Expressway Crash Prediction based on Traffic Big Data

Hailang Meng
School of Electronics and Information Engineering, Tongji University
Shanghai, China
18721017862@163.com

Xinhong Wang
School of Electronics and Information Engineering, Tongji University
Shanghai, China
wang_xinhong@163.com

Xuesong Wang
School of Transportation Engineering, Tongji University
Shanghai, China
wangxs@tongji.edu.cn

ABSTRACT

With the development of society, the number of vehicles increases rapidly. The vehicle plays an important role in people's life, however the problem of traffic safety caused by vehicles has also become increasingly prominent. In China, the high crash rate and casualty rate on expressways have always troubled traffic management department. So crash prediction on expressway becomes vital. Conventionally, crash prediction is based on traffic flow data. These data do not contain all the necessary factors. In this paper, we propose a method of prediction using real-world data, including historical accident data, road geometry data, vehicle speed data, and weather data. We treat the crash prediction problem as a binary classification problem. For classification, sample imbalanced is a great challenge in practice. Modifying sample weights is applied to handle this challenge. Three machine learning classification techniques, namely Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and Xgboost, are considered to carry out the crash prediction task respectively. The best recall and precision rate of these models are respectively 0.764253 and 0.01062. The proposed method can be integrated into urban traffic control systems toward police dispatch and crash prevention.

CCS Concepts

• **Information systems** → **Information systems applications** → **Data mining.**

Keywords

Crash prediction; machine learning; feature extraction and selection; sample imbalance

1. INTRODUCTION

With the development of society, people's living standards have been improved constantly. More and more families own cars, which causes the number of vehicles increases rapidly. According to the statistics of the Ministry of Public Security of China, as of August 2017, the number of motor vehicles in China reached 307 million, and the number of motorists reached 377 million. The number of vehicles in 50 cities exceeded 1 million in China [1]. Vehicles play an important role in people's life and production while the problem of traffic jams and crashes caused by vehicles have also become increasingly prominent. According to Global Status Report on Road Safety, published by World Health Organization in 2015, about 1.25 million people were killed in traffic accidents every year [2]. Therefore, more and more research is devoted to improve traffic conditions and reducing crashes. Expressway crash prediction is one of the important research directions. This paper aims at predicting whether there is a traffic accident on a road section in the next hour. We need to analyze the cause of the accident on expressway and predict the occurrence of accidents. Analyzing the cause of the accident on expressway can provide the reference for traffic safety, accident prevention, and injury reduction [3]. And predicting the occurrence of an accident based on these causes can optimize the dispatch of police. According to the forecast results, the traffic police department can arrange the police force in advance at the time and road section of the accident.

SPML '18, November 28–30, 2018, Shanghai, China

DOI: https://doi.org/10.1145/3297067.3297093

By finding out the causes of traffic crashes, some literature is conducive to the feature extraction when predicting traffic crashes [3-5]. [3] collected 166 serious traffic crashes that occurred in China from 2008 to 2014 and extracted 23 major factors. The comprehensive analysis shows that, among all affecting factors, the main factors are as follows: speeding, ramp, weather. [4] collected traffic crash data of an expressway from 2007 to 2008 and studied the impact on expressway traffic crash of road geometric feature, traffic volume and environment. [5] found that the width of lanes, posted speed limit, nature of pavement, and annual daily traffic (AADT) were correlated with vehicle crashes. These papers gave important factors affecting the occurrence of accidents, and can provide some reference for feature extraction.

To perform the prediction, some literature told out the expected accidents directly or gave an accident possibility based on the causes. According to the dataset they used, the literature can mainly be divided into two parts. One part of the research is in the field of transportation engineering. They only relied on traffic flow data to predict crash traditionally [6-11]. Some of paper collected traffic flow data by using loop detectors and then use these data for accident prediction [6], [9]. [7] was first to use the real-time speed data collected from automatic vehicle identification (AVI) to predict traffic accidents. [8] uses random forest (RF) model to select the significant variables from the data of the traffic flow 5-10 min before the crash occurred. Then they propose a hybrid model combining a support vector machine (SVM) model with a k-means clustering algorithm to predict the likelihood of crashes. [10] developed a prediction model and proved the accuracy and effectiveness of the introduced incremental learning algorithm through comparative experiments. [11] presents an innovative approach to investigate the inner mechanism between traffic status and crash potential based on High Definition Monitoring Systems (HDMS) data. HDMS records delicate vehicle trajectory data and characteristic details.

