Accident Analysis and Prevention 144 (2020) 105610

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Examining imbalanced classification algorithms in predicting real-time traffic crash risk

Yichuan Peng[a], Chongyi Li[a], Ke Wang[a], Zhen Gao[b,*], Rongjie Yu[a]

[a] Tongji University, School of Transportation Engineering, China
[b] Tongji University, School of Software Engineering, China

ARTICLE INFO

Keywords:
Continuous data environment
Real-time crash risk prediction models
Imbalanced data classification
RCSMLP
Rusboost model

ABSTRACT

The Active Traffic Management (ATM) system has been widely used in the United States and the European countries to improve the traffic safety of urban expressways. The accurate real-time crash risk prediction is fundamental to the system running well. Crash data are characterized by small probability, which poses a typical Imbalanced Data Classification problem. Most previous studies mainly improved the prediction methods only in data level or algorithm level, which may be inadequate to predict the crash risk accurately especially in a continuous real-time traffic data environment. The comprehensive imbalanced classification algorithm was examined in this research to build more accurate real-time traffic crash risk prediction model. At the output level, the Youden index method has been proved to be of the best ability to divide the prediction results and Probability Calibration Method was proposed to optimize the prediction results in further. At the data level, Under-sampling and Synthetic Minority Oversampling Technique(SMOTE) methods were compared to solve the imbalanced data classification problem by changing the data distribution. At the algorithm level, the cost-sensitive MLP algorithm and Adaboost algorithm were examined and finally the random sampling cost-sensitive MLP model(RCSMLP) and Rusboost model were constructed by synthesizing the optimization methods from three levels. The sensitivity of the RCSMLP model reached 78.10 % and the specificity of the model reached 81.44 %. The AUC and sensitivity of the Rusboost model reached 0.892 and 0.842 while the specificity of the model reached 0.816, which shows the better performance in dealing with the imbalanced traffic crash risk prediction problem compared to existed prediction models. The proposed method of improving prediction accuracy in this study is universal and can be applied to many other prediction models to predict real-time traffic crash risk.

1. Introduction

Urban expressway is the main artery of urban transportation system, which is characterized by large capacity and high speed limit. In order to improve the traffic safety of urban expressway, Active Traffic Management (ATM) system is adopted by many countries to manage urban expressway (Kurzhanskiy and Varaiya, 2010; Tignor et al., 1999). This kind of system has the characteristics of real-time and active control instead of previous passive control (Hu-pu and Ri-min, 2014) The active traffic management system can effectively predict the real-time traffic crash risk on urban expressway once an accurate real-time traffic crash risk prediction model is built. It can predict when to take intervention measures such as variable speed limit to reduce or even eliminate the crash risk. Sufficient traffic data and reasonable algorithms are fundamental to build the model.

With the development of intelligent transportation technology, the methods of collecting traffic data become more and more diverse. At present, Shanghai has a lot of advanced equipment to collect real-time traffic data collection such as loop detector, video detector and floating car, which have been deployed in the urban expressway network of Shanghai to obtain real-time traffic information. These intelligent transportation technologies provide massive kinds of data including traffic crash data and real-time traffic data about operation status, which provide valuable data basis for creating accurate crash risk prediction model. In addition, crashes have the characteristics of small probability. The data of traffic operation state under non-crash state is obviously more than that under crash state. The monthly sample size of normal traffic status data (non-crash data) is often 6000–7000 times of

* Corresponding author.
E-mail addresses: yichuanpeng1982@hotmail.com (Y. Peng), 1731304@tongji.edu.cn (C. Li), kew@tongji.edu.cn (K. Wang), gaozhen@tongji.edu.cn (Z. Gao), yurongjie@tongji.edu.cn (R. Yu).

https://doi.org/10.1016/j.aap.2020.105610
Received 5 May 2019; Received in revised form 15 March 2020; Accepted 22 May 2020
Available online 16 June 2020
0001-4575/

the crash data. Therefore, the problem of creating real-time traffic crash risk prediction model is a typical imbalanced data classification problem, which means the number of the majority class samples is far greater than that of the minority class samples. The application of traditional classification algorithms often results in a bias of the prediction results towards the majority of classes namely non-crashes. However, the minority classes tend to have a higher classification error cost in the prediction of traffic crashes so that the traditional classification algorithm will lead to a great loss. Most of existing real-time traffic crash prediction models try to solve this imbalanced data classification problem mainly on data level such as adopting under-sampling method named "case-control"(Abdel-Aty et al., 2004, 2012). Although the performance of modeling results using this method is excellent for the discrete data, the performance of this kind of model for predicting continuous real-time traffic data is not ideal.

Considering the imbalanced and continuous characteristics of collected real-time traffic data and the limitation of previous traffic crash risk prediction models, this study tries to build a new kind of real-time traffic crash prediction model to solve the problem of imbalanced data. The imbalanced classification algorithm will be optimized from three aspects including output level, data level and algorithm level. The Youden index method is used to divide the prediction results of the models and Probability Calibration Method is proposed. Under-sampling (random sampling, "Matched Case-control" under-sampling) and SMOTE methods will be compared by changing the data distribution. After that, a cost-sensitive MLP model and a Rusboost model will be constructed. The final constructed two models can be a potentially better option for real-time traffic crash risk prediction and will be beneficial for traffic management to conduct more effective countermeasures to prevent crashes.

2. Literature review

2.1. Real-time traffic crash risk prediction model

Hughes et al(Hughes and Council, 1999)first used loop detector data to discuss the relationship between traffic operation state and crash risk. They proposed that dynamic traffic operation state such as the change of real-time speed is more suitable for traffic crash risk prediction than static traffic characteristics. Oh et al.(Oh et al., 2001)also used loop detector data to construct a real-time traffic crash risk model. They indicated that the speed change in 5 min before the crash could be used to predict the crash risk effectively. Lee et al.(Lee et al., 2003) found that the speed difference between the investigated road segment and the upstream and downstream segment has a great influence on the real-time traffic crash risk.

As for the selection of significant variables included in the prediction model, existing studies have shown that the average speed, traffic flow rate, standard deviation of speed and standard deviation of traffic flow are important factors affecting the crash risk. The traffic conditions of the upstream and downstream of the locations where traffic crashes occurred also have a great impact on the crash risk(Abdel-Aty et al., 2004, 2012; Pande and Abdel-Aty, 2006; Xu et al., 2013). Ahmed et al. (Ahmed and Abdel-Aty, 2012) constructed the real-time traffic crash risk prediction model by aggregating the traffic status data in 2 min, 3 min, 5 min and 10 min, respectively. The results showed that the optimal traffic crash risk prediction model could be obtained by aggregating the data in 5 min.

Abdel - Aty (Abdel-Aty et al., 2004) proposed that the data in five minutes before the crash is too short for real-time traffic management. They analyzed the data in 30 min before the crash and the data was divided into six 5-minute time slice. The traffic parameters including mean speed, traffic volume, occupancy and speed variation have been computed and analyzed. They proposed the use of Case-Control method by selecting traffic status in a crash as a case and corresponding non-crash traffic status as a control. The crash and non-crash ratio 1:1, 1:2 and 1:3, 1:4, 1: 5 has been used to construct traffic crash risk prediction model respectively and the results showed that the ratio of crashes and non-crashes has little influence on the prediction results. Since then, relevant studies have basically adopted the "case control" sampling method to prepare data for prediction models. Different studies have used different crash and non-crash proportions: 1:4(Xie et al., 2016; Xu et al., 2012; Yu and Abdel-Aty, 2013), 1:5(Abdel-Aty et al., 2005, 2004, 2012; Sun and Sun, 2015), 1:10(Pande and Abdel-Aty, 2006; Xu et al., 2013), 1:20(Abdel-Aty and Pande, 2005; Pande and Abdel-Aty, 2006), etc. to select the needed data from the whole sample. This case-control method is based on "case-control" sampling data to create prediction models. However, only a very small fraction of all the non-crash related traffic data has been used in the prediction models for the case-control method. A lot of non-crash related traffic state information could not be included in the model, which leads to the performance of the prediction model is not ideal especially under the continuous traffic data condition.

