Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 708467, 8 pages
http://dx.doi.org/10.1155/2015/708467

Research Article

Data Calibration Based on Multisensor Using Classification Analysis: A Random Forests Approach

Xue Xing,[1,2] Dexin Yu,[1,3] and Wei Zhang[4]

[1] *College of Transportation, Jilin University, Changchun 130022, China*
[2] *College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China*
[3] *State Key Laboratory of Automobile Dynamic Simulation, Jilin University, Changchun 130022, China*
[4] *Shandong Hi-Speed Company Limited, Jinan 250002, China*

Correspondence should be addressed to Wei Zhang; zhangwei_txj@126.com

Received 1 September 2015; Revised 22 October 2015; Accepted 25 October 2015

Academic Editor: Sergio Preidikman

This paper analyzes the problem of meaningless outliers in traffic detective data sets and researches characteristics about the data of monophyletic detector and multisensor detector based on real-time data on highway. Based on analysis of the current random forests algorithm, which is a learning algorithm of high accuracy and fast speed, new optimum random forests about filtrating outlier in the sample are proposed, which employ bagging strategy combined with boosting strategy. Random forests of different number of trees are applied to analyze status classification of meaningless outliers in traffic detective data sets, respectively, based on traffic flow, spot mean speed, and roadway occupancy rate of traffic parameters. The results show that optimum model of random forest is more accurate to filtrate meaningless outliers in traffic detective data collected from road intersections. With filtrated data for processing, transportation information system can decrease the influence of error data to improve highway traffic information services.

1. Introduction

With the constant development in digital image technology and detection technology, traffic state information can be collected by technology of magnetic frequency, wave frequency, video, and GPS which has been installed in most of the vehicles [1]. In addition, RFID technology and mobile signaling technology can also provide such information as a supplementary role. A lot of spatiotemporal data sets are obtained by above technology. For the purpose of efficient traffic state identification and prediction [2–4], the premise is to grasp accurate real-time traffic data. Outliers problem [5] occurs in progress of traffic awareness data sets obtaining traffic information; namely, the traffic information contains some data which are obviously inconsistent data with other data. There are many causes of outliers as follows: (1) short period of collection; (2) imperfect detective devices; (3) loss of data; (4) errors in detective data being transferred; (5) environmental factors. If discriminating process of the traffic state ignores the existence of outlier data, a mixture of meaningless outlier data and traffic events data will be stored. It is a basic question in transportation information: how to effectively distinguish outlier data using the multidimensional characteristics to effectively improve accuracy of the traffic prediction.

In recent years, increasing attention has been given to outlier research in dynamic traffic data. Nam and Drew [6] pointed out that conservation laws for the traffic flow could recognize and process erroneous data. Vanajakshi and Rilett [7] used loop detector data to analyze cumulative flow with adjacent section data. The law of conservation of flow optimization model was established with target of minimizing the sum of the squares of adjacent detection section cumulative flow in order to eliminate the error when several continuous detection sections showed counting errors. Smith et al. [8] proposed a calibrated idea to fix outlier data using exponential smoothing method. Methods of optimum data based on clustering [9–11] and genetic algorithm [12] were presented in view of outlier in the multidimensional characteristic data in recent years.

This paper analyzes the problem of meaningless outliers in traffic detective data sets and researches characteristics about the data of monophyletic detector and multisensor detector based on real-time data on highway. Based on analysis of the current random forests algorithm, which is a learning algorithm of high accuracy and fast speed, new optimum random forests about filtrating outlier in the sample are proposed, which employ bagging strategy combined with boosting strategy. Random forests of different numbers of trees are applied to analyze status classification of meaningless outliers in traffic detective data sets, respectively, based on traffic flow, spot mean speed, and roadway occupancy rate of traffic parameters. The results show that optimum model of random forest is more accurate to filtrate meaningless outliers in traffic detective data collected from road intersections. With filtrated data for processing, transportation information system can decrease the influence of error data to improve highway traffic information services.

2. Random Forest Optimization Model Based on Traffic Data

2.1. Problem Description. The data derived from road traffic detectors contains detecting time, detector type, flow, spot mean speed, occupancy rate, and so on. The following three conditions show the real-time road data from group representation:

(1) Road traffic state detection data is out of line largely with the actual road traffic status value.

(2) Obtained states of road traffic data are error data, because the values from them are beyond the reasonable scope or have violated the relevant law of road traffic.

(3) The data of road traffic state data are missing.

First of all, road detection data of single parameter are compared in Figure 1. Figure 1 represents the scatterplot of spot mean velocity extracted from geomagnetic detection data of freeways in November 2014. There are 1320 groups of discrete detection data collected forming the same section of 165 time points in it. In addition, Figure 2 lays out the difference of sensor data from the same cross section. Figure 2 represents integrated scatterplot of the same section of multisensor data, which contains three parameters of flow, spot mean speed, and occupancy rate, determining location of data point. Two figures of data samples show that outlier data is present in the data, but the proportion of outlier data in the samples is small. In the above, statistics cases which accounted for the largest number of samples are called the most classes, and accounts for the fewest category are called the minority class (nonequilibrium data) [13].