The conventional accident analysis approaches studied in transportation engineering have two drawbacks. Firstly, they rely on traffic flow data, so they need to install many sensors or

detectors to monitor traffic flow, such as loop detectors and automatic vehicle identification system. However, it is difficult to obtain a wide range of accurate traffic flow data by installing sensors or other testing equipment on a large scale. In our scenario, we intend to predict the occurrence of accidents on various sections of the highway with a length of 138km. So this method cannot be used in our scenario because it is costly to lay equipment on such a long highway. Secondly, more factors, such as drivers' driving status, should be considered in the model. These factors have a significant impact on the accidents, while they are difficult to collect [12]. To solve this problem, it is necessary to consider as much as possible of other data or factors that can be collected to make up for this problem.

The second part of the research considered other data besides traffic data [2], [13-16]. In addition to the traffic flow data mentioned above, these papers also considered weather data, holidays and other factors. [13] used logistic regression to predict crashes based on real-time traffic-flow and rain data. [2] collected multiple data sources, including traffic accident, traffic flow, weather condition and air pollution from the same city. They proposed a deep learning model based on recurrent neural network to predict the risk of traffic accident. [14] employed the crash data set maintained by the Nevada Department of Transportation to train a logistic quantile regression model. The result shows that travel speed, signal spacing, and driveway density are significantly influencing factors on crash rate. [15] developed artificial neural network classifiers that can predict accident severity with reasonable accuracy. The overall accuracy of the predictive model for the testing data was 74.6%.

The second part of the research is similar to our study. In this paper, we intend to study expressway accidents prediction problem. Specifically, we predict whether traffic accidents will occur in the next hour on a road section. However, research contents of these papers are different from our work. Firstly, some of these papers predicted the severity of the crash [15-16]. These models are not suitable for predicting whether an accident will occur. Secondly, for the remaining papers that predict whether an accident will occur, the time and space range of the model prediction are different from our study [2], [13], [14]. Some studies predict whether an accident will occur on a certain day, and some studies predict whether an accident will occur in a block in the city, and so on. However, there is no research to predict whether accidents will occur in the next hour on a road section.

Therefore, this paper studies the prediction of expressway accidents in road sections and short time. In this scenario, the number of accidents per hour in the road section is too small (in our dataset, the accident probability is 0.04235%, and the probability of multiple accidents occurring in a single hour is 0.0041%). Therefore, we model the accident prediction problem into a binary classification problem. We modify sample weights to solve the problem of sample imbalance. And we choose three classification methods, namely Random Forest (RF) [17], Gradient Boosting Decision Tree (GBDT) [18] and Xgboost [19]. We use an actual dataset to validate and select the method. We will introduce a complete data analysis process, including data cleaning, feature selection, sample imbalance processing, and application model. Each model finds the optimal parameters through cross-validation, and then uses the optimal model to complete the prediction.

The remainder of this paper is organized as follows: Section 2 introduces the dataset and the process of data. Section 3 describes feature selection. Section 4 introduces how to handle the problem of imbalanced samples, the model selection and evaluation metrics. Section 5 explores the results of experiment. Finally, we draw the conclusions and opportunities for future research.

2. DATA DESCRIPTION AND PROCESSING

In this paper, we intend to predict whether traffic accidents will occur in the next hour of a road section. The problem has the following unique properties so that conventional methods cannot be directly applied. The difference from other accident prediction problems lies in the time scale and spatial scale of prediction. And there is no research to predict whether accidents will occur in the next hour of a road section. To reflect the real-world situation, we collected a dataset that contains hourly recorded expressway accidents (including location, accident type, and severity, etc. information) in a two-month period (from June to July 2017) from the police department (the data is sensitive and confidential, and hence details cannot be released). We have also collected some other data to support accident prediction. Our dataset can be divided into five categories. The details of the data description and data processing are as follows.

