IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier 10.1109/ACCESS.2020.Doi Number

Traffic incident detection method based on factor analysis and weighted random forest

Hui Jiang[1, 2] and Hongxing Deng[1,*]

[1] School of Traffic and Transportation, Northeast Forestry University, Harbin 150040, China
[2] School of Civil engineering and Transportation, Beihua University, Jilin 132013, China

Corresponding author: Hongxing Deng (e-mail:denghongxing1969@126.com).

This work was supported in part by the National Natural Science Foundation of China under Grant 81638004 and 71801149.

ABSTRACT Timely and accurate detection of traffic incidents can effectively reduce personal casualties and property losses, and improve the ability of macro-control and scientific decision-making of traffic. The unbalance of traffic incident data has a great influence on the detection effect. Therefore, a traffic incident detection method based on factor analysis and weighted random forest (FA-WRF) is designed. Through the analysis of the change rule of traffic flow parameters to build the initial incident variable. The factor analysis (FA) method is used to reduce the dimension of the initial incident variables. Using Bootstrap improved algorithm to predetermine the data extraction standard of the training set. The MCC coefficient value is calculated for the classification effect of the decision tree after training, and is assigned to each tree as a weight value, so as to ensure that the trees with better classification ability have more voting power in the voting process, thus improve the overall classification performance of the random forest （RF） algorithm for unbalanced data. The detection performance is evaluated by the common criteria including the detection rate, the false alarm rate, the classification rate and the area under the curve of the receiver operating characteristic (AUC). Based on the location detector data from expressway, the incident data in which accounts for 6.5%, showing a typical unbalance. The experimental results indicate that the model based on FA-WRF has the better classification effect. Meanwhile it is competitive in processing unbalanced data classification compared with Support Vector Machine.

INDEX TERMS Traffic incident detection; weighted random forest; factor analysis; expressway; unbalanced data.

I. INTRODUCTION

As the aorta of urban traffic, urban expressway has the function of urban road and the characteristics of freeway fast passage, which can meet the requirements of high speed and long-time continuous driving. In recent years, due to the sharp increase in the number of vehicles, these advantages of expressway are gradually being lost. Traffic jams and even traffic incident occur frequently, especially during rush hours, once the traffic incident is not dealt with in time, it is very easy to cause traffic congestion and even lead to secondary traffic incident. The traffic situation will deteriorate rapidly in a short time, which will aggravate the difficulty of traffic incident management and bring inconvenience to residents' travel. Therefore, it is very necessary to timely and accurately detect traffic incident, reduce the impact of emergencies on roads and life and property, enhance the macro-control ability of traffic, and improve the scientific decision-making ability of traffic.

The earliest Automatic Incident Detection（AID） is the California algorithm （CA） proposed by Payne et al. The algorithm determines the possibility of traffic incident by comparing occupancy data of adjacent detectors. Payne [1] proposed 10 improved AID algorithms based on CA, among which California # 7 and # 8 algorithm had better performance. The Standard Normal Deviate (SND) was developed by the Texas Transportation Association [2] to realize the identification of sudden traffic incident by judging whether the rate of change of traffic flow parameters is greater than the specified threshold. Cook *et al.* [3] developed the Double Exponential Smoothing (DES) algorithm. The algorithm takes the double exponential smoothing value of traffic flow parameter data as the predicted value, and constructs a tracking signal by comparing the predicted value with the measured value. When the tracking signal exceeds the predetermined threshold value, the sudden traffic incident alarm can be triggered. By constructing 13 different traffic flow parameter variables to test, it is found that the detection

results of traffic flow and occupancy rate are better. Levin *et al.* [4] developed a Bayesian algorithm. The algorithm realizes the discrimination of traffic congestion by calculating the conditional probability of occupancy change. The results show that the algorithm needs longer mean time to detection while achieving higher detection rate and lower false detection rate. Ahmed *et al.* [5] developed an Auto Regressive Integrated Moving Average (ARIMA) algorithm. Using occupancy as input data, the algorithm established a third-order ARIMA (0, 1, 3) model for short-term prediction of occupancy and confidence level. By judging the deviation between the predicted data and the observed data, it is determined whether to start the emergency traffic alarm system. Persaud *et al.* [6] developed McMaster algorithm based on mutation principle. This algorithm uses a large number of historical data of traffic flow parameters to construct the "flow-occupancy" distribution relation curve, and then determines the occurrence of traffic congestion by comparing the relationship between the observed data and the curve. The results show that the detection effect of McMaster algorithm is not affected by the fault of downstream detector.

Artificial Neural Network (ANN) was applied in the study of AID algorithm by Ritchie *et al.* [7]. This algorithm does not need to adjust the original model, but can describe the change rules of traffic flow only through the training of real traffic flow data. Since then, ANN-based AID algorithm research has become a hot spot. On the basis of Ritchie's research [8], Chew et al. proposed an AID algorithm based on Multilayer Perceptrons (MLP) forward feedback neural network. The algorithm based on freeway traffic flow data at three different locations. The results showed that the algorithm could detect most traffic incident and obtain lower false alarm rate compared with the classical AID algorithm. Ritchie *et al.* [9] developed an AID algorithm based on Basic Probabilistic Neural Network (BPNN) by introducing parameters such as the prior probability of traffic incidents, road conditions and misjudgment losses into the AID algorithm. Abdulhai *et al.* [10] proposed an AID algorithm based on Bayesian probabilistic neural network, which realized the discrimination of the probability of traffic incident through statistical distance. Combined with the measured data of SR91 road, the verification showed that the detection rate could be close to 90% in the case of low false alarm rate.

Srinivasan *et al.* [11] proposed an AID algorithm based on hybrid fuzzy logic genetic algorithm. Based on expressway data, the algorithm constructs 11 fuzzy operators to represent the traffic index. The results show that the flexibility and robustness of the fuzzy controller are highly fault tolerant to inaccurate data, and the false alarm rate is only 0.8% when the detection rate is 70%. Teng *et al.* [12] proposed an AID algorithm based on wavelet technology. It uses wavelet technology to extract features of variables constructed from the occupancy data of upstream and downstream and detect traffic incident. The results show that the characteristic variables extracted by the wavelet technique can better present the changing characteristics of traffic flow under the incident condition. Yuan *et al.* [13] applied the Support Vector Machine (SVM) technology to the AID algorithm and tested it on simulated and measured data. The results showed that compared with MLF algorithm and CA algorithm, SVM had obvious advantages in detection effect. Chen *et al.* [14] proposed an AID algorithm based on SVM, verified the effectiveness of the algorithm by using the measured data of urban expressway in I-880 database. Wang *et al.* [15] proposed an AID algorithm based on Feature-weighted SVM. In order to reduce the information redundancy of input variables, using of classification interval method to determine the weight of input variables. The actual test results of I-880 data show that the algorithm get further improved after characteristic weighting of input variables. Liu *et al.* [16] proposed an AID algorithm based on Random Forest (RF), which applied the random forest method in AID algorithm for the first time. Verification of the measured data in PeMS database shows that when the incident detection rate is satisfactory, the acceptable false alarm rate and detection time are obtained. Jiang *et al.* [17] proposed an AID algorithm of freeway traffic incident based on FA-SVM. By designing multiple initial variables with or without traffic incident, FA was introduced into the screening of SVM variables. Simulation data test results show that the detection performance of SVM algorithm is further improved after FA is introduced.

