Accident Analysis and Prevention 51 (2013) 252–259

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Utilizing support vector machine in real-time crash risk evaluation

Rongjie Yu*, Mohamed Abdel-Aty

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32826-2450, United States

ARTICLE INFO

Article history:
Received 24 September 2012
Received in revised form 5 November 2012
Accepted 29 November 2012

Keywords:
Support vector machine model
Bayesian logistic regression
Real-time crash risk evaluation
Mountainous freeway safety

ABSTRACT

Real-time crash risk evaluation models will likely play a key role in Active Traffic Management (ATM). Models have been developed to predict crash occurrence in order to proactively improve traffic safety. Previous real-time crash risk evaluation studies mainly employed logistic regression and neural network models which have a linear functional form and over-fitting drawbacks, respectively. Moreover, these studies mostly focused on estimating the models but barely investigated the models' predictive abilities. In this study, support vector machine (SVM), a recently proposed statistical learning model was introduced to evaluate real-time crash risk. The data has been split into a training dataset (used for developing the models) and scoring datasets (meant for assessing the models' predictive power). Classification and regression tree (CART) model has been developed to select the most important explanatory variables and based on the results, three candidates Bayesian logistic regression models have been estimated with accounting for different levels unobserved heterogeneity. Then SVM models with different kernel functions have been developed and compared to the Bayesian logistic regression model. Model comparisons based on areas under the ROC curve (AUC) demonstrated that the SVM model with Radial-basis kernel function outperformed the others. Moreover, several extension analyses have been conducted to evaluate the effect of sample size on SVM models' predictive capability; the importance of variable selection before developing SVM models; and the effect of the explanatory variables in the SVM models. Results indicate that (1) smaller sample size would enhance the SVM model's classification accuracy, (2) variable selection procedure is needed prior to the SVM model estimation, and (3) explanatory variables have identical effects on crash occurrence for the SVM models and logistic regression models.

1. Introduction

Recently Active Traffic Management (ATM) have been emerging in the US and Europe, its key control strategies such as variable speed limits (VSL) were recognized to have the benefits of improving traffic safety (Mirshahi et al., 2007; Pan et al., 2010; Chang et al., 2011). These implemented systems were mostly designed to reduce speed variations to reduce the crash risk. Moreover, more advanced proactive crash prediction models have showed promising effects on reducing crash occurrence along with the VSL system (Abdel-Aty et al., 2006; Lee et al., 2006b; Lee and Abdel-Aty, 2008). Within these studies, sophisticated real-time crash risk evaluation models were estimated to emulate the crash occurrence probabilities with the real-time traffic data. Crash risks would be evaluated with real-time traffic data and once a certain threshold of crash risk has been reached, the VSL control system would be triggered to smoothen the traffic flow and improve traffic safety. The real-time crash risk evaluation models try to identify the "crash precursor conditions" by comparing the crash occurrence traffic statuses and randomly selected non-crash cases. In the previous studies, both the traditional statistical models and artificial intelligence models have been utilized. Matched case–control logistic regression was one of the widely employed traditional statistical models (Abdel-Aty et al., 2004, 2007; Lee et al., 2006a) while the artificial neural network models (Pande and Abdel-Aty, 2006a; Pande et al., 2011) was another popular modeling technique that has been adopted in previous studies. More recently, as the Bayesian inference technique became popular, Bayesian logistic regression models have been used in real-time crash risk evaluation studies (Ahmed et al., in press-a,b).

Although previous real-time crash risk evaluation models have been proven to be capable of differentiating between crash and non-crash cases, these models have some limitations. Logistic regression models assumed a linear relationship between the dependent and independent variables while neural network models work as a black-box and may have over-fitting issues. Support vector machine (SVM), a newly introduced pattern classifier based on statistical learning theory (Vladimir and Vapnik, 1995) was introduced in this study to formulize the real-time crash risk evaluation model. Data from a 15-mile mountainous freeway (I-70) in

* Corresponding author. Tel.: +1 407 823 0300.
E-mail addresses: rongjie.yu@knights.ucf.edu, rongjie.yu@gmail.com (R. Yu).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.11.027

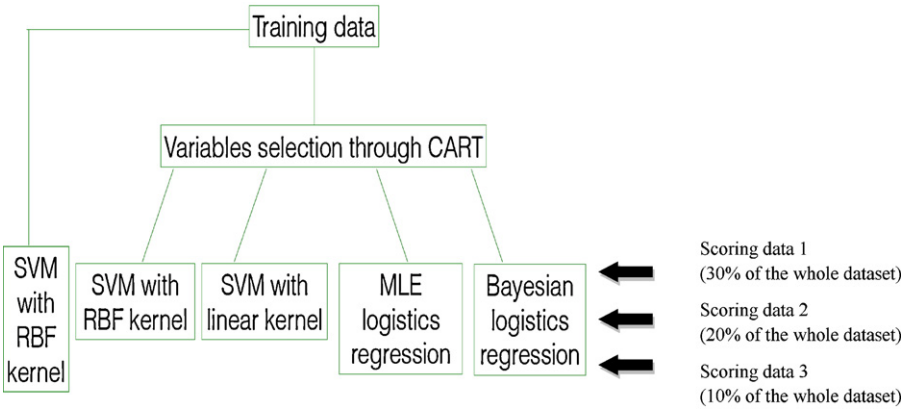

Fig. 1. Modeling procedure in this study.