2.1 Data description

Historical crash data: two months of crash data on the expressway were collected by the Traffic Management Bureau. The road is in two directions, each direction is 69km. The road is divided into 30 road sections with average length of 4.6km. The location and time of each crash have been recorded.

Road geometry data: the road geometry information of the 138km expressway was collected. The properties of each road section are shown in Table 1.

Table 1. The properties of road sections

Feature	Meaning
MICRO_ROAD_ID	ID of each road section, chosen from 1 to 30.
CD	Length of each road section, and the value ranges from 1.99km to 5.14km.
LANES	Number of lanes, and the value is 3, 4 or 5.
RAMP	Condition of ramps on the road sections. It is a category type variable to record the number of ramps and the status of the entry and exit.

Time factor: because traffic crash patterns may change drastically in the time of the day and week. For example, traffic crash is more frequent at rush hours than that at off-peaks. Therefore, day of week and time information is taken into account.

Weather data: because the weather will affect the accident [13], so we crawled weather data from the Internet.

Speed data: in addition to the above data, we also obtained vehicles' speed data for each section of the expressway in July 2017 from the Traffic Management Bureau. This speed data records the average speed of all cars passing through the road section in one hour, not the instantaneous speed of each vehicle.

2.2 Data processing

Historical crash data: according to the location information in the raw data, the accident was associated with the corresponding road section. And the number of accidents per hour on each section can be got. Since we want to predict whether there will be accidents in

the next hour, so we processed all the number of accidents to 0 (no accidents) or 1 (accidents).

Time factor: if it is divided into 24 hours a day, excessive features will be generated after one-hot encoding. And some time periods and working days are not much different. So according to Chinese lifestyle and frequency of historical accidents, we divide the time of one day into 4 blocks:00:00-05:59 and 23:00-23:59 (mid-night to dawn), 06:00-08:59 and 19:00-22:59 (morning hours and nighttime), 09:00-13:59 and 17:00-18:59 (normal traffic hours), 14:00-16:59 (afternoon rushing hours).

Weather data: because the raw data is unstructured and only bad weather has a greater impact on the occurrence of traffic crash. The original text data was digitized and combined to generate features. We divide the weather into 0 (sunny, cloudy or light rain) and 1 (Heavy rain or thunderstorm).

Speed data: 24.8% of the raw speed data is missing. Generally, if a small amount of data is missing, we use interpolation to fill. However, in our data, the speed data is missing too much. To facilitate analysis accurately, we discard the data in those sections and times which has no speed data.

After data processing, we can get the following features: *Avg_speed*, *CD*, *Hour*, *Weekday*, *RAMP*, *LANES* and *Weather*. Among these features, *Avg_speed* and *CD* are continuous variables. The rest are category variables.

3. FEATURE SELECTION

After these features are generated, feature selection is required. For continuous variables, we use Pearson correlation coefficients to select features [20]. The Pearson correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

In this formula, $cov(X,Y)$ represents the covariance of the variables X and Y. σ_x and σ_y represents the standard deviation of X and Y respectively. μ_x and μ_y representing the mean of X and Y respectively. The Pearson correlation coefficient varies from -1 to 1, and the coefficient value of 0 means that there is no linear relationship between the two variables. The larger the absolute value of the coefficient, the stronger the correlation. The Pearson correlation coefficients for each continuous variable and crash (here the accident data is taken from the original number of accidents, rather than the processed 0 or 1.) are shown in Table 2.

Table 2. Correlation of continuous variable

Variable	Pearson correlation coefficient
Avg_speed5	-0.069
CD	-0.013

From the Table 2, we can find *Avg_speed5* has a relatively strong correlation with the occurrence of crashes. And -0.069 can be understood as: when the speed is decreased, it means that the traffic density is large and the vehicle is more likely to cause congestion, so it is easy to cause a crash. However, *CD* have a low correlation with the occurrence of crashes. And it is difficult to explain why they are negatively related to the occurrence of crashes. Logically, the longer the road, the more accidents will occur. We suspect that this is caused by the length of the road section being almost the same. Therefore, this study will not consider this variable.

