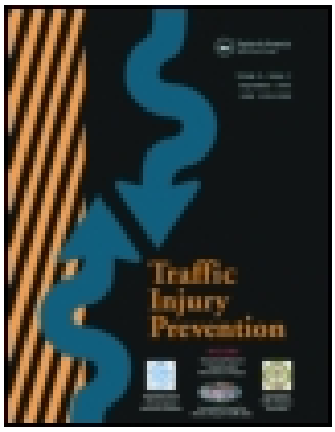
Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data

Peer-reviewed author version

PIRDAVANI, Ali; DE PAUW, Ellen; BRIJS, Tom; DANIELS, Stijn; Magis, Maarten; BELLEMANS, Tom & WETS, Geert (2015) Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data. In: Traffic Injury Prevention, 16 (8), p. 786-791.

DOI: 10.1080/15389588.2015.1017572
Handle: http://hdl.handle.net/1942/18445



Traffic Injury Prevention
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gcpi20

Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data

Ali Pirdavani[ab], Ellen De Pauw[a], Tom Brijs[a], Stijn Daniels[ab], Maarten Magis[a], Tom Bellemans[a] & Geert Wets[a]

[a] Transportation Research Institute (IMOB), School for Transportation Sciences, Hasselt University, Wetenschapspark 5, BE-3590 Diepenbeek, Belgium. (E-mail:), (E-mail:), (E-mail:) (E-mail:), (E-mail:), (E-mail:)

[b] Research Foundation - Flanders (FWO), Egmontstraat 5, B-1000 Brussels, Belgium

Accepted author version posted online: 20 Mar 2015.

To cite this article: Ali Pirdavani, Ellen De Pauw, Tom Brijs, Stijn Daniels, Maarten Magis, Tom Bellemans & Geert Wets (2015): Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data, Traffic Injury Prevention, DOI: 10.1080/15389588.2015.1017572

To link to this article: http://dx.doi.org/10.1080/15389588.2015.1017572

Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data

Ali Pirdavani[ÿ*] (E-mail: ali.pirdavani@uhasselt.be)

(Corresponding author) - Tel: +32 11 26 91 39

Fax: +32 11 26 91 99

Ellen De Pauw[ÿ] (E-mail: ellen.depauw@uhasselt.be)

Tom Brijs[ÿ] (E-mail: tom.brijs@uhasselt.be)

Stijn Daniels[ÿ*] (E-mail: stijn.daniels@uhasselt.be)

Maarten Magis[ÿ] (E-mail: maarten.magis@student.uhasselt.be)

Tom Bellemans[ÿ] (E-mail: tom.bellemans@uhasselt.be)

Geert Wets[ÿ] (E-mail: geert.wets@uhasselt.be)

ÿ Transportation Research Institute (IMOB)

School for Transportation Sciences

Hasselt University

Wetenschapspark 5

BE-3590 Diepenbeek

Belgium

∗ Research Foundation ï Flanders (FWO)

Egmontstraat 5

ACCEPTED MANUSCRIPT

B-1000 Brussels

Belgium

ABSTRACT

Objectives: There is a growing trend in development and application of real-time crash risk prediction models within dynamic safety management systems. These real-time crash risk prediction models are constructed by associating crash data with the real-time traffic surveillance data (e.g. collected by loop detectors). The main objective of this paper is to develop a real-time risk model that will potentially be utilized within traffic management systems. This model aims to predict the likelihood of crash occurrence on motorways.

Methods: In this study, the potential prediction variables are confined to traffic related characteristics. Given that the dependent variable (i.e. traffic safety condition) is dichotomous (i.e. ñno-crashò or ñcrashò), a rule-based approach is considered for model development. The performance of rule-based classifiers is further compared with the more conventional techniques like binary logistic regression and decision trees. The crash and traffic data used in this study were collected between June 2009 and December 2011 on a part of the E313 motorway in Belgium between Geel-East and Antwerp-East exits, on the direction towards Antwerp.

Results: The results of analysis show that several traffic flow characteristics such as traffic volume, average speed and standard deviation of speed at the upstream loop detector station, and the difference in average speed on upstream and downstream loop detector stations significantly contribute to the crash occurrence prediction. The final chosen classifier is able to predict 70% of

ACCEPTED MANUSCRIPT

crash occasions accurately while it correctly predicts 90% of no-crash instances, indicating a 10% false alarm rate.