Colorado was used in this study. With the merit of the Remote Traffic Microwave Sensor (RTMS) radars implemented along the freeway, real-time traffic data (speed, occupancy and volume) was captured and matched with the historical crash data.

The data has been split into training and scoring datasets. The training dataset was utilized to estimate the models and the scoring dataset was meant to test the prediction powers of different models. Due to SVM models lack of the capability of selecting significant variables and the use of all the variables as input would make the model cumbersome, a classification and regression tree (CART) was first estimated to select the most significant contributing variables. Then based on the chosen explanatory variables, three candidates Bayesian logistic regression models have been estimated with accounting for different levels unobserved heterogeneity. Then SVM model with Radial-basis kernel function and linear kernel function have been estimated and compared to Bayesian logistic regression model. Comparisons have been made based on the areas under the ROC curves (AUC). Moreover, SVM models without the variable selection procedure have also been estimated and investigated. Furthermore, the scoring datasets were divided into different sample sizes to test the sample size issue on these models prediction abilities. Finally, sensitivity analyses have been conducted to reveal the effects of the explanatory variables. Fig. 1 presents the flowchart of the main modeling procedures for this study.

2. Background

Support vector machine (SVM) models have been employed in some aspects of transportation research studies. Yuan and Cheu (2003) introduced SVM in incident detection and they compared the results from SVM models with the multi-layer feed forward neural network (MLFNN) and probabilistic neural network models. They concluded that SVM models provided a lower misclassification rate, higher correct detection rate and lower false alarm rate. Later on, Chen et al. (2009) also constructed SVM models to detect traffic incidents. Instead of building one SVM, different SVM models have been estimated and combined using various ensemble methods (bagging, boosting, cross-validated). The authors utilized ensemble methods to overcome the variability of a single SVM and improve the model's accuracy. Besides the incident detection, SVM has been utilized in crash frequency studies. Li et al. (2008) estimated safety performance functions for motor vehicle crashes with support vector machine models. SVM models have been estimated and compared to the traditional negative binomial models, and the results demonstrated that SVM models provide better goodness-of-fit than the negative binomial models. Moreover, a sensitivity analysis method has been used to analyze the effects of explanatory variables. Ren and Zhou (2011) employed the SVM technique along with a particle swarm optimization method to predict crash frequencies. Li et al. (2011) utilized SVM models for crash injury severity analysis. SVM models have been compared with the frequently used severity analysis method: ordered probit models. It was concluded that SVM models outperformed the ordered probit models. Besides, SVM models were also employed in traffic flow prediction studies (Cheu et al., 2006; Zhang and Xie, 2008). Lv et al. (2009) have tried to use SVM models in real-time highway accident prediction, however, the data used in their study was from simulation software rather than the real-field data and the SVM was not compared with the traditional real-time crash prediction models. This study would fill the gap by utilizing SVM models in real-time crash risk evaluation and comparing the results to the frequently adopted methods. Moreover, sensitivity analysis would also be conducted to observe the effects of the chosen explanatory variables.

Real-time crash prediction models were estimated with the purpose of unveiling the crash precursors and the results have been utilized in traffic management systems. With the advanced traffic surveillance system (loop detectors, speed radars, automatic vehicle identification systems), traffic statuses prior to crash occurrence could be identified and matched with the crashes to be analyzed. Abdel-Aty et al. (2004) employed the matched case–control logistic regression modeling technique to predict freeway crashes based on loop detector data. The matched case–control logistic regression was introduced to predict crash potential and it was found that the coefficient of variation of speed at the downstream station and the average occupancy of the upstream station are significant in the final models. Pande and Abdel-Aty (2006b) built a disaggregate model for rear-end crashes on I-4 (Orlando area) based on loop-detectors' data at 5-min aggregated level. A variable selection technique has been performed to select the most significant variables (traffic flow and geometric design parameters) for multi-layer perceptron and normalized Radial-basis function neural networks models. Moreover, Ahmed et al. (in press-a) utilized AVI data along with real-time weather information and roadway geometric characteristics to formulate a real-time crash occurrence model. Logistic regression was estimated with the Bayesian inference technique.

