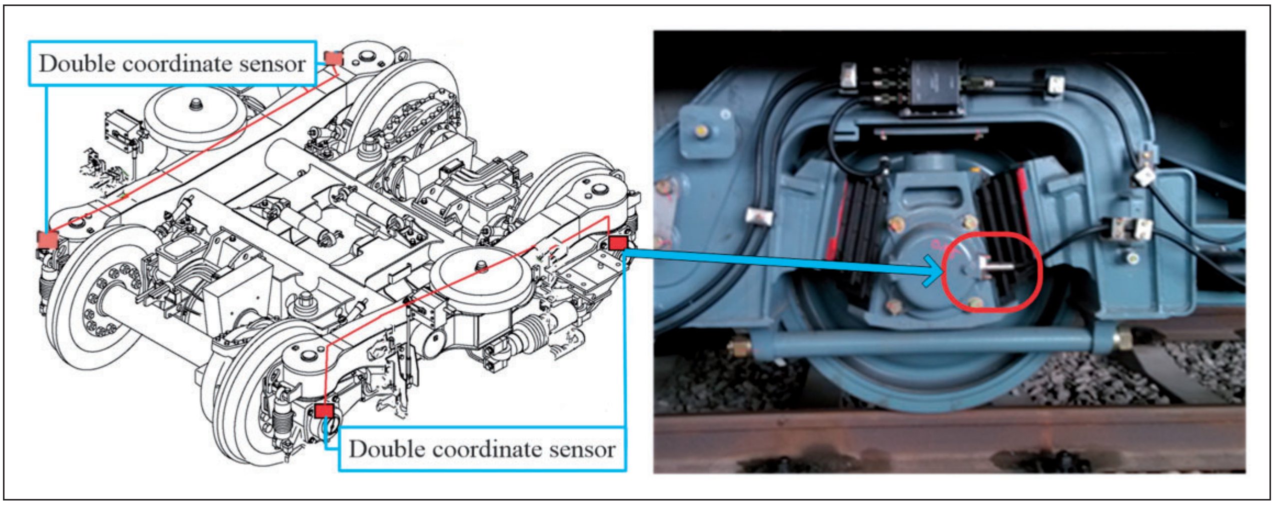

Figure 6. Schematic of the bogie and locations of the vibration sensors.