For categorical variables, feature selection is performed using information gain ratio [21]. The information gain ratio is an important way to select features. We introduce the information gain before introducing the information gain ratio. For an original data set *D*, its information entropy is defined as:

$$H(D) = -\sum_{k=1}^{m} P_k \, log_2 \, P_k \quad (2)$$

m represents the number of categories of predicted labels in the dataset, P_k represents the ratio of the number of samples in this category to the total number of samples. When the feature f is used to divide the sample, the information entropy of the divided dataset is:

$$H(D,f) = -\sum_{v=1}^{n} \frac{|D^v|}{|D|} H(D^v) \quad (3)$$

n represents the number of all possible values of the feature. v represents the value of the feature. $|D^v|$ represents the number of samples whose feature f takes v. $H(D^v)$ represents the information entropy of the sample set when the feature takes v. Information gain is expressed as:

$$G(D,f) = H(D) - H(D,f) \quad (4)$$

Because when the information gain is used as a criterion for feature selection, there is a problem that the feature having more value is selected. So the information gain ratio is multiplied by a penalty parameter based on the information gain. When the number of feature values is large, the penalty parameter is small; when the number of feature values is small, the penalty parameter is large. The penalty parameter is defined as:

$$penalty = \frac{1}{H_f(D)} = \frac{1}{-\sum_{v=1}^{n} p_v \, log_2 \, p_v} \quad (5)$$

$H_f(D)$ represents the entropy of feature f. p_v represents the ratio of the number of samples with feature f taking v to the total number of samples. Therefore, the formula for the information gain ratio is:

$$G_R(D,f) = \frac{G(D,f)}{H_f(D)} \quad (6)$$

The larger the information gain ratio, the more effective the feature. Table 3 shows the result of feature selection using information gain ratio.

Table 3. Selection of category variables

Variable	Information gain ratio
Hour	0.000616
LANES	0.0002
Weekday	0.000156
RAMP	0.000138
Weather	0.000096

We can find that information gain ratio of *Hour* is the highest, which is 0.000616. And information gain ratio of *Weather* is the lowest, which is 0.000096. However, relying solely on the information gain ratio, it is not possible to determine whether the feature with a lower information gain ratio should be removed. Therefore, this paper also uses the chi-square test [22] to filter the categorical variables. At the same time, the results of the chi-

square test can be used to support the results of the information gain ratio. The basic idea of the chi-square test is to infer whether there is a significant difference between the overall distribution and the expected distribution based on the sample data, or to determine whether the two categorical variables are related or independent. The general null hypothesis is that the two variables are independent of each other. We firstly assume that the null hypothesis is true and calculate the chi-square value, which represents the degree of deviation between the observed value and the theoretical value. According to the chi-square distribution, the chi-square statistic and the degree of freedom, the probability P of obtaining the current statistic under the assumption of zero hypothesis can be determined. If the probability P is small, the deviation between the observed value and the theoretical value is large, and the null hypothesis should be rejected. Otherwise the original hypothesis cannot be rejected. The formula for calculating the chi-square statistic is as follows:

$$\chi^2 = \sum_{v=1}^{n} \frac{(A_v - T_v)^2}{T_v} \quad (7)$$

A is the actual frequency of accidents, and T is the theoretical value of accident frequency.

Table 4. Selection of category variables

Variable	P-value
Hour	1.368386e-15
Weekday	3.326015e-5
RAMP	0.006765
LANES	0.02615
Weather	0.17232

Table 4 shows the result of feature selection using information gain ratio. When using chi-square test, we reject the null hypothesis if P-value is less than 0.05, because we believe that this variable is related to the occurrence of an accident when P-value is less than 0.05. From Table 4, we can see that the result of feature selection using information gain ratios is basically the same as using chi-square test. Also, P-value of *Weather* is greater than 0.05. This verifies the correctness of the results. The importance of *Weather* variables is not very significant, we think this has a lot to do with the dataset. The weather data we collected was in June and July. In China, there are very few extreme weather conditions in June and July, such as fog and snow. Therefore, we think five variables are effective features.

4. MODEL SELECTION AND EVALUATION METRICS

In the following, we will describe the models which we use and the evaluation metrics which we use to evaluate the performance of the model. Also, we discuss how to select model and parameter.