2.2. Model Based on Traffic Data

2.2.1. Random Forest. RF is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual tree [14]. RF using bagging resampling strategy form sample sets combines the tree predictors by majority voting. Each tree grows using a new bagging training set.

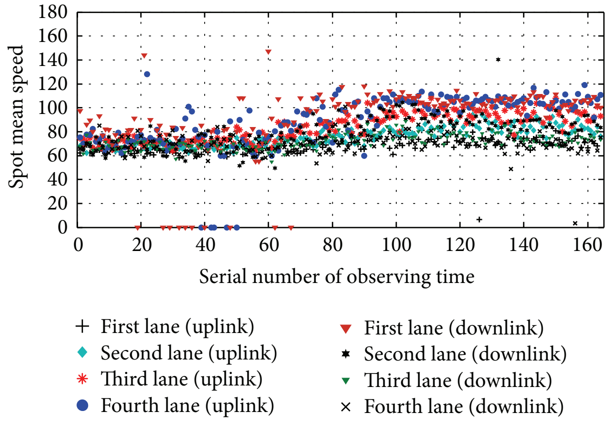

FIGURE 1: Road detection data scatterplot of single parameter.

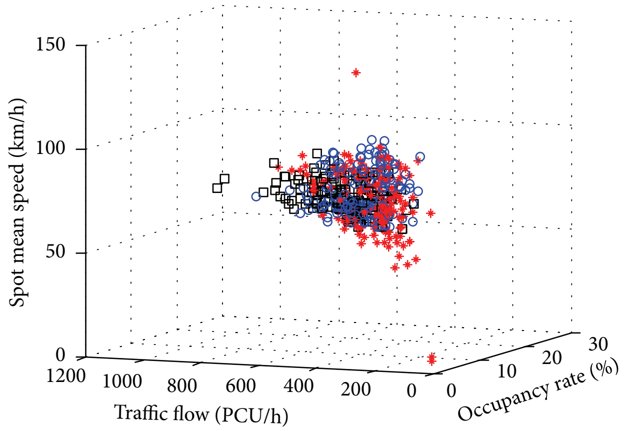

FIGURE 2: Scatterplot of multisensor parameters data in the same section.

RF is one of the most accurate leaning algorithms available. For many data sets, it produces a highly accurate classifier and runs efficiently on large databases. A significant advantage of RF is that it can generate an internal unbiased estimate of the generalization errors within the forest building progress. Furthermore, RF is less prone to overfit. It is widely applied to many domains, such as computer vision, information retrieval, data mining, and pattern recognition. Mathematical description of random tree classification model is as follows.

Definition 1. A random forest is a classifier consisting of a collection of tree structured classifiers $\{h(x, \Theta_k), k = 1, 2, \ldots, m\}$, where $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. Voting model equally weighted is presented as in the following formula:

$$H(x) = \arg\left\{\max \frac{1}{m}\sum_{i=1}^{m} I\left(h\left(X; \Theta_i\right) = y_j\right)\right\}. \quad (1)$$

Given an ensemble of classifiers $\{h_1(x), h_2(x), \ldots, h_m(x)\}$, each of these can get a classification. A classifier $h_k(x)$ is a common way of $h(x, \Theta_k)$. With the training set drawn at random from the distribution of the random vector Y, X, define the margin function as

$$\begin{aligned} \operatorname{mg}(X, Y) = {} & av_n I\left(h_k(X) = Y\right) \\ & - \max_{j \neq Y} av_k I\left(h_k(X) = j\right), \end{aligned} \tag{2}$$

where $I(\cdot)$ is the indicator function. The margin measures the extent to which the average number of votes at X, Y for the right class exceeds the average vote for any other class. The larger the margin is, the more the confidence is in the classification. The generalization error is given by

$$PE^* = P_{X,Y}\left(\operatorname{mg}(X, Y) < 0\right). \tag{3}$$

The strength of the set of classifiers $\{h(x, \Theta)\}$ is

$$\begin{aligned} s = E_{X,Y} \Bigg(& P_\Theta\left(h(X, \Theta) = Y\right) \\ & - \max_{j \neq Y} P_\Theta\left(h(X, \Theta) = j\right) \Bigg). \end{aligned} \tag{4}$$

An upper bound for the generalization error is given by

$$PE^* \leq \frac{\overline{\rho}\left(1 - s^2\right)}{s^2}, \tag{5}$$

where $\overline{\rho}$ is the correlation between two members of the forest averaged over different distribution.

