ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 154

6th International Conference on Measurement, Instrumentation and Automation (ICMIA 2017)

Multi-source Traffic Data Calibration with Optimized Adaboost

Xue XING[1, a], Ciyun LIN[2,b] and Zhuorui WANG[2,c]

[1] College of Information and Control Engineering, Jilin Institute of Chemical Technology, 132022, Jilin, China

[2] College of Transportation, Jilin University, 130022, Jilin, China

[a]patricia_xx@126.com, [b]linciyun@jlu.edu.cn, [c]wenchangflower@hotmail.com

Keywords: data of measurement; AdaBoost ; outlier data detection; traffic data.

Abstract. A large amount of real-time traffic data supports the processing requirements of traffic state discrimination and prediction. Therefore, accurate real-time traffic information can be grasped for effective detection of outliers. In this paper, an optimized AdaBoost model for screening abnormal traffic samples is proposed based on the multi-source features of the detected data. Considering the unbalanced characteristics of traffic data, AdaBoost is optimized by cost-sensitive method, which avoids the problem that classification performance is degraded by non-equilibrium detection data. The accuracy, false alarm rate and false alarm rate of the model test are verified by the example of expressway test data set. The experimental results show that the AdaBoost model is 5.547% higher than the AdaBoost method in screening traffic samples. The algorithm can effectively adjust the classification error caused by unbalanced data.

Introduction

With the intelligent high-speed road network in the actual continuous improvement, a large number of traffic space-time data set for the basis of intelligent transportation coordination. There are some outliers in the process of obtaining the traffic-aware datasets [1,2], which is obviously inconsistent with other data. In order to protect the efficiency of traffic intelligent coordination, effectively stripping the outliers from the data set by multidimensional data features has become the basic problem of traffic information processing nowadays. The researchers proposed the method of ear-neighbor clustering and genetic algorithm [3], and some scholars put forward the method of evaluating the abrupt data in abnormal traffic flow [4] and the detector data evaluation based on rough set fuzzy recognition method[5] recently. The author of this paper also starts with the method of random forest [6] to give out the corresponding method to verify the outlier data. This paper proposes a new method based on AdaBoost optimization, which is an iterative classification algorithm with high accuracy and fast computation speed, to filter the outliers of the outliers, which is meaningful for traffic data detection.

Decision Tree Construction for Outlier Detection

In traffic detection data set, each testing point can get many sensory data composed of a variety of detection sources. Suppose there are n sources, each data source is through multiple traffic parameters to describe detected objects, and then each time all can get a set of multi-sensor data. The data collected from the detector of the road cross section is convenient to analyze the traffic flow parameters such as flow rate, velocity and occupancy rate, which commonly used in three kinds of data (induction coil data, geomagnetic data and bayonet data).

For instance, 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, volume q_{Ui} spot mean speed v_{Ui} and occupancy o_{Ui} from magnetic, 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,…,n in which L_i value belongs to the {1,1}, indicates that testing calibration set evaluation data information is normal data or outlier.

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

The formal description of matrices X and Y can be written as follows:

$$X=[x_1 \quad x_2 \quad \cdots \quad x_{10}]=\begin{bmatrix} t_1 & q_{C1} & v_{C1} & o_{C1} & q_{U1} & v_{U1} & o_{U1} & q_{T1} & v_{T1} & o_{T1} \\ t_2 & q_{C2} & v_{C2} & o_{C2} & q_{U2} & v_{U2} & o_{U2} & q_{T2} & v_{T2} & o_{T2} \\ \cdots & & \cdots & & \cdots & & \cdots & & \cdots & \\ t_n & q_{Cn} & v_{Cn} & o_{Cn} & q_{Un} & v_{Un} & o_{Un} & q_{Tn} & v_{Tn} & o_{Tn} \end{bmatrix} \quad (1)$$

where *X*i is a set of data elements, and *n is* the number of input samples;

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_n \end{bmatrix} \quad (2)$$

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

Traffic Data Detection Optimization Model

Characteristics of outliers in real - time traffic detection data

The data acquired by the road traffic detector include data attributes such as traffic data acquisition time, detector type, flow rate, spot mean speed, and time occupancy rate and so on.