3. Data preparation

The 15-mile mountainous freeway is located on I-70 in Colorado and the studied segment starts from the Mile Marker (MM) 205 and ends at MM 220. There were two datasets utilized in this study, (1) crash data from October 2010 to October 2011 provided by Colorado Department of Transportation (CDOT) and (2) real-time traffic data detected by 30 RTMS radars. There were 265

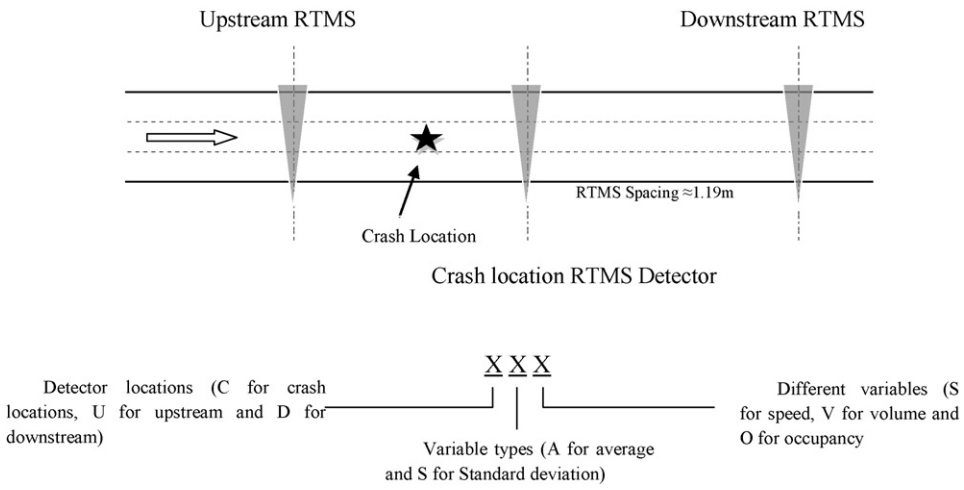

Fig. 2. Arrangements of RTMS detectors and nomenclature method for traffic variables.

crashes documented and matched with the traffic data and 1017 non-crash cases that were matched with the crash cases. The RTMS radars archived speed, volume and occupancy information at 30-s intervals. The real-time traffic data that corresponds to each crash was prepared by first aggregating the raw data into 5-min intervals and then each crash was assigned to the nearest downstream radar detector, the traffic data 5–10 min prior to the crash time was selected to represent the traffic condition. The 5–10 min traffic variables prior to the reported crash time were extracted in order to avoid confusing pre and post crash conditions. For example, if a crash happened at 15:25, at the Mile Marker 211.3. The corresponding traffic status for this crash is the traffic condition of time interval 15:15 and 15:20 recorded by RTMS radar at Mile Marker 211.8. Similarly, upstream and downstream traffic statuses were also extracted for each crash case. For each observation, average and standard deviation values of the speed, occupancy and volume have been calculated for the three detectors. So there are 18 (3 traffic flow parameters × 2 measures × 3 detectors) explanatory variables for each observation. Fig. 2 shows how the upstream, downstream and crash location RTMS detectors were assigned. Moreover, these traffic variables are named in a specific way also shown in Fig. 2. For example, DAO stands for the average occupancy captured by radar located at downstream of the crash location. The matched case–control design was adopted in this study, it was frequently utilized in the disaggregate crash occurrence studies since the confounding factors can be controlled for by matching (Breslow and Day, 1980). For each specific crash case, four non-crash cases were identified and matched. The non-crash cases were selected based on the following procedure: for example, a crash happened on Tuesday (May 24, 2011) then the four non-crash cases will be selected for the exact same time two weeks before and two weeks after the crash time (May 10, May 17, May 31, and June 7) at the exact location of crash occurrence. This matched case–control structure would eliminate the geometric characteristics' influences on crash risk evaluation. Moreover, seasonal random parameter and segment level random effects have been introduced to the Bayesian logistic regression model to capture the possible unobserved heterogeneity.

4. Methodology

4.1. Support vector machine

Support vector machine was originally designed based on statistical learning theory and the structural risk minimization. The algorithm tries to find a separating hyperplane by minimizing the distance of misclassified points to the decision boundary. For the binary classification problem in this study (crash and non-crash), given the training data

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_i, y_i), \quad y_i \in \{-1, 1\}$$

assuming that for the crash cases $y_i = 1$ and $y_i = -1$ for the non-crash cases; $\boldsymbol{x}_i$ represent the matrix for explanatory variables selected by the CART model.

The SVM tries to find the function $f(\boldsymbol{x}, \alpha_0) \in f(\boldsymbol{x}, \alpha)$, which best approximates the unknown function $y = f(\boldsymbol{x})$. The hyperplanes could be written as

$$f(x, \alpha) = (\omega_\alpha \cdot \boldsymbol{x}) + b$$

Among these hyperplanes there is one with the maximum margin, which is regarded as the optimal separating hyperplane. This hyperplane is uniquely determined by the vectors on the margin, the support vectors (Vapnik and Chervonenkis, 1974). Sometimes the data is not linearly separable, and then the SVM model becomes the following optimization problem (Cortes and Vapnik, 1995):

$$\text{minimize} \quad (\omega \cdot \omega) + C\sum_i \varepsilon_i^\delta, \quad \delta \geq 0$$

$$\text{under} \quad \text{constraints} \quad y_i[(\omega \cdot \boldsymbol{x}_i) + b] \geq 1 - \varepsilon_i, \quad \varepsilon \geq 0$$

where the ε_i allow for some error.