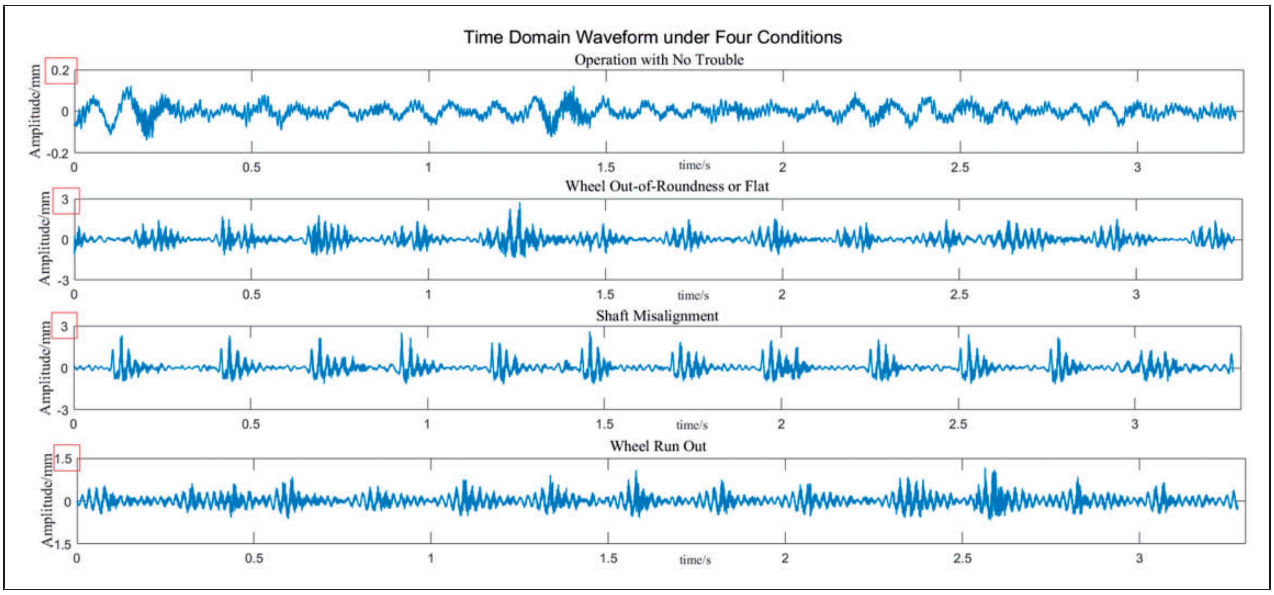

Figure 7. Time-domain waveform of the vibration data under four conditions.

3. As faults occur occasionally, fault data are much smaller in size than those data in normal condition. The typical imbalance characteristic makes misclassification more prone.
4. Determination of vehicle failure and seizing the alarm substance in a very short period of time under such circumstances are also required for the real-time safety operation. Feature dimension should be reduced to get a high accuracy and reduced time consumption in classification. However, dimension reduction algorithm itself costs time.

The complex relations between features, non-stationary of the signal, the class imbalance problem, together with the high accuracy and efficiency requirement make online bogie multi-fault detection very difficult. Experiments are conducted in the next section to show how efficient the integrated model is.

Experiments

To test the effectiveness and robustness of the proposed ADASYN–GBDT method, this section comprehensively evaluates the performance of KNN, SVM, GaussianNB methods under different proportions of the test dataset.

KNN is commonly used as a classifier method[30] because of its simplicity and computational efficiency. It is also a lazy algorithm that does not use the training data points to do any generalization. In other words, it keeps all the training data during the testing phase,[31] which has been widely applied in many areas.[32]

SVM is one of the best working classifiers, due to its excellent generalization ability.[33] It reduces the claim on data scale and distribution by structural description of data distribution with margin concept. SVM showed state-of-the-art performance in real-world applications such as text categorization[34] and bogie fault detection.[11,35]

Naïve Bayes is totally different from other classifiers. Compared to most of the other classification methods – like decision tree, KNN, logistic, SVM, etc., which are all discriminating methods that learn the relationship between feature X and output Y directly, or the decision function $Y = f(X)$, or the conditional distribution $P(Y|X)$ – Naïve Bayes is a generated method. It tries to find out the joint distribution $P(X, Y)$ of feature X and output Y, then gets the result of $P(Y|X) = P(X, Y)/P(X)$. Naïve Bayes is intuitive, and not sensitive to missing data, with small computation volume.[36] It works pretty well with respect to accuracy, precision, and specificity in the instructors' performance evaluation.[37]

Evaluation index

Confusion matrix is chosen as the evaluation index in this study. Confusion matrix is a visualization tool typically used in the field of machine learning and specifically for the problem of statistical classification. Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one, also known as an error matrix.[38] Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

In binary classification, a confusion matrix is a table with two rows and two columns that report the number of false positives, false negatives, true positives, and true negatives (shown in Table 1). This allows a more detailed analysis than mere proportion of correct guesses (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the dataset is unbalanced (i.e. when the number of samples in different classes vary greatly). The four outcomes can be formulated in a 2×2 confusion matrix, as follows.

A confusion matrix describes how many results were correctly classified (TP and TN) and how many were incorrectly classified (FN and FP) for each of the categories.

Comparison and discussion

Analyses were performed with the Python system (version 3.5) and the scikit-learn (version 0.17.1) for statistical computing on a desktop computer with Intel(R) Core(TM) i3-3240 CPU @ 3.40 GHz and 8.0 GB RAM.