Conclusions: The findings of this study can be used to predict the likelihood of crash occurrence on motorways within dynamic safety management systems.

KEYWORDS

Rule-based classifiers; Real-Time Crash Risk Prediction; Dynamic Safety Management Systems; Traffic Surveillance Data.

ACCEPTED MANUSCRIPT

INTRODUCTION

Over the past decade, proactive traffic management systems have increasingly attracted researchers and policy makersô attention. These systems, which are mainly implemented on motorways, are meant to improve traffic safety by smoothing the traffic flow. In such dynamic safety management systems, real-time crash prediction models are major elements. These models estimate the likelihood of crash occurrence by using real-time traffic flow characteristics that are collected by traffic surveillance systems such as loop detectors. These models can dynamically evaluate the traffic safety condition of motorways and identify occasions that would potentially lead to crash occurrence. When a risky condition is identified, proactive safety countermeasures can be implemented to alleviate crash occurrence possibility. Among others, variable speed limits (Lee et al. 2004, 2006c; Abdel-Aty et al. 2006a; Jo et al. 2012; Li et al. 2013), ramp metering (Abdel-Aty et al. 2006b; Lee et al. 2006b) and intelligent speed adaptation (Chen et al. 2002; Carsten and Tate 2005; Servin et al. 2008; Lai et al. 2012) are effective measures that are known to improve traffic safety on motorways. These measures are intended to smoothen the traffic flow by increasing average time headways, reducing speed variation and subsequently improving traffic safety. For instance, safety benefits will be gained by simultaneously lowering the speed upstream and increasing the speed downstream of the location where the real-time crash prediction models identify an unsafe condition.

Over the last decade, several studies were conducted where real-time crash prediction models were developed by associating real-time traffic flow data with crash data (Lee et al. 2002, 2003; Chang and Chen 2005; Oh et al. 2005; Abdel-Aty et al. 2006a; Lee et al. 2006c; Oh et al. 2006; Pande and Abdel-Aty 2006a; Abdel-Aty et al. 2007; Zheng et al. 2010; Pande et al.

2011; Xu et al. 2012; Abdel-Aty et al. 2012; Xu et al. 2013; Ahmed and Abdel-Aty 2013; Xu et al. 2014). In these studies, several traffic flow related attributes, including, but not limited to, ñabsolute speedò, ñvariation of speedò, ñspeed difference between upstream and downstream loop detectorsò, ñtraffic densityò, ñaverage occupancyò, and etc. that would be linked with crash occurrence on motorways are inverstigated. Due to the dichotomous nature of the dependent variable (i.e. ñno-crashò or ñcrashò), different approaches such as statistical modeling (e.g. logistic regression modeling technique) or machine learning (e.g. decision tree algorithm) have been commonly applied in developing real-time crash prediction models.

Besides application of the logistic regression technique (Abdel-Aty et al. 2004, 2005; Brijs et al. 2006; Hourdos et al. 2006; Lee et al. 2006a; Golob et al. 2008; Xu et al. 2013), other approaches were adopted to associate crash likelihood and real-time traffic flow characteristics. Oh et al. (2005) developed a Bayesian model, Abdel-Aty and Pande (2005) applied a probabilistic neural network model, Chang et al. (2005) employed a classification and regression tree technique, Pande et al. (2006b) also developed a crash risk prediction model based on the classification tree and neural network while a random forests technique was used by Pande et al. (2011).

Despite the considerable application of different machine learning techniques (e.g. decision trees) in the traffic safety domain (Abdel-Aty and Haleem 2011; Chung 2013; Yan and Radwan 2006; Yu and Abdel-Aty 2013), rule-based classifiers have rarely been used in traffic safety analyses, particularly in the context of real-time crash prediction modeling. These classifiers have certain advantages; they are easy to generate and interpret, straightforward in classifying new instances, and are simpler but as highly accurate and expressive as decision

trees. Perhaps the most important advantage for practitioners who would use the models in traffic management systems is the easiness to interpret these classifier results. Rule-based classifiers are expressed in a natural language format of IF-THEN rules and, therefore, are easily understood by practitioners. Therefore, the primary objective of this research is set to develop real-time crash prediction models following a rule-based approach and based on the traffic flow characteristics that are collected by double loop detector stations. To this end, different rule-based classifiers are tested for model development. Moreover, the performance of the selected rule-based classifier will be evaluated by comparing the results with other conventional modeling techniques. This model development effort is considered as the first step in the realization of a proactive highway safety management system. When the developed models predict the crash conditions appropriately, the transportation authorities are enabled to implement crash prediction algorithms within available dynamic traffic management systems in order to improve the traffic safety conditions of motorways.

