An integrated model of kNN and GBDT for fault diagnosis of wheel on railway vehicle

Linlin Kou[1], Yong Qin[1], Xuejun Zhao[1]
1. State key laboratory of rail traffic control and safety.
Beijing Jiaotong University
Beijing, China

***Abstract*—Wheelsets are critically important for the safety operation of rail vehicle. The real vibration data of railway vehicle wheel, obtained from a Chinese subway company, is used in this article. Principal component analysis (PCA) is conducted to reduce the dimension of the feature indexes. We propose an integrating algorithm with k-Nearest Neighbours (kNN) and gradient boosting decision tree (GBDT) to deal with the typical imbalance and divergent characteristics and satisfy the high requirement of fault detection accuracy. The results show that the classification accuracy of kNN-GBDT reaches 94.66%, 82.35%. While, the kNN and the SVM miss classified all fault samples into normal condition, and GBDT got an accuracy of 56.82% in fault detection. The entire process of the proposed model finished in about 0.35s. Our kNN-GBDT integrated algorithm satisfies the requirements of real-time performance and accuracy for online fault detection.**

Keywords—wheel; rail vehicle; fault diagnosis; imbalance, KNN; gradient boosting decision tree

I. Introduction

Wheelsets are the main supportive parts of rail vehicle, which is also one of the parts easily prone to trouble. Research on fault diagnosis has important practical significance. Railway data shows a typical imbalance characteristic, i.e. compared with data in normal condition, fault data is much smaller in size. In railway transit system, the main attention was paid on fault conditions, which are the minority classes in dataset, and misclassification of those classes comes at a high price. Learning from imbalanced data sets is a challenge for many of today's data mining applications [1]. There are two approaches to handle imbalanced datasets, over-sampling [2, 3] and under-sampling techniques. Over-sampling technique may increase the overlap between classes, destroy the relationship between feature indexes, impact the data distribution, or generate useless samples, which induce more problems in classification. While under-sampling techniques like random under-sampling [4], etc., abandon some samples in majority class. It loss information of the data, which is very precious for railway vehicle fault diagnosis in real world. To the authors' best knowledge, research on imbalanced data in rail vehicle fault diagnosis has not been fully documented to date.

Because the commonly used classification methods, like support vector machine (SVM) [5, 6], naive Bayes (NB) [7, 8], decision tree [9] are easily overfitted when it comes to the actual running data based on our experiments. We propose an integrating algorithm with PCA, KNN and GBDT methods to overcome the imbalance problem and get a relatively efficient classifier, where PCA is used for dimension reduction, KNN for per-classification and GBDT for advanced fault diagnosis on difficult samples.

This paper is arranged as follows, Description of wheel and definitions of the fifteen feature index is given in section 2. The proposed kNN-GBDT model is in section 3. We introduced the kNN-GBDT algorithm to wheel fault detection of rail vehicle, compared with other classifiers SVM, KNN, and GBDT is in section 4. The last part is the conclusion.

II. Description of Wheel and its Vibration Data

A. Wheel Description

Wheelsets are the main force components of rail vehicle, they keep the vehicle in place of track to ensure safe operation. Moreover, they are also of significant in transferring vehicle load, traction and braking force. Wheelset in normal condition is shown in Fig. 1.

Peeling, scrape, crack and sloughing on the surface of wheel tread are collectively called wheel flats [10, 11]. Wheel tread peeling is a phenomenon that metal on the surface exfoliates or flakily lifts, which happened more frequently than other faults on wheelset. (shown in Fig. 2).

Fig. 1. Wheel in normal state

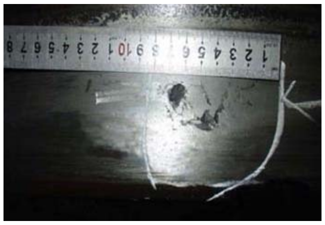
Fig. 2. Wheel tread peeling

B. Data Acquecision

We got some real-running vibration data from a bullet-train carriage of type A vehicle. The locations of vibration sensors are shown in Fig. 3. Those sensors are parts of a signal acquisition system, which was supported by the National High Technology Research and Development Program of China.