To summarize, the principle of the early AID methods such as CA, SND, DES, Bayesian, ARIMA and McMaster is relatively simple, and which is designed by deducing, assuming and simplifying the changing rules of traffic flow parameters. Although the practicability is good, but the overall detection effect is general. With the deepening of the research, more and more experts and scholars is spending a lot of time and effort to develop new algorithms or improve existing models, such as ANN, MLP, BPNN, fuzzy theory, wavelet technique, SVM, RF and all sorts of fusion algorithm based on the above algorithm, *etc.*, which makes the AID method become increasingly complex.

Among the above algorithms, RF, as an integrated algorithm with excellent classification ability, has been widely used in the classification and identification fields. It has also begun to attract the attention of scholars in the study of traffic incident. But the noise /redundancy of input incident variables or the unbalance of traffic incident data will affect the classification effect of RF algorithm.

In summary, It is necessary to propose a method that can not only extract the effective information of incident variables, but also reduce the impact of unbalanced data on the classification effect. To do so, this work proposes AID method based on factor analysis and weighted random forest. Compared with the existing work, we make two contributions: 1) In order to better extract the effective information of incident variables, FA method is proposed to reduce the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3023961, IEEE Access

dimension of the initial incident variables. 2) In order to reduce the impact of unbalanced data for RF algorithm, we used the improved bootstrap algorithm to predetermine the data extraction standard of the training set. The MCC coefficient value is calculated for the classification effect of the decision tree, and is assigned to each tree as a weight value, so as to ensure that the trees with better classification ability have more voting power in the voting process. By comparing with the prior method, we validate the effectiveness of the proposed methodology.

The rest of this is paper is organized as follows. Firstly, the traffic flow parameter variation characteristics and construction of initial incident variables is explained in Section II. Then, the FA-WRF algorithm is proposed in Section III. Where the results and discussion on the findings are elaborated in Section IV. Finally, Section V concludes the paper.

II. Analysis of traffic flow parameter variation characteristics and construction of initial incident variables

In-depth and effective analysis of expressway traffic flow parameters, mining the characteristics of traffic flow parameters before and after the occurrence of traffic incidents from multiple angles, and designing the input variables of the AID algorithm are the foundation and key to the research of expressway AID algorithm[28,29]. Based on the data of traffic flow, speed and occupancy collected by the expressway location detector, this section conducts an in-depth analysis of its variation characteristics and designs incident variables that can highlight the variation characteristics of traffic flow parameters before and after a traffic incident, and constitutes the initial input incident variable set of expressway traffic incident detection.

A. Analysis of traffic flow parameter variation characteristics

The traffic state of expressway section can be described directly by the change of traffic flow parameters.

As shown in Fig. 1 and Fig. 2, during the normal (non-traffic incident) period, the change process of traffic flow parameters (flow, speed and occupancy rate) is relatively gentle. During the period of traffic incident, the traffic flow parameters will fluctuate violently in a short time under the influence of traffic incident.

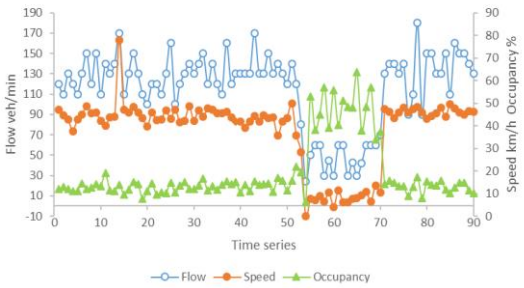

FIGURE 1. The trend diagram of traffic flow parameters upstream of the location of traffic incidents

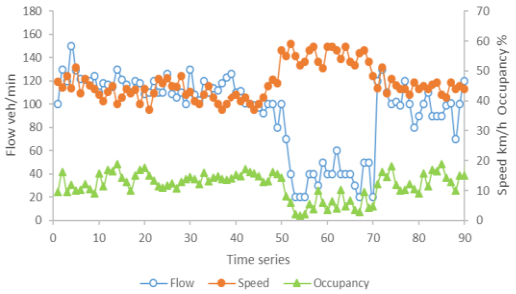

FIGURE 2. The trend diagram of traffic flow parameters downstream of the location of traffic incidents

In the actual road environment, the upstream and downstream of traffic incident are affected by compression wave and expansion wave respectively, and the changes of traffic flow parameters detected are as follows: the upstream occupancy increases rapidly, while the flow and speed decrease rapidly; the downstream occupancy and flow rate decrease rapidly, and the speed may increase or decrease [18]. Therefore, according to the changes of traffic flow parameters detected by the upstream and downstream detectors, traffic incidents can be judged.

B. Construction of initial input incident variables

In this paper, the variables that can represent the traffic flow variation characteristics affected by traffic incidents are named traffic incident characteristic variables, or incident variables for short. When there is a traffic incident, the incident variable presents a different range of variation. Compared with the normal state, the larger the range of variation, the more sensitive it is to the occurrence of a traffic incident, and the better the detection effect of the expressway AID algorithm based on the highly sensitive incident variable will be. Therefore, in the design process of AID algorithm, the effective construction of incident variables is the foundation and a key step.

On the basis of fully mining the law of change of traffic flow parameters of expressway before and after traffic incidents, the set of incident variables is constructed in the way of basic traffic flow parameters and their multi-angle combination, and is used as the initial input incident variable of AID algorithm. The specific construction will be described from the following three perspectives:

1) The combination of predicted and measured traffic flow parameters

The measured value of traffic flow parameters refers to the actual detected value of traffic flow parameters, and the predicted value of traffic flow parameters is the moving average of the first a moments of the current moment. As shown in Fig. 3, take flow as an example. During the normal period, as the predicted value of the flow is close to the measured value, the ratio of the two should oscillate slightly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3023961, IEEE Access

around the value of "1". During the period of occurrence of traffic incidents, the measured value of traffic will change greatly (become smaller) than that of the normal period under the influence of incident, while the predicted value of traffic will still contain the data information under the normal state. In this case, the ratio of the two has a trend of significant deviation from the value of "1". Therefore, the ratio of traffic flow parameter predicted value to measured value is highly sensitive to the occurrence of traffic incidents.