METHODS

Data Preparation

The study area in this research is a part of the European motorway E313 in Belgium between Geel-East and Antwerp-East exits, on the direction towards Antwerp. The total length of the studied road segment is about 42.5 km. The primary crash dataset includes all crashes that occurred in the study area between June 2009 and December 2011. Due to the necessity of having precise crash occurrence time (i.e. in the order of 1 minute) and since the crash data (gathered by the police) were obtained from a different authority than the one which provides traffic flow data (i.e. Ministry of Mobility and Public Works, Flemish Traffic Center), the accuracy of crash occurrence time was double checked by matching these two datasets. The

detailed process of verifying the authenticity of data is described in Appendix A (see online supplement). The next step in data preparation is the data aggregation. The 1-minute raw data are commonly noisy data and, therefore, the primary raw data could be combined into - for instance - 5-minute level (Abdel-Aty et al. 2012). The extracted raw data were then aggregated to three different aggregation levels; namely 5-minute, 10-minute and 15-minute intervals, all covering 5 minutes prior to crash occurrence time and earlier. All the three aggregation levels will be investigated to identify the best level that will result in better crash prediction.

In the next step and for preparing the complete dataset, for each crash record four non-crash cases were also chosen from the same location, the same day of the week and the same time given the condition that no crash had occurred within a period of one hour around the targeted time. To eliminate the seasonal effects and to avoid possible bias resulting from dissimilar traffic patterns on different days of the week, non-crash cases were extracted from exactly one and two weeks before and after the crash occurrence time. All non-crash cases matched the condition that no crash had occurred within a one-hour period around the targeted time. This results in utilizing traffic flow data for each location and for the following times:

- Exactly two weeks before crash occurrence
- Exactly one week before crash occurrence
- Crash occurrence
- Exactly one week after crash occurrence
- Exactly two weeks after crash occurrence

To summarize, the final dataset consists of the traffic flow data corresponding to each crash record and four matched non-crash records. This dataset includes 390 observations (i.e. 78 crashes and 312 non-crash records).

Variable Selection

As can be observed from the literature, there exist several traffic flow characteristics that are potentially associated with crash occurrence, such as speed, speed variation, traffic density and occupancy. These traffic flow variables are generally collected by double loop detectors. However, in order to clean the available data, a pre-analysis is performed to identify the most relevant explanatory variables. To this end, firstly the non-parametric Spearmanôs correlation test was performed to investigate which variables have significant correlations with the dependent variable (i.e. safety condition). Hence, all uncorrelated variables at the 90% confidence level (e.g. occupancy and average speed on the first downstream loop detector station) are removed from the dataset. Moreover, due to the existence of inter-relationship among the remaining variables, the variance inflation factor (VIF) test (Kutner et al. 2004) is performed to ensure that no collinearity exists among the remaining explanatory variables (e.g. all variables at the US2 and US3 stations are eliminated). Final variables that will be considered for model development have VIF values of less than five, as recommended in the literature (Kutner et al. 2004). These variables are listed in Table A1 of Appendix B together with their descriptive statistics (see online supplement).

Methodological Framework

This study aims to predict the traffic safety condition of motorways by associating crash data with traffic flow characteristics that are collected by traffic loop detectors. Due to the dichotomous nature of the dependent variable Y (i.e. dependent variable can only take two

values; Y=1 for crash condition and Y=0 for no-crash condition), application of different classifiers is possible. In the context of this study, we aim to develop different rule learning schemes such as C4.5 and its derivative method called PART (stands for PArtial Decision Tree), and RIPPER scheme (stands for Repeated Incremental Pruning to Produce Error Reduction) (Frank and Witten 1998; Witten et al. 2011) and further evaluate their prediction performance against more conventional statistical models like binary logistic regression. An elaborate description of these methods can be found in Appendix C (see online supplement).

