




Real-time crash prediction on freeways using data mining and emerging techniques

Jinming You¹  · Junhua Wang¹ · Jingqiu Guo¹

Received: 6 November 2016 / Revised: 30 March 2017 / Accepted: 31 March 2017 / Published online: 26 April 2017
© The Author(s) 2017. This article is an open access publication

Abstract Recent advances in intelligent transportation system allow traffic safety studies to extend from historic data-based analyses to real-time applications. The study presents a new method to predict crash likelihood with traffic data collected by discrete loop detectors as well as the web-crawl weather data. Matched case–control method and support vector machines (SVMs) technique were employed to identify the risk status. The adaptive synthetic over-sampling technique was applied to solve the imbalanced dataset issues. Random forest technique was applied to select the contributing factors and avoid the over-fitting issues. The results indicate that the SVMs classifier could successfully classify 76.32% of the crashes on the test dataset and 87.52% of the crashes on the overall dataset, which were relatively satisfactory compared with the results of the previous studies. Compared with the SVMs classifier without the data, the SVMs classifier with the web-crawl weather data increased the crash prediction accuracy by 1.32% and decreased the false alarm rate by 1.72%, showing the potential value of the massive web weather data. Mean impact value method was employed to evaluate the variable effects, and the results are identical with the results of most of previous studies. The emerging technique based on the discrete traffic data

and web weather data proves to be more applicable on real-time safety management on freeways.

Keywords Crash prediction · Real time · Discrete loop detectors · Web-crawl data · Support vector machines

1 Introduction

In recent years, dynamic safety management systems for freeways have been emerging. There is a growing trend to investigate the relationship between crash mechanism and traffic operating characteristics such as traffic state, road environment and weather condition. Several data mining techniques were utilized to integrate historical operating data on freeways into crash risk prediction models. Lee et al. [1] proposed the concept ‘crash precursor’ and developed an aggregate log-linear model with crash data of the upstream detectors, showing that high-speed variation and high traffic density play a key role in the crash occurrence. Various traffic datasets from different traffic surveillance systems were obtained to estimate the crash likelihood, such as dual loop detectors [1–5], single loop detector [6–8] and automatic vehicle identification system [9–11].

Although previous models have been proven to be capable of predicting crash likelihood in order to proactively improve traffic safety on freeways, various modeling techniques result in different prediction accuracies. For instance, Hossain et al. [4] developed a Bayesian belief network for real-time crash prediction and achieved a crash prediction accuracy 66% with the false alarming rate less than 20%. Ahmed et al. [9] employed a Bayesian updating approach and increased the crash prediction accuracy up to 72% with a relatively high false alarming rate 42.01%. Xu et al. [12] utilized the Bayesian inference method and the

✉ Jinming You
1210702@tongji.edu.cn
Junhua Wang
benwjh@163.com
Jingqiu Guo
hobartmc@gmail.com

¹ Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China

developed model achieved 36.8% crash prediction accuracy with a low false alarm rate of 5%. Moreover, these models show various limitations. Traditional generalized linear models such as logistic regression model [2, 5, 7, 13] could evaluate each contributing factors effectively and efficiently. Even though, some studies [14] find that a generalized linear model-based approach may lead to biased estimates when the independent variables demonstrate strong nonlinear features. The commonly used techniques in predicting real-time crash likelihood are Neural Network (NN) [3, 15, 16] and Support vector machines (SVMs) [17, 18].