Most real-time traffic crash risk prediction models are constructed based on statistical regression or machine learning algorithms. Common statistical models include "case control" logistic regression model(Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2012), Bayesian logistic regression model(Hossain and Muromachi, 2012; Yu and Abdel-Aty, 2013) etc. Machine learning models include neural network(Abdel-Aty and Pande, 2005; Pande and Abdel-Aty, 2006), random forest(Pham et al., 2010; Katrakazas et al., 2019), support vector machine(Yu and Abdel-Aty, 2013), convolutional neural network(Baheti et al., 2018),boosting trees(Kocsis et al., 2013) and genetic algorithm(Xu et al., 2013). These methods can effectively solve multiclass problem. However, the problems of handling prediction of imbalanced datasets still exist in these models. A threshold value is needed to classify the division for these prediction models for the probability of crashes. The traditional binary classification problems usually choose "0.5″ as the classification threshold, but it is not suitable for the real-time traffic crash risk prediction. The real-time traffic crash risk prediction is a typical imbalanced data classification problem. The non-crash cases are far greater than crash cases. For this typical imbalanced data classification problem, "0.5″ will make almost all prediction results are classified as a non-crash case. Traditional classification algorithms tend to consider the overall performance of the prediction model. All minority samples crash cases can be wrongly divided into majority class non-crash cases but the prediction model still remains high prediction accuracy in this traditional way. Therefore, it is necessary to find a better way to determine the classification threshold. Some studies employed the cross point method that selects the cross point of Sensitivity curve and Specificity curve to determine the classification threshold (Abdel-Aty et al., 2005).

2.2. Imbalanced data classification algorithm

The imbalanced data classification means that the number of samples in one class of the dataset is far greater than that of other classes, among which the samples that account for the majority are called majority classes, while the samples that account for only a small part are considered as minority classes or rare classes. Traditional classification methods such as logistic regression algorithm only consider the optimal overall accuracy but do not consider the distribution among different classes. Classification algorithms often assume the same data distribution for different classes, which results in the preference to majority classes. However, the minority classes tend to have a higher classification error cost in the imbalanced data classification problem so that the traditional classification algorithm will lead to a great loss. There are mainly three kinds of methods to solve the imbalanced data classification problem. The classification threshold needs to be adjusted to make the classification results more partial to the minority categories at the output level. The data distribution can be changed to make the data more balanced on the data level. The traditional classification

algorithm needs to be revised to make the algorithm more biased to minority classes at the algorithm level.

At the output level, the prediction results are biased towards the minority classes by adjusting the classification threshold without changing the algorithm(Maloof, 2003; Schlögl, et al., 2007). Krawczyk et al.(Krawczyk and Woźniak, 2015), Sun et al.(Sun et al., 2007)used cost matrix to determine the classification threshold. Zhou et al.(Zhou and Liu, 2006) pointed out that the Moving Threshold method can effectively solve the problem of imbalanced data classification. Furthermore, the rebalancing of classes by sampling method always yields a distortion in the predicted probabilities. Alexandru and Caruana (2005) proposed two calibration ways of correcting the biased probabilities :Platt Scaling and Isotonic Regression. Pozzolo et al. (2015) applied Bayes Minimum Risk theory to adjust the probabilities after undersampling.

In terms of data level, sampling method is often used to change the sample distribution. There are mainly two kinds of methods. First, oversampling method is used to increase the number of the minority class samples. Second, under-sampling method is used to reduce the number of majority class samples. Sampling methods include Randomly Oversampling, Randomly Under-sampling, Informatively Sampling and SMOTE. The shortcoming of under-sampling method is that it will lose a lot of useful information. Random oversampling method is to repeat a number of data in minority classes, which will lead to the increase of difficulty in calculation and even the occurrence of overfitting (Longadge and Dongre, 2013). Therefore, Chawla et al.(Chawla et al., 2002) proposed SMOTE algorithm which uses k proximity algorithm and linear interpolation to generate new samples for the minority classes. Cost sensitive methods are also commonly used in the data level. They make the data distribution tend to be more balanced by adding different weights to the data according to the different costs of different categories of data(Sun et al., 2007).

At the algorithm level, the integrated algorithm is a way to solve the problem of imbalanced data classification. The integrated algorithm integrates multiple weak classifiers into a strong classifier through different election methods to improve the prediction accuracy of the model for the majority classes. The selection method of Ensemble method is divided into two categories(Zhou and Liu, 2006; Theofilatos et al., 2019), namely soft-ensemble and hard-ensemble. The Cost – Sensitive method is often used to improve traditional algorithm(Sun et al., 2007; Zhou and Liu, 2006). This method is created on the basis of the traditional classification algorithm by joining the Cost matrix, which makes the minimum misclassification cost for the whole dataset. Then, the algorithm is biased towards the minority class. In addition, some scholars have proposed the method “Rusboost”(Seiffert et al., 2010), which combines Random undersampling method and Boost algorithm. This method undersamples the majority class data in each boost and works well with imbalanced data classification problem. This method has not been examined in the field of traffic crash risk prediction to the best knowledge of authors.

Although most current imbalanced data classification algorithms claim to achieve good classification performance, the advantages and disadvantages of different imbalanced data classification algorithms are highly dependent on the characteristics and distribution of original dataset. Therefore, the imbalanced data classification algorithms should be improved according to the specific continuous traffic data environment. In addition, most of previous studies optimized the prediction accuracy of traffic crash risk only from one aspect such as data level. None of previous researches have examined the imbalanced classification problem from all the above mentioned three aspects. It is necessary to conduct a new method to optimize the predictive modeling comprehensively for the imbalanced and continuous real-time traffic data.

3. Data collection and preparation

The Shanghai expressway system mainly consists of six urban expressways: Inner ring expressway, central expressway, Yanan expressway, Humin expressway, North-south expressway and Yixian expressway. The six urban expressways were further divided into 237 road segments according to the distribution of ramp. The average length of the segment was 949.2 m. Each segment included one or more groups of loop detectors to collect traffic operation data. The data collection interval was 20 s. Some important traffic variables like traffic speed, traffic volume and occupancy have been collected. There will be some anomalies in the collected data including data loss and abnormal standard deviation of variables due to equipment failure. Therefore, it is necessary to preprocess the original data including the deletion of repeated, unknown and invalid abnormal data.

Previous studies shown that mean speed, traffic flow, standard deviation of speed and traffic flow are important factors on traffic crash risk (Yu *et al.*, 2018; Yu *et al.*, 2020). Therefore, this study uses relevant indicators of speed and traffic flow as independent variables. To describe the traffic flow at a certain time, the original traffic flow data set was counted to 5 min. The traffic data of the 20 min before this time point were extracted and divided into 4 time slices to describe the traffic status of certain time. These four time slices were named time slice 1 – 4. Time slice 1 was the closest to the current time. The mean value, standard deviation and sum of the traffic speed and traffic flow of the current road segment as well as those of the upstream and downstream segments within each time segment were calculated. A total of 72 variables were included in the final traffic dataset and the structure of variable naming was shown in Fig. 1.