Model Validation Technique

In this study the k-fold cross-validation technique is employed to validate the accuracy and ensure the robustness of the prediction models (Olson and Delen 2008). The k-fold cross-validation technique minimizes the possible bias caused by the random sampling of the training and testing datasets. In the k-fold cross-validation technique, the complete dataset is equally divided into k subsets. In each step of the model development, one subset is considered as the validation dataset while the other k-1 subsets are used as the training dataset. The cross-validation process is then repeated k times, where each of the subsamples will be used only once as the validation data. In this study, a 10-fold cross-validation approach is followed.

Model Development

For developing the real-time prediction models, the final variables (see Table A1 in online supplement) were considered separately for each time interval. In other words, different modeling frameworks including rule-based classifiers, decision trees and the binary logistic regression models were employed using explanatory variables of each time interval separately. For model development, the Waikato Environment for Knowledge Analysis (WEKA) software is utilized (Hall et al. 2009). WEKA is an open source software with a collection of machine

learning algorithms. WEKA explorer consists of several useful tools including data pre-processing, classification, clustering, association rules, attribute selection, and visualization.

RESULTS

Model Performance Evaluation

The classifications of results are shown in Table 1. Table 1 consists of several boxes that are commonly referred to as contingency table or confusion matrix. In a binary prediction problem, the outcomes are labeled either as positive or negative. In the context of this study and since the ultimate objective is to predict crash conditions, a positive outcome is set to be a crash condition while predicting a no-crash condition is displayed as a negative outcome. Hence, there will be four possible outcomes by which the prediction accuracy of the model can be evaluated. If the outcome of a prediction is positive and the observed value is also positive, then this condition is considered as true positive (TP) while if the observed value is negative then it is stated to be a false positive (FP). Similarly, a true negative (TN) will occur when both the prediction outcome and the observed values are negative, and false negative (FN) is when the prediction outcome is negative while the observed value is positive. Model outcomes that are labeled with this convention are shown in Table 1. Numbers reported in Table 1 represent each modelôs performance based on the 10-fold cross-validation results.

<< Please insert Table 1 here >>

While the options to evaluate the outcome of a confusion matrix by a single number are very limited, Matthews correlation coefficient (MCC) is known to be a good measurement to reveal the quality of binary classifiers (Powers 2011). The MCC takes into account true and false positives and negatives and is considered as a balanced measure. This is an important advantage of this measurement in comparison to other measures (e.g. accuracy), since they are not very

useful when they correspond to datasets consisting of two classes of very different sizes. The MCC is a correlation coefficient between the observed and predicted classifications. Hence, the larger the MCC measure is, the better the prediction performance of a classifier will be.

The MCC is formulated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Moreover, it is important to distinguish between different costs imposed by different false or true predictions. In many cases optimizing the classification rate without considering the costs of errors often leads to misleading interpretations. A very well-known example of this would be applicable in loan decisions. Evidently, the cost of lending to a defaulter is far greater than the lost-business cost of refusing a loan to a non-defaulter (Witten et al. 2011). The same rule is applicable in the context of this study where the cost of predicting crash occasions as no-crash (i.e. false negatives) is greater than predicting no-crash events as crash (i.e. false positives). Hence, a classifier that minimizes false negatives is considered as a good classifier. However, it is quite possible that by choosing a model which minimizes false negatives, false positive rate (also referred to as false alarm rate and is formulated as FP/(FP+TN)) will be increased. In the context of this research, it would be beneficial to minimize the false negatives, although we might increase the false alarm rate and consequently predict more crash occasions (even if they are not observed as crash occasions). In practice and in the case of predicting a crash condition, different countermeasures can be implemented (e.g. intelligent speed adaptation or variable speed limits) to avoid a potential crash occurrence. Although this crash condition might not be correctly predicted (i.e. it might not eventually lead to crash occurrence), implementation of

safety countermeasures would be anyhow beneficial because at least they are expected to reduce traffic flow disturbance. To conclude, model selection and performance evaluation should be carried out based on both MCC measure and the condition of minimizing the false negatives.