However, the NN models work as a black box, and this strategy may raise over-fitting and local extremum issues [17]. Furthermore, the traditional method to select samples in case-control studies often applies a crash/non-crash ratio as 1:4. This method would create an imbalanced dataset. Mujalli et al. [19] investigated that data mining algorithms tend to produce lower prediction accuracy over the minority class in an imbalanced dataset. Meanwhile, the SVMs approach works better than NN when dealing with small sample size [17]. The SVMs solve the over-fitting issues by introducing kernel function and try to get the global optimal solution by solving the convex optimization problems. Despite the convincing results by utilizing SVMs to evaluate real-time crash risk [17], the SVMs have difficulty in dealing with the imbalanced dataset as well, and additional efforts should be made to optimize the function parameters of the model and preprocess the raw data to achieve a higher prediction accuracy. In general, with the requirement of high-quality traffic flow data, majority of existing models cannot be applied in other regions where limited detectors or surveillance devices are installed on the freeways, though transferability of the models has been validated [8]. Moreover, without consideration of the human factors and traffic patterns, the validity of the models cannot be verified.

In China, the use of traffic flow detectors is generally not as common as that in the USA and in Europe. The detectors or surveillance devices are often installed in the road sections with frequent congestion or between two interchanges. Due to neglecting the potential value of the data collected by these devices, the limited data have not been fully utilized. A previous study conducted by You et al. [20] shows that it is applicable to predict real-time crash with discrete ultrasonic detectors and it achieved a crash prediction accuracy of 61.9%.

Due to the limited data sources in many developing regions, proactive safety management cannot be implemented to alleviate crash risk. Multiple data sources of crash occurrence and traffic flow have shown promising effects on dynamic real-time crash prediction. Recent studies present great practicability to evaluate traffic

incidents by data mining of social media data on the web [21, 22] or the mobile phone usage data [23]. Schulz et al. [21] proposed a supervised learning technique and the trained SVMs models for specific event types. Results indicated that the method could detect multiple labels with a match rate of 84.35%. It provides a significant opportunity for researchers to investigate the crash mechanism with the diverse and complex datasets with the rapid development of information technology. However, these studies mostly focused on the incident detection and evaluation after occurrence, but seldom investigated proactive safety countermeasures to optimize the traffic condition prior to the incidents.

The weather factor significantly impacts the road safety, especially bad weather condition such as snow and heavy fog. However, it is common in China that real-time weather information is not available in the Department of Traffic Management due to the lack of weather detectors installed on the freeways. This work tries to crawl the history weather data from the Internet with the web-crawling method. Meanwhile, once the method proves to be valid and helpful to predict crash risk, the real-time dynamic weather data could be crawled from the website of the weather institute or department and it can be utilized to evaluate the real-time crash risk on certain freeway segments by developing a real-time traffic managing system.

The objective of this paper is to develop a comprehensive real-time crash prediction model based on the data mining and emerging techniques. This paper includes five sections. Section 2 discusses the data preparation and the web-crawling process. Section 3 explains the methods including the over-sampling technique, the SVMs modeling technique and Random Forest technique. The estimation results and further discussion are presented in Sect. 4.

2 Data collection and preparation

The test area in this study is a part of mainline on the G60 Freeway in Shanghai, China. The total length of the road segment is 48.7 km with 6–10 lanes (3–5 lanes for each direction). As factors causing crashes mainly include human factors, vehicle factors, road geometric factors, traffic factors and weather factors, it is necessary to obtain as much as related data as possible to explore the crash mechanism. However, human factors and vehicle factors cannot be detected in real time due to the lack of data feedback mechanism. Road geometric data can be obtained from the geographic information system such as Google Earth with a brief description. As the alignment of the study area is relatively flat, the descriptive data were of little use in this study. Historical traffic data and crash data are provided by the Department of Freeway Operation and

Management. The historical weather data are collected from the massive web data by web data crawling technique.

2.1 Traffic data

The primary traffic dataset includes data of nine pairs of loop detectors along the G60 Freeway. Five pairs are located on the mainline (as shown in Fig. 1), and four pairs are installed on the ramps. The average distance between the loop detectors on the mainline is approximately 6.6 km. Several traffic flow characteristics associated with crash occurrence were collected by the loop detectors, such as vehicle type, vehicle speed for corresponding vehicle type and vehicle occupancy every 20 seconds on each lane. The database also stores the device working state, data validity and timing record, etc.