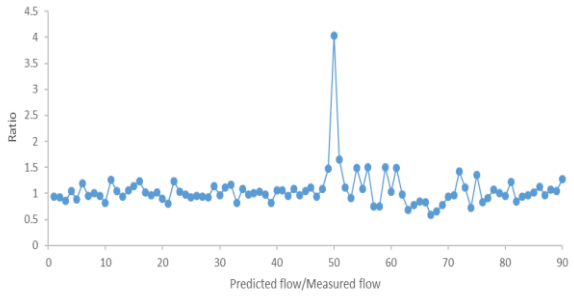

FIGURE 3. The changes of Predicted flow /measured flow under the influence of the traffic incident

2) The combination of different traffic flow parameters collected by the same detector

As can be seen from the previous analysis, when traffic incidents occur, traffic flow parameters themselves will change to a certain extent. In order to further amplify this change amplitude and increase the sensitivity to the incident state, the form of ratio between different traffic flow parameters collected by the same detector is intended to be constructed. As shown in Fig. 4, take "occupancy/speed" as an example.

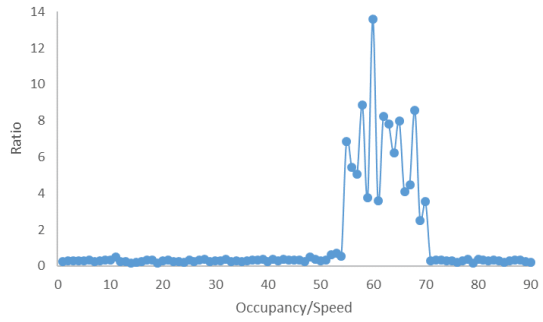

FIGURE 4. The changes of Occupancy/Speed under the influence of the traffic incident

In normal period, the change of occupancy and speed is relatively stable, and the ratio of them remains at a level with a small amplitude of shock. Affected by the occurrence of traffic incidents, the occupancy value collected by the upstream detector gradually increases, and the velocity value gradually decreases. The ratio of the two values will change greatly compared with the normal period, and this change is highly sensitive to the occurrence of traffic incidents.

3) The combination of same traffic flow parameters collected by the adjacent detector.

Initial incident variables are constructed by collecting the combination of the same traffic flow parameters from the upstream and downstream adjacent detectors. As shown in Fig. 5, taking occupancy as an example, in the normal period, the change of occupancy rate collected by the upstream and downstream adjacent detectors is relatively stable, which can maintain small amplitude of oscillation at a level. Under the influence of the occurrence of traffic incidents, the ratio will have a large change compared with the normal period, and this change is highly sensitive to the occurrence of traffic incidents.

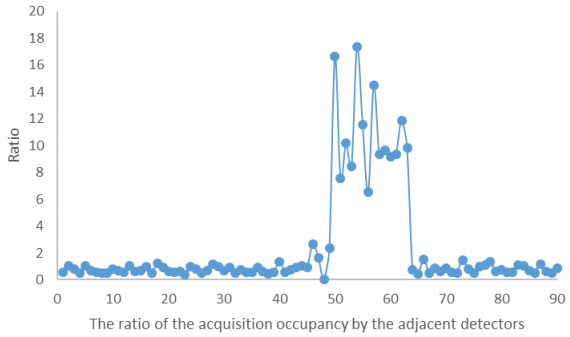

FIGURE 5. The changes of Occupancy ratio of adjacent detectors under the influence of the traffic incident

In conclusion, the final construction contains 21 sets of initial input variables for traffic incidents, as shown in Table I.

TABLE I
SET OF INITIAL INPUT VAIRABLES

	Meaning		Meaning
x_1	upstream flow	x_{12}	downstream occupancy
x_2	upstream speed	x_{13}	downstream “occupancy/flow”
x_3	upstream occupancy	x_{14}	downstream “occupancy/speed”
x_4	upstream “occupancy/flow”	x_{15}	downstream “flow/speed”
x_5	upstream “occupancy/speed”	x_{16}	downstream ”predicted flow/measured flow”
x_6	upstream “flow/speed”	x_{17}	downstream ”predicted occupancy /measured occupancy”
x_7	upstream ”predicted flow/measured flow”	x_{18}	downstream ”predicted speed /measured speed”
x_8	upstream ”predicted occupancy /measured occupancy”	x_{19}	the ratio of the acquisition flow by the adjacent detectors
x_9	upstream ”predicted speed /measured speed”	x_{20}	the ratio of the acquisition speed by the adjacent detectors
x_{10}	downstream flow	x_{21}	the ratio of the acquisition occupancy by the adjacent detectors
x_{11}	downstream speed		

III. Expressway traffic incident detection based on FA-WRF

A. An input variable extraction method for expressway traffic incident detection based on FA

Theoretically, the more initial incident variables in Table 1, the more sensitive it is to the occurrence of traffic incidents, and the more favourable it is to the later detection of traffic incidents.

From the perspective of application, too many incident variables will affect the running speed of AID algorithm, which will lead to the decrease of the detection efficiency of traffic incidents.

In addition, the information contained in too many incident variables will be redundant and repetitive, which will lead to the decrease of the detection effect of traffic incidents. Therefore, in order to improve the detection efficiency and effect of subsequent traffic incidents, factor analysis method is proposed to effectively extract the initial incident variables.

1) Principle of FA

In 1904, Charles Spearman et al. proposed factor analysis method (FA), which simplified the complexity of analysis and research of multivariable problems with the idea of dimensionality reduction. At present, it has been widely used in many fields such as medicine, economy and industry [19].

The factor analysis model is:

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + ... + a_{1n}F_n + \varepsilon_1 \\ x_2 = a_{21}F_1 + a_{22}F_2 + ... + a_{2n}F_n + \varepsilon_2 \\ ..... \\ x_m = a_{m1}F_1 + a_{m2}F_2 + ... + a_{mn}F_n + \varepsilon_m \end{cases} \quad (1)$$

where $x_1, x_2, x_3, ..., x_m$ is the actual variable; $F_i (i = 1,2,...,m)$ is the dominant factor; The loading coefficient of $a_{ij} (i = 1,2,...m; j = 1,2,...,n)$ principal factor represents the correlation coefficient between the i variable and the j factor. The larger a_{ij} is, the higher the correlation between them will be. $\varepsilon_i (i = 1,2,...,m)$ is the error or special factor.