The total sample size of traffic data on Shanghai expressway in May 2014 was 7559,352, including 1210 traffic crashes and 7,558,142 non-crashes. The sample size of traffic data in June 2014 was 7870,211, including 1128 traffic crashes and 7,869,083 non-crashes. The statistics of the full sample size are shown in Table 1. The full sample in May 2014 was used for creating prediction models, while the data of June 2014 was used for testing and verifying the performance of models. It can be shown from the Table1 that the prediction of traffic crash risk based on continuous traffic variables is a typical problem related to imbalanced data.

The summary statistics for all the dependent and significant independent variables included in this study have been shown in the Table 2.

4. Methodology

This study mainly examined the following methods to optimize the prediction results of traffic crash risk from output level, data level and algorithm level considering the imbalanced and continuous characteristic of traffic datasets.

4.1. Youden index method

Adjustment of classification threshold is a good way to solve imbalanced data problem at the output level, which means only adjusting

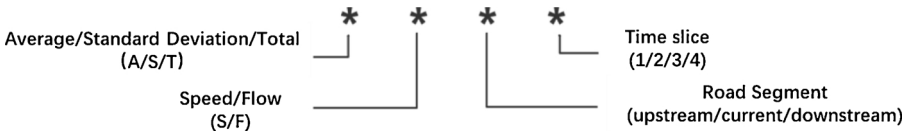

Fig. 1. Structure of Variable Naming.

Table 1
Summary Statistics for Samples.

Time	Number of crash cases	Number of non-crash cases	Proportion of crashes
May 2014	1210	7,558,142	0.016%
June 2014	1128	7,869,083	0.014%

the classification threshold while keeping the sample data and modeling algorithm unchanged. Existing studies(Abdel-Aty et al., 2005; Sun and Sun, 2015) adjusted the classification threshold by using crash ratio method or intersection point method which use the intersection point of Sensitivity curve and Specificity curve as to select the threshold.

Based on the Liu's research (Liu, 2012), Youden Index of ROC curve was used as the criteria to select the threshold value, which was called the Youden Index method. The Youden Index is defined as the distance between the points on the ROC curve and the diagonal crossing point (0,0) and (1,1), as shown in Fig. 2. The calculation formula of Youden Index is shown in Eq.1.

$$Youden = Sensitivity(n) + Specificity(n) - 1 \tag{1}$$

Where n represents the set of all points in the ROC curve, the sensitivity and specificity of each point in the ROC curve are calculated and the Youden index of each point is then calculated. The threshold corresponding to the greatest Youden index in the training data is selected as the classification threshold of the model

4.2. Probability calibration method

On the output level, the Probability Calibration method is also proposed for predicting posterior probabilities in addition to Youden Index method. The Platt Scaling calibration method (Hereinafter referred to as the calibration method 1 with Sigmoid) is proposed to get calibrated probabilities through a Sigmoid :

$$\mathrm{P(y{=}1f)} = \frac{1}{1 + \exp(Af + B)} \tag{2}$$

Where the parameters A and B are fitted using the gradient descent from a fitting training dataset (f_i, y_i):

$$\underset{A,B}{argmin}\{-\sum_i y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\}$$

$$p_i = \frac{1}{1 + \exp(Af_i + B)} \tag{3}$$

In order to satisfy the use of different algorithms, the basic assumption of Isotonic calibration problem (Hereinafter referred to as the calibration_method 2 with Isotonic) was proposed such that

$$y_i = \mathrm{m}(f_i) + \in_i \tag{4}$$

Where f_i is the prediction from a model and y_i is the true targets, and m

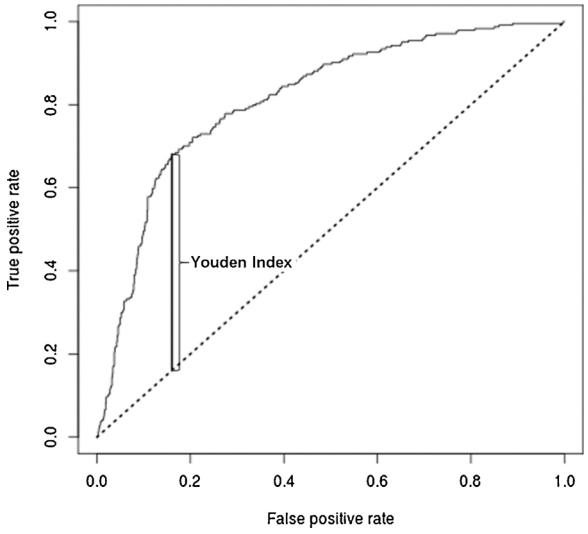

Fig. 2. Relationship between Youden Index and ROC.

is a monotonic increasing function, which is named Isotonic. Then, the target is using a fitting training dataset (f_i, y_i) to find the isotonic function such that:

$$m^{'} = \underset{z}{argmin}\sum (y_i - z(y_i))^2 \tag{5}$$

And in the Pozzolo's research, the calibrated probabilities equation (Hereinafter referred to as calibration method 3 with Bayes Minimum Risk) was designed for the sampling model such that :

$$p_s = \frac{p}{p + \beta(1 - p)} \tag{6}$$

Where p is the probabilities of selecting a negative instance with undersampling, and β denotes the proportion of the positive and negative instances.

4.3. Under-sampling and over-sampling method

The case control under-Sampling method and random under-sampling methods are applied and compared in this study. The "case control" under-sampling method is applied to extract the whole sample data of May 2014, where "case" is the traffic flow condition before the traffic crash and "control" is the normal traffic flow condition without traffic crash. Non-crash control data corresponding to case data were extracted and 5 conditions were considered in the selection of the control group: (1) the date of the control group was different from the date of the corresponding case group (2) the time is the same as the case (3) the place is the same (4) the crash occurred on the same day of the week; (5) there are not any crashes within 1 h for the selected control group on the same location. Random under-sampling method is a

Table 2
Summary Statistics for Dependent and Independent Variables.

Variable	Definition	Mean	Std. dev.	Min	Max
Crashes	Number of Crashes	0.0016	0.0016	0	1
ASC2	Average speed for current segment during the 2nd time slice	51.09	20.73	0	97.58
AFC2	Average volume for current segment during the 2nd time slice	42.43	30.14	0	293.6
SSC2	Std. Dev. of speed for current segment during the 2nd time slice	5.084	3.96	0	53.03
SFC2	Std. Dev. of flow for current segment during the 2nd time slice	5.525	3.68	0	91.92
SSU2	Std. Dev. of speed for upstream segment during the 2nd time slice	5.126	3.99	0	71.42
SFU2	Std. Dev. of flow for upstream segment during the 2nd time slice	5.539	3.69	0	91.92
TFU2	Total volume for upstream segment during the 2nd time slice	671.1	482.71	0	4697.7
SSD2	Std. Dev. of speed for downstream segment during the 2nd time slice	5.018	3.86	0	53.03
SFD2	Std. Dev. of flow for downstream segment during the 2nd time slice	5.386	3.60	0	91.92
TFD2	Total flow for downstream segment during the 2nd time slice	674.3	482.79	0	4697.7

simple under-sampling method by adjusting original dataset, aiming to balance data distribution by reducing the sample size of the majority class randomly. This sampling method is applied in this study to extract non-crash data in the whole sample data according to a certain proportion of crashes and non-crashes, so as to reduce the number of non-crash samples and make the data distribution more balanced.