The property of a decision tree in random forests is bagging random sampling. Input data for random forest is a process of resample from training set, and the sampling of sample collection may be duplicate samples. Compared to another common boosting method, in terms of sampling, bagging is uniform sampling. Boosting is sampling according to the error rate; thus the classification accuracy of boosting is better than bagging; the choice of the training set of bagging is random, and it is independent of the training set. The choice of boosting is related to each previous sampling, and it gets learning results. The prediction function of bagging cannot be weighted and can be generated in parallel. The prediction function of boosting can be weighted and can only be generated sequence. Both methods can effectively improve the accuracy of classification, and thus the paper using advantages of the two methods proposes an integrated method to optimize traffic field data classification.

2.2.2. Training Set and Testing Set. When data sets are generated in random forest model, the initial training of some samples could not be extracted from all collected data. The data which could not be sampled are called OOB (out of the bag). The whole data set is divided into two parts: a set of training and a set of testing. The former one is used to build the model; the latter one is used to test capability of the model.

In traffic detection data set, each testing point can get a lot of sensory data composed of a variety of detection sources. Suppose there are n sources; each data source can get multiple traffic parameters of detected section, and then each time all can get a set of multisensor data. Define a perception data set consisting of time t, n different types of data sources to the monitoring object, and attributes of m, represented by $\{d, t_i, \mathrm{DN}_i, p_{1,1}, p_{1,2}, p_{1,3}, \ldots, p_{k,1}, p_{k,2}, p_{k,3}, L_i\}$, in which DN indicates detector number, d indicates the day, t_i indicates the data acquisition time, $p_{j,m}$ indicates the mth parameter of the jth traffic detector, and L is quality mark.

For the convenience of analyzing detector data collected in cross section of road, three fundamental traffic parameters, namely, flow, spot speed, and occupancy rate, which are extracted from data sets commonly, using three kinds of detection equipment of data (induction loop data, magnetic data, and monitoring data). Data calibration about traffic parameters for some detector of a certain acquisition time needs to extract the spatial correlation data from other detectors. For instance, a detection equipment at acquisition time, t_i, gets the flow q_{Ci}, spot mean speed v_{Ci}, and occupancy o_{Ci} from traffic induction loop. If the data need be calibrated, properties should be selected, such as traffic data collection time t_i, flow q_{Ci}, spot mean speed v_{Ci}, and occupancy o_{Ci} from induction loop data, volume q_{Ui}, spot mean speed v_{Ui}, and occupancy o_{Ui} from magnetic data, and volume q_{Ti}, spot mean speed v_{Ti}, and occupancy o_{Ti} from monitoring data. And traffic data quality mark L_i, $i = 1, 2, \ldots, n$, in which L_i value belongs to $\{1, 1\}$, indicates that testing calibration set evaluation data information is normal data or outlier.

The number of X-variables is 10. This means that the matrix X used in training the model has the size of 22093 × 10. The test data X forms a matrix with a size of 22093 × 10. The formal description of matrices X and Y can be written as follows:

$$\begin{aligned} X &= \begin{bmatrix} x_1 & x_2 & \cdots & x_{10} \end{bmatrix} \\ &= \begin{bmatrix} t_1 & q_{C1} & v_{C1} & o_{C1} & q_{U1} & v_{U1} & o_{U1} & q_{T1} & v_{T1} & o_{T1} \\ t_2 & q_{C2} & v_{C2} & o_{C2} & q_{U2} & v_{U2} & o_{U2} & q_{T2} & v_{T2} & o_{T2} \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\ t_n & q_{Cn} & v_{Cn} & o_{Cn} & q_{Un} & v_{Un} & o_{Un} & q_{Tn} & v_{Tn} & o_{Tn} \end{bmatrix}, \end{aligned} \tag{6}$$

where X_i is a set of data elements and n is the number of input samples; consider

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_n \end{bmatrix}, \tag{7}$$

where $y_i \in \{-1, 1\}$ and y_i represent the results of data quality assessment.

In the training of random tree building, not every sample is selected in input sample of the decision tree. The number of choices is m from M features. A decision tree builds by completely split process, such as from x_i to y_i. Processes of the decision tree terminate, when each leaf node cannot continue to split or all samples are pointing to the same classification.

2.2.3. Random Forest Optimization Model. Based on realistic significance of identifying the nonequilibrium data in traffic information, and the decline performance of random forest classification method for scarce and extreme value [15], this paper focuses on a few samples which are given greater weight in each independent decision tree of random forests to avoid unrepresentative training of the decision tree rules by the amount of data being trained. The character of training forces the classifier to pay more attention to the minority class samples and improve the accuracy of the class with less training data. Proposed model can solve the problem of nonequilibrium data sets classification. Eventually research gets the vote result on nonequilibrium samples with higher accuracy.