By introducing the Lagarange multiplier, this optimization problem has the form of

$$W(\alpha) = \sum_i^N \alpha_i - \frac{1}{2}\sum_{i,j}^N \alpha_i \alpha_j y_i y_j (h(\boldsymbol{x}_i), h(\boldsymbol{x}_j))$$

Subject to the constraints $\sum_i^N \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$

And the separating hyperplane has the form

$$f(x) = \sum_{i=1}^N \alpha_i y_i (h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)) + \beta_0$$

Both the two aforementioned equations only involve $h(\boldsymbol{x})$ through inner products. Then instead of the specific transformation $h(\boldsymbol{x})$, only the knowledge of the kernel function

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (h(\boldsymbol{x}_i), h(\boldsymbol{x}_j))$$

Table 1
Variable selection results by classification and regression tree.

Variable	Description	# of splitting rules	# of surrogate rules	Importance
DAS	Downstream average speed	5	10	1
CAS	Crash location average speed	7	9	0.992
CSO	Crash location standard deviation of occupancy	3	13	0.949
CSV	Crash location standard deviation of volume	2	11	0.849

that computes the inner products in the transformed space (Friedman et al., 2001).

In this study, the Radial-basis function (RBF) kernel and linear kernel were considered:

Linear kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$

Radial-basis function kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma|\boldsymbol{x}_i - \boldsymbol{x}_j|^2)$

4.2. Bayesian logistic regression

Bayesian inference technique was utilized to estimate the probability of crash occurrence. Suppose the crash occurrence has the outcomes $y = 1$ or $y = 0$ with respective probability p and $1 - p$. The Bayesian logistic regression can be setup as follows:

$$y \sim \text{Binomial}\quad (p)$$

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{X}\boldsymbol{\beta_t} + \alpha_{j[i]}$$

where β_0 is the intercept, $\boldsymbol{X}$ is the vector of the explanatory variables, for $t = 1, 2$, $\boldsymbol{\beta_t}$ is the vector of coefficients for the explanatory variables for snow season ($t = 2$) and dry season ($t = 1$); $j[i]$ indexes the segment where observation i occurs and $\alpha_{j[i]}$ is the random effects variable defined in the model, which represents the segment specific random effects in this study (the freeway section has been split into 120 homogenous segments according to the geometric characteristics (Yu et al., 2013)):

$$\alpha_j \sim N(0, \sigma_\alpha^2), \quad \text{for} \quad j = 1, \ldots, 120$$

where σ_α is the standard deviation of the unexplained segment-level errors.

Full Bayesian inference was employed in this study. For each model, three chains of 15,000 iterations were set up in WinBUGS (Lunn et al., 2000), 5000 iterations were used in the burn-in step. Convergences of the models have been checked by monitoring the MCMC trace plots for the model parameters: if all values are within a zone without strong periodicities and tendencies, then the model would be concluded as convergence.

5. Modeling results and discussion

5.1. Variable selection

Due to the SVM models lack of capability of selecting significant variables from the 18 explanatory variables, a classification and regression tree (CART) has been estimated to do the variable selection work. CART models have frequently used in traffic safety studies for their classification capability. For example, Chang and Wang (2006) utilized a CART model to analyze crash injury severity and Chang and Chen (2005) employed a CART model to predict crash frequency. Moreover, Kuhnert et al. (2000) used logistic regression, CART and multivariate adaptive regression splines (MARS) to conduct motor-vehicle injury analysis and the authors suggested that CART can be used as precursor to a more detailed logistic regression model. Variable selection through decision trees is kind of target dependent method and the variable importance is calculated based on the number of times a variable appeared and its relative position in the tree. This procedure was conducted in SAS Enterprise Miner (SAS Institute, 2004) with the following settings in the program: Splitting Criterion: Gini; Maximum Depth: 10; Leaf Size: 10; Split Size: 20; and Number of Surrogate: 3. The final variable selection results are presented in Table 1. Moreover, before estimating any models using the selected variables, multicollinearity test has been carried using the PROC CORR procedure in SAS and the results can be found in Table 2. The results suggest that all the four explanatory variables are not highly correlated and they can be used in the following logistic regression and SVM models.

5.2. Model results

Based on the results of variable selection, firstly three candidate Bayesian logistic regression models have been estimated: (1) Bayesian fixed parameter logistic regression, (2) Bayesian random parameter logistic regression considering the seasonal variations, and (3) Bayesian random effects logistic regression accounting for the unobserved segment level heterogeneity. The purpose of developing multi-level logistic regression models is to investigate the necessity of accounting for the observed heterogeneity and their effects on parameter estimations and model fit. Therefore the best logistic regression model will be compared to SVM model with linear kernel function and SVM with Radial-basis kernel function. The whole dataset was split into training and scoring datasets. For the training dataset, 70% of the original dataset was used at the models' training session which contains 179 crash cases and 691 non-crash cases. The descriptive statistics of variables can be found in Table 3.

The two SVM models were formulized in the SAS Enterprise Miner while the Bayesian logistic regression models have been calculated with the freeware WinBUGS. Analyses results for the three Bayesian logistic regression models are presented in Table 4. In addition, the MLE logistic regression model has been estimated and it turned out that it has comparable results with the fixed parameter Bayesian logistic regression model regarding the coefficient estimations.