In addition to using under-sampling method to reduce the samples of the majority class, the over-sampling method can also be considered by keeping samples of the majority class and generating a certain proportion of the minority class samples to change the data distribution. Commonly used sampling methods include Random Over-sampling and SMOTE(Chawla et al., 2002).

4.4. Rusboost algorithm

The Rusboost algorithm is proposed by combining Random under-sampling and the Adaboost.M2 algorithm. In each iteration of Adaboost, Random under-sampling method is introduced to balance the dataset gradually. Instead of removing valuable information of majority samples, Rusboost creates diversed training data through each iteration's random under-sampling, which make the algorithm more generalized. The algorithm procedure is shown in the following Table3.

5. Comparison of modeling results on output level

For the purpose of optimizing modeling results on output level, the section first compared the performance of traditional logistic regression model, the random forest model and MLP model to predict the real-time traffic crash risk by using full sample data in May 2014. The full sample data in June 2014 was used as data for validation. Three different threshold classification methods were examined to compare the performance of prediction results. The sensitivity, specificity and geometric mean of the prediction results were calculated. The prediction results are shown in Table4.

It can be seen from the Table4 that the sensitivity of the logistic regression model using full sample is the highest when the crash ratio method was used to determine the classification threshold. The specificity is the highest when the intersection method was used. The classification threshold determined by Youden index is the best to predict the crash risk by comparing the geometric mean. The calculated intersection points and Youden index thresholds are exactly the same for the MLP model using full sample. The classification threshold determined by Youden index shows the optimal prediction results for the MLP model. As for the random forest model, the sensitivity can reach 81.29 % and the false alarm rate can be controlled at 26.54 % by using the Youden index to calculate the classification threshold. The classification threshold determined by Youden index is also the best to predict the crash risk by comparing the geometric mean. In summary, the classification threshold calculated by the Youden index method can obtain better comprehensive prediction accuracy than the other two threshold calculation methods in the continuous traffic data environment. Therefore, The Youden index method was used to determine the classification threshold to classify crashes and non-crashes in the following analysis.

In addition to Youden index method, the abovementioned three different kinds of calibration methods were examined for traditional logistic regression model, the random forest model, MLP model and Rusboost model. The prediction results were compared with that without using any calibration methods, which are shown in the following Figs. 3–6.

The prediction results show that using calibration methods can optimize the predicted probabilities closer to the real probabilities. The reliability diagram is in the top row and the histogram of the predicted values is in the bottom row of each figure. It can be seen from the Fig. 3 that the predicted probabilities of the models without any calibration methods can not represent the true probabilities well and the reliability curve lies below the perfectly calibrated line. The histogram in Fig. 3 showed that partial predicted probabilities lies in the central region, especially for the Rusboost model. In the Figs. 4–6, the predicted probabilities of all the models are optimized closer to perfectly calibrated line in general and it can be seen from histogram that more predicted probabilities lies near 0. This is beneficial to the traffic crash risk prediction in the real world. For example, when using the Bayes Minimum Risk calibration method, we can get a predicted probabilities more closer to the true probabilities in fixed region and then take some more effective prevention countermeasures. The difference of calibration methods can also be shown in the length of reliability curve. The longer the curve, the more diverse the probabilities are. It can be seen from the Fig. 6 that models with the Bayes Minimum Risk calibration method will reserve more diversity of predicted probabilities.

Table 3
Procedure of Rusboost Algorithm.

1、Initialize $w_1(i) = 1/n$, i = 1,2,…,n

2、Do for m = 1,2,…,M:

2.1、 Create temporary training dataset s'_m and its distribution w'_m with the Random Under-sampling method. And then training $T^{(m)}(x)$ *with* s'_m *and* w'_m

2.2、Get back a hypothesis h_y: X × Y → [0,1]

2.3、Calculate the pseudo-loss (for S and S_m):

$\epsilon_t = \sum_{(i,y):y_i \neq y} w_m(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$

2.4、Calculate the weight update parameter:

$\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}$

2.5、Update w_m

$w_{m+1}(i) = w_m(i)\,\alpha_t^{\frac{1}{2}(1+(x_i,y_i)-h_t(x_i,y:y\neq y_i))}$

2.6、Normalize $w_{m+1}(i)$: Let $Z_t = \sum_i w_{m+1}(i)$

$w_{m+1}(i) = \frac{w_{m+1}(i)}{Z_t}$

3、Output the final hypothesis:

$C(x) = \arg\max \sum_{m=1}^{M} h_k^{(m)}(x, y) * log\frac{1}{\alpha_t}$

Note: Set Training data S of Samples (x_i, y_i), i ∈ n , $y_i \in Y$; Number of iteration, M ; Number of label, $|\ Y\ | = 2$; The Weak learner used in No.m iteration, $T^{(m)}(x)$.

Table 4
(a) Comparison of Prediction Results for Logistic Regression Model.

Threshold Classification Method- >	Youden Index	Crash Ratio	Intersection Point
Classification Threshold	0.0001625	0.00016	0.000165
Sensitivity	77.04 %	77.22 %	76.51 %
Specificity	76.21 %	75.91 %	76.51 %
Geometric Mean	0.7794	0.7656	0.7651

(b) Comparison of Prediction Results for Random Forest Model.

Threshold Classification Method- >	Youden Index	Crash Ratio	Intersection Point
Classification Threshold	0.037	0.00016	0.0406
Sensitivity	81.29 %	100 %	76.33 %
Specificity	73.46 %	0.5 %	76.93 %
Geometric Mean	0.7728	0.0224	0.7662

(c) Comparison of Prediction Results for MLP Model.

Threshold Classification Method- >	Youden Index	Crash Ratio	Intersection Point
Classification Threshold	0.0002643	0.00016	0.0002643
Sensitivity	73.40 %	81.56 %	73.40 %
Specificity	83.87 %	76.77 %	83.87 %
Geometric Mean	0.7846	0.7313	0.7846

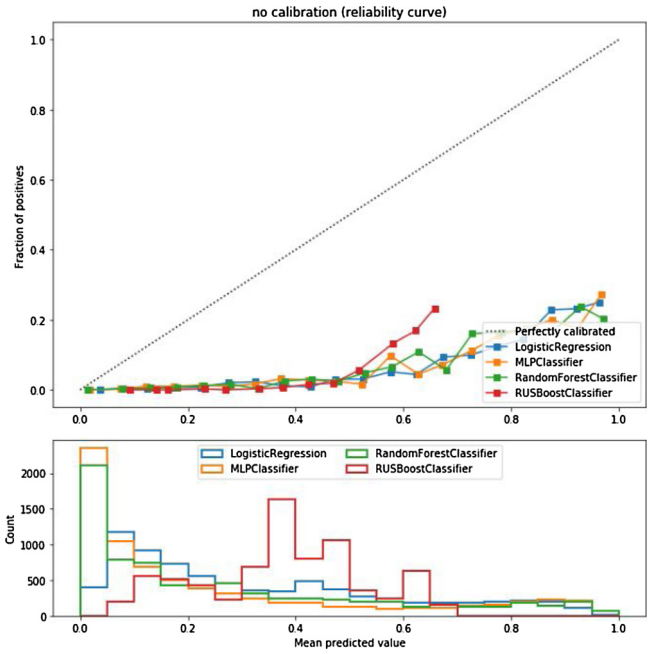

Fig. 3. Reliability Curve without Calibration.