When using bagging in a method of randomly selecting sampling, the original training set of the minority class is less and probability of selected nonequilibrium samples is very low. This section proposes a method to optimize random forest to put particular emphasis on nonequilibrium samples. The basic idea of optimization algorithm is loading some established characteristics of the tree in the process of building a new tree. The core of the optimization algorithm steps combines formal bagging strategy with boosting strategy. First of all, according to the original algorithm for random sampling, the resampling n numbers of instances (when the initial set is training, this algorithm keeps the original bagging method), then, adjust the data according to the principle of boosting. An algorithmic principle maintains original effective randomization process and selection of the random properties and improves the random forest adaptive accuracy. Therefore, except for using the bagging to build the first tree, an evaluation of the current forest added to data induction; that is, it estimates the prediction error of data to weight random selection of training examples. It is necessary to improve the classification ability of the sample and contribute to the subsequent decision tree. The bag outside data are independent of generating sampled data. With continuation of the underlying principle of random forest (using OOB to estimate the error), an estimation function only about the bag outside data is given in the following definition, as the following formula:

$$\varepsilon(X,Y) = \frac{\sum_{h_i \in h_{\text{oob}}} I\left(h\left(X;\Theta_i\right) = y_1\right)}{\sum_{h_i \in h_{\text{oob}}} I\left(h\left(X;\Theta_i\right) = y_j\right)}, \tag{8}$$

where $I(\cdot)$ is the indicator function. X is given by the independent variable; y is an actual classification. $h(X,\Theta_i)$ represents output of ith decision tree; h_{oob} represents outside the bag set of X. Less the value of $\varepsilon(X,Y)$ means that the more the current forest error classification tree exists, the more the attention should be paid to the subsequent instance X. Therefore, the design of the weighting function should increase with decrease of the corresponding $\varepsilon(X,Y)$. By analyzing a typical example, this section gives a corresponding weight distribution formula, as shown in the following formula:

$$W(\varepsilon(X,Y)) = 1 - \varepsilon(X,Y). \tag{9}$$

In order to clearly describe the random forest optimization model (RFOM), the required explanation is as follows: a given N represents the individual number of training sets (X,Y) in the individual number. M is the category of the classification characteristics. And K represents the number of decision trees in the "forest." The optimization method for traffic outlier data is as follows:

(1) The original samples are on the training set, given the initial distribution $D_1(x_i, y_i) = 1/N$.

(2) Train the decision tree. Randomly sample with replacement for the first random sample set T_1, and then train the decision tree based on sampling. Unsampled samples form the first bag outside data.

(3) Add a weight value of D_k to each instance of set T_k $(k = 2, 3, \ldots, n)$.

(4) For every tree, randomly sample m characteristics $(m < M)$. Calculate Gini coefficient of each sample and the Gini coefficient of each division, such as formula (10) and formula (11):

$$\text{Gini}\left(T_k\right) = 1 - \sum_{i=1}^{n} P_i^2, \tag{10}$$

where P_i represents probability of class Cj and n the sample set in T_k, and

$$\text{Gini}_s\left(T_k\right) = \frac{\left|T_{k1}\right|}{\left|T_{k2}\right|}\text{Gini}\left(T_{k1}\right) + \frac{\left|T_{k2}\right|}{\left|T_{k1}\right|}\text{Gini}\left(T_{k2}\right). \tag{11}$$

Then based on the principle of minimum Gini index, select a variable to split. Finally through a recursive form, train classification rules of a decision tree. Maximize each tree without clipping.

(5) Merge decision trees into a forest.

(6) The normalized variable is $Z = 0$.

(7) For each training set x_i, if the number of the bag outside the tree is not empty, $D_{k+1}(x_i, y_i) = W(\varepsilon(x_i, y_i))$, where $W(\varepsilon(x_i, y_i))$ calculate as formula (9); else the original weight remains the same, and sum $D_{k+1}(x_i, y_i)$ with normalized variable Z.

(8) Consider $D_{k+1}(x_i, y_i) = D_{k+1}(x_i, y_i)/Z$.

(9) If the current k of a decision tree is less than K, according to the boosting, weight the new distribution of random sampling T_k with replacement. Unsampled samples form the bag outside data and return to (3).

(10) After merging decision trees into forests, classify new data with random forests. The vote of tree classifier depends on classification results.

Table 1: Properties Description of the Non-equilibrium Datasets.

Data Set	Data	Sample size	Outlier size	Normal sample	Non-equilibrium rate
Jibei Station Data Set	Induction loop data 1	4824	182	4642	3.921%
	Magnetic data 1	6399	443	5956	7.452%
	Monitoring data 1	10870	659	9619	6.851%
Jiaozhou Station Data Set	Induction loop data 2	5475	201	5217	3.671%
	Magnetic data 2	6001	342	5633	5.699%
	Monitoring data 2	12801	779	11939	6.085%
Gaotang Station Data Set	Induction loop data 3	5013	198	4762	3.949%
	Magnetic data 3	6512	531	5972	8.154%
	Monitoring data 3	11194	625	10478	5.583%