4.1 Sample imbalance processing

Crash is a small probability event. In our dataset, the accident probability is 0.04235%, and the probability of multiple accidents occurring in a single hour is 0.0041%. And our task is to predict whether a traffic crash will occur or not during next hour on a road section. Explain here, as mentioned earlier, we divide the time of one day into 4 blocks. That means when the time feature is extracted, the time is segmented. In our dataset, most of the samples are negative samples (that means no traffic crash occurs on a road section in one hour). And a very small number of samples are positive (that means traffic crashes occur on a road section in one hour).

For imbalanced dataset, the learning ability of the model is weak. They will tend to predict all samples as negative samples because they can already have a relatively high accuracy in this way. However, our goal is to predict positive samples as accurately as possible. Many methods exist to handle data with imbalanced distributions. The paper [23] gave a solution for the problem of sample imbalance. The methods are as following:

Random over-sampling: this method adds a set which is randomly sampled from the minority class. The purpose of reaching the balanced degree is achieved by replicating the examples and increasing the overall number of minority examples. In regards to over-sampling, since there are multiple copies of the same sample, the model trained may be too specific and over-fitting [23-24].

Random under-sampling: this method picks a part of the examples from majority class randomly and removes the rest of the examples of the majority class. It achieves a balance by reducing a large number of samples from majority class. The disadvantage of this method is relatively obvious: removing examples from the majority class may cause the classifier to miss important concepts pertaining to the majority class [23].

Modify sample weights: this method assigns weights to positive and negative samples respectively. It will assign a larger weight to the samples from minority class rather than the samples from majority class. In other words, the larger the number of samples in majority class is, the smaller the weight coefficient is. Similarly, the smaller the number of samples in majority class, the larger the weight coefficient is. In this way, the model will pay more attention to the samples from minority class [23]. In our study, we find that better model effects can be obtained by modifying sample weights.

4.2 Model description

The problem we are studying is to predict whether traffic accidents will occur on a certain road section in the next hour. It is a binary classification problem. The performance of machine learning model is related to the matching degree between internal algorithm and data set. It is desirable to find an optimal model to process our data. The models for solving the binary classification problem generally include logistic regression, tree models, neural networks, and so on. Logistic regression has a weak learning ability. And the neural network is less explanatory. So in this paper, three tree-based models are used, which are Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and Xgboost respectively. The introduction of these three models is as follows.

4.2.1 Random Forest

RF is an ensemble learning algorithm. Ensemble learning combines multiple learners. It is often possible to achieve generalization performance that is significantly superior to a single learner. RF fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. RF has the following advantages. Firstly, since each tree being trained is parallel, it runs efficiently on large data bases. Secondly, it is relatively robust to outliers and noise. Third, RF can prevent over-fitting to some extent. Because each subset selected using bagging to train each individual tree usually contains 2/3 of the dataset, and each time the node selects the optimal feature for splitting, it can only select from a fixed proportion of features [17].

4.2.2 Gradient Boosting Decision Tree

GBDT is also an ensemble learning algorithm. Unlike RF, RF is based on bagging and GBDT is based on Boosting. The classification result of this algorithm is the accumulation of the result of each tree. Therefore, GBDT is more concerned to reduce the deviation during training, and RF is more concerned with reducing the variance during training. GBDT has some advantages, including the ability to find non-linear transformations and handle skewed variables without requiring transformations. Also, GBDT has good robustness and high scalability [18].

4.2.3 Xgboost

Xgboost is also a tree-based model. Unlike RF and GBDT, Xgboost supports linear classifiers in addition to supporting trees as base classifiers. Besides this, Xgboost also has some other advantages. Traditional GBDT only uses first-order derivative information in optimization. But Xgboost performs second-order Taylor expansion on cost function. Xgboost also adds regularization to the cost function to control the complexity of the model [19].

4.3 Model selection

To make the learned model have better generalization performance, the five-fold cross validation [25] will be used to determine the model's real prediction performance. Five-fold cross validation means that the sample is divided into five parts. Each time we select one part as test set, and the rest for the training set. The model is trained and tested a total of 5 times. And the average accuracy of five results is recorded as the performance of the model. When the average accuracy is high, we can conclude that the model performs well. And we use grid search method of scikit-learn [26] to choose parameters of models, i.e. RF, GBDT and Xgboost.