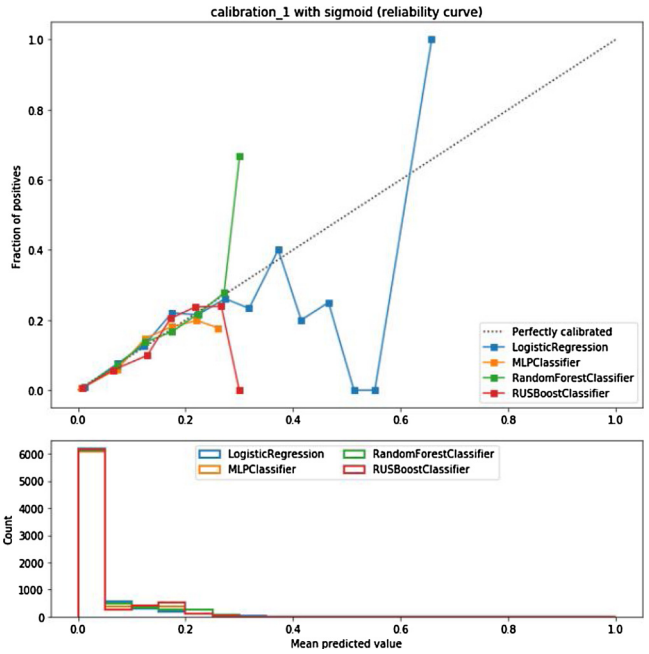

Fig. 5. Reliability Curve with Calibration Method 2 (Sigmoid).

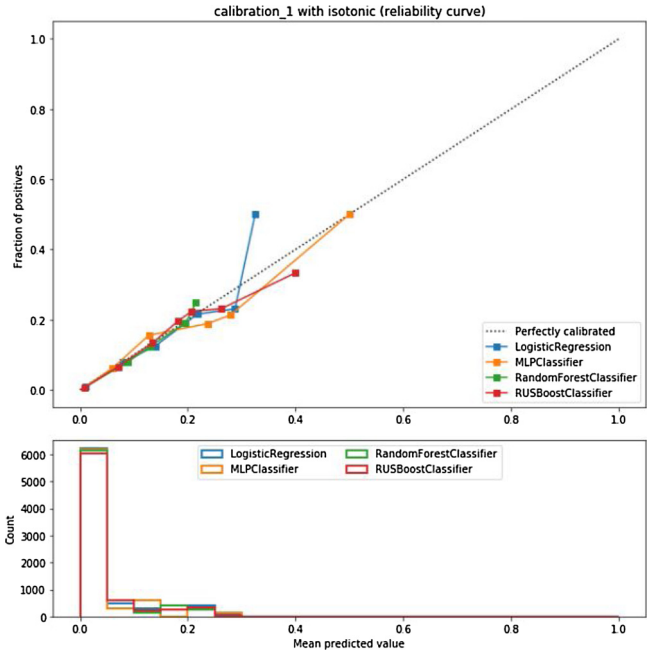

Fig. 4. Reliability Curve with Calibration Method 1 (Isotonic).

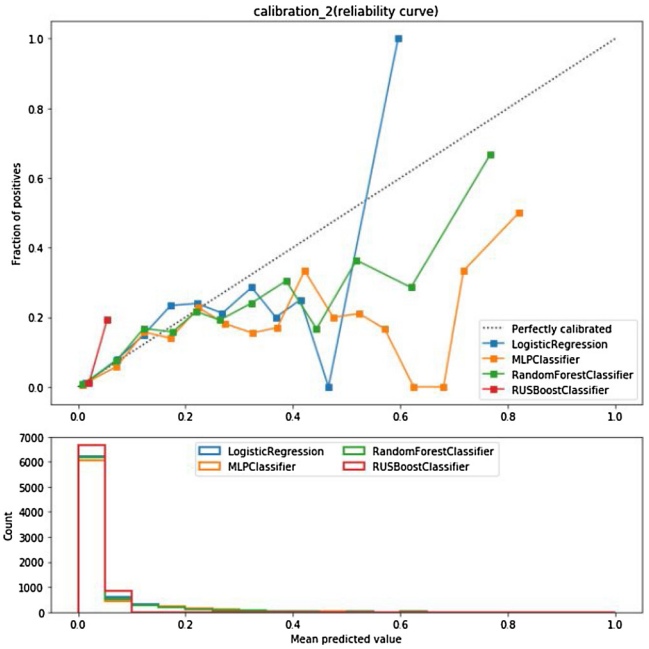

Fig. 6. Reliability Curve with Calibration Method 3 (Bayes Minimum Risk).

5.1. Comparison of modeling results on data level

As has been mentioned in Section 2.2, sampling method is often used to change the sample distribution for the purpose of solving imbalanced data problem on data level. There are mainly two kinds of sampling methods. Over-sampling method is used to increase the number of the minority class samples while under-sampling method is used to reduce the number of majority class samples. In this study, the full sample data of May 2014 were reconstructed and processed to solve the problem of imbalanced data classification from the data level. The under-sampled "case control" data, random sampling data and SMOTE over-sampled data were constructed respectively and compared with the full sample data. Logistic regression model is applied to compare the performance. The full sample data of June 2014 is used to testify the modeling results. The classification threshold is calculated by using the Youden index method, and the AUC, sensitivity and specificity are used as the evaluation indexes of the model to compare the effects of different kinds of sampling methods on modeling results.

5.2. Case control under-sampling method

This section uses the "case control" under-sampling method to extract the whole sample data of May 2014, where "case" is the traffic flow condition before the traffic crash and "control" is the normal traffic flow condition without traffic crash. A total of 1210 crash records and 3808 non-crash records on Shanghai expressway in May 2014 were selected by "case control" method. The full sample data in June 2014 was used to verify the performance of the modeling. The comparison of predicted

Table 5
Comparison Results of Case Control Model and Full Sample Model.

Sample Data	AUC	Youden Index Threshold	Sensitivity	Specificity
Case Control Sample	0.7336	0.2718	75.71%	62.17 %
Full Sample	0.8342	0.0001625	77.04 %	76.21 %

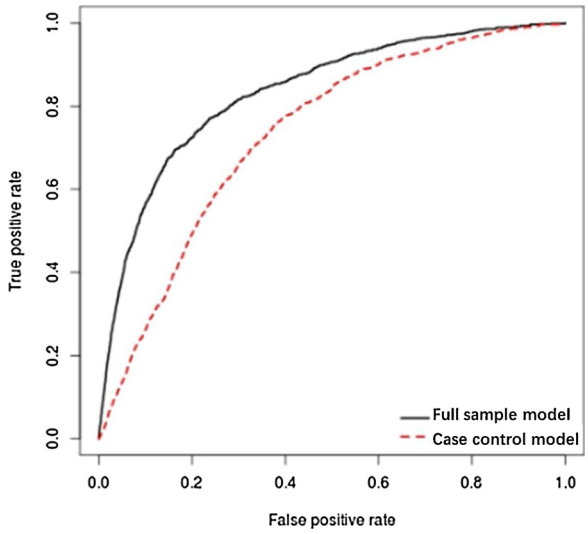

Fig. 7. Comparison of ROC for Case Control Model and Full Sample Model.

results between this model and the model using full sample data is shown in the Table5 and the ROC curve is shown in the Fig. 7.