Results from the logistic regression models can be interpreted as: crashes are likely to happen during congestion periods (especially when the queuing area has propagated from downstream areas); high variations of occupancy and/or volume would increase the probability of crash occurrence. Moreover, the random parameter model was estimated to capture the seasonal variation effects: the most substantial difference lies at the CSO variable which is significant for snow seasons but not significantly during dry seasons. Furthermore, segment level random effects were introduced to account for unobserved heterogeneity caused by the various geometric characteristics. DIC was selected as the model comparison evaluation criterion for the Bayesian three logistic regression models: The DIC, recognized as Bayesian generalization of AIC (Akaike information criterion), is a combination of the measure of model

Table 2
Correlation matrix for the selected explanatory variables.

	CAS	DAS	CSO	CSV
CAS	1.00	0.35	−0.22	0.03
DAS	0.35	1.00	−0.21	0.13
CSO	−0.22	−0.21	1.00	0.11
CSV	0.03	0.13	0.11	1.00

Table 3
Summary of variables descriptive statistics for the training dataset.

Variables	Description	Mean	Std. dev	Min	Max
Crash	Binary index for crash occurrence (1 for crash and 0 for non-crash cases)	0.206	0.405	0	1
CAS	Crash locations average speed	54.36	13.08	7.00	77.32
DAS	Downstream average speed	55.61	13.42	3.32	77.86
CSO	Crash locations standard deviation of occupancy	1.84	1.42	0	22.23
CSV	Crash locations standard deviation of volume	7.33	4.72	0	25.78

Table 4
Estimations for the explanatory variables in the Bayesian logistic regression models.

Variables	Bayesian fixed parameter		Bayesian random parameter		Bayesian random effects	
	Mean	95% credible interval	Mean	95% credible interval	Mean	95% credible interval
Intercept	1.72	(0.86, 2.59)	2.57[1][a] 1.42[2]	(0.80, 4.48) (0.39, 2.43)	1.88	(0.99, 2.79)
CAS	−0.026	(−0.04, −0.009)	−0.025[1] −0.024[2]	(−0.057, 0.007) (−0.042, −0.006)	−0.027	(−0.043, −0.011)
DAS	−0.043	(−0.058, −0.029)	−0.058[1] −0.039[2]	(−0.092, −0.027) (−0.056, −0.023)	−0.046	(−0.062, −0.031)
CSO	0.11	(0.036, 0.19)	0.067[1] 0.13[2]	(−0.14, 0.24) (0.04, 0.23)	0.12	(0.038, 0.19)
CSV	0.045	(0.0076, 0.082)	0.039[1] 0.047[2]	(−0.034, 0.11) (0.002, 0.092)	0.054	(0.014, 0.093)
Random error					0.47	(0.33, 0.65)
DIC	774.629		783.953		786.786	
Number of observations: 870						

[a] [1] dry seasons; [2] snow seasons.

fitting and the effective number of parameters. The smaller DIC indicate a better model fitting and according to Spiegelhalter et al. (2003), differences of more than 10 might definitely rule out the model with higher DIC. Comparisons of DIC values demonstrate that Bayesian fixed parameter model is superior to the other complex models and it was selected to represent the ordinal approach for the real-time crash risk evaluation and compare to the SVM models.

After the training sessions for the four models, scoring datasets with variety of sample sizes have been employed to test the prediction power of the proposed models. Areas under the ROC curve (AUC) was chosen to evaluate and compare these models, the larger AUC values indicate a better goodness-of-fit and classification power. Chen et al. (2009) employed AUC to compare different SVM models while Ahmed et al. (in press-a,b) adopted AUC to evaluate real-time crash risk evaluation models with Bayesian logistic regression models. Table 5 lists the AUCs of the three models for the training and scoring datasets. One point that needs to be mentioned is that for the scoring datasets, to overcome the small datasets evaluation drawback, a cross-validation technique has been adopted here (Kohavi, 1995).

From the models' goodness-of-fit results for the training and scoring datasets, it can be seen that SVM with RBF kernel models outperformed the traditional models while the linear SVM models have identical results with the two logistic regression models. These phenomena demonstrated the existence of some non-linear relationships between the dependent variables and explanatory variables in the real-time crash risk evaluation models which haven't been captured by the linear formulation models.

Table 5
AUC values of the SVM and logistic regression models.

	SVM with RBF	Linear SVM	Bayesian logistic regression
Training data (70% of the whole dataset)	0.81	0.78	0.78
Scoring data 1 (30% of the whole dataset)	0.74	0.73	0.73
Scoring data 2 (20% of the whole dataset)	0.75	0.74	0.74
Scoring data 3 (10% of the whole dataset)	0.77	0.73	0.73