2166-5656/18/$31.00 ©2018 IEEE
DOI 10.1109/PHM-Chongqing.2018.00080

Our data was downloaded from service host on that vehicle. Details of related signal devices and experimental setup are as follows, the rail vehicles were running on the track in a metro depot at speed of 35 ± 5 km/h; The load on that wheel is about 4.75 tones; The sensor sample frequency is set to be 10k Hz; Each sample contains 32768 points.

The dataset consists of 2 types of wheel conditions, operation with no trouble (Normal), wheel out-of-roundness or flat (Fault). We got 2098 samples - 2002 samples of operation with no trouble, 96 samples of wheel out-of-roundness or flat.

There are obvious periodic feature in wheel vibration data from its time domain waveform in Fig. 4. Amplitude of vibration in abnormal conditions was significantly larger than that in normal ones (shown in red rectangle of Fig. 4.). Signals possess a clear sinusoidal property with random noise under condition of operation with no trouble, and shock characteristic in abnormal situation.

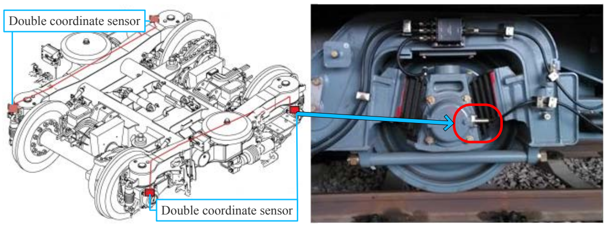

Fig. 3. Locations of vibration sensors

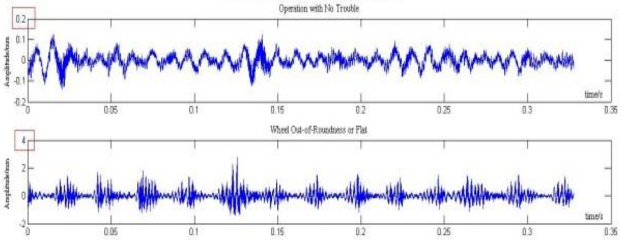

Fig. 4. Time domain waveforms of vibration data under two conditions

C. *Feature Analysis*

Principal component analysis [12] is a method to reduce the dimensionality of a set of data from a large number of interrelated variables [13], and variables are sorted by their weights[12]. Feature dimension reduction can improve the model generalization ability and avoid over-fitting. PCA [14, 15] is a commonly used feature extraction method, and we also take its advantages there.

In this paper, we calculate fifteen indexes as feature vector, which are Root Mean Square (RMS), Peak, Root of Amplitude (RA), Absolute Mean Value (AMV), Skewness, Kurtosis, Skewness factor, Kurtosis factor, Crest factor, Shape factor, Impulse factor, K factor, Energy, Energy moment, and Shannon entropy, more detail are in [16]. They show good performance in fault diagnosis individually or in combination [17-20]. After dimension reduction by PCA, Cumulative explained variance ratio of the first two components reached 1.0. Therefore, we got a relatively good indicator in two dimensions instead of in 15 dimensions. Dataset of wheel in two conditions is shown in Fig. 5.

Data in normal conditions distributes divergently in the whole space except for the two outliers in red rectangles. Density of the distribution is decreasing from left to right. While fault data gathers in a relatively small area of two subspaces. However, samples in two conditions are mixed together, and cannot be easily divided into two parts.

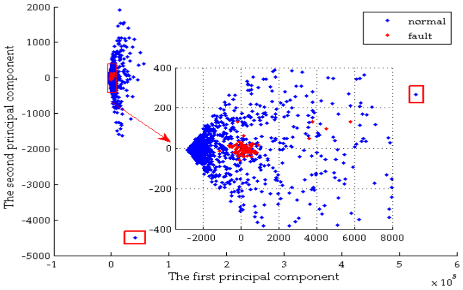

Fig. 5. Dataset of wheel in two conditions after PCA

D. *Summary*

Based on the above analysis and the real situation, difficulties of on-line wheel fault diagnosis of railway vehicle can be summarized below.