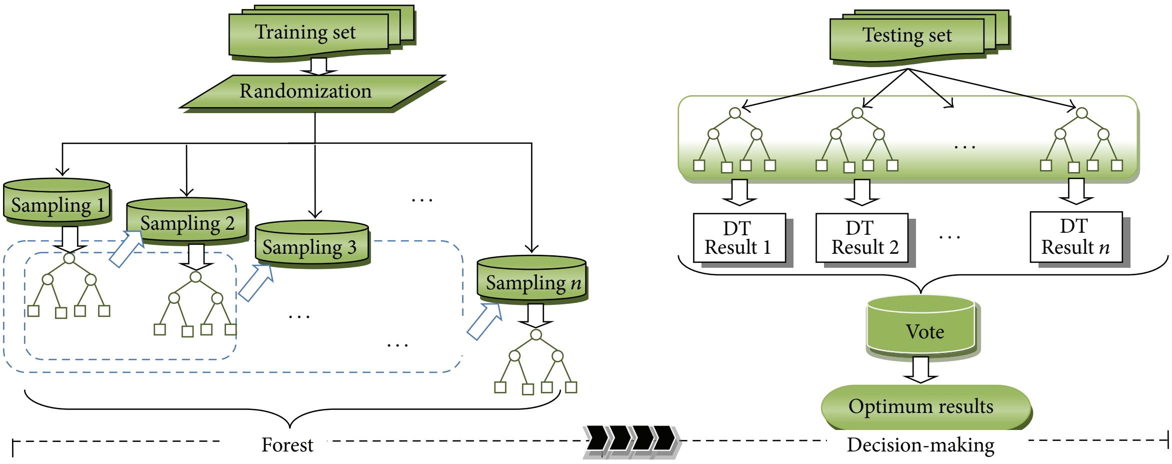

Figure 3: Schematic of random forest optimization model.

The optimization model for induced random trees shows a schematic of the algorithm in Figure 3. The first tree is trained in the traditional way, namely, in the process of training decision tree with equal consideration to each sample; then algorithm modifies the weight of some samples, namely, adding corresponding sample weight of correct classification. New training set is sampled under the condition of second tree weighted; "forecast" of the first and second tree is calculation to get the updated weights in the third iteration. By analogy, the new weighted data set is trained. The optimization process based on the efficiency of random feature selection approach [16, 17] could retain characteristics of the random forest algorithm and constant prior probability. According to the actual data need, induction from samples of the minority class is improved to train a random forest.

3. Experimental Validations

3.1. Data Collection. The data was collected by Shandong Hi-Speed Group Co., Ltd., at the freeways in Shandong province, China. The comparison result is obtained by using data sets in monitoring stations data selected on November 13, 2014. Data contains traffic parameters, such as flow, spot speed, and occupancy rate from induction loops, magnetoresistive sensors, and monitoring devices at Jibei, Jiaozhou, and Gaotang monitoring stations. The properties of the nonequilibrium datasets are described in Table 1.

3.2. Performance Indicator. The performance indexes of classification accuracy, detection rate, false positive rate, and precision rate are used to evaluate performance of algorithm classification. Classification accuracy, detection rate, false positive rate, and precision rate are defined as follows:

$$\begin{aligned} \text{Acc} &= \frac{\text{CN}+\text{CG}}{\text{CN}+\text{CG}+\text{EN}+\text{EG}}, \\ \text{DR} &= \frac{\text{CN}}{\text{EG}+\text{CN}}, \\ \text{FPR} &= \frac{\text{EN}}{\text{EN}+\text{CG}}, \\ \text{PR} &= \frac{\text{CN}}{\text{EN}+\text{CN}}, \end{aligned} \tag{12}$$

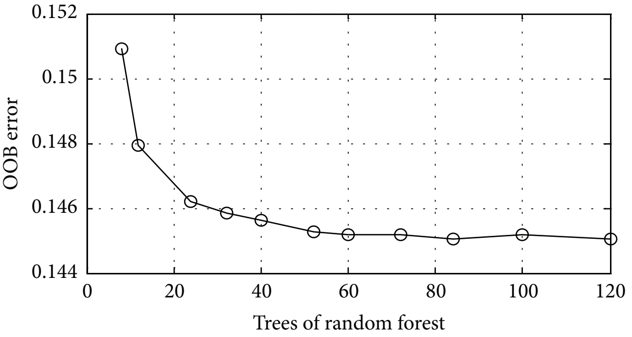

Figure 4: Diagram of OOB estimate with number of trees in forests.

Table 2: Comparison of Algorithm Performance.

	Acc (%)	DR (%)	FPR (%)	PR (%)	F_m	G_m
CART	83.37	76.26	0.90	86.13	0.9411	0.7687
RF_100	84.23	72.73	0.92	91.46	0.9105	0.7343
RFOM_60	90.89	89.76	0.89	92.55	0.9449	0.8227
RFOM_80	**91.93**	90.01	**0.86**	92.78	0.9528	**0.8489**
RFOM_100	91.81	**90.17**	0.90	**93.90**	**0.9635**	0.8333

where CN represents the number of detected outliers; EG represents the number of undetected outliers; CG represents the number of detected normal data; EN represents the number of undetected normal data.