In factor analysis, each factor can be expressed as a linear combination of variables, and then the actual value of variables can be used to estimate the corresponding principal factor value (factor score). Its mathematical model is as follows:

$$Fscore_i = b_{i1}x_1 + b_{i2}x_2 + ... + b_{in}x_n \quad (2)$$

where $Fscore_i$ is the score value corresponding to the i factor, and $b_{i1} \sim b_{in}$ represents the variance contribution rate corresponding to i factor on the variables from the first to the n. The mathematical model of the score of the comprehensive factor is as follows:

$$C - Fscore = (\lambda_1 Fscore_1 + \lambda_2 Fscore_2 + ... + \lambda_i Fscore_i)/(\lambda_1 + \lambda_2 + ... + \lambda_i) \quad (3)$$

where $C - Fscore$ is the score value of comprehensive factor; λ_i is the eigenvalue corresponding to the i factor.

Although FA is simple and easy to operate. But, incident variables processed by FA need to be carried out in the same dimension, and FA requires high correlation between data.

2) Step of extracting the initial input incident variable of expressway based on FA.

The steps are as follows：

(1) On the premise that the initial incident variable has a strong correlation, principal component analysis method is used to extract the initial incident variable factor, calculate the variance contribution rate of the incident variable factor, and draw a gravel diagram to determine the extraction number of the initial incident variable main factor.

(2) In order to define the specific meaning of the main factor of incident variable, it is necessary to determine whether the load matrix needs to be rotated;

(3) Calculate the score of main factor of incident variable and output the result.

Draw the work flow chart, as shown in Fig. 6.

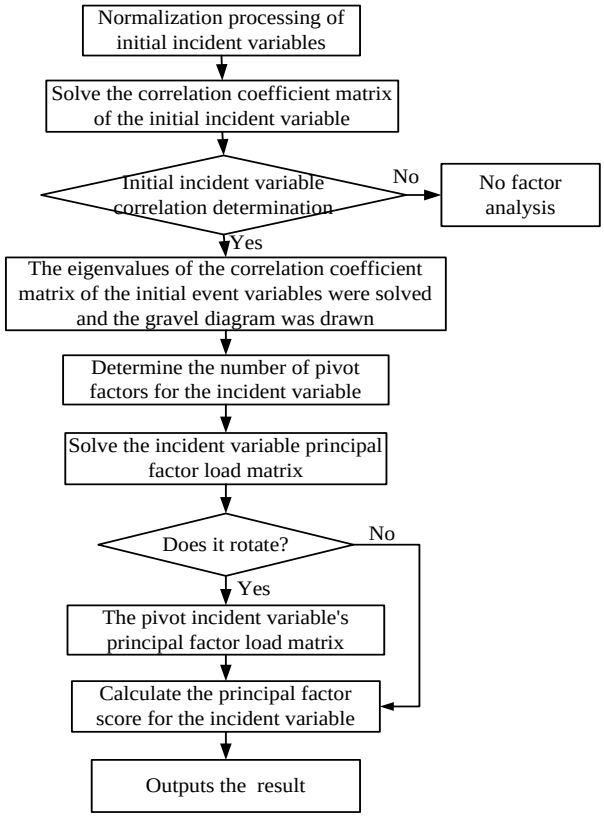

FIGURE 6. The workflow of extraction of input variables based on FA

B. Expressway traffic incident detection based on WRF

In the process of traffic incident detection, each decision tree in the random forest algorithm has equal voting power. However, because each tree has different training sets and completes the training independently, the classification accuracy of decision trees will be different, the imbalance of traffic accident data will further increase the difference of classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3023961, IEEE Access

accuracy among decision trees, and finally, the classification effect of random forest will be affected by the decision trees with poor classification effect.

Therefore, in order to improve the effect of classification of traffic incident data, we proposed the AID based on WRF: First of all, using the improved bootstrap algorithm to determine the training set to extract data standards. In the stage of decision tree training, MCC coefficient was used to evaluate the classification effect of the decision tree, which was assigned to each tree as a weight value to participate in the final voting, so as to avoid the impact of low classification effect on the voting results of the random forest model. Ultimately achieve the goal of improve the effect of detection.

1) Principle of RF

In 2001, Leo Breiman [20] first proposed Radom Forests (RF) algorithms. Compared with other machine learning algorithms, RF algorithms take multiple training sample subsets to enhance the difference between classification models, thus improving the generalization ability and prediction ability of the algorithms. It is characterized by high classification accuracy, few optimization parameters and stable operation performance [21].

Assuming that the original data set D contains n data bars and the sample feature dimension is M, p decision trees are established. The workflow to realize classification based on RF method mainly includes two parts: training process and decision classification process. As shown in Fig. 7, the specific process is described as follows:

Step 1: Training process

The Bootstrap method was used to randomly select D data samples from the original data set n and repeat p times to generate p training sets. Data that is not selected in each extraction process constitutes out-of-bag (OOB) data and forms a test set;

Select M features as candidate features randomly from m feature dimensions ($m < M$), maximize the growth of each decision tree according to CART algorithm[23], repeat the above operation for p training sets respectively, and finally obtain p decision trees and constitute random forest.

Step 2：Decision classification Process

Use p decision trees in RF to make decisions on test set (OOB) and obtain classification results; Use the voting method to summarize and output the classification results.

In the above process, the OOB test set made p decisions respectively to obtain a classification result sequence: $\{h_1(x), h_2(x), h_3(x), ..., h_p(x)\}$, and then made a voting decision to obtain the final classification result, as shown in formula (4).

$$\mathrm{H}(x) = \arg\max_{\mathrm{Y}} \left\{ \sum_{i=1}^{p} I\left(h_i(x) = Y\right) \right\} \tag{4}$$

where $H(x)$ represents the random forest algorithm model, x represents the test set, $h_i(x)$ represents the classification result of the i decision tree, Y represents the output target variable, $I(\bullet)$represents the indicative function. When the parameter in the function is true, the output of the function value is 1, otherwise it is 0.