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 multi-sensor 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 accounted for the largest number of samples is called the most classes, and accounts for the fewest category called the minority class (non-equilibrium data) [7]. The multi-source synchronization feature of traffic data is used to separate the outlier data from the multi-source traffic data without affecting the traffic anomaly data analysis, to effectively guarantee the traffic incident judgment and forecast the traffic state efficiency.

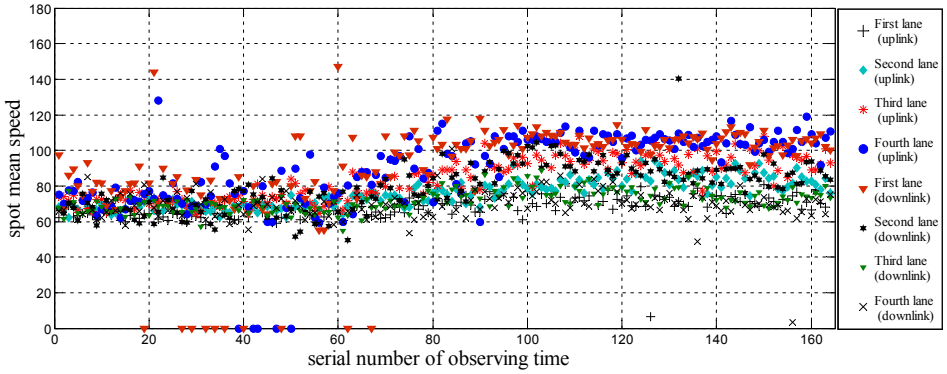

Figure 1: Road Detection Data Scatterplot of Single Parameter

An Optimized AdaBoost Model for Outlier Data in Traffic Data

The identification of non-equilibrium data in traffic data has practical significance, and the data scarcity and extreme values can lead to the performance degradation of AdaBoost classification method. In this paper, we propose to give more weight to a small class of samples in a weak decision-maker. In order to avoid the drawback that the decision tree rule is not representative due to the small amount of data in the original training set, the classifier is forced to pay more attention to the minority class samples. This method can improve the classification accuracy of a small number of samples, and can solve the problem of unbalanced data set classification.

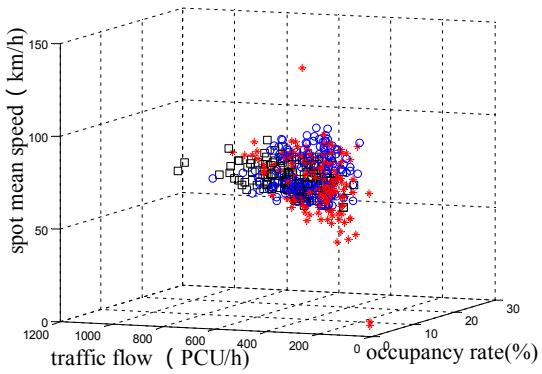

Figure 2: Scatterplot of Multi-sensor Parameters Data in the Same Section

The AdaBoost algorithm [8] does not have any direct dependency classification on the exponential error bound, and the literature [9-10] focuses on the class-conditional direct modification of the weight update rule. The model improves the sensitivity of classification to unbalanced data characteristics, as shown in (3) (4).

$$J(f) = E([y=1]e^{-C_P f(x_i)} + [y=-1]e^{-C_N f(x_i)}) \tag{3}$$

$$f(x) = \frac{1}{C_P + C_N} \log \frac{C_P P(y=1|x)}{C_N P(y=-1|x)} \tag{4}$$

Where C_P and C_N represent the Cost of Positive and Negative Error Classification

In order to clearly describe the optimal AdaBoost model, we give the number of individuals in the training set (X,Y), where y_i of each (x_i, y_i) of the training set is given by formula (5)

$$y_i = \begin{cases} 1 & 1 \le i \le m \\ -1 & m < i \le n \end{cases} \tag{5}$$

The AdaBoostOM description for traffic outlier data is described as follows (where F weak classifiers $h_t(x)$ with cost parameters C_P and C_N).