As described in previous section, nonequilibrium feature of traffic instance data is a significant problem to be solved. Because it would increase, risk of classification error increases. So F_m and G_m are often used to measure the classification of this situation. F_m index is defined, such as formula (13). Parameter G_m is geometric average of two kinds of classifying accuracy, such as formula (14):

$$F_m = \left(1+\beta^2\right) \times \mathrm{DR} \times \mathrm{PR} \times \left(\beta^2 \times \mathrm{DR} + \mathrm{PR}\right), \tag{13}$$

where β represents proportionality coefficient of precision rate and detection rate,

$$G_m = \sqrt{\left[\frac{\mathrm{CN}}{(\mathrm{CN}+\mathrm{EG})}\right] \times \left[\frac{\mathrm{CG}}{(\mathrm{EN}+\mathrm{CG})}\right]}. \tag{14}$$

3.3. Comparison. In this section, we perform experiments for comparison: the first comparison is used to choose the optimal number of trees for random forest optimum model performance; the second comparison compares performance of decision tree, random forest, and random forest optimum model; the last comparison compares ROC curve of random forest and random forest optimum model. The experiments are performed on Shandong freeways real data to investigate the performance of random forest optimum model. Evaluation indicators include classification accuracy, detection rate, false positive rate, precision rate, and ROC.

Random forest optimum model (RFOM) uses OOB error estimation as indexes to select appropriate number of decision trees. Through repeated experiments about different forests, we take average standard deviation of each tree in a forest as OOB estimate of the forest. The experimental results are shown in Figure 4. OOB estimated gradually reduce with the increase of the trees. The classification accuracy of algorithm increases with the increasing number of trees in the forest and then keeps stabilization. When the number of trees increases to over a certain degree, a limiting value of OOB error appears; namely, classification accuracy of RFOM algorithm tends to being stable.

In addition, the number of decision trees is related to accuracy, detection rate, false positive rate, and precision rate of RFOM algorithm. Figures 5(a)–5(d) show boxplots of error rates. Horizontal lines inside the boxes are median error rates. Figures 5(a)–5(d) are detection indexes, which are different degrees of growth except for FPR. When the number of trees is fewer than 60, Acc, DR, and PR grow relatively fast. Through repeated experiments about forests of different trees, we take average standard deviation of each tree in a forest as OOB estimate of the forest. The experimental results are shown in Figure 5. According to the two aspects, parameter num is selected in [60, 100].

This experiment using different size of training set builds RFOM, respectively. Algorithm performance of classifications is compared. The number of trees is from 60 to 100, adding 20 every time. We increase the number of trees in order to obtain a greater difference. The optimization random forests with 60 trees, 80 trees, and 100 trees and tree optimization random forest are named RFOM_60, RFOM_80, and RFOM_100. In order to compare algorithm performance, CART, RF, and RFOM are used to classify outlier for *Jibei Station Data Set*, respectively. Six performance measures, such as Acc, DR, FPR, PR, F_m, and G_m, are computed for different situations, which are shown in Table 2.

It is observed that different numbers of trees yield similar classification accuracy, and RFOM obtains a better performance than CART or RF. The ACC of RFOM_80 is 91.93%, which is the best. The false positive rate of RFOM_80 is 0.86%, which is the best. As F_m for RFOM is concerned 0.9635 of detection rate yield by RFOM_100 is the best one. G_m of RFOM_80 is 0.8489, which is the best. RFOM_60 obtains performance lowest in RFOM algorithms. Among five comparisons, RFOM_100 and RFOM_60 outperform the other methods. In the Shandong freeways real data, when the tree number is 80, it can obtain some improvement and save time of calculation.

In addition, RFOM is superior to other algorithms, as shown in Figure 6. It uses ROC curve to evaluate the detection method. Comparison of the ROC curve is more intuitive, which is as false positive rate for the horizontal axis, with detection rate for the vertical axis.

Three data sets, including *Jibei* Station Data Set, *Jiaozhou* Station Data Set, and *Gaotang* Station Data Set, are computed for performance measures, which are shown in Table 3. The number of trees is 80 in the comparison with different