The results reported in Table 1 reveal that for all model types, classifiers that are developed on the 5-minute interval dataset are better in comparison to other classifiers. This can be explained by the fact that the 5-minute dataset contains information of a closer proximity to crash occurrence time and, therefore, is more representative than the other two datasets. Comparing different model results for the 5-minute interval reveals that the PART classifier outperforms the other models by producing minimum false negatives and maximum MCC measure. The final chosen classifier correctly predicts 70% of crash and 90% of no-crash instances, signifying a better predictive performance compared with many of the real-time crash risk models in previous studies reported by Xu et al. (2013). Moreover, the false alarm rate produced by this classifier is around 10% which is significantly lower in comparison to the results of previous studies reported by Xu et al. (2013).

Model Results

Based on the discussion of the previous section, the PART classifier for the 5-minute interval is selected as the best prediction model. There are three rules derived from this classifier that are as follows:

1) IF SP_US1 > 78.8 THEN: Crash=0 (247.0/8.0)
2) IF TV_US1 > 17.8 AND
   STDEV_SP_US1 Ò22.8 AND
   Diff_SP_US1-DS1 Ò47.6 THEN: Crash=0 (43.0/3.0)

3) ELSE: Crash=1 (100.0/33.0)

The numbers in the brackets stand for coverage/errors in the training dataset (i.e. number of covered instances/number of misclassified instances), 0 stands for no-crash and 1 stands for crash conditions. As it was mentioned earlier, PART classifier is a post-pruned partial decision tree based on the C4.5 scheme. To show the superiority of this classifier over the conventional decision tree, the corresponding decision tree for the 5-minute interval is depicted in Figure 1. As can be seen in Figure 1, three rule sets of the PART classifier are indeed parts of the complete decision tree. These rule sets classify instances simpler and with a high accuracy level.

<< Please insert Figure 1 here >>

To provide a better comparison between different classifiers, the other two classifiersô formulations (i.e. RIPPER classifiers and the binary logistic regression model) are reported in Appendix D (see online supplement).

Interpreting the results is a bit different for each method. In statistical formulation, positive and negative signs of estimate coefficients correspond with direct and inverse relationship with the probability of crash occurrence, respectively. In the rule-based methods, there are the thresholds that define these associations. For instance, in the rules of the PART classifier, STDEV_SP_US1 values smaller than 22.8 contribute to probability of no-crash condition. All estimate conditions and signs are in line with intuitive expectations. As can be seen from the results of all model types, speed and traffic volume at the upstream loop detector station have negative signs. Higher speed and traffic volume correspond with lower density that implies a lower likelihood of crash occurrence. Moreover, standard deviation of speed upstream the crash location has a positive association with crash occurrence. This implies that fluctuations

in speed at a location will potentially increase the risk of crash occurrence somewhere downstream. Another significant variable is the difference in average speeds at upstream and downstream stations. This interesting result reveals the importance of speed and its derivative variables in crash likelihood prediction. This signifies that if the difference in average travel speed of two consecutive locations becomes greater, there will be a higher probability of crash occurrence in between those two locations.

DISCUSSIONS

The main objective of this study was to evaluate the traffic safety conditions on motorways by means of traffic flow characteristics collected by loop detector stations. Various variables such as traffic volume, occupancy, average speed, standard deviation of speed, difference between average speeds on two consecutive loop detector stations were among the potential predictor variables that were considered for model development. The raw data were at 1-minute level of aggregation, which would potentially bias the results due to their random noise. To avoid this problem, the primary data were aggregated into three different levels, namely 5-minute, 10-minute and 15-minute intervals. This enables us to identify the best level of aggregation that will result in better crash prediction accuracy. All of these three aggregation level data were used to develop individual prediction models.

The results of analysis showed that all 5-minute models outperform the other two intervalsô models by means of greater MCC measures. Among different classifiers, the PART classifier appears to be the best performing classifier by correctly predicting 70% and 90% of crash and no-crash instances respectively. The false alarm rate (i.e. false positive rate in this study) that resulted from the 5-minute PART classifier is 10%, which is significantly lower than

false alarm rates reported in the literature. This classifier predicts the traffic safety condition in a simpler way and yet more accurate compared with other conventional methods such as binary logistic regression model or decision tree.

An important prerequisite in developing a successful prediction model is the existence of good quality data. Among others, presence of missing values is considered as a crucial threat for classification assignments. In this regard, rule-based classifiers and decision trees are superior to statistical models (e.g. binary logistic regression model). Simply put, instances with missing values are dealt with by rules that involve other not-missing attributes. In general, classification rules have several advantages such as easy to generate and interpret, easy and quick to classify new instances, as highly accurate and expressive as decision trees and etc. PART classifier, in particular, is superior to other rule-learning schemes because of its simplicity and equal or even higher level of performance.