In addition, we also need to choose the evaluation criteria of the model. In the binary classification problem, the samples can be divided into true-positive (TP) cases, false-positive (FP) cases, true-negative (TN) cases, and false-negative (FN) cases according to the combination of the true category of the sample and the classifier prediction category. Precision and recall rate are usually used to judge the quality of model classification. The precision and recall rate formulas are as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

Precision and recall rate are two contradictory measures. In general, the higher the precision rate, the lower the recall rate, and vice versa. In this paper, the recall rate refers to the numeric ratio between correctly predicted positive samples and all positive samples, while the precision rate refers to the numeric ratio between correctly predicted positive samples and all predicted positive samples. Since the cost of incorrectly predicting positive samples is much higher than the cost of incorrectly predicting negative samples, we pay more attention to the improvement of the recall rate. Consequently, in this paper, we mainly used recall rate to measure the performance of the model.

5. EXPERIMENTS AND RESULTS

In our dataset, there are 16651 negative samples and 37 positive samples. As mentioned earlier, imbalanced samples need to be processed. We found that the effect of modifying sample weights are significantly better than under-sampling and over-sampling. Also, under-sampling can cause a problem that samples are insufficient and over-sampling may cause over-fitting, so the performance of under-sampling and over-sampling is not described here. In this paper, we modify the weight of the sample by setting the input parameter which is called sample_weight when the model is fitted. The result of the models after processing the sample by modifying the sample weights is showed in Table 5 respectively.

The result is the best result of the three models. As mentioned earlier, the higher the precision and recall rate, the better the results of the model. The recall rate indicates the proportion of positive samples (samples of accidents) that were predicted to be correct in the test set. And the precision rate represents the prediction accuracy rate of a set that is predicted to be a positive sample. From

Table 5. The result of the model using modifying the sample weights

Models	Precision	Recall
RF	0.0088	0.717803
GBDT	0.01062	0.764253
Xgboost	0.00991	0.739242

the Table 5, we can find the precision and recall rate of GBDT are higher than these of RF and Xgboost. The recall rate of GBDT is 0.764253, and the recall rate of RF and Xgboost is 0.717803 and 0.739242 respectively. The precision rate of RF, GBDT and Xgboost are 0.0088, 0.01062 and 0.00991. Therefore, GBDT has the best performance. The specific classification results of GBDT are shown in Table 6.

Table 6. The classification result of GBDT

Real category	forecast result	
	0(non-crash)	1(crash)
0(non-crash)	0.839922	0.15786
1(crash)	0.000523	0.001694

Because there is no research to predict whether accidents will occur in the next hour on a highway section. So we can only compare the results of the model with the results of guessing based on statistical accident rates. That is to say, in our dataset, 37 out of 16688 samples are positive samples. Then for an unknown sample, the probability that it is a positive sample is $37/16688 \approx 0.002217$, and the probability that it is a negative sample is $16651/16688 \approx 0.997783$. The results of guessing in this way are shown in the Table 7.

Table 7. The classification result of guessing

Real category	forecast result	
	0(non-crash)	1(crash)
0(non-crash)	0.995571	0.002212
1(crash)	0.002212	4.91581e-6

In this way, the recall and precision rate are both 0.002217. We can find that for negative samples, the result of guessing is better

than the model. But for the recall and precision rate of positive samples, the result of the model is much higher than the result obtained by guessing. Since the cost of incorrectly predicting positive samples is much higher than the cost of incorrectly predicting negative samples, we pay more attention to the improvement of the recall rate and precision rate of positive samples. Consequently, the results of our model are significantly better than the results of guessing based on statistical accident rates.

6. CONCLUSION

The paper mainly describes the prediction of accidents on the expressway in the next hour through machine learning. It introduces the process of feature extraction and selection from traffic data, weather data, and road geometry data. The imbalanced sample is balanced by different methods. And finally, the model achieves better performance. The performance of the GBDT is: the precision rate is 0.01062, and the recall rate is 0.764253. The occurrence of traffic accidents is a small probability event and there are many factors that affect the accident. But the factors similar to the driver's driving state are still difficult to obtain. So the result of this method is already quite good.