5.3. SVM predictive performance on small sample data

As crash occurrence is a small probability event, traffic safety researchers have to deal frequently with small samples. For example, to analyze traffic safety for a highway with newly installed ITS system, crash and real-time traffic sample size would be very restricted. In order to analyze the crash occurrence mechanism efficiently, a small sample may be used to develop safety performance functions and real-time crash risk evaluation models. However, a small sample may cause many issues about the model fit and coefficient estimation; e.g., Lord and Mannering (2010) concluded that low sample-mean and small sample size could cause errors in parameter estimates in developing safety performance functions. In this section, the predictive performance on small sample datasets of the SVM models and ordinal logistic regression models were compared by goodness-of-fit three different scoring datasets for the models (Table 5). It can be concluded that scoring datasets generally have lower AUC values than the training datasets; while the RBF SVM models performed better as the data size become smaller and the other models showed stable performance on different sample size datasets. For the purpose of confirming that smaller a dataset would achieve better performance with the RBF SVM models, those scoring datasets in Table 5 were split into datasets according to the crash types (multi-vehicle crashes and single-vehicle crashes). As stated by Pande and Abdel-Aty (2006a,b) it is important to analyze crashes by type, particularly when it comes to real-time crash risk assessment. AUC values have been calculated by applying the RBF SVM models to the different datasets separately. Table 6 compares the predictive ability of the SVM models and the results affirmed that RBF SVM models have better predictive power for the smaller datasets.

Table 6
AUC values for the RBF SVM models with datasets by crash type.

	All crashes	Multi-vehicle crashes	Single-vehicle crashes
Scoring data 1	0.74	0.80	0.75
Scoring data 2	0.75	0.77	0.75
Scoring data 3	0.77	0.79	0.77

5.4. The importance of variable selection

The aforementioned analysis results were achieved by the SVM models with the four chosen explanatory variables, in order to evaluate the SVM models, comparisons have been made between the RBF SVM models with and without variable selection. Same model configurations will be held and the only difference between the models is that one SVM model with only the four selected explanatory variables and the other SVM model has the whole 18 explanatory variables as input. Table 7 provides the AUC values for the two models with both the training and scoring datasets. From the results it can be concluded that for the model without variable selection, the SVM models for the training dataset have the over-fitting problem while for the scoring datasets it performed even worse than the traditional logistic regression models. So it is highly recommended that before formalizing SVM models, variable selection methodology would be implemented.

5.5. Sensitivity analysis for variable effects

As an emerging machine learning method, SVM was blamed for being a black-box technique whereas the effects of explanatory variables on the dependent variable could not be seen. However, with the benefits of sensitivity analysis, the relationships between crash occurrence and the chosen four explanatory variables could be analyzed. Fish and Blodgett (2003) suggested using sensitivity analysis to explore the effects of explanatory variables; this approach has been adopted in Li et al. (2008) and Li et al.'s (2012) SVM studies to evaluate the effects of explanatory variables. Two-stage sensitivity analysis was employed in this study: First, sensitivity analysis was conducted by changing each explanatory variable by a user-defined amount while the other variables remain at the original values. SVM models were re-calculated with the new datasets and comparisons have been made for the mean predicted probabilities for crash occurrence. This analysis can be used to detect the positive and negative relationships between the explanatory variables and crash occurrence probabilities. Results of the sensitivity analysis are depicted in Table 8.

Effects of the four explanatory variables can be unveiled from the changes of mean predicting probabilities for crash occurrence. It can be concluded that lower values of CAS and DAS are probable to increase the crash hazardousness and higher values of CAO and CAV would likely result in a higher probabilities of crash occurrence. In addition, it seems that crash occurrence is more associated with the downstream congestion conditions. These conclusions are identical to the logistic regression models' result. Moreover, more thorough investigation into the functional relationship between the explanatory variables and crash occurrence probabilities has been conducted by assigning a set of the changing units (for example, for the CAS variable, changing units have been applied to it from 1 mph to 10 mph instead of 5 mph in Table 7; while keeping other variables at the original values). Figs. 3 and 4 show the results of the sensitivity analysis for the CAS and DAS variables separately, the changing unit varies from 1 mph to 10 mph. It can be seen that as the CAS values for the crash cases decrease, the mean crash occurrence probability increased; and the relationships between crash occurrence probabilities and CAS changing unit has a quadratic functional form. Furthermore, from Fig. 4 it can be discovered that as the DAS values of crash cases decrease, the mean crash occurrence probability increased; and a cubic relationship was found between the crash occurrence probability and DAS changing units.

Table 7
AUC values for the SVM models with and without variable selection.

	SVM RBF model with variable selection	SVM RBF model without variable selection
Training data	0.81	0.98
Scoring data 1	0.74	0.73
Scoring data 2	0.75	0.66
Scoring data 3	0.77	0.69

Table 8
Sensitivity analysis of the explanatory variables in the SVM models.

Variable	Changing unit	Original mean probability	New mean probability
CAS	−5 mph for crash cases; +5 mph for non-crash cases	0.295232	0.316409
DAS	−5 mph for crash cases; +5 mph for non-crash cases	0.295232	0.317412
CAO	+0.3% for crash cases; −0.3% for non-crash cases	0.295232	0.300565
CAV	+0.3 vph for crash cases; −0.3 vph for non-crash cases	0.295232	0.299218

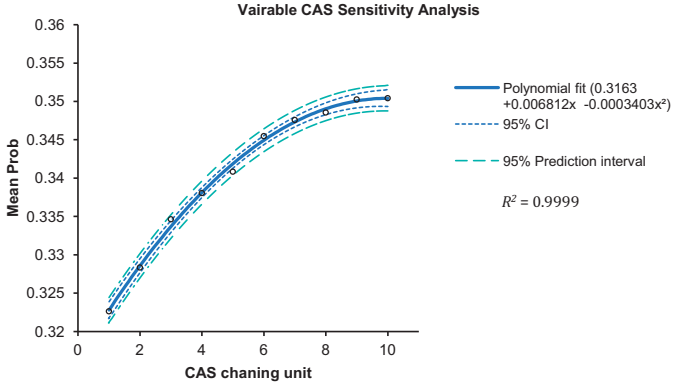

Fig. 3. Sensitivity analysis for the variable CAS.