The first requirement in realization of a proactive highway safety management system is having accurate real-time risk prediction models. The performance of the developed prediction model in this study (i.e. the PART classifier based on 5-minute interval data) appropriately fulfills this condition by predicting an acceptable rate of crash and no-crash conditions. Flemish government aims to use these prediction models to warn drivers of the possible upcoming dangerous situations through variable message signs. Having said that, there is always room for improving the accuracy level of developed model by enriching the crash and traffic data. This would improve model accuracy and robustness and subsequently would increase the acceptability of the prediction model by traffic authorities who are willing to utilize these models in their dynamic safety management systems. Another extension for future research would be the

transferability check of the developed prediction model. To this end, the model should be validated against crash and traffic data collected from various motorways. This should be carried out to ensure that the final model is able to predict traffic safety conditions on any motorway correctly, under the same jurisdiction and with the same infrastructural basis (e.g. speed limit, traffic volume order or geometric conditions).

ACKNOWLEDGEMENTS

The authors thank the Flemish Traffic Centre for their support during the data gathering and for their collaborated efforts in conducting the survey. This research was carried out within the framework of the Policy Research Centre on Traffic Safety with the support of the Flemish government and was partly supported by a grant from the Research Foundation Flanders (FWO). The content of this paper is the sole responsibility of the authors.

REFERENCES

Abdel-Aty M, Hassan HM, Ahmed M, Al-Ghamdi AS. Real-time prediction of visibility related crashes. *Transp Res Part C Emerg Technol.* 2012; 24: 288ï 298.

Abdel-Aty M, Dilmore J, Dhindsa A. Evaluation of variable speed limits for real-time freeway safety improvement. *Accid Anal Prev.* 2006a; 38: 335ï 345.

Abdel-Aty M, Haleem K. Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accid Anal Prev.* 2011; 43: 461ï 470.

Abdel-Aty M, Haleem K, Cunningham R, Gayah V. Application of variable speed limits and ramp metering to improve safety and efficiency of freeways. Paper presented at: The 2nd International Symposium on Freeway and Tollway Operations. June 21-24, 2009; Honolulu, Hawaii.

Abdel-Aty M, Pande A. Identifying crash propensity using specific traffic speed conditions. *J Safety Res*. 2005; 36: 97ï 108.

Abdel-Aty M, Pande A, Lee C, Gayah V, Santos CD. Crash Risk Assessment Using Intelligent Transportation Systems Data and Real-Time Intervention Strategies to Improve Safety on Freeways. *J Intell Transp Syst*. 2007; 11: 107ï 120.

Abdel-Aty M, Uddin N, Pande A. Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. *Transp Res Rec J Transp Res Board.* 2005; 1908: 51ï 58.

Abdel-Aty M, Uddin N, Pande A, Abdalla F, Hsia L. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transp Res Rec J Transp Res Board.* 2004; 1897: 88ï 95.

Ahmed M, Abdel-Aty M. A data fusion framework for real-time risk assessment on freeways. *Transp Res Part C Emerg Technol.* 2004; 26: 203ï 213.

Brijs T, Wets G, Krimpenfort R, Offermans C. Impact of Hourly Measured Speed on Accident Risk in the Netherlands: Results from Exploratory Study Using Geographic Information Systems. *Transp Res Rec J Transp Res Board.* 2006; 1972: 85ï 93.

Carsten OMJ, Tate FN. Intelligent speed adaptation: accident savings and costï benefit analysis. *Accid Anal Prev.* 2005; 37: 407ï 416.

Chang LY, Chen WC. Data mining of tree-based models to analyze freeway accident frequency. *J Safety Res.* 2005; 36: 365ï 375.

Chen G, Meckle W, Wilson J. Speed and safety effect of photo radar enforcement on a highway corridor in British Columbia. *Accid Anal Prev.* 2002; 34: 129ï 138.

Chung YS. Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. *Accid Anal Prev.* 2013; 61: 107ï 118.

Frank E. Witten IH1998. Generating Accurate Rule Sets Without Global Optimization. *Morgan Kaufmann*; 1998: 144ï 151.