The next step in data preparation is the data aggregation. Considering the random noise issue, Ahmed et al. [9] recommended to aggregate the raw traffic data to 5-min level. In this study, the extracted raw data were selected 5–10 min prior to crash occurrence time in order to avoid confusing pre- and post-crash conditions. The original 20-s traffic data, flow (Q_n), speed (V_n), occupancy (C_n) on each lane (L_n) are aggregated into a 5-min level.

2.2 Crash data

The target crash dataset includes 913 crashes that occurred in the study area between January 2014 and September

2015. Only rear-end crashes and sideswipe crashes (the total number is 551) were utilized in this study.

2.3 Weather data

In this study, a web data crawling strategy was conducted with a python script based on the PyCharm software. Web-crawling method can retrieve data faster and in greater depth. Many historical weather data pages hide in the deep or invisible web. These pages are typically only accessible by submitting dynamic queries of certain regions and a certain time to a background database. The data were extracted with regular expressions by parsing the structured HTML pages. The extracted data columns include time, region, temperature and weather. As the obtained weather data record the text information such as sunny, cloudy, rainy and snowy with different degrees, effect coding was applied to allow for nonlinear effects in the levels of attributes. For instance, sunny weather was coded as 0 and cloudy weather was coded as 1. Moreover, the weather data extracted by the web crawlers were stored into the Mysql database as a separate dataset for the following data emerging procedure.

2.4 Matched case-control method and data filtering

To eliminate the seasonal factors and day factors, the matched case-control method was utilized to avoid possible bias resulting from dissimilar traffic patterns on different days of the month and the week. A 4:1 control-case

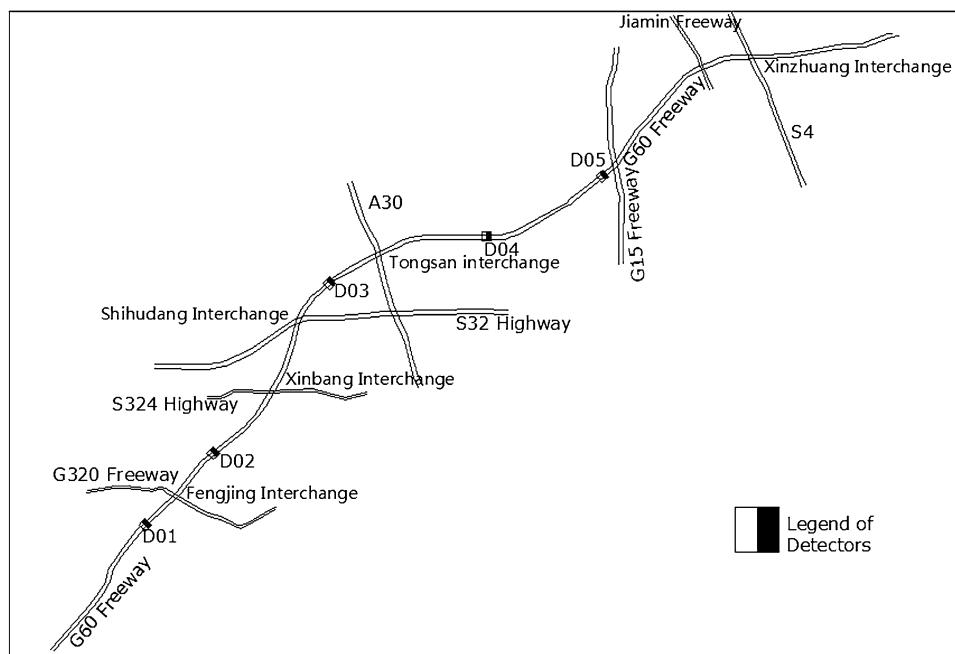


Fig. 1 Locations of the loop detectors on the mainline of G60 Freeway in Shanghai