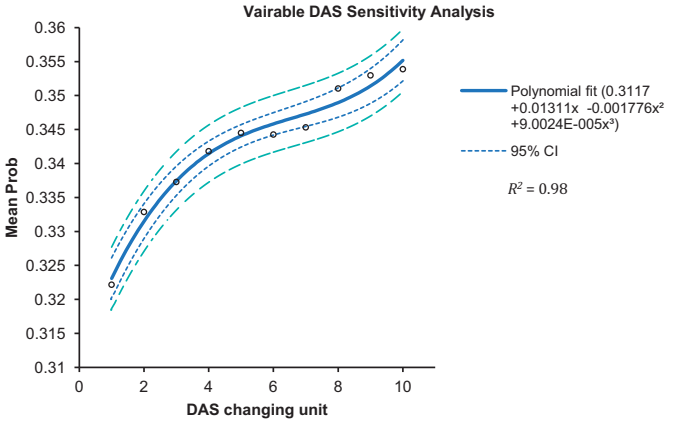

Fig. 4. Sensitivity analysis for the variable DAS.

6. Conclusion

Active Traffic Management (ATM) concepts are gaining momentum around the world. Improving traffic safety is expected to be a

major component of ATM. Thus efficient and accurate real-time crash prediction models are required. Previous studies that have focused on this topic adopted both the traditional statistical (logistic regression model) and the artificial neural network techniques. Due to limitations of these models (linear function forms and overfitting problems), SVM models have been proposed here to evaluate the crash risk at the disaggregate level. Previous studies that applied SVM models in the transportation area have proved that SVM models provide superior or at least comparable results as the neural network models (Yuan and Cheu, 2003; Li et al., 2008; Ren and Zhou, 2011). This study compared SVM models to the Bayesian logistic regression models.

CART model was employed to select critical variables which contribute to crash occurrence. With the variable selection results, three Bayesian logistic regression models have been estimated: (1) Bayesian fixed parameter logistic regression model, (2) Bayesian random parameter logistic regression model meant for capturing the seasonal variation effects, and (3) Bayesian random effects logistic regression which is capable of accounting unobserved segment level heterogeneity. However, comparisons of the goodness-of-fit and parameter estimations suggested that the three models are very comparable and the Bayesian fixed parameter logistic regression model has the lowest DIC. Then the Bayesian logistic regression models have been compared to the SVM models with both linear and RBF kernels.

Model comparisons' results showed that the SVM model with RBF kernel provided the best goodness-of-fit. While the SVM models with linear kernel have similar results as the logistic regression models. The findings of including the RBF kernel would enhance the model classification capability indicated that there are some nonlinear relationships that exist between the dependent variable and explanatory variables which could not be captured by the logistic regression models.

Moreover, by applying different small sample sizes to test the predictive capability of the SVM and ordinary logistic regression models, it has been revealed that the RBF SVM models' performance improved as the sample size decreased while the logistic regression models hold a stable goodness-of-fit. This conclusion has been further confirmed by applying the SVM models to different datasets by crash types (scoring datasets were split by crash types). With this merit, SVM models would have promising applications in traffic safety studies for newly built roadways or freeways with newly implemented ITS systems.

Two-stage sensitivity analyses have been conducted to unveil the effects of the chosen explanatory variables on crash occurrence; by changing one explanatory variable with a pre-defined value and keeping the other variables with the original values, the effects of each explanatory variable can be revealed by comparing the mean crash occurrence probabilities. Similar results with the logistic regression models have been achieved: crashes are more likely to happen within the congested area, especially for the queuing area that propagates from downstream; large variation of occupancy and volume indicates turbulent and stop-and-go traffic scenarios also have relatively high risk of crash occurrence. In addition, thorough investigation has been done by varying a specific variable with different units while keeping the other variables at the same value. Results indicated that linear functions cannot fully describe the relationship between the mean crash occurrence probabilities and the explanatory variables.

Moreover, in order to decide whether the variable selection procedure is needed for developing the SVM models, SVM models with the same configurations have been developed with all the explanatory variables and then only the selected four explanatory variables. Results demonstrated that SVM models would have an over-fitting issue for the training dataset and perform poor on the scoring datasets if all variables are used (no selection procedure). Thus it is highly recommended that variable selection procedures precede SVM modeling.