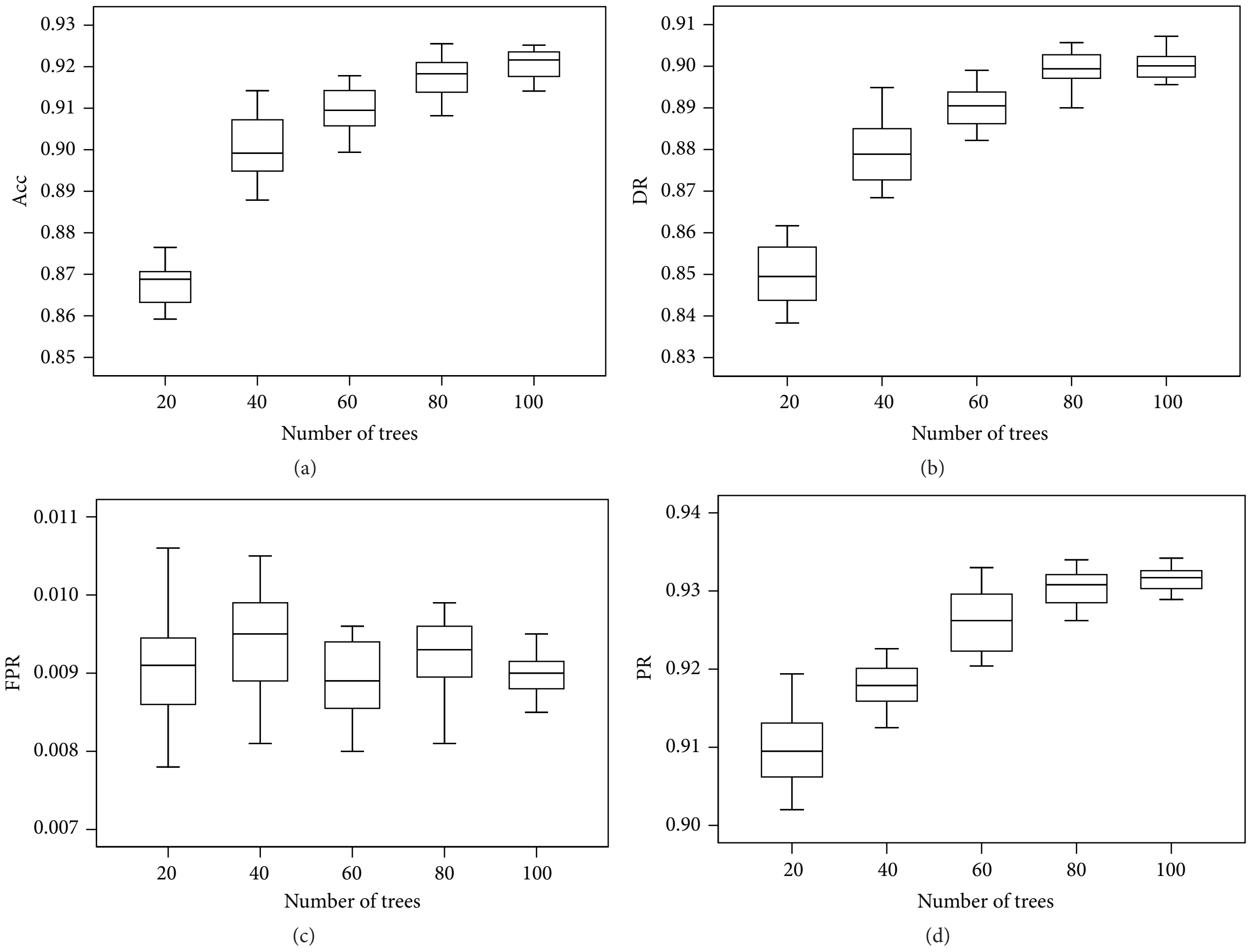

Figure 5: Boxplots of testing of Jibei Station Data Set: (a) Acc; (b) DR; (c) FPR; (d) PR.

Table 3: Comparison of Algorithm Performance with Different Data Set.

Data Set	Acc (%)	DR (%)	FPR (%)	PR (%)	F_m	G_m
Jibei Station Data Set	91.93	90.01	0.86	92.78	0.9528	0.8489
Jiaozhou Station Data Set	89.86	87.25	0.82	93.27	0.9142	0.8417
Gaotang Station Data Set	90.26	89.36	0.89	91.54	0.9394	0.8623

datasets. It is indicated that the algorithm performance is similar in data sets based on different detection regions.

4. Conclusions

This paper proposes a random forest optimum model of traffic samples calibration using multisource features to separate outliers from real-time data. The model optimizes the training of random forests and decision-making process using bagging and boosting simultaneously based on nonequilibrium feature of outliers in the traffic data. According to the actual data need, induction from samples of the minority class is improved to train a random forest.

The optimized RF model has the following characteristics: (1) the advantage of the randomization process in the original RF algorithm remained; (2) the boosting method is introduced to strengthen the "induction" of the decision trees. The experimental results show that the optimized RF effectively separates outliers from traffic data by test of Shandong freeways samples. By the algorithmic verification, it compares index of classification accuracy, detection rate, false positive rate, precision rate, and ROC, which evaluates the detection method. Compared with the previous method 2, RFOM has advantageous properties such as high generalization performance and high accuracy. However, it can only measure nonequilibrium sample set of traffic data. So there are several restrictions concerning nonequilibrium feature of detection data sets. Further research will focus on the improvement of limitations.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