Table 1 Symbols and variable description

Symbol	Variable description	Formulation*
Q	The total flow during 5 min	$\sum_{t=1}^{15} \sum_{n=1}^n Q_{nt}$
C_m	The mean value of occupancy of all lanes during 5 min	$E\left(\sum_{t=1}^{15} \sum_{n=1}^n C_{nt}\right)$
V_m	The mean value of speed of all lanes during 5 min	$\frac{\sum_{t=1}^{15} \sum_{n=1}^n (Q_{nt} \times V_{nt})}{\sum_{t=1}^{15} \sum_{n=1}^n Q_{nt}}$
Q_D	The accumulated standard deviation of flow within the lanes during 5 min	$\sum_{t=1}^{15} D_t(Q)$
C_D	The accumulated standard deviation of occupancy within the lanes during 5 min	$\sum_{t=1}^{15} D_t(C)$
V_D	The accumulated standard deviation of speed within the lanes during 5 min	$\sum_{t=1}^{15} D_t(V)$
Q_{DL}	Sum of the accumulated standard deviation of flow within 5 min for each lane	$\sum_{n=1}^n D_n(Q)$
C_{DL}	Sum of the accumulated standard deviation of occupancy within 5 min for each lane	$\sum_{n=1}^n D_n(C)$
V_{DL}	Sum of the accumulated standard deviation of speed within 5 min for each lane	$\sum_{n=1}^n D_n(V)$
Q_{MDL}	The maximum value of the accumulated standard deviation of flow within 5 min for each lane	$\text{Max}(D_n(Q))$
O_{MDL}	The maximum value of the accumulated standard deviation of occupancy within 5 min for each lane	$\text{Max}(D_n(C))$
V_{MDL}	The maximum value of the accumulated standard deviation of speed within 5 min for each lane	$\text{Max}(D_n(V))$
L_{cd}	The distance from the crash to the detector	$Lc - Ld$
W_{ea}	Weather condition code	Weather code

* In the formulations, 't' denotes time slice, 'n' denotes the number of lanes

Table 2 Summary statistics of variables

Variables	Average	SD	First quartile	Third quartile
Q	498.76	220.74	301.50	662.88
C_m	8.89	7.64	5.35	10.41
V_m	82.60	13.96	76.42	92.87
Q_D	65.44	18.52	54.28	74.24
C_D	36.84	23.86	28.78	37.18
V_D	409.80	303.57	119.65	737.29
Q_{DL}	51.07	21.38	34.87	64.95
C_{DL}	71.47	159.79	30.61	61.19
V_{DL}	11,40.21	11,39.31	271.71	17,55.27
Q_{MDL}	4.58	0.98	3.94	5.15
O_{MDL}	4.65	2.77	3.53	4.88
V_{MDL}	22.54	12.88	10.19	33.62
L_{cd}^*	-3.00	2.42	-5.03	-1.00
W_{ea}	2.87	1.73	2.00	4.00

* $L_{cd} > 0$ means that the crash locates upstream of the detectors on a certain segment, vise versa

ratio was recommended by some existing study [24]. For each specific crash case, four non-crash samples were selected. The four control samples were selected by the crash recorded time, respectively, 14 days before, 7 days before, 7 days after and 14 days after. The control samples with invalid traffic data would also be removed from the control dataset.

Due to the discrete loop detectors, traffic data from a freeway segment between one on-ramp and the next off-ramp were utilized to predict the crashes based on the hypothesis that the crash potential was highly relevant with

the traffic condition of a certain segment. Several variables were collected by loop detectors and may be relevant to the model. A pre-analysis was performed to minimize the number of potential explanatory variables. The variable definitions and formulations are listed in Table 1.

A few data filtering rules were applied to avoid possible bias. 'No data' or 'invalid data' are defined if there is no loop detector on certain segments on the G60 freeway, or the loop detectors fail. The dataset consists of the traffic flow data corresponding to each crash record. To summarize, the final crash dataset includes 138 observations and the final control dataset includes 549 non-crash samples. The statistics of variables are listed in Table 2.