It can be seen from the Table5 that the threshold value of the model prediction results has been improved, which means that the "case control" sampling method can effectively solve the problem that the prediction results of imbalanced data classification are biased towards the minority class. However, it is obviously can be seen from Fig. 7 that the overall prediction capability (AUC) of this under-sampling model decreased compared with the model using full sample data.

5.3. Random under-sampling method

Random under-sampling method is applied in this section to extract non-crash data in the whole sample data according to a certain proportion of crashes and non-crashes, so as to reduce the number of non-crash samples and make the data distribution more balanced. All crash data still have been retained in the dataset. Five crash and non-crash ratios (1:1, 1:4, 1:5, 1:10 and 1:20) were used to form the modeling dataset. The description of the dataset is shown in Table6.

The logistic regression model is still used to construct the real-time traffic crash risk prediction model. The prediction results of the model are shown in Table 7. As can be seen from the Table 7, the random sampling data model constructed by crash and non-crash ratio 1:4 has

Table 7
Prediction Results of Random Sampling Models.

Data Proportion	AUC	Youden Index Threshold	Sensitivity	Specificity
1 : 1	0.7912	0.3826	79.79%	64.61 %
1 : 4	0.8434	0.1955	76.68%	80.11 %
1 : 5	0.8429	0.1726	75.89%	81.27 %
1 : 10	0.8418	0.0947	75.27%	81.31 %
1 : 20	0.8407	0.0499	75.27%	81.21 %
Full sample	0.8342	0.0001625	77.04 %	76.21 %

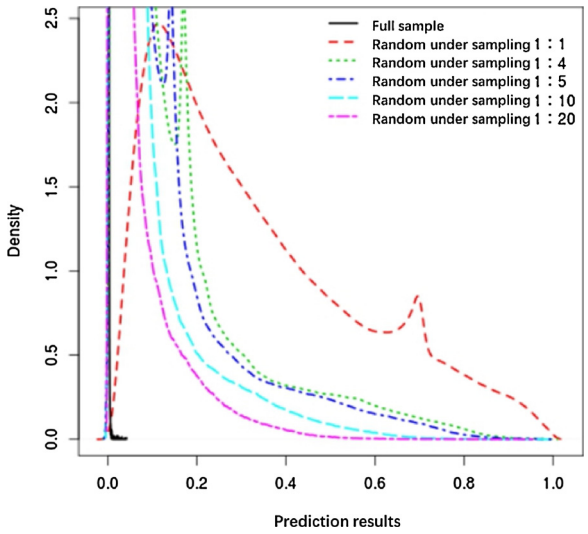

Fig. 8. Kernel Density Distribution of Different Models with Different Data Proportions.

the highest AUC, which is about 0.01 higher than the logistic regression model using full sample size. Compared with the logistic regression model using full sample, the threshold values of all the random sampling data model are improved. In addition, the sensitivity is improved with the increase of data proportion on crash cases. Fig. 8 shows the kernel density of the predicted results of different models with different data proportions. It can be seen that the distribution of prediction results is related to the proportion of crash and non-crash samples. The prediction results will be more balanced when the distribution of crash and non-crash data is more balanced, which solves the problem that the prediction results tend to be the majority class that is non-crash cases.

5.4. Comparison of under-sampling methods

It can be seen from the above two sections that the under-sampling method can effectively solve the problem that the prediction results in imbalanced data classification are biased towards the majority class. The random sampling method improves the overall prediction ability (AUC) of the model to a certain extent, while the AUC of the model using "case control" data decreases significantly. The comparison

Table 6
Summary Statistics for Random Under-Samplings.

Dataset	Prediction data			Verification data		
	Number of crash	Number of non-crash	Proportion of crashes	Number of crash	Number of non-crash	Proportion of crashes
Full sample	1210	7,559,352	0.016%	1128	7,869,083	0.014%
Random under sampling1 : 1	1210	1210	50%	1128	7,869,083	0.014%
Random under sampling 1 : 4	1210	4840	20%	1128	7,869,083	0.014%
Random under sampling 1 : 5	1210	6050	16.67%	1128	7,869,083	0.014%
Random under sampling 1 : 10	1210	12,100	9.09%	1128	7,869,083	0.014%
Random under sampling 1 : 20	1210	24,200	4.76%	1128	7,869,083	0.014%

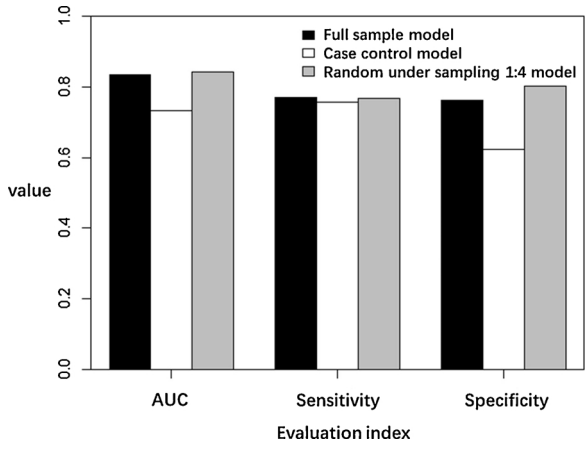

Fig. 9. Comparison Between Three Kinds of Models.

between the prediction results of the model using full sample data and the model using under-sampling data is shown in Fig. 9.

It can be seen from the Fig. 9 that the prediction effect of "case-control" model constructed under the continuous traffic data environment is not as good as random under sampling model. The main reason is that "case-control" restrictions on the controlled non-crash data such as specific time and locations cannot reflect the continuous data environment very well. Compare to "case-control" data, random sampling data can reflect the continuous data environment better.

5.5. Over-sampling method

The SMOTE algorithm was applied in this section to retain all non-crash samples while each crash sample was regenerated by using SMOTE algorithm. The regenerated crash and non-crash ratio is 1:1, 1:4, 1:5, 1:10 and 1:20, respectively. The description of samples for prediction models are shown in Table 8.

The above regenerated over-sampling data were used to construct a logistic regression model to predict the real-time traffic crash risk respectively and the full sample data in June 2014 were used to simulate the continuous data environment. The prediction results of the models are shown in Table 9. The optimal model is SMOTE over-sampling model with crash and non-crash ratio 1:1.

Similar to the results of random over-sampling, the classification threshold of SMOTE model is improved compared with the model using full sample. Fig. 10 shows the prediction results of different models with different data proportion in the continuous data environment. It can be seen from the prediction results that the performance of prediction results will increase when the distribution of crash and non-crash data is more balanced. The sensitivity will also increase when the distribution is more balanced. Therefore, SMOTE method can effectively solve the problem of imbalanced data classification in real-time traffic crash risk prediction where the prediction results are biased towards the non-crash class.

Table 9
Prediction of SMOTE Models with Different Proportions.

Data Proportion	AUC	Youden Index Threshold	Sensitivity	Specificity
1 : 1	0.8479	0.5103	75.35%	80.84 %
1 : 4	0.8447	0.2168	74.29%	81.49 %
1 : 5	0.8442	0.1831	73.94%	81.63 %
1 : 10	0.8431	0.1035	73.05%	82.07 %
1 : 20	0.8423	0.0616	70.30%	84.37 %

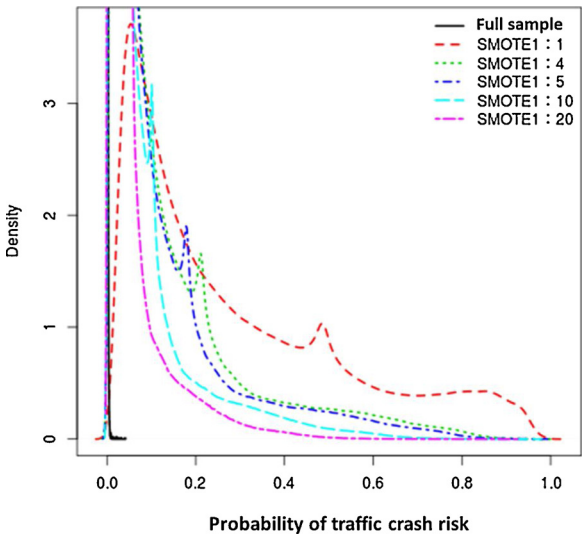

Fig. 10. Kernel Density Distribution of Different SMOTE Models.