Step 1. For the samples on the original training set, give the initial distribution of each classification as

$$D(i) = \begin{cases} \dfrac{1}{2(n\text{-}m)} & 1 \le i \le m \\ \dfrac{1}{2m} & m < i \le n \end{cases} \tag{6}$$

Step 2. Initialize the number of rounds $t = 1$

Step 3. Calculate T_P and T_N， as formula（11）and（12）.

$$T_P = \sum_{i=1}^{m} D(i) \tag{7}$$

$$T_N = \sum_{i=m+1}^{n} D(i) \tag{8}$$

Step 4. Initialize the classifier variable $f = 1$.

Step 5. Calculate $D(i)$ in the fth weak classifier h $h_f(X)$.

$$D(i) = \begin{cases} \sum_{i=1}^{m} D(i) \left\| y_i \ne h_f(x_i) \right\| \\ \sum_{i=m+1}^{n} D(i) \left\| y_i \ne h_f(x_i) \right\| \end{cases} \tag{9}$$

Step 6. Calculate $\alpha_{t,f}$ in the equation (10) that satisfies the equation.

$$\begin{aligned} &2C_P B \cosh(C_P \alpha_{t,f}) + 2C_N D \cosh(C_N \alpha_{t,f}) \\ &= C_1 T_P e^{-C_P \alpha_{t,f}} + C_2 T_N e^{-C_N \alpha_{t,f}} \end{aligned} \tag{10}$$

Step 7. Calculate the loss of the weak learner

$$L_{t,f} = B(e^{C_P\alpha_{t,f}} - e^{-C_P\alpha_{t,f}}) + T_P e^{-C_P\alpha_{t,f}} + D(e^{C_N\alpha_{t,f}} - e^{-C_N\alpha_{t,f}}) + T_N e^{-C_N\alpha_{t,f}} \quad (11)$$

Step 8. Accumulate $f = f + 1$, if $f <= F$, repeat Step5.

Step 9. The minimum loss weak classifier $(h_t(X), \alpha_t(X))$, which is compared in this round, is $\arg\min_f [L_{t,f}]$.

Step 10. Update the $D(i)$ weight to $D(i) = \begin{cases} D(i)e^{-C_P\alpha_t h_t(X_i)} & 1 \le i \le m \\ D(i)e^{C_N\alpha_t h_t(X_i)} & m < i \le n \end{cases}$.

Step 11. Accumulate $t = t + 1$， if $t <= T$， repeat Step3.

Step 12. Determine the classifier as $H(x) = sign(f(x)) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

The performance indexes of classification accuracy, false positive rate and false negative rate are used to evaluate performance of algorithm classification. Classification accuracy, false positive rate and false negative rate are defined as follow:

$$Acc = \frac{CN + CG}{CN + CG + EN + EG} \quad (12)$$

$$FPR = \frac{EN}{EN + CG} \quad (13)$$

$$FNR = \frac{EG}{EG + CN} \quad (14)$$

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

In addition, the definition of the given probability cost function and the normalized expected cost is defined as shown in (15) and (16).

$$PCF = \frac{p(+)EG}{p(+)EG + p(-)CG} \quad (15)$$

$$NEC = CG * PCF + EG \quad (16)$$

where $p(+)$ and $p(-)$ are the prior probabilities of detection of outlier samples and detected general traffic samples.

Experimental results and analysis

Experimental data acquisition

In order to test the performance of the AdaBoostOM model, firstly, we compare the accuracy of the proposed model and the classical algorithm with the probability cost function (*PCF*), classification accuracy(*Acc*), false positive rate(*FPR*) , false negative rate (*FNR*)and indicator of normalized expected cost (*NEC*). The data of induction coil, geomagnetic data and bayonet processing data of 13 monitoring points on November 5, 2014 are selected to examine the data of the detector. The characteristics of the collected data set are shown in Table 1.