3 Methodology and modeling technique

3.1 Over-sampling technique

The data mining algorithms often find it difficult to deal with imbalanced dataset. Under-sampling and over-sampling are data analysis techniques used to adjust the class distribution. As under-sampling technique often leads to the loss of the potential information of the samples, in this study over-sampling technique was utilized to create artificial samples. The adaptive synthetic sampling technique (ADASYN) is one of commonly used over-sampling techniques. It uses a weighed distribution for different minority class samples according to their levels of difficulty in learning. More synthetic data are generated for minority class samples that are harder to learn, thus

reducing the bias introduced by the imbalanced data distribution [25]. The final crash dataset includes 537 samples, and the ratio of crash/non-crash approaches 1:1.

3.2 Support vector machines

Support vector machines (SVMs) modeling technique has been widely applied in text classification, image recognition, voice recognition in machine learning. The method can often be employed for data with high dimensions and linearity problems. SVMs models have also been employed in some aspects of transportation field, such as traffic flow prediction, incident detection and crash frequency studies.

Based on structural risk minimization theory, SVMs generate an optimal classification hyperplane, which maximizes the margin between the hyperplane and the nearest samples of the classified sample categories and sets an equal margin. SVMs attempt to achieve the global optimal solutions, which help the classifiers obtain better generalization ability. Meanwhile, SVMs outperform other machine learning techniques when dealing with small sample size. The C-SVC (C-Support Vector Classification) model was employed in this study with the punitive coefficient C , and RBF (Radial Basis Function) kernel function was utilized to deal with the high-dimension variables. The RBF function form is shown as the following:

$$K(x_i, x) = e^{-\gamma \|x_i - x\|^2}, \quad (1)$$

where γ denotes the width parameter, x_i a point in the space, and x the central point.

The function of the decision function based on the RBF kernel function can be deduced as

$$f(x) = \text{sgn}\left(\sum_{i=1}^l y_i \alpha_i e^{-\gamma \|x_i - x\|^2} + b\right), \quad (2)$$

where y_i denotes the classified label, α_i the Lagrange multiplier, and b the intercept of the hyperplane.

Grid search method was employed to select the optimized parameters (C , γ) for the decision function. The modeling process was based on the LIBSVM tool developed by Chang and Lin [26] in the MATLAB software.

3.3 Random forest

Despite the advantages, SVMs models lack the capability of detecting the contributing factors and the use of all the variables as input makes the estimation inefficient. As suggested by Yu et al. [17], variable selection procedure is needed prior to the SVMs estimation. Meanwhile, by selecting variables it is able to solve the over-fitting issues. Hence, random forest was employed to select the contributing factors, as it is well known for selecting

significant contributing variables from a set of factors [27]. The strategy of random forest is that every tree is built with several factors, so a particular tree grows from a bootstrap aggregate sample, part of the cases is discarded and they will not be used in the development of the trees. The left-out cases are called Out-Of-Bag (OOB) data. The OOB data turn to validate the built trees with an unbiased error estimate as well as the important level estimations of variables. To test whether the attempted numbers of trees are sufficient to reach relatively stable results, the plot of OOB error rate against various tree numbers is developed. The optimal number of trees is the one having the minimum OOB error rate along with a nearly constant error rate. A wrapper MATLAB file interface to C code used in R package random forest [28] was employed to select the contributing factors. The tool provides the 'mean decrease in Gini index' method to select contributing variables. A higher magnitude implies a higher variable importance. Hassan et al. [27] chose several variables with higher scores (approximately 50% of the scores for all variables in total) than the remaining variables. In this paper, the variables which score higher than the mean score of all the variables were selected for the modeling process.

4 Results and conclusions

4.1 Contributing factors by random forest

Random Forest technique was employed to select the contributing factors before the first SVMs modeling process. The results are shown in Fig. 2.