5.6. Comparison of different sampling methods

The experimental results show that both under-sampling and over-sampling methods can effectively solve the problem that the prediction results of imbalanced data classification are biased towards the majority class. Table 10 shows the comparison of prediction results of the optimal models in above mentioned different kinds of sampling methods.

It can be seen from AUC that the model constructed using SMOTE over sampled data with crash and non-crash ratio 1:1 shows the best performance. However, the SMOTE over sampled data regenerates crash data in a certain proportion to balance the crash and non-crash data on the basis of retaining all non-crash data. As a result, the whole sample size is too large and the efficiency of training model is low. The improvement of performance is also very limited. The "case control" under-sampling model needs a moderate amount of data. However, the data used to create the model cannot fully reflect the whole continuous dataset due to factors such as controlling the time and space to extract non-crash data. The overall prediction ability of the model is poor. For the continuous data environment, the AUC of the random under-sampling model with crash and non-crash ratio 1:4 is only slightly smaller

Table 8
Summary Statistics of Samples for SMOTE Prediction models.

Dataset	Prediction data			Verification data		
	Number of crash	Number of non-crash	Proportion of crashes	Number of non-crash	Number of crash	Proportion of crashes
Full Sample	1210	7,559,352	0.016%	1128	7,869,083	0.014%
SMOTE1 : 1	7,558,142	7,558,142	50%	1128	7,869,083	0.014%
SMOTE1 : 4	1,889,536	7,558,142	20%	1128	7,869,083	0.014%
SMOTE1 : 5	1,511,629	7,558,142	16.67%	1128	7,869,083	0.014%
SMOTE1 : 10	755,815	7,558,142	9.99%	1128	7,869,083	0.014%
SMOTE1 : 20	377,908	7,558,142	4.76%	1128	7,869,083	0.014%

Table 10
Comparison of Optimized Models Using Different Sampling Methods.

Data	AUC	Youden Index Threshold	Sensitivity	Specificity
Full sample data	0.8342	0.000163	77.04 %	76.21 %
"case-control" under-sampling data	0.7336	0.2718	75.71%	62.17 %
1 : 4 random sampling	0.8434	0.1955	76.68%	80.11 %
1 : 1SMOTE data	0.8479	0.5103	75.35%	80.84 %

than that of SMOTE model and the amount of data analyzed is greatly reduced. Overall, the model constructed using random under-sampling method with crash and non-crash ratio 1:4 can obtain the optimal prediction results.

6. RCSMLP prediction model

It can be concluded from above modeling results that the classification threshold calculated by the Youden method at the output level can be a good choice to divide the predicted results of the model. In terms of data level, random under-sampling data with crash and non-crash ratio 1:4 are selected to construct the optimal predictive modeling. At the algorithm level, the cost sensitive MLP model is built by applying cost matrix to improve the traditional MLP model, and the overall prediction ability of the model is the best when the threshold cost matrix is applied. Therefore, the random sampling cost-sensitive MLP model (RCSMLP) model is constructed finally by synthesizing the conclusions from three levels.

The comparison of modeling results between RCSMLP model that optimized from three levels and the other MLP models can be seen in the following Table 11.

The AUC of four kinds of models in the Table11 are similar. The RCSMLP model has the highest AUC, which is about 0.002 higher than the full-sample cost-sensitive MLP model. The threshold value of Youden index in RCSMLP model is closer to 0.5 compared to other models. The sensitivity of the RCSMLP model reached 78.10 %, which means that 78.10 % of the crashes in the testing dataset can be effectively predicted. The specificity of the model reached 81.44 %, which means that 81.44 % of the non-crash data can be effectively predicted. The ability of RCSMLP to predict crashes has been significantly improved compared with the full-sample cost-sensitive MLP model. Although the ability to predict non-crash has been reduced to a certain extent, the accuracy of predicting crash is much more important. Moreover, the training time of the MLP model with full sample data is around 10 min, while the RCSMLP model can complete the model training very fast because the sample size is greatly reduced, which greatly improves the training efficiency of the model. In conclusion, RCSMLP model is one of the most effective model for real-time traffic crash risk prediction.

7. Rusboost prediction model

At the alogorithm level, we also examine the result of Rusboost models with various data proportion. Bayesian optimization was used to find the optimal hyperparameters of the Rusboost and Bayes Minimum Risk calibration method was applied. It can be seen in the Table 12 that the AUC of model in different proportion is similar. The Rusboost model has the highest AUC when the data proportion is 1:20. The Sensitivity and specificity are basically kept above 0.8.

Table 12
Rusboost Prediction Results with Bayes Minimum Risk Calibration Method.

Data Proportion	AUC	Sensitivity	Specificity
1 : 1	0.757	0.760	0.754
1 : 4	0.891	0.848	0.800
1 : 5	0.890	0.851	0.801
1 : 10	0.889	0.840	0.811
1 : 20	0.892	0.842	0.816
Full sample	0.888	0.834	0.816

Pande et al.(Pande and Abdel-Aty, 2006) also used MLP algorithm to predict the risk of real-time traffic crashes. This model could only predict 57 % of crashes and 71.17 % of non-crashes. Sun et al.(Sun and Sun, 2016) using the SVM model to construct real-time traffic crash risk prediction model. The experimental results show that the sensitivity and specificity were both more than 77 %. Wang et al.(Wang et al., 2017) integrated crash frequency model and crash risk prediction model to predict real-time traffic crash risk prediction. The AUC of the model reached 0.801 and the model proved to be better than most of the traditional model, Basso et al.(Basso et al., 2018)applied the SVM model to get 75.3 % sensitivity and 77.53 % specificity. The summary of comparison of existing traffic crash predictive modeling is shown in Table 13. It can be shown from the table that the overall prediction accuracy of the Rusboost Model and RCSMLP model proposed in this research are better compared to the other types of prediction models. The prediction accuracy can be improved by using the methods proposed in this study.

7.1. Conclusions and discussion

Considering the imbalanced and continuous characteristics of real-time traffic crash data and the limitation of previous traffic crash risk prediction models to solve the typical problem of imbalanced data, this study proposed two new kinds of real-time prediction models to comprehensively optimize the prediction results from output level, data level and algorithm level.

It can be concluded from this study that the classification threshold calculated by the Youden index method at the output level can be a good choice to divide the predicted results of the model. In addition, at the output level it also can be seen by comparing various calibration methods that the predicted probabilities with proper calibration will be much closer to true probabilities in fixed region and more practical for traffic engineers. In terms of data level, random under-sampling data with different crash and non-crash ratio should be examined to construct the optimal predictive modeling. At the algorithm level, the cost sensitive MLP model can be built by applying cost matrix to improve the traditional MLP model, and the overall prediction ability of the model is the best when the threshold cost matrix is applied. Finally, the RCSMLP model and Rusboost model are constructed by synthesizing the conclusions from three levels. The sensitivity of the RCSMLP model reached 78.10 % and the specificity of the model reached 81.44 %. The

Table 11
Comparison of MLP models.