7. ACKNOWLEDGMENTS

This work was supported by the key project of Science and Technology of Shanghai (Grant No. 18DZ1200200) and the National Key R&D Program of China (Grant No. 2018YFB0105101).

8. REFERENCES

[1] http://www.mps.gov.cn/n2255079/n5590589/n5747791/n5778470/c5776516/content.html

[2] Ren, H. et al. 2017. *A Deep Learning Approach to the Prediction of Short-term Traffic Accident Risk*. (2017).

[3] Yuan, Q. et al. 2017. Cluster and factor analysis on data of fatal traffic crashes in China. *International Conference on Transportation Information and Safety* (2017), 211-224.

[4] Chang, L.Y. et al. 2012. Analysis of Freeway Accident Frequency using Multivariate Adaptive Regression Splines. *Procedia Engineering*. 45, 2 (2012), 824-829.

[5] Gill, G. et al. 2017. Investigation of Roadway Geometric and Traffic Flow Factors for Vehicle Crashes Using Spatiotemporal Interaction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. XLII-2/W7, (2017), 1163-1166.

[6] Huang, Z. et al. 2017. Utilizing latent class logit model to predict crash risk. Ieee/acis *International Conference on Computer and Information Science* (2017), 161-165.

[7] Ahmed, M.M. and Abdel-Aty, M.A. 2012. The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction. *IEEE Transactions on Intelligent Transportation Systems*. 13, 2 (2012), 459-468.

[8] Sun, J. and Sun, J. 2016. Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model. *Iet Intelligent Transport Systems*. 10, 5 (2016), 331-337.

[9] Abdel-Aty, M. et al. 2004. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record Journal of the Transportation Research Board*. 1897, 1 (2004), 88-95.

[10] Sun, P. et al. 2017. Traffic crash prediction based on incremental learning algorithm. *IEEE International Conference on Big Data Analysis* (2017), 182-185.

[11] You, J. et al. 2017. Real-time crash prediction based on high definition monitoring systems. *IEEE International Conference on Intelligent Transportation Engineering* (2017), 208-211.

[12] Chen, Q. et al. 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (2016), 338-344.

[13] Abdel-Aty, M.A. and Pemmanaboina, R. 2006. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Transactions on Intelligent Transportation Systems*. 7, 2 (2006), 167-174.

[14] Xu, X. and Duan, L. 2017. Predicting Crash Rate Using Logistic Quantile Regression with Bounded Outcomes. *IEEE Access. PP*, 99 (2017), 1-1.

[15] Alkheder, S. et al. 2016. Severity Prediction of Traffic Accident Using an Artificial Neural Network. *Journal of Forecasting*. 36, 1 (2016).

[16] Najada, H.A. and Mahgoub, I. 2016. Big vehicular traffic Data mining: Towards accident and congestion prevention. *Wireless Communications and Mobile Computing Conference* (2016).

[17] Rodriguez-Galiano, V.F. et al. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *Isprs Journal of Photogrammetry & Remote Sensing*. 67, 1 (2012), 93-104.

[18] Wang, Y. et al. 2016. A mobile recommendation system based on logistic regression and Gradient Boosting Decision Trees. *International Joint Conference on Neural Networks* (2016), 1896-1902.

[19] Chen, T. and Guestrin, C. 2016. XGBoost:A Scalable Tree Boosting System. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 785-794.

[20] Ly, A. et al. 2018. Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*. 72, 1 (2018), 4-13.

[21] Dai, J. and Xu, Q. 2013. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Elsevier Science Publishers B. V.*

[22] Plackett, R.L. 1983. Karl Pearson and the Chi-Squared Test. *International Statistical Review*. 51, 1 (1983), 59-72.

[23] He, H. and Garcia, E.A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge & Data Engineering*. 21, 9 (2009), 1263-1284.

[24] Holte, R. et al. 1989. Concept Learning and the Problem of Small Disjuncts. *University of Texas at Austin*.

[25] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence* (1995), 1137-1143.

[26] Pedregosa, F. et al. 2013. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, 10 (2013), 2825-28