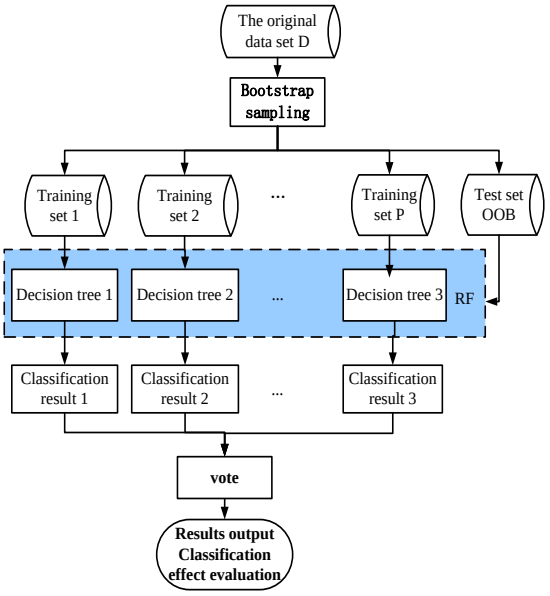

FIGURE 7. The workflow of RF algorithms

2) Design of AID algorithm based on WRF

In this section, on the premise of completing the preliminary design of incident variables and effective extraction, aiming at the typical unbalance of traffic incident data, the AID algorithm based on WRF is designed.

(1) Improved bootstrap algorithm

Bootstrap resampling algorithm is an important guarantee for RF algorithm. By repeatedly extracting 2/3 of the sample size from the original data set with put back, a plurality of training sets are formed. Theoretically, when the sampling frequency reaches a certain degree, the obtained training sets can approach to the original data set samples indefinitely. However, it has strong randomness, especially for unbalanced data, which will aggravate the imbalance of sampling data.

Take the unbalance coefficient $U = \frac{S_{\max}}{S_{\min}}$, where $S_{\max}$ and S_{min} respectively represent the majority and minority sample data in the original data set D. The sampling results of unbalanced data sets with Bootstrap resampling algorithm can be summarized as follows:

First, there is no minority sample data in the training set, and the unbalance coefficient U' in the training set cannot be calculated;

Secondly, there are a small number of minority sample data in the training set, and $U' > U$;

Third, there are a large number of minority sample data in the training set, and $U' \le U$.

Due to the randomness of data extracted by Bootstrap algorithm, the probability of occurrence of the above three cases is phase. However, for the non-balanced data processing, the existence of the first two conditions will aggravate the imbalance of the training set, so it can be regarded as invalid extraction. The existence of the third situation will help the follow-up work.

Based on this, we proposes a Bootstrap resampling algorithm with additional constraints. By judging the sampling results, invalid extraction is eliminated to ensure $U' \le U$.

Assuming that the amount of sample data contained in the original data set D is S, $S'_{\max}$ and S'_{min} respectively represent the majority and minority sample data extracted from the training set, then:

$$S_{\max} + S_{\min} = S \Rightarrow S_{\max} = S - S_{\min} \tag{5}$$

$$S'_{\max} + S'_{\min} = \frac{2}{3}S \Rightarrow S'_{\max} = \frac{2}{3}S - S'_{\min} \tag{6}$$

$$U' = \frac{S'_{\max}}{S'_{\min}} \tag{7}$$

$$U \ge U' \Rightarrow \frac{S_{\max}}{S_{\min}} \ge \frac{S'_{\max}}{S'_{\min}} \tag{8}$$

Substitute (5) and (6) into (8) to further derive:

$$\frac{S - S_{\min}}{S_{\min}} \ge \frac{\frac{2}{3}S - S'_{\min}}{S'_{\min}} \Rightarrow S'_{\min} \ge \frac{2}{3}S_{\min} \tag{9}$$

Therefore, the additional constraints of the Bootstrap resampling algorithm can be described as follows: the number of minority samples extracted each time should account for at least 2/3 of the number of minority samples in the original data set.

(2) Weighted random forest model

According to the previous analysis, in the voting stage, the traditional random forest model assigns the same weight to the output results of each decision tree. Aiming at unbalanced data, in order to improve the role of decision trees with strong classification ability in the final voting link of random forest, at the same time, reduce the role of decision trees with weak classification ability in the final voting link of random forest. This paper proposes the Weighted Random Forest (WRF) model.

Firstly, the bootstrap resampling algorithm with additional constraints was used to extract multiple training sets to form the random forest and complete the training;

Secondly, OOB data is used to evaluate the classification effect of each decision tree, and a certain weight is assigned to each tree based on the classification effect;

Finally, in the voting stage, the weighted decision tree is used to complete the classification voting on the final test set;

Therefore, the WRF model can be modified as:

$$\mathrm{H}(x) = \arg\max_{Y}\left\{\sum_{i=1}^{p} I\left(h_i(x) = Y\right) * w_i\right\} \tag{10}$$

where w_i is the weight value corresponding to the decision tree i;

Research in literature [23] has proved that: if there is a group of independent classifiers, and the classification accuracy of each classifier is $p_1, p_2, ..., p_n$, the relationship between the weight w_i of the i classifier and the classification accuracy p_i of the classifier is as follows:

$$w_i = \lg\frac{p_i}{1 - p_i} \qquad 0 < p_i < 1 \tag{11}$$

Formula (10) can be further revised as follows:

$$\mathrm{H}(x) = \arg\max_{Y}\left\{\sum_{i=1}^{p} I\left(h_i(x) = Y\right) * \left(\lg\frac{p_i}{1 - p_i}\right)\right\} \tag{12}$$

The selection criteria of classification accuracy p_i for unbalanced data will be explained later.

(3) Working steps of WRF-based AID for expressway traffic incident

On the premise of completing the construction of incident variables and extracting them effectively, a WRF-based AID for expressway traffic incidents is designed. The implementation process of this model includes the following steps, and draw the workflow as shown in Fig. 8.

First, prepare data.

Based on the data collected by the expressway site detector and combined with Table 1 and the FA method, the score value of the main factor of the incident variable was extracted and calculated, and the data set of the input variable for the AID was formed. The data set was divided into the initial training data set and the final test data set.

The main factor of the incident variable with the occurrence of traffic incidents in the data set is identified as +1, and the others are identified as -1.

Second, extract the training sample set and construct the decision tree.

The additional constraints of the Bootstrap resampling algorithm is used to put back and extract the initial training data set, and the OOB data outside the bag formed the test set. Determine the value of m and p.

Third, determine the classification accuracy and calculate the weight value.

Each decision tree is used to classify OOB data and calculate classification accuracy. The corresponding weight value of each decision tree is further calculated.

Fourth, output of result and evaluation of classification effect.