(1) The data was collected on the same type vehicle, but not the same one; Due to the route, switch and over-speed protection, vehicle velocity is hard to stay stable while testing, as well as the driver's behavior; In addition, to get different fault type dataset, we detected vehicles in a long period. The data was easily interfered by the changeable environment. The characteristic of multi-formity and divergence affects method's forecast accuracy and suitability of application.

(2) The huge data volume caused by the high sampling frequency, the non-stationary of the signal, the imbalance sample problem, together with the high accuracy and efficiency requirement make online wheel fault detection very difficult.

(3) 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.

In order to overcome these challenges, an integrating synthetic model called kNN-GBDT algorithm was proposed in next section.

III. KNN-GBDT INTEGRATED METHOD

The k-Nearest-Neighbours (kNN) is a non-parametric classification method, which is simple but effective in many cases [21]. For a data to be classified, its k nearest neighbours are retrieved, which forms a neighbourhood. Majority voting among the data records in the neighbourhood is usually used to decide the classification with or without consideration of distance-based weighting. Gradient boosting decision tree (GBDT) [22] 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 an accurate model.

The proposed synthetic model integrates kNN and GBDT to timely and precise diagnostic performance of wheel fault with respect to the imbalanced and divergent characteristics. The kNN method is used to per-classify each class in data. It divides datasets into non-overlap subclasses, and overlapped subclasses. The overlapped subclasses which are difficult-to-learn samples is gathered to form a new training dataset for advanced classification by GBDT. Procedure of kNN-GBDT based wheel fault detection method is as follows.

(1) Feature extraction. We calculate 12 time-domain parameters, root mean square (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 wheel vibration dataset.

(2) Feature dimension reduction. The PCA is used in this step. In our cases, the first two components were chosen, as the cumulative explained variance ratio of them have reached 1.0.

(3) Per-classification. Each class in data was firstly divided into subclasses through kNN cluster method. Notations are as follows: *Subclass*(i_j): the subclass after kNN; *Center*(i_j): the center of *Subclass*(i_j); R(i_j): the radius of *Subclass*(i_j), where, $i = I, II, III$ ... is the class number, j = 1, 2, 3, … is the subclass number in i_{th} class. In our case, wheelsets are only in two conditions, normal and fault, which means we have class I and class II ($i = I, II$). The sketch is shown in Fig. 6, Per-classification of Class I and Class II are conducted separately.

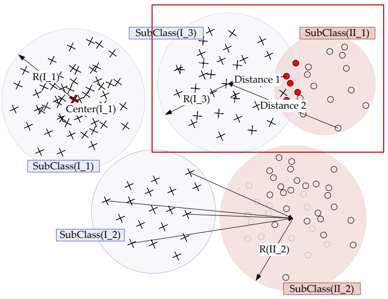

Fig. 6. Sketch of per-classification and subclasses selection

(4) To calculate the overlap between subclasses belonging to different classes, the distances of samples in one subclass were compared with another subclass's radius. If there any distance is less than the radius of the compared subclass, both two subclasses, (named overlapped subclasses; Otherwise, it's named non-overlapped subclass) are put into the further training dataset.

We compute the distances of samples in Subclass(II_1) to Center(I_3) (the center of Subclass(I_3)) (as shown in Fig. 6, Distance 1, Distance 2 …), and compare them with R(I_3) (the radius of SubClass(I_3)). We can see that Distance 1 < R(I_3), which means there are overlap between SubClass(I_3) and Subclass(II_1) (samples of Subclass(II_1) marked in red), therefore, both SubClass(I_3) and Subclass(II_1) which are in red rectangle shown in Fig. 6, are selected into new training dataset for further classification. Meanwhile, SubClass(I_1), SubClass(I_2) and Subclass(II_2) are finished in our algorithm.

(5) Further classification with the GBDT method. The new training dataset obtained from previous step is the input of GBDT model. Cross-validation is conducted here in order to limit problems like overfitting. The processed dataset is divided into two segments for training and validating the model respectively.

(6) New sample identification. The new sample is firstly judged to which subclasses it belongs, with the trained kNN method. Then checking the non-overlapped subclass and mark it. Otherwise, the new sample was sent to the trained GBDT model for further classification. The overall integrating model can be represented as shown in Fig. 7.