Parameters in each classifier are set as below. As with the KNN method, we set the number of neighbors used to be 10, and all points in each neighborhood are weighted equally. Euclidean distance is used as the decision function of KNN. Support vector machine is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. It is widely used in railway vehicle fault diagnosis. In this paper, RBF kernel function and the shrinking heuristic are used, and penalty parameter C of the error term is set to be 1.0. Tolerance for stopping criterion is $1e-3$. All classes are supposed to have weight one. Parameters in GaussianNB method are default, where the prior probabilities of each class are adjusted according to the input data, and the weight applied to individual samples is 1. As to GBDT method, the number of boosting stages to perform, subsample values is 0.8,[39] the deviance loss function is chosen as the deviance for classification with probabilistic outputs, and learning rate is 0.1, while the number of boosting stages to perform and the maximum depth of the individual regression estimators are 100 and 4, respectively. The Friedman mean squared error with improvement is used here. Other parameters are set as default.

1. Performance comparison is taken between ADASYN–GBDT and KNN, SVM, GaussianNB. We are trying to present the great performance of ADASYN–GBDT in multi-class classification and the superiority on imbalanced dataset. The proportion of the test dataset is set to be 0.4.

The grid color represents the percentage of the sample number that is classified into the corresponding classes, as shown in the stripe on the right of each confusion matrix (Figure 8). For example, in the first

Table 1. Confusion matrix.

		Predicted condition	
	Total population	Prediction positive	Prediction negative
True condition	Condition positive	True positive (*TP*)	False negative (*FN*)
	Condition negative	False positive (*FP*)	True negative (*TN*)

Note: $TP + FN = 1$, $FP + TN = 1$ in normalized confusion matrix.

column of Figure 8(a), 0.9200 is diagnostic accuracy of normal class, and 0.0200, 0.0050, and 0.0550 is the percentages of normal samples which are misclassified into Fault 1, Fault 2, and Fault 3 class, respectively. Values that lie along the diagonal from top left to bottom right of each matrix are the precisions of the given four bogie conditions.

From Figure 8(a) and (b), different methods are sensitive to different failure modes, KNN and SVM are more likely affected by the imbalanced dataset, especially SVM, and samples in minority classes are more likely to be misclassified to the majority class. KNN works by accounting the number of samples that belong to certain class of the K-nearest neighbor among the objective. It is obvious that the likelihood of target sample classified into the majority class is greater than that into the minority ones. Furthermore, SVM is overfitting in our case due to the disparity of dataset, as the total penalty coefficient to samples in majority class is much greater than that to minority classes.

GaussianNB method gets a high accuracy of 94.59% only in Fault 1 class as shown in Figure 8(c). However, results of other three categories are far from acceptable, in particular with regard to the bogie normal state. Samples in normal state are more likely misclassified into the Fault 2 and Fault 3 classes with the high errors possibilities of 0.2375 and 0.5088. And about 25.97% of samples in Fault 3 condition were misclassified as Fault 1. That may be caused by the assumption that features are independent between each other, as well as the feature distribution of normal, Fault 2 and Fault 3 samples in each dimension is not subjected to Gaussian distribution which is the basic requirement of GaussianNB algorithm.