Fürnkranz J. Separate-and-Conquer Rule Learning. *Artif Intell Rev.* 1999; 13: 3ï 54.

Golob TF, Recker W, Pavlis Y. Probabilistic models of freeway safety performance using traffic flow data as predictors. *Saf Sci.* 2008; 46: 1306ï 1333.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* 2009; 11: 10ï 18.

Hosmer Jr. DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. John Wiley & Sons; 2013

Hourdos J, Garg V, Michalopoulos P, Davis G. Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations. *Transp Res Rec J Transp Res Board.* 2006; 1968: 83ï 91.

Jo Y, Yoon K, Jung I. Variable Speed Limit to Improve Safety near Traffic Congestion on Urban Freeways. *Int J Fuzzy Syst.* 2012: 14: 278ï 288.

Kutner MH, Nachtsheim C, Neter J. *Applied linear regression models.* McGraw-Hill/Irwin, Boston; New York; 2004

Lai F, Carsten O, Tate F. How much benefit does Intelligent Speed Adaptation deliver: An analysis of its potential contribution to safety and environment. *Accid Anal Prev.* 20012; 48: 63ï 72.

Lee C, Abdel-Aty M, Hsia L. Potential Real-Time Indicators of Sideswipe Crashes on Freeways. *Transp Res Rec J Transp Res Board.* 2006a; 1953: 41ï 49.

Lee C, Hellinga B, Ozbay K. Quantifying effects of ramp metering on freeway safety. *Accid Anal Prev.* 2006b; 38: 279ï 288.

Lee C, Hellinga B, Saccomanno F. Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. *Transp Res Rec J Transp Res Board.* 2003; 1840: 67ï 77.

Lee C, Hellinga B, Saccomanno F. Assessing Safety Benefits of Variable Speed Limits. *Transp Res Rec J Transp Res Board.* 2004; 1897: 183ï 190.

Lee C, Hellinga B, Saccomanno F. Evaluation of variable speed limits to improve traffic safety. *Transp Res Part C Emerg Technol.* 2006c; 14: 213ï 228.

Lee C, Saccomanno F, Hellinga B. Analysis of Crash Precursors on Instrumented Freeways. *Transp Res Rec J Transp Res Board.* 2002; 1784: 1ï 8.

Li Z, Wang W, Chen R, Liu P, Xu C. Evaluation of the Impacts of Speed Variation on Freeway Traffic Collisions in Various Traffic States. *Traffic Inj Prev.* 2013; 14: 861ï 866.

Oh C, Park S, Ritchie SG. A method for identifying rear-end collision risks using inductive loop detectors. *Accid Anal Prev.* 2006; 38: 295ï 301.

Oh JS, Oh C, Ritchie SG, Chang, M. Real-Time Estimation of Accident Likelihood for Safety Enhancement. *J Transp Eng.* 2005; 131: 358ï 363.

Olson DL, Delen D. *Advanced data mining techniques.* Springer, Berlin; 2008

Pande A, Abdel-Aty M. Comprehensive Analysis of the Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways. *Transp Res Rec J Transp Res Board.* 2006a; 1953: 31ï 40.

Pande A, Abdel-Aty M. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accid Anal Prev.* 2006b; 38: 936ï 948.

Pande A, Das A, Abdel-Aty M, Hassan H. Estimation of Real-Time Crash Risk. *Transp Res Rec J Transp Res Board.* 2011; 2237: 60ï 66.

Powers DM. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *J Mach Learn Technol.* 2011; 2: 37ï 63.

Servin O, Boriboonsomsin K, Barth MJ. Preliminary Design of Speed Control Strategies in Dynamic Intelligent Speed Adaptation System for Freeways. Paper presented at: the Transportation Research Board 87th Annual Meeting. January 13-17, 2008; Washington D.C., USA.

Witten IH, Frank E, Hall MA. *Data mining practical machine learning tools and techniques, third edition.* Morgan Kaufmann Publishers, Burlington, Mass; 2011

Xu C, Liu P, Wang W, Li Z. Evaluation of the impacts of traffic states on crash risks on freeways. *Accid Anal Prev.* 2012; 47: 162ï 171.

Xu C, Tarko AP, Wang W, Liu P. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accid Anal Prev.* 2013; 57: 30ï 39.