The WRF was used to classify the final test data set, and the classification effect was evaluated.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3023961, IEEE Access

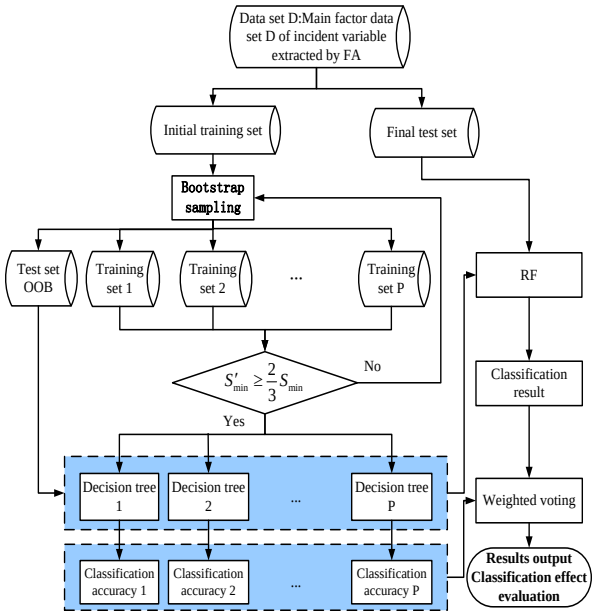

FIGURE 8. The workflow of WRF-based AID

3) AID algorithm evaluation index

In addition to Detection Rate (DR), False Alarm Rate (FAR) and Mean Time to Detection (MTTD), the evaluation index of AID algorithm also includes indexes for evaluating the classification performance of unbalanced data sets: accuracy rate, F-Measure, MCC coefficient, ROC curve and AUC value etc.

As shown in Table II, the original traffic incident data set contains incident (Positive) data and non-incident (Positive) data. The corresponding confusion matrix as shown in table II, in which the TP (True), TN (Negative) and FP (False Positive) and FN (False Negative) represent the number of rightly classified incidents, the number of rightly classified non-incidents, the number of wrongly classified non-incidents, the number of wrongly classified incidents.

TABLE II
TRAFFIC INCIDENT DETECTION DICHOTOMY ACCURACY CONFUSION MATRIX

	Classified Positive	Classified Negative
Positive	TP	FN
Negative	FP	TN

(1)Accuracy rate

Accuracy rate (AR) is an important index to measure the performance of data set classification. It comprehensively considers the common classification effect of majority and minority samples.

$$AR = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

(2)F-Measure

F-Measure(F) is a comprehensive performance evaluation index, its calculation is shown in Formula (14).

$$\mathrm{F} = \frac{\left(1+\alpha^2\right)\times TP^2}{\alpha^2 \times (2TP^2 + TP*FP + TP*FN)} \tag{14}$$

where α usually takes the value of 1. For the classification of unbalanced data, F is an effective evaluation index. $F \in [0,1]$, the greater the value of F, the better the classification effect of the model.

(3)Matthews Correlation Coefficient

Matthews Correlation Coefficient (MCC) is an important indicator to evaluate the performance of dichotomies. It is an index to describe the correlation between actual classification and predicted classification, and its calculation is shown in Formula (15).

$$M_i = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{15}$$

M_i represents the MCC to the Decision tree i, whose value interval is (-1,1). Relevant studies show that [24], the M_i is a relatively balanced index, which is stable even when there is a large difference in the number of two types of samples in the unbalanced data set.

Therefore, we chooses M_i as the measurement index of algorithm classification accuracy, and uses M_i substitution p_i in formula (11) to calculate the weight value. Since the value interval of p_i is (0,1) and the value interval of M_i is (-1,1), formula (11) can be further changed to:

$$w_i = \lg\frac{1+M_i}{1-M_i} \qquad -1 < M_i < 1 \tag{16}$$

The higher the Value of M_i corresponding to the i decision tree, the greater the weight w_i assigned to the voting link in the RF algorithm, and the greater the impact on the final voting results.

Thus, the WRF model can be further modified as:

$$\mathrm{H}(x) = \arg\max_{Y}\left\{\sum_{i=1}^{p} I\left(h_i(x) = Y\right) * \left(\lg\frac{1+M_i}{1-M_i}\right)\right\} \qquad -1 < M_i < 1 \tag{17}$$

(4) ROC curve and AUC value

The ROC curve takes FAR as the horizontal axis and DR as the vertical axis to form a curve in the two-dimensional coordinate plane. And the length of the horizontal axis and the vertical axis is unit 1. The area bounded by the ROC curve and the two axes is the AUC value. The higher the AUC value is, the better the classification performance will be. For the classification of unbalanced data, the AUC value is still a good metric [25, 30, 34, 35].

IV. Results and Discussion

A. Data sources

The main section of an urban expressway is selected as the area to be analyzed. The section includes ten detection sections with four lanes and a total of 40 coil detectors. The data that will be collected by the detectors include detector number, attribute of detection data, detection time, flow, speed and

occupancy. The data was sampled from August 26, 2018, to August 30, 2018, with a sampling interval of 5min. The original data set contains 28,800 groups of data, and 63 traffic incidents were obtained through manual screening [31], corresponding to 1899 incident data in total. Therefore, the incident data in the original data set accounts for 6.5%, and the incident data is significantly less than the normal data, showing a typical unbalance.

The first 3/5 data of the original data set was taken as the initial training set (including 1324 incident data for 40 traffic incidents) and the other 2/5 data (575 incident data for 23 traffic incidents) as the final test set.

B. Analysis of the Experimental Results

Combined with SPSS22.0 software, for the 21 initial incident variables designed in Table III, FA is used to extract the main factors of incident variables:

The correlation of the initial incident variables is determined by KMO test and Bartlett Sphericity test[19], as shown in table 3: KMO value is 0.707 (>0.5), BTS value is 64080.956, and the significance probability value of the statistical value is 0.000 (<0.01). The relevant index values all indicate that the initial incident variables are strongly correlated and suitable for FA[19].

TABLE III
CORRELATION TEST RESULTS FOR INITIAL INCIDENT WARIABLES

Inspection method	Relative index values	
KMO test	KMO value	0.707
Bartlett Sphericity test	BTS value	64080.956
	Degree of freedom	210
	Sig.	.000

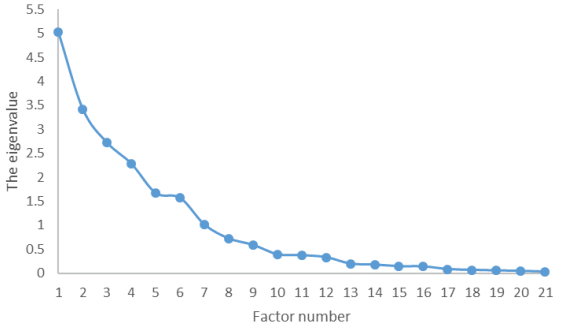

FIGURE 9. The main factor extraction gravel figure

Fig. 9 is the extraction of gravel diagram of the main factor of the initial incident variable. It can be seen that the first 7 eigenvalues have a significant description of the initial incident variable and can be used as the main factor of the incident variable to participate in the follow-up AID algorithm research.