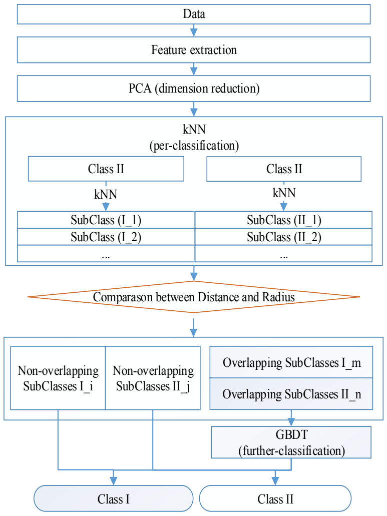
Data
Feature extraction
PCA (dimension reduction)
kNN
(per-classification)
Class II
Class II
kNN
kNN
SubClass (I_1)
SubClass (I_2)
...
SubClass (II_1)
SubClass (II_2)
...
Comparason between Distance and Radius
Non-overlapping SubClasses I_i
Non-overlapping SubClasses II_j
Overlapping SubClasses I_m
Overlapping SubClasses II_n
GBDT
(further-classification)
Class I
Class II

Fig. 7. The overall integrating kNN-GBDT algorithm

IV. EXPERIMENTS

To test the effectiveness of the proposed kNN-GBDT integrated method, this subsection comprehensively evaluates performance of the SVM and the GBDT methods by confusion matrix. Scope and backwards of the integrated method are also given out.

A. Confusion matrix

Confusion matrix is a visualization tool typically used in the field of machine learning and specifically 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 [23]. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The four outcomes can be formulated in a 2 × 2 confusion matrix, as TABLE I.

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.

TABLE I. 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)

where, *TP* + *FN* = 1, *FP* + *TN* =1 in normalized confusion matrix.

B. *Per-classification*

In the kNN procedure, the cluster number of normal data is set to be 20, because normal samples are 20 times to fault ones in size. While fault's cluster number is 2, based on its own distribution characteristic. Parameters in GBDT procedure is the same as GBDT classifier.

We magnified a portion that in red rectangle (displayed in Fig. 8). The kNN clustered data are in normal and fault condition separately. Without consideration of outliers, there is a subclass of normal data with only one sample, which is located at the very right of the whole space. Samples with the same color are of one subclass. It's very clear that, only a few subclasses of normal data are overlapped with fault ones, which will be selected for further classification.

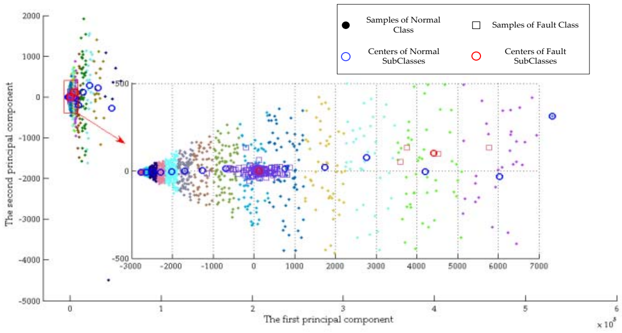

Fig. 8. Per-classification with kNN

Then procedure (4) in subsection 2.2.2 was conducted, and the result was shown in Fig. 9. The current ratio is 421 normal samples for 96 fault samples, declined from about 20:1 to 4:1.

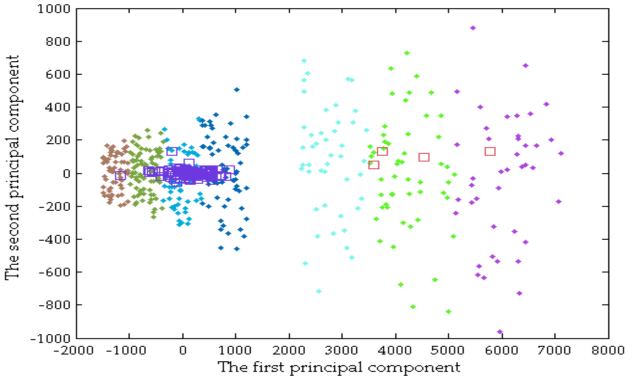

Fig. 9. Remained dataset for further classification

The Further classifier only learns information from these selected samples. Outliers make no difference in our further classification.