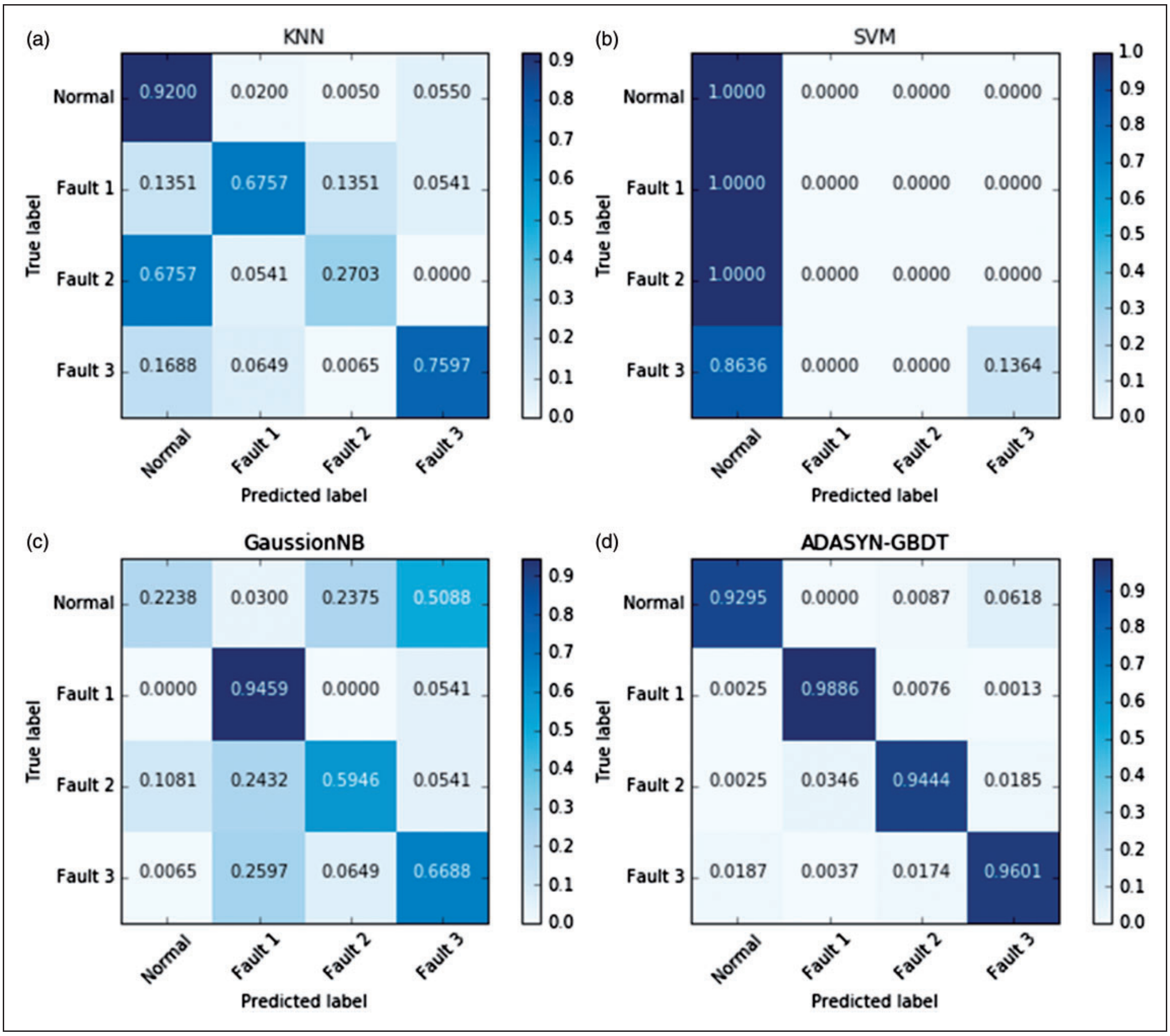

Figure 8. Normalized confusion matrix of (a) KNN, (b) SVM, (c) GaussianNB and (d) ADASYN–GBDT.
KNN: K-nearest neighbor; GaussianNB: Gaussian naïve Bayes; ADASYN–GBDT: adaptive synthetic sampling approach–gradient boosting decision tree.

As illustrated in Figure 8(a) to (c), KNN, SVM and GaussianNB methods have difficulties in identifying the bogie condition and will cause high false alarm rate and missed alarm rate, which would result in unnecessary inspection and cost in time and money.

The best result is obtained by our proposed method, ADASYN–GBDT. The integrated synthetic ADASYN–GBDT method took advantages of both ADASYN and GBDT method. ADASYN is set to solve the class imbalance problem in practical railway operation. It adaptively generates new samples of minority classes. While GBDT derives feature contributing information from the fitted regression trees which intrinsically perform feature selection by choosing appropriate split points, it is able to handle different types of predictor variables with fit complex nonlinear relationship. In addition, GBDT is an ensemble method. By combining weak classification models, typically decision tree, with ‘poor’ performance, it usually produces high classification accuracy. In contrast to other machine learning methods that have been treated as black-boxes, gradient boosting method provides interpretable results, while requiring little data preprocessing, and is able to handle different types of predictor variables with fit complex nonlinear relationship. It works pretty well in bogie multi-

Table 2. The average executed time.