Figure 2b indicates that the optimal number of trees is approximately 330 for the forest with 330 trees, which has the minimum OOB error rate along with a constant error rate 0.16. From Fig. 2a, six key factors (Q , C_M , Q_D etc.) can be selected by the mean magnitude value, which will be utilized in the SVMs modeling process afterward.

4.2 SVMs classifier performance

In the SVMs modeling process, the full sample was divided into two parts randomly, training data and test data with a ratio of 7:3. Three indexes were employed to evaluate the classifier performance, accuracy, True Positive Ratio (TPR) and False Positive Ratio (FPR). Additionally, the area under the Receiver Operating Characteristic (ROC) curve (AUC) was utilized to evaluate the classifier performance. The larger the AUC value is, the better the classifier performs.

To evaluate the potential value of the web-crawl weather data, a comparative study was conducted. First, the modeling process was conducted with the selected

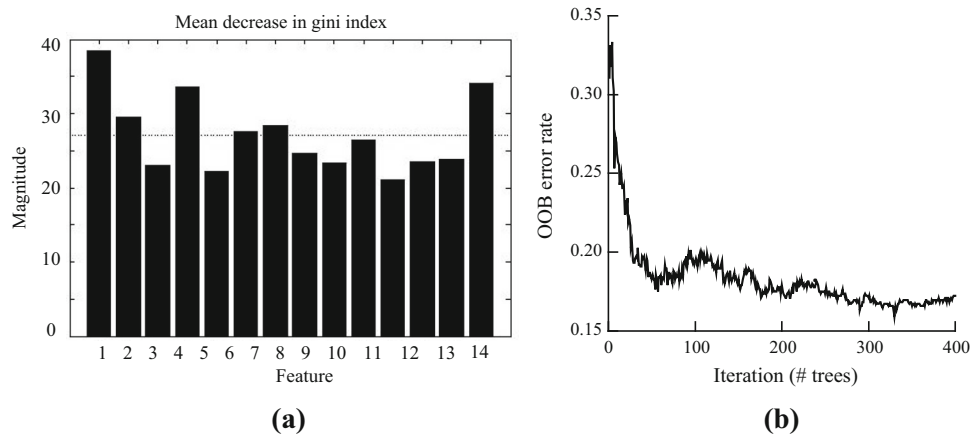


Fig. 2 Results of random forest. a Importance plots, b OOB error rate

Table 3 SVMs classifier performance with web-crawl weather data

	TPR (%)	FPR (%)	Accuracy (%)	AUC
Training dataset	91.95 (354/385)	23.47 (88/375)	84.34 (641/760)	0.8037
Test dataset	76.32 (116/152)	33.91 (59/174)	70.86 (231/326)	
Overall dataset	87.52 (470/537)	26.78 (147/549)	80.29 (872/1086)	

The proportion of the support vectors is 56.84% (432/760). The optimized parameters (C, γ) for the SVMs decision function value (12,417, 0.00,003)

Table 4 SVMs classifier performance without web-crawl weather data

	TPR (%)	FPR (%)	Accuracy (%)	AUC
Training dataset	88.05 (339/385)	23.47 (88/375)	82.37 (626/760)	0.7852
Test dataset	75.00 (114/152)	35.63 (62/174)	69.33 (226/326)	
Overall dataset	84.36 (453/537)	27.32 (150/549)	78.45 (852/1086)	

The proportion of the support vectors is 57.50% (437/760). The optimized parameters (C, γ) for the SVMs decision function value (12,417, 0.00,003)

variables including the weather data. Next, the modeling process was conducted without the weather data. The SVMs classifier performances of both methods are listed in Tables 3 and 4.