In order to clarify the meaning of the seven incident variables, the load matrix of the incident variables was rotated according to the Varimax method. Table IV is the score matrix of the incident variables. The score value of the incident variables was calculated according to formula (2).

TABLE IV
THE SCORE MATRIX OF THE INCIDENT VARIABLES

	Incident variable principal factor score value						
	F_1	F_2	F_3	F_4	F_5	F_6	F_7
x_1	.002	.015	.001	-.330	.030	.027	-.005
x_2	.004	.015	.013	-.407	.095	.027	-.024
x_3	-.003	.006	-.014	.397	-.069	-.069	.017
x_4	.001	.004	.350	-.007	-.032	.005	-.011
x_5	-.001	-.019	.002	-.045	.380	-.024	-.026
x_6	.007	-.018	-.103	-.061	.414	-.028	-.082
x_7	.000	.003	.346	-.018	-.023	.009	-.003
x_8	-.013	-.008	.024	-.069	.348	-.035	.070
x_9	.003	.002	.357	-.006	-.041	.011	-.013
x_{10}	.009	.023	-.017	.093	-.024	-.317	-.087
x_{11}	.015	.086	-.007	.016	.064	-.457	-.036
x_{12}	-.011	-.078	.001	-.005	-.059	.436	-.042
x_{13}	.315	-.057	-.006	-.009	.037	-.008	-.295
x_{14}	.006	.279	-.014	.003	.045	-.036	-.181
x_{15}	-.108	.320	-.014	.006	.039	-.050	-.085
x_{16}	.279	-.015	.006	.000	-.027	-.016	.098
x_{17}	.012	.299	.009	-.017	-.029	-.069	.117
x_{18}	.288	-.075	-.011	.001	.030	.009	-.076
x_{19}	.244	.010	.012	.000	-.048	-.023	.129
x_{20}	-.032	.319	.027	-.019	-.090	-.055	.096
x_{21}	-.044	-.015	-.015	.010	-.011	.029	.867

C. Detection effect analysis of AID algorithm based on WRF

The experimental operation environment is based on Matlab R2014A. The relevant parameters of the model are obtained through cross validation [26, 27, 33, 35, 36], the final selection is $p = 600$ and $m = 3$.

Three algorithms, namely RF-AID algorithm [16] (Algorithm 1), WRF-AID algorithm (Algorithm 2), and the FA-SVM AID algorithm [17] (Algorithm 3), are compared with the algorithm proposed in this paper (Algorithm 4). Among them, algorithm 1 is the RF algorithm with better detection effect, algorithm 3 is the most widely used artificial intelligence AID algorithm. Due to the data does not include the start time of the incident, the MTTD cannot be used to evaluate the classification effect. The corresponding confusion matrix and the detection effect are respectively shown in Table V and VI, the specific analysis is as follows.

TABLE V
The CORRESPONDING CONFUSION MATRIX

Compare Algorithm	The parameter values	
Algorithm 1	487	88
	200	10745
Algorithm 2	520	55
	155	10790
Algorithm 3	512	63
	121	10824
Algorithm 4	550	25
	100	10845

TABLE VI
AID ALGORITHM CLASSIFICATION EFFECT

Compare Algorithm	Evaluation Index				
	DR(%)	FAR(%)	AR(%)	F	AUC
Algorithm 1	84.7%	1.83%	97.5%	0.771	0.813
Algorithm 2	90.43%	1.42%	98.18%	0.832	0.897
Algorithm 3	89.04%	1.11%	98.4%	0.847	0.859
Algorithm 4	95.65%	0.91%	98.9%	0.897	0.923

The classification results are shown in Table VI. On the whole, the four algorithms can achieve better detection effect, and the design algorithm in this paper has the incident overall detection effect, indicating that WRF model is effective for the classification of unbalanced data, and the extraction of incident feature variables by using FA method in this paper can improve the detection effect of the algorithm;

On the premise of effective extraction of initial incident variables, the detection effect of the designed algorithm in this paper is better than that of algorithm 3, which indicates that the WRF-AID model has better classification effect than the SVM algorithm, which is recognized to have good classification performance.

Comparing three RF-based AID, the evaluation indexes of the algorithm designed in this paper are all better, the value of AUC is 0.11 and 0.026 higher than that of algorithm 1 and algorithm 2 respectively, and the F value is also higher. It indicating that weighted processing of each decision tree in the RF model can effectively improve the overall detection effect when detecting unbalanced traffic incident data. At the same time, it is further verified that the effective extraction of initial incident variables designed from a multi-dimensional perspective can also improve the detection effect of traffic incidents.

From the perspective of each evaluation index, the accuracy of each algorithm is high. For example, algorithm 1 can obtain a high accuracy even when DR and F all are low, FAR is high, indicating that it is inappropriate to use the accuracy alone to evaluate the detection effect of unbalanced traffic incident data.

V. CONCLUSION

This work solves the problem of the influence of unbalanced traffic incident data on the classification effect of RF algorithm. Its goal is to get better classification result. A RF-WFR method is designed to solve the proposed problem. The results show that the overall detection effect of FA-WRF-based AID algorithm has the better detection effect, which is competitive in processing unbalanced data classification compared with RF-AID algorithms. And compared with the classical SVM algorithm, the advantages are obvious. The results can be adopted to guide traffic managers to make wise decisions in traffic control, and provide a new analysis method for expressway traffic incident detection.

Although the efficacy of the proposed method has been tested, this work has some limitations. 1) The realization of the method in this paper depends on the quantity and density of the location traffic parameter acquisition equipment on the expressway, and it is only applicable to the freeway and urban expressway with continuous traffic flow, and has certain limitations for urban roads with signal control. 2) There are many initial incident variables set by the method in this paper, and incident variables need to be extracted, which leads to the need to further improve the real-time performance and computing efficiency. Therefore, we need to develop more advanced disassembly planning models and approaches in the future.

REFERENCES

[1] H.J. Payne and S.C. Tignor, "Freeway incident detection algorithms based on decision trees with states," *Transportation Research Record 682*, TRB, National Research Council, Washington D.C., USA, 1978, pp. 30-37.

[2] C.L. Dudek and G.M. Messer, "Incident detection on urban freeways," *Transportation Research Record 459*, TRB, National Research Council, Washington D.C., USA, 1974, pp. 12-24.

[3] A.R. Cook and D.E. Cleveland, "Detection of freeway capacdty reducing incidents by traffic stream measurement," *Transportation Research Record*, TRB, pp. 1-11, 1974.

[4] M. Levin and G.M. Krause, "Incident Detection: A Bayesian Approach," *In Transportation Research Record 682*, TRB, National Research Council, Washington, D.C., USA, 1978, pp. 52–58.