Methods	Time
KNN	0.0177 s
SVM	2.6763 s
GaussianNB	0.0164 s
ADASYN–GBDT	2.4231 s

KNN: K-nearest neighbor; SVM: support vector machine; GaussianNB: Gaussian naïve Bayes; ADASYN–GBDT: adaptive synthetic sampling approach–gradient boosting decision tree.

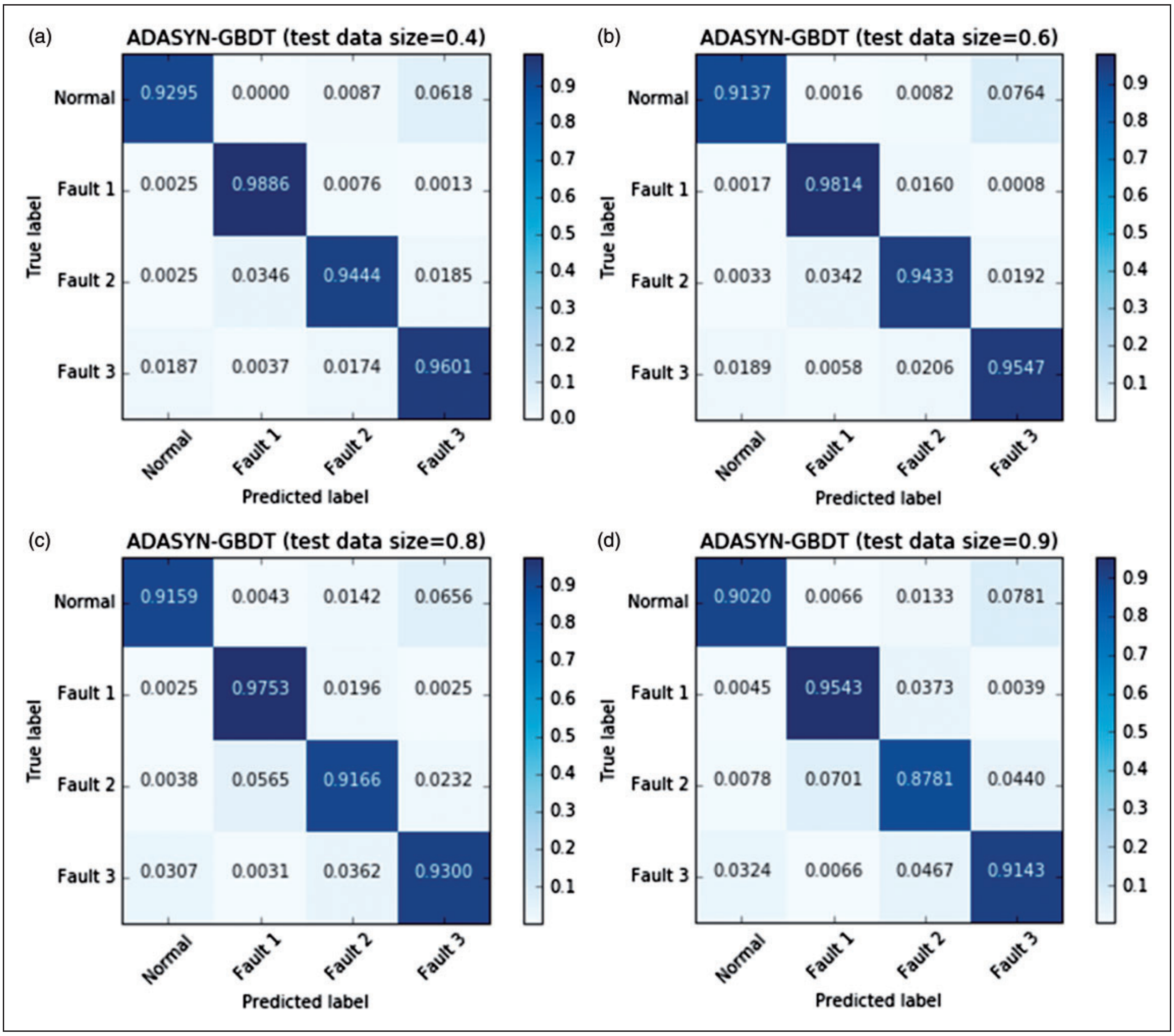

Figure 9. Normalized confusion matrix of ADASYN–GBDT at a proportion of the test dataset (a) 0.4, (b) 0.6, (c) 0.8, and (d) 0.9. ADASYN–GBDT: adaptive synthetic sampling approach–gradient boosting decision tree.

fault detection of railway vehicle in both accuracy and operating implicity.

We ran 100 trials and calculated the average execute time of each method as illustrated in Table 2. Without consideration of diagnosis accuracy, KNN and Gaussian are the fastest two methods, and they have the same order of magnitude in consuming time. However, much more effort should be taken to improve the precision, which is the top-priority in bogie fault diagnosis. Although ADASYN–GBDT is not the fastest method, it is faster and more accurate than SVM. Consuming time in detection procedure averages 0.0027 s. As the sample frequency is 10 k/s and a sample contains 32,768 data points, it takes about 3.3 s to obtain a sample and then input to our algorithm, which means our proposed method satisfies the requirements of real-time performance and accuracy for online fault detection of bogies.

2. Performance comparison is taken under different proportions of the test dataset to show the generalization ability of our ADASYN–GBDT method. Proportion of the test dataset is set to be 0.4, 0.6, 0.8, 0.9 in this regard.

The overall performance of the proposed ADASYN–GBDT method in fault detection is pretty good; more than 94% of faults are detected successfully when the test data size = 0.4 (shown in Figure 9(a)). Even in the case of test data size = 0.9, the lowest fault detection rate is still 87.81% (detection of Fault 2, shown in Figure 9(d)). False positive rates of normal state are 0.0705, 0.0863, 0.0841 and 0.0980 in the four cases, respectively, which is indeed acceptable. The robustness of ADASYN–GBDT method makes it more suitable for multi-class fault diagnosis, especially when the training data size is relatively small.