C. *Results and analysis*

To test the effectiveness of the proposed kNN-GBDT integrated model, this section comprehensively evaluates performance of SVM, kNN and GBDT methods through confusion matrix.

Analyses were performed with the Python system for statistical computing on a desktop computer with Intel(R) Core(TM) i3-3240 CPU @ 3.40GHz.

Support Vector Machine is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. It's widely used in railway vehicle fault diagnosis. In this paper, the 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. Classes are supposed to have weight one. About the kNN method, we set the number of neighbors used to be 20, and all points in each neighborhood are weighted equally. Other parameters are default. As to gradient boosting decision tree method, the number of boosting stages to perform, subsample values is 0.8 [24], other parameters are set default. Parameters in GBDT procedure of the integrated model are the same as compared GBDT method.

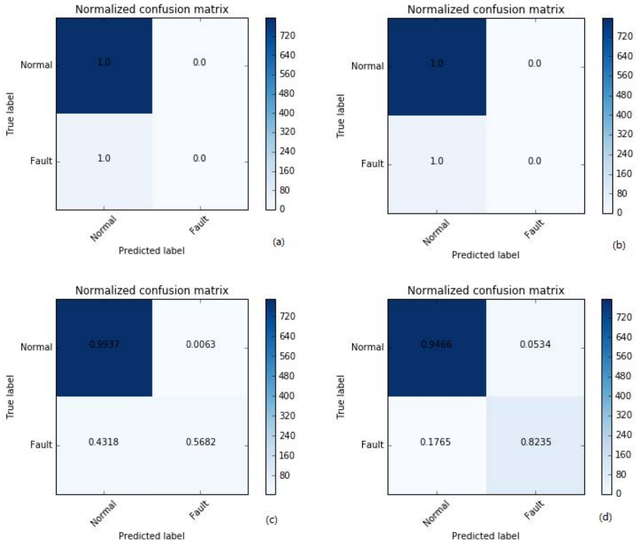

Fig. 10. Confusion matrixes of (a) SVM, (b) kNN, (c) GBDT, and (d) kNN-GBDT

As we can see from Fig. 10, the kNN and the SVM suffer a lot from over-fitting problem. All fault samples have been miss classified into normal condition. While, performance of GBDT method is a little slightly better than that of kNN and SVM. However, it's only a bit over half of the fault data correctly classified. The high imbalance ratio does have a very negative impact on classification, which will cost a lot in time and money in our rail industry and is totally inacceptable.Our proposed kNN-GBDT integrated model shows the best result among those classifiers. True positive and true negative rate have reached 0.9466 and 0.8235 respectively, which means the fault diagnosis got a great improvement on the accuracy which is 82.35%. Meanwhile, only 5.34% of normal samples is wrongly classified.

In addition, the total execute time from the very beginning feature calculation to the final fault detection is recorded to be 0.35s in average. It satisfies the requirements of real-time performance and accuracy for online fault detection.

V. Conclusion

In this study, an integrating model of the kNN and the GBDT is investigated for fault diagnosis of railway vehicle wheel. Comparisons of the SVM, kNN, and the GBDT methods indicate that kNN-GBDT model can be successfully used for real-time fault detection as the relatively high accuracy and short time consuming.

The provided novel model overcame challenges caused by imbalance dataset. Compared with models that use resampling methods, it does not create non-existent information, which may increase the overlap between classes and destroy the distribution of original dataset. In addition, it does not loss any information which is precise for wheel research in real world project.

This work was driven by the unique characteristic of data distribution. Except the imbalance property, majority data distributes divergently in the whole space, and its density is not the same in each area. While minority class gathers in a relatively small area. It makes our model performs better in that case.

Acknowledgment

This paper is supported by the National Key Research and Development Program of China (2016YFB1200401) and the Open Research Fund Program of the State Key Laboratory of Rail Traffic Control and Safety(Contract No.RCS2016K005)

References

[1] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks2008. p. 1322-8.

[2] G. V. The class imbalance problem in pattern classification and learning. Tamida2007. p. 115-6.