Model		Training time	AUC	Youden Index Threshold	Sensitivity	Specificity
Cost sensitive models	Random under-sampling data with crash and non-crash ratio 1:4(RCSMLP)	17s	0.8744	0.5179	78.10 %	81.44 %
	Full sample	11min18s	0.8723	0.3851	75.79%	84.31 %
Traditional models	Random under-sampling data with crash and non-crash ratio 1:4	16s	0.8735	0.2436	74.46%	84.04 %
	Full sample	9min47s	0.8708	0.0002643	73.4%	83.87 %

Table 13
Comparison of Traffic Crash Predictive Modeling Results.

Model	Algorithm	AUC	Sensitivity	Specificity
Rusboost model	Adaboost.M2	0.89	84.21%	81.62 %
RCSMLP model	cost sensitive MLP	0.87	78.10 %	81.44 %
Full sample cost sensitive MLP models	cost sensitive MLP	0.87	75.79%	84.31 %
(Pande et al, 2006)	MLP	–	57 %	71.17 %
(Sun et al, 2016)	SVM	–	77.9 %	79.3 %
(Wang et al., 2017)	Bayesian	0.801	–	–
(Basso et al., 2018)	SVM	–	75.3 %	77.53 %

AUC and sensitivity of the Rusboost model reached 0.892 and 0.842 while the specificity of the model reached 0.816, which shows the even better performance in dealing with the imbalanced traffic crash risk prediction problem. It is noted that the proposed method about improving prediction accuracy in this study is universal and can be applied to many other prediction models besides RCSMLP and Rusboost model to predict real-time traffic crash risk.

Although the constructed RCSMLP and Rusboost models have shown good performance in this study, this kind of real-time traffic crash risk prediction model can still be further improved in the following two aspects. The independent variables selected in this study are common traffic parameters easily calculated such as average speed, standard deviation of speed and other variables. Some deep learning methods can be applied in further to figure out hidden variables to improve the overall prediction ability of the model. The current constructed model only considers information of traffic flow status. Other variables including road type, road geometric characteristics and weather information can be added to the model and improve the prediction ability in further.

CRediT authorship contribution statement

Yichuan Peng: Conceptualization, Methodology, Formal analysis, Supervision, Funding acquisition, Writing - review & editing. **Chongyi Li:** Formal analysis, Software, Data curation, Visualization. **Ke Wang:** Software, Investigation. **Zhen Gao:** Supervision, Writing - review & editing. **Rongjie Yu:** Software, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was jointly sponsored by the National Key R&D Program of China (2018YFB1201403), Chinese National Science Foundation (71601143 and 71771174), and the "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission. All opinions and results are solely those of the authors.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105610.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Safety Res. 36 (1), 97–108.
Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M.F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec. 1897 (1), 88–95.
Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Transp. Res. Rec. 1908 (1), 51–58.
Abdel-Aty, M.A., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C Emerg. Technol. 24, 288–298.
Ahmed, M.M., Abdel-Aty, M.A., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE trans. Intell. Transp. Syst. 13 (2), 459–468.
Niculescumizil, Alexandru, Caruana, Rich, 2005. Predicting Good probabilities with supervised learning[C]// machine learning. In: Proceedings of the Twenty-Second International Conference (ICML 2005). Bonn, Germany, August 7-11, 2005.
Baheti, B., Gajre, S., Talbar, S., 2018. Detection of distracted driver using convolutional neural network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE.
Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transp. Res. Part C Emerg. Technol. 86, 202–219.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.
Hughes, R., Council, F., 1999. . On establishing relationship (s) between freeway safety and peak period operations: performance measurement and methodological considerations. In: 78th Annual Meeting of Transportation Research Board. Washington, DC..
Hu-pu, L.U., Ri-min, L.I., 2014. Developing trend of ITS and strategy suggestions [J]. J. Eng. Stud. 1 (6), 6–19.
Katrakazas, C., Antoniou, C., Yannis, G., 2019. In: Time Series Classification Using Imbalanced Learning for Real-Time Safety Assessment. Transportation Research Board 98th Annual Meeting. Washington, DC..
Kocsis, L., Gyrgy, András, Bán, Andrea N., 2013. Boostingtree: parallel selection of weak learners in boosting, with application to ranking. Mach. Learn. 93 (2-3), 293–320.
Krawczyk, B., Woźniak, M., 2015. Cost-sensitive neural network with roc-based moving threshold for imbalanced classification. International Conference on Intelligent Data Engineering and Automated Learning 45–52.
Kurzhanskiy, A.A., Varaiya, P., 2010. Active traffic management on road networks: a macroscopic approach. Philos. Trans. 368 (1928), 4607–4626.
Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. Transp. Res. Rec. 1840 (1), 67–77.
Liu, X., 2012. Classification accuracy and cut point selection. Stat. Med. 31 (23), 2676–2686.
Longadge, R., Dongre, S., 2013. Class Imbalance Problem in Data Mining Review. arXiv Prepr. arXiv1305.1707.
Maloof, M.A., 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. ICML-2003 Workshop on Learning From Imbalanced Data Sets II. pp. 1–2.
Oh, C., Oh, J.-S., Ritchie, S., Chang, M., 2001. Real-time estimation of freeway accident likelihood. In: 80th Annual Meeting of the Transportation Research Board. Washington, DC..
Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38 (5), 936–948.
Pham, M.-H., Bhaskar, A., Chung, E., Dumont, A.-G., 2010. Random forest models for identifying motorway rear-end crash risks using disaggregate data. 13th International IEEE Conference on Intelligent Transportation Systems 468–473.
Pozzolo, A.D., Caelen, O., Johnson, R.A., et al., 2015. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE.
Seiffert, C., Khoshgoftaar, T., Van Hulse, J., Napolitano, A., 2010. Rusboost: a hybrid approach to alleviating class imbalance", IEEE Trans. IEEE Trans. Syst. Man Cybern. A. Syst. Hum. 40 (1), 185–197 January.
Sun, Jie, Sun, Jian, 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transp. Res. Part C Emerg. Technol. 54, 176–186.
Sun, Jie, Sun, Jian, 2016. Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model. IET Intell. Transp. Syst. 10 (5), 331–337.
Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit. 40 (12), 3358–3378.
Theofilatos, A., Chen, C., Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Journal of Transportation Research Board(0).
Tignor, S.C., Brown, L.L., Butner, J.L., Cunard, R., Davis, S.C., Hawkins, H.G., Fischer, E.L., Kehrli, M.R., Rusch, P.F., Wainwright, W.S., 1999. Innovative traffic control technology and practice in Europe. Ite J. 70 (1), 45–49.
Wang, L., Abdel-Aty, M., Lee, J., 2017. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. Accid. Anal. Prev. 104, 58–64.
Xie, W., Wang, J., Ragland, D.R., et al., 2016. Utilizing the eigenvectors of freeway loop data spatiotemporal schematic for real time crash prediction. Accid. Anal. Prev. 94, 59–64.
Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171.
Xu, C., Wang, W., Liu, P., 2013. A genetic programming model for real-time crash prediction on freeways. IEEE trans. Intell. Transp. Syst. 14 (2), 574–586.
Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
Zhou, Z.-H., Liu, X.-Y., 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 1, 63–77.