Samples of normal state are more likely to be misclassified to Fault 3, as there is a tolerance margin of the natural attrition for wheel before it is decommissioned. Fault 1 is easier to be identified than the other three conditions, because signal has more pronounced response to the impact caused by wheel out-of-roundness or flat. Manifestation and amplitude of axle misalignment are similar to that of wheel out-of-roundness or flat in time-domain signal, and most features used in this paper are in time domain and reflect the signal energy, so it is no surprise that the higher ratio of Fault 2 was misclassified into Fault 1 than into normal and Fault 3 conditions.

With the decrease of training data size, a slight drop of diagnosis accuracy for all conditions shows up. The performance of our proposed method is still subjected to sample size which is a common failing of machine learning algorithms.

In addition, the accuracy of three faults' detection is a little higher than that of samples in normal state. It indicates that the integrated synthetic ADASYN–GBDT method not only overcomes the class imbalance problem in the original datasets by artificially generating data samples in minority classes, but also pays more attention to fault detection by shifting its decision boundary to focus on those hard-to-learn examples which are the fault samples in our cases. That is really significant in our rail vehicle fault diagnosis to reduce operational risk and to avoid sudden failure.

Conclusion

In this study, an ADASYN–GBDT approach is investigated to detect bogie fault, with the consideration of imbalanced data of multi-classes which is very common in fault diagnosis of railway vehicle. The proposed ADASYN–GBDT model produced great results wherein the fault detection rate is at least 0.94 with false positive rates of 0.0583 (when the proportion of the test dataset is 0.4); the total executed time is 2.4231 s, detection time is 0.0027 s – on average – which is much shorter than the time of a sample, indicating that the ADASYN–GBDT model can be successfully used for real-time bogie fault detection.