[5] S.A. Ahmed and A.R. Cook, "Analysis of freeway traffic time-series data by using Box-Jenkins technique," *Transportation Research Record*, pp. 3-11, 1979.

[6] B.N. Persaud and F.L. Hall, "Catastrophe theory and patterns in 30-second freeway traffic data-implication for incident detection," *Transportation Research Record*, vol. 23, no. 2, Pages 104-110, Mar.1989.

[7] S.G. Ritchie and R.L. Cheu, "Simulation of freeway incident detection using Artificial Neural Networks," *Trasportation Research Part C*, vol.1, no. 3, pp. 313-331, Sep.1993.

[8] S.G. Ritchie and R.L. Cheu, "Neural network models for automatically detection of non-recurring congestion," *Transportation Research Record*, vol.1, no. 3, pp. 9-17, Jan.1993.

[9] S.G. Ritchie, B. Abdulhai, "Development testing and evaluation of advanced techniques for freeway incident detection," *California Partners for Advanced Transit and Highways (PATH)*, Institute of Transportation Studies (UCB), UC Berkeley, pp. 45-53, Jan.1997.

[10] B. Abdulhai and S.G. Ritchie, "Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network," *Transportation Research Part C: Emerging Technologies*, vol.7, no.5, pp. 280—287, May.1999.

[11] D. Srinivasan, R. L. Cheu and P. Poh, "Development of an intelligent kcimique for traffic network incident detection," *Engineering Applications of Artificial Intelligence*, vol.13, no. 3, pp. 311-322, Feb.2000.

[12] H. Teng and Y. Qi, "Application of wavelet technique to freeway incident detection," *Transportation Research Part C: Emerging Technologies*, vol.11, no. 3, pp. 289-308, Jun. 2003.

[13] F. Yuan and R.L. Cheu, "Incident detection using support vector machines," *Transportation Research Part C*, vol.11, no. 3, pp. 307-325, May.2003.

[14] W.R. Chen, P. Guan and Y.X. Zou, "SVM-based traffic incident detection technology," *Journal of Southwest Jiaotong University*, vol.46, no. 1, pp. 63-67,Jun. 2011.

[15] C. Wang and X.H. Fan, "Research on traffic incident detection based on feature weighting," *Microelectronics and computers*, vol.29, no.10, pp.121-123, Feb.2012.

[16] Q.C. Liu, L. Jian and C. Chen, "Design and analysis of traffic incident detection based on Random Forest." *Journal of Southeast University: English Edition* , vo1.30, no.1, pp.88-95, Mar.2014.

[17] H. Jiang, "Freeway traffic incident detection based on FA-SVM," *Journal of Beihua University (Natural Sciences)*, vol.20, no.1, pp. 103-108, Jan.2019.

[18] Q.C. Bing, Z.S. Yang and X.Y. Zhou, "An automatic traffic incident detection algorithm based on factor analysis and minimum and maximum probability machine," *Traffic information and safety*, vol.33, no.2, pp. 74-78, Aug.2015.

[19] W.Z. Yang, S.K. Chen and R. Liu, "SPSS statistical analysis from introduction to mastery (The fourth Edition)," *Tsinghua University Press*. vol.31, no.9, pp. 55-62, Sep.2018

[20] L. Breiman, "Random forests," *Machine Learning*, vol.41, no.1, pp. 5–32, 2001.

[21] C. Wang and R. Gao, "Research on stochastic forest improvement algorithm based on feature reduction," *Computer technology and development*, vol.30, pp. 40-45, May.2020.

[22] Z.T. Wei and R. Gao, "Improvement of random forest classification algorithm for unbalanced data," *Journal of Chongqing University*, vol.41, pp. 54-62, Oct.2018.

[23] L.I. Kuncheva and J.J. Rodríguez, "A weighted voting framework for classification ensembles," *Knowledge and Information Systems*, vol. 38, no.2, pp. 259-275, Nov.2014.

[24] J. Supper, C. Spieth, and A. Zell, "Reconstructing linear gene regulatory networks," *European Conference on Evolutionary Computation*.2007.

[25] R.M. Haralick, S.R. Sternberg and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans on Pattern Anal Machine Intell*, vol.9, pp. 532-550, Dec.1987.

[26] M. Liu, R.L. Lang and Y.B. Cao, "Number of trees in random forest," *Computer Engineering and Applications*, vol.51, no.5, pp. 126-131, May.*2015*.

[27] G. Tian, M. Zhou, and P. Li, "Disassembly sequence planning considering fuzzy component quality and varying operational cost," *IEEE Transactions on Automation Science and Engineering*, vol. 15, pp. 748-760, 2018.

[28] Y. Li, "Traffic engineering," *People's Communications Publishing House*, China, 2019.

[29] L. Yu, "Urban traffic flow theory," *Beijing Jiaotong University Press*, China, 2016.

[30] C. Zhang, G. Tian, A. M. Fathollahi-Fard, W. Wang, P. Wu, Z. Li, "Interval-Valued Intuitionistic Uncertain Linguistic Cloud Petri Net and Its Application to Risk Assessment for Subway Fire Accident", *IEEE Transactions on Automation Science and Engineering*, doi: 10.1109/TASE.2020.3014907, 2020.

[31] G.Y. Jiang, "Technology and application of road traffic state discrimination," *People's Communications Publishing House*,China, 2004.

[32] W.J Wang, G.D Tian, and M.N Chen *et al.*, "Dual-objective program and improved artificial bee colony for the optimization of energy-conscious milling parameters subject to multiple constraints," *Journal of Cleaner Production*, vol. 245, no. 1, pp. 124-135, Feb.2020.

[33] M Tang, Z.W Li, and G.D Tian *et al.*, "A Data-Driven-Based wavelet support vector approach for passenger flow forecasting of the metropolitan hub," *IEEE Access*, vol. 7, no. 2, pp. 7176–7183, Jan.2019.

[34] Z.F. Chen , L.L Zhang, and G.D Tian *et al.* , "Economic Maintenance Planning of Complex Systems Based on Discrete Artificial Bee Colony Algorithm," *IEEE ACCESS*, vol. 8, no. 8, pp. 108062-108071, Mar. 2020.

[35] S. Gao, M. Zhou, Y.R. Wang and J.J. Cheng *et al.*, " Dendritic neuron model with effective learning algorithms for classification, approximation and prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 601-614, Feb. 2019.

[36] C.C. Leng, H. Zhang and G.R. Cai *et al.*, "Graph regularized Lp smooth non-negative matrix factorization for data representation," *IEEE/CAA J*, vol. 6, no. 2, pp. 584-595, Mar. 2019.