Table1 Properties Description of the Non-equilibrium Datasets

Data set	Data attribute	Sample size	Outlier size	Normal sample	Non-equilibrium rate
Jibei Station data set	Induction loop data1	4824	182	4642	3.921%
	Magnetic data1	6399	443	5956	7.452%
	Monitoring data1	10870	659	9619	6.851%
Gaotang Station data set	Induction loop data2	5013	198	4762	3.949%
	Magnetic data2	6512	531	5972	8.154%
	Monitoring data2	11194	625	10478	5.583%

In this paper, the test results of different methods on different highway data sets are analyzed according to the probability cost function (*PCF*), classification accuracy(*Acc*), false positive

rate(*FPR*) , false negative rate (*FNR*)and indicator of normalized expected cost (*NEC*). The experimental results of the Gaotang station data set are shown in Fig.3. The comparison of the detection indexes based on the highway test data set (Jibei station) and the detection index comparison chart of the highway detection data set (Gaotang station). Figure (a), (b), (c), (d) in Fig. 3 and Fig. 4 compare each detection index with the probability cost function (*PCF*) as abscissa.

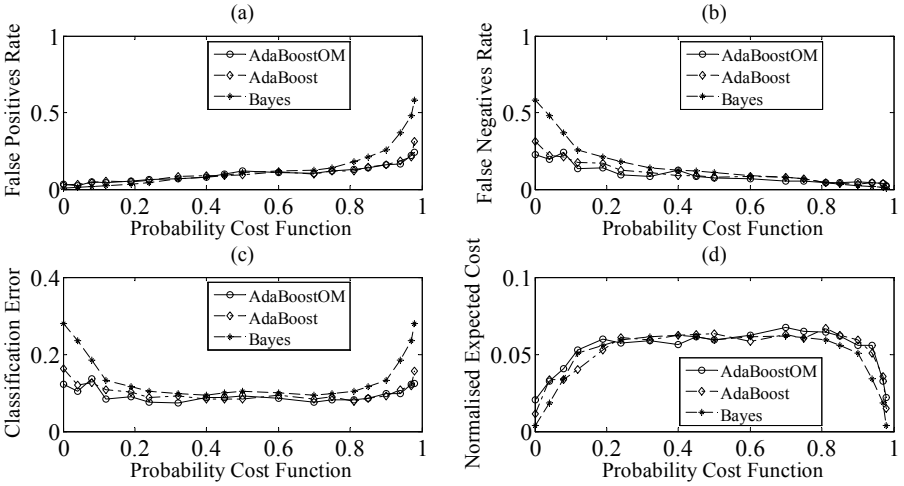

Fig.3 Index comparisons of Jibei Station data set. (a) FPR;(b) FNR;(c)1-Acc;(d)NEC.

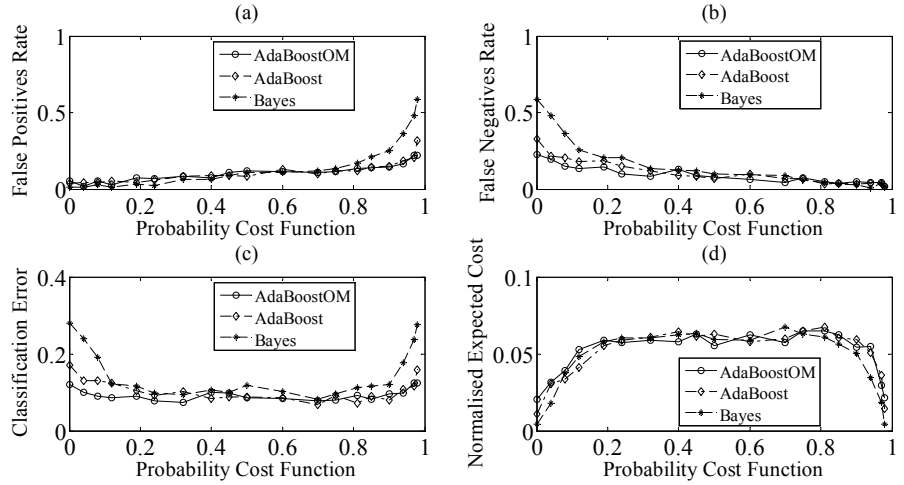

Fig.4 Index comparisons of Gaotang Station data set. (a) FPR;(b) FNR;(c)1-Acc;(d)NEC.