The GBDT model can automatically select relevant variables, fit accurate models, and identify and model parameter interactions. More importantly, different from other machine learning algorithms which are working as a 'black-box', this model provides clear internal functioning mechanism, which is critical for engineering control. In addition, a comparison between KNN, SVM, GaussianNB and the proposed model indicates that the ADASYN–GBDT model is more suitable for multi-class imbalanced data classification with regard to its effectiveness and robustness in fault detection.

Our future work will explore the signal characteristics to extract more effective features. The classifier on a small sample size is also a research priority, because such a problem is more prominent in diagnosing a bogie fault in a practical situation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the National Key Research and Development Program of China (2016YFB1200401), the Open Research Fund Program of Beijing Key Laboratory of Performance Guarantee on Urban Rail Transit Vehicles and the State Key Laboratory of Rail Traffic Control and Safety (Contract No. RCS2016K005).

ORCID iD

Linlin Kou http://orcid.org/0000-0003-2296-6239

References

1. Guo L, Wang Y and Zhu L. Failure mode and effect analysis of the EMU bogie system. *Railway Locomot Car* 2013; 33: 97–100.
2. Ribeiro D, Calcada R, Delgado R, et al. Finite-element model calibration of a railway vehicle based on experimental modal parameters. *Veh Syst Dyn* 2013; 51: 821–856.
3. Shahidi P, Maraini D, Hopkins B, et al. Railcar bogie performance monitoring using mutual information and support vector machines. In: *2015 annual conference of the prognostics and health management society*, Coronado, CA, 2–5 October 2015, pp.363–372. Washington, DC, USA: IEEE.
4. Japkowicz N and Stephen S. The class imbalance problem: A systematic study. *Intell Data Anal* 2002; 6: 429–449.
5. Haibo H and Garcia EA. Learning from imbalanced data. *IEEE Trans Knowledge Data Eng* 2009; 21: 1263–1284.
6. Liu X-Y and Zhou Z-H. The influence of class imbalance on cost-sensitive learning: An empirical study. In: *2006 ICDM'06 sixth international conference on data mining*, Hong Kong, China, 18–22 December 2006, pp.970–974. Washington, DC, USA: IEEE Computer Society.
7. Yin QY, Zhang JS, Zhang CX, et al. An empirical study on the performance of cost-sensitive boosting algorithms with different levels of class imbalance. *Math Probl Eng* 2013; 2013: 1256–1271.
8. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–357.
9. He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *IEEE international joint conference on neural networks*, Hong Kong, China, 1–8 June 2008, pp.1322–1328. Washington, DC, USA: IEEE.
10. Hu h, Tang B, Gong X-j, et al. Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks. *IEEE Trans Ind Informat* 2017; 13: 2106–2116.
11. Liu J, Li Y-F and Zio E. A SVM framework for fault detection of the braking system in a high speed train. *Mech Syst Signal Process* 2017; 87: 401–409.
12. Na Q, Peng J, Yongkui S, et al. Fault diagnosis of high speed train bogie based on EEMD and permutation entropy. *J Vibr Meas Diagn* 2015; 5: 015.
13. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.
14. Yang T, Chen W and Cao G. Automated classification of neonatal amplitude-integrated EEG based on gradient boosting method. *Biomed Signal Process Control* 2016; 28: 50–57.
15. Son J, Jung I, Park K, et al. Tracking-by-segmentation with online gradient boosting decision tree. In: *IEEE international conference on computer vision*, Santiago, Chile, 7–13 December 2015, pp.3056–3064. Washington, DC, USA: IEEE.