Results in Table 3 display that the SVM approach has a satisfying performance since the AUC value is relatively high and the TPR value of the overall dataset is 87.52%, which indicates better crash prediction performance. From the results of the test dataset, it can be concluded that despite the limited data from discrete loop detectors, an available approach for real-time crash prediction can be implemented by the SVMs modeling technique. The TPR value is 76.32%, which indicates more than two-thirds of the crashes can be detected with a reasonable FPR value 33.91%. The results are relatively satisfactory when compared with the results of the previous studies. Results of

several previous studies are listed in Table 5. The results show the potential value of traffic data from discrete loop detectors.

Results in Tables 3 and 4 show that the SVMs classifier with web-crawl weather data has a relatively better performance than that without web-crawl weather data. From the results of the test dataset, the TPR value increased 1.32% and the FPR value decreased 1.72% after the SVMs modeling with the added weather parameter. The same conclusion can be deduced as demonstrated by the AUC value. In general, the weather factors have a certain impact on the safety performance on the G60 Freeway. It illustrates the potential value of the web-crawl data as well. To implement the technique, massive real-time weather data can be crawled from the website of meteorological bureau. Thus, the comprehensive analysis on the real-time crash

Table 5 Comparison of the modeling results with the previous studies

Authors	TPR (%)	FPR (%)	Accuracy (%)	AUC
Abdel-Aty et al. [2]	69.4	47.2	55.7	–
Hossain et al. [4]	66	<20	78	–
Ahmed et al. [9]	75.93	45.83	58.52	–
Pande et al. [16]	57.1	28.8	70.6	–
Yu et al. [17]	–	–	–	0.75
This study	76.32	33.91	70.86	0.80

‘–’ Indicates that this index has not been mentioned

Table 6 MIV of each variable

Variables	Q	C_M	Q_D	V_D	Q_{DL}	W_{ea}
MIV	–0.050	–0.015	0.045	0.035	–0.011	0.006

risk on certain segments can be done for proactive safety management based on the emerging techniques of the traffic data and weather data.

4.3 Importance analysis for variable effects

SVMs was blamed for being a black-box technique as the variable effects cannot be evaluated. To unveil the variable effects, the mean importance value (MIV) method has been employed, which has been widely employed in neural network to evaluate the relative effects of the variables. The method calculates the value of the changed probability of each variable by changing the value of variables with 10% increase and 10% decrease. The results are listed in Table 6.

From Table 6, it can be concluded that the lower values of Q , C_M and Q_{DL} are probable to increase the crash risk and the higher values of Q_D , V_D and W_{ea} will lead to higher crash possibility as well. It seems that crash risk is more associated with the flow of the corresponding segment. In real traffic operation environment, the drivers tend to be more aggressive when the traffic is smooth, thus leading to higher rear-end or sideswipe crash risk. Meanwhile, several other conclusions can be made such as the crash risk increases when the deviation of flow and speed increases. These conclusions are identical with the results of most of previous studies.

The study presents a comprehensive SVM model with the traffic data collected by discrete loop detectors and the web-crawl weather data. It is common in China that real-time weather information is not available in the Department of Traffic Management due to the lack of weather detectors installed on the freeways. A new method was proposed to crawl the real-time weather data from the Internet, where massive weather forecasting data and real-

time weather data are available. Once the potential crash segment is identified in real time, measures for traffic guidance may be implemented to warn the drivers of the potential risk ahead.