[3] Hanifah FS, Wijayanto H, Kurnia A. SMOTEBagging Algorithm for Imbalanced Dataset in Logistic Regression Analysis (Case: Credit of Bank X). Applied Mathematical Sciences. 2015;9:6857-65.

[4] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis. 2002;6:429-49.

[5] Ao H, Cheng JS, Li KL, Truong TK. A Roller Bearing Fault Diagnosis Method Based on LCD Energy Entropy and ACROA-SVM. Shock and Vibration. 2014.

[6] Li YJ, Zhang WH, Xiong Q, Lu TW, Mei GM. A Novel Fault Diagnosis Model for Bearing of Railway Vehicles Using Vibration Signals Based on Symmetric Alpha-Stable Distribution Feature Extraction. Shock and Vibration. 2016.

[7] Duan LX, Yao MC, Wang JJ, Bai TB, Zhang LB. Segmented infrared image analysis for rotating machinery fault diagnosis. Infrared Physics & Technology. 2016;77:267-76.

[8] Flett J, Bone GM. Fault detection and diagnosis of diesel engine valve trains. Mechanical Systems and Signal Processing. 2016;72-73:316-27.

[9] Amarnath M, Sugumaran V, Kumar H. Exploiting sound signals for fault diagnosis of bearings using decision tree. Measurement. 2013;46:1250-6.

[10] Wang Y. Review of dynamic detecting methods for railway wheel flat. Rolling Stock. 2002;40.

[11] Qin N, Jin W-D, Huang J, Li Z-M. Ensemble empirical mode decomposition and fuzzy entropy in fault feature analysis for high-speed train bogie. Kongzhi Lilun Yu Yingyong/Control Theory and Applications. 2014;31:1245-51.

[12] Brereton RG. Chemometrics Data Analysis for the Laboratory and Chemical Plant. 2003. England: John Wiley & Sons CrossRef Google Scholar.

[13] Kumar N, Bansal A, Sarma G, Rawal RK. Chemometrics tools used in analytical chemistry: An overview. Talanta. 2014;123:186-99.

[14] Liu HM, Zhang JC, Cheng YJ, Lu C. Fault diagnosis of gearbox using empirical mode decomposition and multi-fractal detrended cross-correlation analysis. Journal of Sound and Vibration. 2016;385:350-71.

[15] Tian J, Morillo C, Azarian MH, Pecht M. Motor Bearing Fault Detection Using Spectral Kurtosis-Based Feature Extraction Coupled With K-Nearest Neighbor Distance Analysis. Ieee Transactions on Industrial Electronics. 2016;63:1793-803.

[16] yuan Z. Service status identification and method prediction research based on safety region estimation for key experiments in rail vehicles: Beijing Jiaotong University; 2014. (in Chinese)

[17] Borghesani P, Pennacchi P, Chatterton S. The relationship between kurtosis- and envelope-based indexes for the diagnostic of rolling element bearings. Mechanical Systems and Signal Processing. 2014;43:25-43.

[18] Su L, Shi TL, Liu ZP, Zhou HD, Du L, Liao GL. Nondestructive diagnosis of flip chips based on vibration analysis using PCA-RBF. Mechanical Systems and Signal Processing. 2017;85:849-56.

[19] Ai YT, Guan JY, Fei CW, Tian J, Zhang FL. Fusion information entropy method of rolling bearing fault diagnosis based on n-dimensional characteristic parameter distance. Mechanical Systems and Signal Processing. 2017;88:123-36.

[20] Asr MY, Ettefagh MM, Hassannejad R, Razavi SN. Diagnosis of combined faults in Rotary Machinery by Non-Naive Bayesian approach. Mechanical Systems and Signal Processing. 2017;85:56-70.

[21] Hand DJ, Mannila H, Smyth P. Principles of data mining: MIT press; 2001.

[22] Chirici G, Scotti R, Montaghi A, Barbati A, Cartisano R, Lopez G, et al. Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery. International Journal of Applied Earth Observation & Geoinformation. 2013;25:87-97.

[23] Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment. 1997;62:77-89.

[24] Jain A. Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python. FEBRUARY 21, 2016.