16. Xie J and Coggeshall S. Prediction of transfers to tertiary care and hospital mortality: A gradient boosting decision tree approach. *Stat Anal Data Mining* 2010; 3: 253–258.
17. Ding C, Wang D, Ma X, et al. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability* 2016; 8: 1100.
18. Borghesani P, Pennacchi P and Chatterton S. The relationship between kurtosis- and envelope-based indexes for the diagnostic of rolling element bearings. *Mech Syst Signal Process* 2014; 43: 25–43.
19. Su L, Shi TL, Liu ZP, et al. Nondestructive diagnosis of flip chips based on vibration analysis using PCA-RBF. *Mech Syst Signal Process* 2017; 85: 849–856.
20. Ai YT, Guan JY, Fei CW, et al. Fusion information entropy method of rolling bearing fault diagnosis based on n-dimensional characteristic parameter distance. *Mech Syst Signal Process* 2017; 88: 123–136.
21. Asr MY, Ettefagh MM, Hassannejad R, et al. Diagnosis of combined faults in rotary machinery by non-naive Bayesian approach. *Mech Syst Signal Process* 2017; 85: 56–70.
22. Zhou J, Qin Y, Kou L, et al. Fault detection of rolling bearing based on FFT and classification. *J Adv Mech Des Syst Manuf* 2015; 9. DOI: 10.1299/jamdsm.2015jamdsm0056.
23. Koller D and Sahami M. Toward optimal feature selection. In: *Thirteenth international conference on international conference on machine learning*, Bari, Italy, 3–6 July 1996, pp.284–292. San Mateo, CA: Morgan Kaufmann Publishers Inc.
24. Hall MA and Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans Knowledge Data Eng* 2003; 15: 1437–1447.
25. Chirici G, Scotti R, Montaghi A, et al. Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery. *Int J Appl Earth Observ Geoinform* 2013; 25: 87–97.
26. Dong XM. *How high-speed EMUs works and their structure characters.* Beijing, China: China Railway Publishing House, 2007.
27. Wang Y. Review of dynamic detecting methods for railway wheel flat. *Rolling Stock* 2002; 40: 9–12.
28. Qin N, Jin W-D, Huang J, et al. Ensemble empirical mode decomposition and fuzzy entropy in fault feature analysis for high-speed train bogie. *Kongzhi Lilun Yu Yingyong/Control Theory Appl* 2014; 31: 1245–1251.
29. Shi H, Wang J, Dai H, et al. Car body vibration analysis subject to coupling misalignment in traction system of metro vehicle. *Zhendong Ceshi Yu Zhenduan/J Vibr Meas Diagn* 2015; 35: 626–631.
30. Yigit H. A weighting approach for KNN classifier. In: *2013 international conference on electronics, computer and computation (ICECCO)*, Ankara, Turkey, 7–9 November 2013, pp.228–231. Washington, DC, USA: IEEE.
31. Song Y, Liang J, Lu J, et al. An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing* 2017; 251: 26–34.
32. Zhang S, Li X, Zong M, et al. Learning k for kNN classification. *ACM Trans Intell Syst Technol (TIST)* 2017; 8: 43.

33. Gu B, Sheng VS, Tay KY, et al. Cross validation through two-dimensional solution surface for cost-sensitive SVM. *IEEE Trans Pattern Anal Mach Intell* 2017; 39: 1103–1121.
34. Li B, Chen N, Wen J, et al. Text categorization system for stock prediction. *Int J u- e-Serv Sci Technol* 2015; 8: 35–44.
35. Li C, Chen D and Yang L. Research on fault detection of high-speed train bogie. In: Li X and Xu X (eds). *Proceedings of the fourth international forum on decision sciences*, Qingdao, China, 15–18 July 2016, pp.253–260. Singapore: Springer Singapore.
36. Roiger RJ. *Data mining: A tutorial-based primer*. Boca Raton, USA: CRC Press, 2017.
37. Patil PS, Choudhary N and Amalnerkar A. Predicting instructor performance using Naïve Bayes classification algorithm in data mining technique: A survey. *Int J Adv Electron Commun Syst* 2017; 6: 9–12.
38. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. *Rem Sens Environ* 1997; 62: 77–89.
39. Jain A. Complete guide to parameter tuning in gradient boosting (GBM) in Python, 21 February 2016.