In this experiment, different training rules use to construct the decision rules. For the same data set, the effect of different algorithms is obvious. For the unreasonable data set synthesis, the same algorithm can continue the characteristics. The AdaBoost method and the AdaBoostOM method are close to and superior to the Bayesian method in the comparison of the Bayesian method, AdaBoost method and AdaBoostOM method in Fig.3. The three methods the difference is not significant in Fig. 3 (d). In Figure 4, the AdaBoost method and the AdaBoostOM method are superior to the AdaBoost method in Fig.4 (b) and (c), and are superior to the AdaBoost method. Bayesian method, while the three methods in Figure 4 (d), the difference are not significant. On the other hand, it can be found from the two data sets that AdaBoostOM method is better than AdaBoost method in the two indexes of $1-Acc$ and FPR , from the detection data set in Fig.3 to the detection data set in Fig.4 , The former than the latter average low 5.547% and 6.792%. The reason is that the proportion of the sample samples is not balanced, AdaBoost focuses on the unbalanced data characteristics, the misclassified outliers reduce the detection rate, while the AdaBoostOM algorithm reduces the false detection rate, which fully reflects the The cost parameter improves the detection accuracy.

Conclusions

In this paper, we propose an AdaBoostOM model with data analysis aiming at the problem of stripping datasets from non-equilibrium outliers in traffic detection data. After analyzing the non-equilibrium characteristics of the outlier samples, the advantage of the cost-sensitive method combine to optimize the AdaBoost decision-making process and avoid the problem of classification performance degradation caused by the non-equilibrium detection data. The model indexes of Acc , FPR , FNR and NEC were validated by the highway traffic inspection data set. Experimental results show that the improved AdaBoost filter can provide a more reliable sorting result and can effectively adjust the classification error caused by unbalanced data.

AdaBoostOM has the following two characteristics, the first one is to retain the original weighting advantage, and the second one is the introduction of cost-sensitive methods to strengthen the non-equilibrium characteristics. By comparing the performance of the algorithm with the algorithm of the traffic dataset, the AdaBoostOM algorithm is proposed to reduce the test error rate in the outlier detection of the traffic detection dataset. However, this algorithm is based on the unbalanced traffic data sample set, so there are certain limitations on the test data set. Further research will focus on improving the limitations of the method.

Acknowledgements

This research has been jointly supported by National Natural Science Foundation of China（Grant No. 51308248）.

References

[1] Barnett V, Lewis T. Outliers in Statistical Data[M]. NewYork: John Wiley&Sons, 1994:95-98.

[2] Payne HJ, Helfenbein E D, Knobel H C.Development and testing of incident detection algorithms[R]. Research methodology and detailed results , 1976.

[3] Pu Shilin, Li Ruimin,Shi Qixin, Study on Auto-Identification Algorithm of Traffic Flow State Based on Rough Set and Fuzzy Theory[J]. Journal of Wuhan University of Technology (Transportation Science& Engineering) ,2010, 34(6):154-158.

[4] Zhou J,Chen H,Zhao J,Zeng H,Li X. Regional O-D Survey Method by Vehicle License Plate Recognition Technology[J]. Multimodal Transportation Systems-Convenient, Safe, Cost-Effective, Efficient,2012: 218-228.

[5] Weiss G M, Mining with Rarity: A Unifying Frameworks[J]. SIGKDD Explorations, 2004, 6(1):7-19.

[6] Xue X,Dexin Y,Wei Z. Data Calibration Based on Multisensor Using Classification Analysis: A Random Forests Approach[J].Mathematical Problems in Engineering,Volume 2015.http://dx.doi.org/10.1155/2015/708467

[7] Karakoulas G,Shawe-Taylor J. Optimizing classifiers for imbalanced training sets[J].Advancesin Neural Information Processing Systems, 1999,12:253–259.

[8] Fan W, Stolfo W,Zhang J,Chan P. AdaCost :misclassification cost-sensitive boosting[C].Proceedings of the 16th International Conferenceon Machine Learning, 1999:97–105.

[9] Ting K. Acomparative study of cost-sensitive boosting algorithms[C]. Proceedings of the17th International Conferenceon Machine Learning,2000:983–990.

[10] H.Masnadi-Shirazi,N.Vasconcelos.Cost-sensitive boosting[J].IEEE Trans.Pattern Anal. Mach. Intell. 2011,33:294–309.