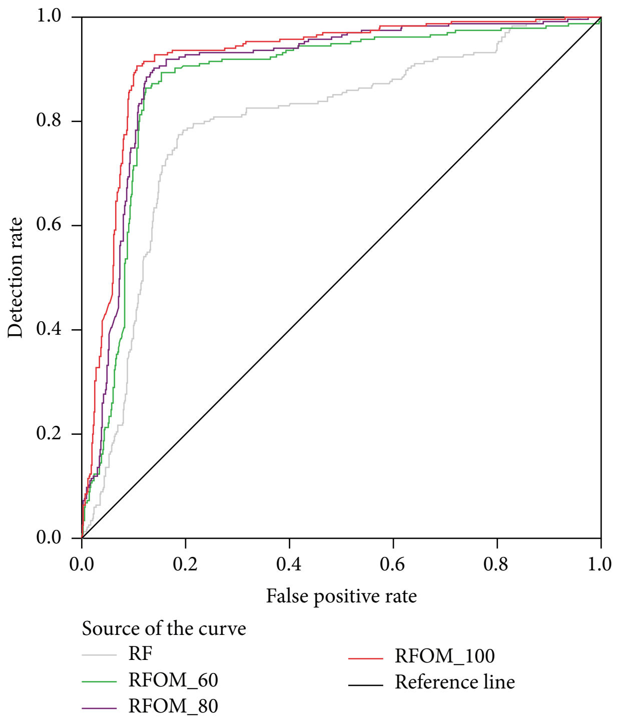

Figure 6: Comparison of ROC curves with Jibei Station Data Set.

Acknowledgment

This research has been jointly supported by National Key Technology Support Program (Grant no. 2014BAG03B03).

References

[1] B. R. Kadali, N. Rathi, and V. Perumal, “Evaluation of pedestrian mid-block road crossing behaviour using artificial neural network,” *Journal of Traffic and Transportation Engineering*, vol. 1, no. 2, pp. 111–119, 2014.

[2] A. M. Semeida, “Impact of highway geometry and posted speed on operating speed at multi-lane highways in Egypt,” *Journal of Advanced Research*, vol. 4, no. 6, pp. 515–523, 2013.

[3] S. C. Wong, N. N. Sze, B. P. Y. Loo, A. S. Y. Chow, H. K. Lo, and W. T. Hung, “Performance evaluations of the spiral-marking roundabouts in Hong Kong,” *Journal of Transportation Engineering*, vol. 138, no. 11, pp. 1377–1387, 2012.

[4] G.-Y. Jiang, A.-D. Chang, S.-F. Niu, Y.-L. Cong, D.-M. Cheng, and Q.-L. Wang, “Dynamic predictability analysis for traffic data serials based on BP neural network,” *Journal of Beijing University of Technology*, vol. 37, no. 7, pp. 1019–1026, 2011.

[5] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley & Sons, New York, NY, USA, 1994.

[6] D. H. Nam and D. R. Drew, “Traffic dynamics: method for estimating freeway travel times in real time from flow measurements,” *Journal of Transportation Engineering*, vol. 122, no. 3, pp. 185–191, 1996.

[7] L. Vanajakshi and L. R. Rilett, “Loop detector data diagnostics based on conservation-of-vehicles principle,” *Transportation Research Record*, vol. 1870, pp. 162–169, 2004.

[8] B. L. Smith, W. L. Scherer, and J. H. Conklin, “Exploring imputation techniques for missing data in transportation management systems,” *Transportation Research Record*, vol. 1836, pp. 132–142, 2003.

[9] A. M. Semeida, “Application of artificial neural networks for operating speed prediction at horizontal curves: a case study in Egypt,” *Journal of Modern Transportation*, vol. 22, no. 1, pp. 20–29, 2014.

[10] L. Xin, X. Xin, and D. Bin, “Vision-based long-distance lane perception and front vehicle location for full autonomous vehicles on highway roads,” *Journal of Central South University*, vol. 19, no. 5, pp. 1454–1465, 2012.

[11] Y. M. Yuan and W. Guan, “Offline handover location positioning for map matching of mobile probes,” *Journal of Central South University*, vol. 19, no. 7, pp. 2067–2072, 2012.

[12] J. Zhou, H. Chen, J. Zhao, H. Zeng, and X. Li, “Regional O-D survey method by vehicle license plate recognition technology,” in *CICTP 2012: Multimodal Transportation Systems Convenient, Safe, Cost-Effective, Efficient*, pp. 218–228, ASCE, 2012.

[13] G. M. Weiss, “Mining with rarity: a unifying framework,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 7–19, 2004.

[14] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[16] M. Kubat, R. C. Holte, and S. Matwin, “Machine learning for the detection of oil spills in satellite radar images,” *Machine Learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

[17] N. Poolsawad, L. Moore, C. Kambhampati, and J. G. F. Cleland, “Issues in the mining of heart failure datasets,” *International Journal of Automation and Computing*, vol. 11, no. 2, pp. 162–179, 2014.