While this study showed that SVM models have great application potential in real-time crash risk evaluation studies, the scale parameter γ in the RBF kernel function was kept constant at 1. Future investigation can focus on tuning the scale parameter's value, such as introducing parameter searching algorithms like Genetic Algorithms to find the best γ fit in the model. Based on this, further improvements of the model classification accuracies can be achieved.

References

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accident Analysis and Prevention 38, 335–345.

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., Dos Santos, C., 2007. Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. Journal of Intelligent Transportation Systems 11, 107–120.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F.M., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case–control logistic regression. Transportation Research Record: Journal of the Transportation Research Board 1897, 88–95.

Ahmed, M., Abdel-Aty, M., Yu, R. Assessment of the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data. Journal of Transportation Research Board, in press-a.

Ahmed, M., Abdel-Aty, M., Yu, R. A Bayesian updating approach for real-time safety evaluation using AVI data. Journal of Transportation Research Record, in press-b.

Breslow, N., Day, N., 1980. Statistical Methods in Cancer Research, vol. 1. The Analysis of Case–Control Studies Distributed for IARC by WHO, Geneva, Switzerland.

Chang, L., Chen, W., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research 36, 365–375.

Chang, G., Park, S., Paracha, J., 2011. Its field demonstration: integration of variable speed limit control and travel time estimation for a recurrently congested highway. In: Proceedings of the TRB 2011 Annual Meeting, Washington, DC.

Chang, L., Wang, H., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

Chen, S., Wang, W., Van Zuylen, H., 2009. Construct support vector machine ensemble to detect traffic incident. Expert Systems with Applications 36, 10976–10986.

Cheu, R., Xu, J., Kek, A., Lim, W., Chen, W., 2006. Forecasting shared-use vehicle trips with neural networks and support vector machines. Transportation Research Record: Journal of the Transportation Research Board 1968, 40–46.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Fish, K., Blodgett, J., 2003. A visual method for determining variable importance in and artificial neural network model: an empirical benchmark study. Journal of Targeting Measurement and Analysis for Marketing 11, 244–254.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer, New York, USA.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1145.

Kuhnert, P., Do, K., McClure, R., 2000. Combing non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics and Data Analysis 34, 371–386.

Lee, C., Abdel-Aty, M., 2008. Testing effects of warning messages and variable speed limits on driver behavior using driving simulator. Transportation Research Record: Journal of the Transportation Research Board 2069, 55–64.

Lee, C., Abdel-Aty, M., Hsia, L., 2006a. Potential real-time indicators of sideswipe crashes on freeways. Transportation Research Record: Journal of the Transportation Research Board 1953, 41–49.

Lee, C., Hellinga, B., Saccomanno, F., 2006b. Evaluation of variable speed limits to improve traffic safety. Transportation Research Part C 14, 213–228.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis and Prevention 40, 1611–1618.

Li, Z., Liu, P., Wang, W., Xu, C., 2011. Using support vector machine models for crash injury severity analysis. Accident Analysis and Prevention 45, 478–486.

Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accident Analysis and Prevention 45, 478–486.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44, 291–305.

Lunn, D., Thomas, A., Best, N., Spiegelhalter, D., 2000. Winbugs—a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325–337.

Lv, Y., Tang, S., Zhao, H., Li, S., 2009. Real-time highway accident prediction based on support vector machines. In: Control and Decision Conference (CCDC), pp. 4403–4407 (in Chinese).

Mirshahi, M., Obenberger, J., Fuhs, C., Howard, C., Krammes, R., Kuhn, B., Mayhew, R., Moore, M., Sahebjam, K., Stone, C., 2007. Active Traffic Management: The Next Step in Congestion Management.

Pan, S., Jia, L., Zou, N., Park, S., 2010. Design and implement of a variable speed limit system with travel time display—a case study in Maryland, USA. In: 17th ITS World Congress Busan 2010, Busan, Korea.

Pande, A., Abdel-Aty, M., 2006a. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transportation Research Record: Journal of the Transportation Research Board 1953, 31–40.

Pande, A., Abdel-Aty, M., 2006b. Assessment of freeway traffic parameters leading to lane-change related collisions. Accident Analysis and Prevention 38, 936–948.

Pande, A., Das, A., Abdel-Aty, M.A., Hassan, H., 2011. Real-time crash risk estimation: are all freeways created equal? Transportation Research Record: Journal of the Transportation Research Board 2237, 60–66.

Ren, G., Zhou, Z., 2011. Traffic safety forecasting method by particle swarm optimization and support vector machine. Expert Systems with Applications 38, 10420–10424.

SAS Institute, 2004. SAS/Graph (r) 9.1 reference. SAS Institute.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2003. Winbugs User Manual. MRC Biostatistics Unit, Cambridge.

Vapnik, V., Chervonenkis, A., 1974. Pattern Recognition Theory. Statistical Learning Problems, Nauka, Moskva.

Vladimir, V., Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Yuan, F., Cheu, R., 2003. Incident detection using support vector machines. Transportation Research Part C: Emerging Technologies 11, 309–328.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis and Prevention 50, 371–376.

Zhang, Y., Xie, Y., 2008. Forecasting of short-term freeway volume with v-support vector machines. Transportation Research Record: Journal of the Transportation Research Board 2024, 92–99.