Acknowledgements This research is supported by the National Natural Science Foundation (71301119) and the Shanghai Natural Science Foundation (12ZR1434100).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Lee C, Hellinga B, Saccomanno F (2002) Analysis of crash precursors on instrumented freeways. *Transp Res Rec J Transp Res Board* 1784(1):1–8
- Abdel-Aty M, Uddin N, Pande A (2004) Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transp Res Rec J Transp Res Board* 1897(1):88–95
- Abdel-Aty M, Pande A (2006) Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transp Res Rec J Transp Res Board* 1953(1):31–40
- 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(8):373–381
- Xu C, Liu P, Wang W et al (2012) Evaluation of the impacts of traffic states on crash risks on freeways. *Accid Anal Prev* 47(1):162–171
- Golob TF, Recker WW, Alvarez VM (2004) Freeway safety as a function of traffic flow. *Accid Anal Prev* 36(6):933–946
- Golob TF, Recker W, Pavlis Y (2008) Probabilistic models of freeway safety performance using traffic flow data as predictors. *Saf Sci* 46(9):1306–1333
- Shew C, Pande A, Nuworsoo C (2013) Transferability and robustness of real-time freeway crash risk assessment. *J Saf Res* 46(9):83–90
- Ahmed M, Abdel-Aty M, Yu R (2012) Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. *Transp Res Rec J Transp Res Board* 2280(1):60–67
- Ahmed M, Abdel-Aty M (2013) A data fusion framework for real-time risk assessment on freeways. *Transp Res Part C Emerg Technol* 26(1):203–213
- Shi Q, Abdel-Aty M, Yu R (2016) Multi-level Bayesian safety analysis with unprocessed Automatic Vehicle Identification data for an urban expressway. *Accid Anal Preven* 88(1):68–76
- Xu C, Wang W, Liu P, Li Z (2015) Calibration of crash risk models on freeways with limited real-time traffic data using Bayesian meta-analysis and Bayesian inference approach. *Accid Anal Prev* 85:207–218
- Kwak H, Kho S (2016) Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. *Accid Anal Prev* 88:9–19
- Lao Y, Zhang G, Wang Y et al (2014) Generalized nonlinear models for rear-end crash risk analysis. *Accid Anal Prev* 62(1):9–16

15. Abdel-Aty M, Pande A (2005) Identifying crash propensity using specific traffic speed conditions. *J Saf Res* 36(1):97–108
16. Pande A, Abdel-Aty M (2006) Assessment of freeway traffic parameters leading to lane-change related collisions. *Accid Anal Prev* 38(5):936–948
17. Yu R, Abdel-Aty M (2013) Utilizing support vector machine in real-time crash risk evaluation. *Accid Anal Prev* 51(1):252–259
18. Dong N, Huang H, Zheng L (2015) Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. *Accid Anal Prev* 82:192–198
19. Mujalli RO, López G, Garach L (2016) Bayes classifiers for imbalanced traffic accidents datasets. *Accid Anal Prev* 88:37–51
20. You J, Wang J, Fang S (2016) Real-time freeway crash prediction model by using single ultrasonic detector lane-level data: the 4th Chinese European workshop on functional pavement design. Delft University of Technology, Delft
21. Schulz A, Loza Mencía E, Schmidt B (2016) A rapid-prototyping framework for extracting small-scale incident-related information in microblogs: application of multi-label classification on tweets. *Inf Syst* 57:88–110
22. Gu Y, Qian ZS, Chen F (2016) From Twitter to detector: real-time traffic incident detection using social media data. *Transp Res Part C Emerg Technol* 67:321–342
23. Steenbruggen J, Tranos E, Rietveld P (2016) Traffic incidents in motorways: an empirical proposal for incident detection using data from mobile phone operators. *J Transp Geogr* 54:81–90
24. Ghosh M, Chen M (2002) Bayesian inference for matched case-control studies. *Sankhyā: Indian J Stat Ser B* 64(2):107–127
25. He H, Bai Y, Garcia EA et al (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. *IEEE Int Joint Conf Neural Netw (IEEE World Congress on Computational Intelligence)* 2008:1322–1328
26. Chang C, Lin C (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):1–27
27. Hassan HM, Abdel-Aty MA (2013) Predicting reduced visibility related crashes on freeways using real-time traffic flow data[J]. *J Saf Res* 45:29–36
28. Andy Liaw. Breiman and Cutler's Random Forests for Classification and Regression [EB/OL], October 2015. <http://cran.r-project.org/web/packages/randomForest/randomForest.pdf>. Accessed July 10, 2016