Xu C, Wang W, Liu P, Zhang F. Development of a Real-Time Crash Risk Prediction Model Incorporating the Various Crash Mechanisms Across Different Traffic States. *Traffic Inj Prev.* 2014; 16: 28ï 35.

Yan X, Radwan E. Analyses of Rear-End Crashes Based on Classification Tree Models. *Traffic Inj Prev.* 2006; 7: 276ï 282.

Yu R, Abdel-Aty M. Utilizing support vector machine in real-time crash risk evaluation. *Accid Anal Prev.* 2013; 51: 252ï 259.

Zheng Z, Ahn S, Monsere CM. Impact of traffic oscillations on freeway crash occurrences. *Accid Anal Prev.* 2010; 42: 626ï 636.

Table 1. The Confusion Matrix of Developed Models

			5-minute interval		10-minute interval		15-minute interval	
			Predicted traffic safety condition					
			Crash	No-crash	Crash	No-crash	Crash	No-crash
Binary Logistic Regression	Observed traffic safety condition	Crash	45 (TP)	33 (FN)	39 (TP)	39 (FN)	37 (TP)	41 (FN)
		No-crash	25 (FP)	287 (TN)	21 (FP)	291 (TN)	20 (FP)	292 (TN)
		MCC	0.518		0.480		0.465	
Decision tree (C4.5)		Crash	49 (TP)	29 (FN)	51 (TP)	27 (FN)	14 (TP)	64 (FN)
		No-crash	30 (FP)	282 (TN)	44 (FP)	268 (TN)	5 (FP)	307 (TN)
		MCC	0.530		0.478		0.304	
RIPPER		Crash	50 (TP)	28 (FN)	42 (TP)	36 (FN)	34 (TP)	44 (FN)
		No-crash	33 (FP)	279 (TN)	26 (FP)	286 (TN)	29 (FP)	283 (TN)
		MCC	0.523		0.480		0.373	
PART		Crash	56 (TP)	22 (FN)	61 (TP)	17 (FN)	15 (TP)	63 (FN)
		No-crash	32 (FP)	280 (TN)	49 (FP)	263 (TN)	5 (FP)	307 (TN)
		MCC	0.589		0.556		0.320	

TP: true positive

FP: false positive

FN: false negative

TN: true negative

MCC: Matthews correlation coefficient

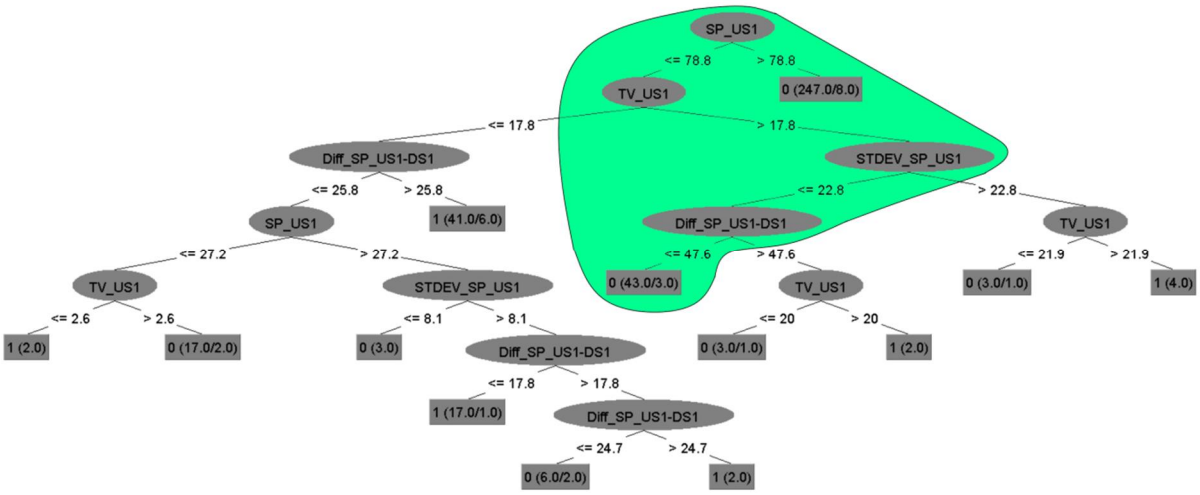


Fig. 1. Decision tree algorithm based on C4.5 scheme and the derivable PART rule sets for 5-minute interval.