Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Predicting motor vehicle crashes using Support Vector Machine models

Xiugang Li*, Dominique Lord, Yunlong Zhang, Yuanchang Xie

Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States

ARTICLE INFO

Article history:
Received 4 February 2008
Received in revised form 29 April 2008
Accepted 30 April 2008

Keywords:
Highway
Crash
Support Vector Machine
Negative binomial model
Neural network

ABSTRACT

Crash prediction models have been very popular in highway safety analyses. However, in highway safety research, the prediction of outcomes is seldom, if ever, the only research objective when estimating crash prediction models. Only very few existing methods can be used to efficiently predict motor vehicle crashes. Thus, there is a need to examine new methods for better predicting motor vehicle crashes. The objective of this study is to evaluate the application of Support Vector Machine (SVM) models for predicting motor vehicle crashes. SVM models, which are based on the statistical learning theory, are a new class of models that can be used for predicting values. To accomplish the objective of this study, Negative Binomial (NB) regression and SVM models were developed and compared using data collected on rural frontage roads in Texas. Several models were estimated using different sample sizes. The study shows that SVM models predict crash data more effectively and accurately than traditional NB models. In addition, SVM models do not over-fit the data and offer similar, if not better, performance than Back-Propagation Neural Network (BPNN) models documented in previous research. Given this characteristic and the fact that SVM models are faster to implement than BPNN models, it is suggested to use these models if the sole purpose of the study consists of predicting motor vehicle crashes.

1. Introduction

Crash prediction models (sometimes referred to as safety performance functions or SPFs) have been very popular in highway safety analyses. These models can be used for numerous applications, such as identifying the relationship between the dependent variable and a series of explanatory variables or screening variables. In highway safety, explanatory variables include traffic volume, lane and shoulder widths as well as signal timing at signalized intersections among others. Another important application of crash prediction models consists of estimating or predicting crashes, as its name implies, on facilities from which they were not used in the model development. Predicted crashes can be used to identify hazardous sites and compare the safety performance of different highway design alternatives for hypothetical scenarios both in the present or the future (non-existing facilities). Depending on the study objectives, they can also be used in before-and-after studies using the empirical Bayes (EB) method (Hauer, 1997).

Accurately predicting crashes cannot be done without an effective crash prediction model. A model that fits the data very well, does not necessarily mean that it will be able to predict crashes successfully (The MathWorks, Inc., 2007). In addition, as noted by Lord et al. (2007), prediction of outcomes (crashes) is seldom, if ever, the only research objective when estimating crash prediction models. Usually, transportation safety analysts report and compare various statistical models based on goodness-of-fit statistics (e.g., Shankar et al., 1997; Miaou and Lord, 2003) to determine which model fits the data the best. Only recently did researchers in highway safety focus their effort on developing models for the sole purpose of predicting motor vehicle crashes (see Oh et al., 2003; Xie et al., 2007). Given this new line of research activity, very few existing methods can be used to efficiently predict motor vehicle crashes. For those available, they can sometimes be complex to implement and include important limitations, such as issues related to the overfitting of data. Thus, there is a need to examine whether other new methods could be used for better predicting motor vehicle crashes.

The objective of this study is to evaluate the application of Support Vector Machine (SVM) for predicting motor vehicle crashes. This research is, in essence, a continuation of recent work preformed by some of the authors on crash prediction methodologies (Xie et al., 2007). SVM models, which are based on the statistical learning theory, are a new class of models that can be used for

* Corresponding author. Tel.: +1 225 266 5221.
E-mail addresses: li_xiugang@tamu.edu (X. Li), d-lord@tamu.edu (D. Lord), yzhang@civil.tamu.edu (Y. Zhang), ycxie@tamu.edu (Y. Xie).

0001-4575/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2008.04.010

predicting values. They have recently been introduced for other transportation applications (see Zhang and Xie, 2007). To accomplish the objective of this study, Negative Binomial (NB) regression and SVM models were developed and compared using data collected on rural frontage roads in Texas. In order to compare the performance of SVM models with Back-Propagation Neural Network (BPNN) models, the same dataset used by Xie et al. (2007) was employed in this research. Several models were developed for different sample sizes. The study will show that SVM models predict crash data more effectively and accurately than traditional NB models. In addition, SVM models do not over-fit the data, and offer better or comparative performance than BPNN models.

This paper is divided into seven sections. The second section presents a literature review on crash models and prediction methods. The third section describes the characteristics of NB regression and SVM models. The fourth section describes the methodology used for estimating the models. The fifth section covers the application of models to the dataset collected on frontage roads in Texas. The sixth section describes the results of the analysis and comparison. The last section summarizes the key results of this study and provides recommendations for further work.

2. Background

The most common types of model that have been used in highway safety remain the traditional Poisson and Poisson-gamma (or Negative Binomial) models. These models are used for modeling discrete, independent, and non-negative events. The Poisson and Poisson-gamma models have been used both for examining or screening variables (e.g., Wong et al., 2007) and for predicting motor vehicle crashes (e.g., Donnell and Mason, 2006; Lord, 2008). Because crash data often exhibit over-dispersion (even conditional upon the mean), where the variance exceeds the mean, NB models are usually the model of choice utilized by transportation safety analysts. For this kind of data, Poisson models should not be used for developing crash prediction models. For instance, if a Poisson distribution is assumed in estimating the expected number of crashes, larger discrepancies between the observed and the predicted crashes may be observed (Hauer, 2001). Similarly, a mis-specified Poisson model may lead to the inclusion of covariates that have been erroneously identified as being significant when, in fact, they are not (Park and Lord, 2007; Hilbe, 2007). It should be pointed out that Poisson and other mixed-Poisson models (i.e., Poisson-gamma, Poisson-lognormal, etc.) have usually been developed under a univariate modeling framework. Over the last 2 years, multivariate models, in which each variable is a vector (often describing different crash severity levels), have started being used in highway safety research (Miaou and Song, 2005; Ma and Kockelman, 2006; Park and Lord, 2007; Ma et al., 2008). Very recently, Lord et al. (2008a) have proposed the use of the Conway-Maxwell-Poisson (COM-Poisson) generalized linear model (GLM) as an alternative to the NB model for modeling crash data. This new model is more flexible, since it can account for under-dispersed data, and preliminary results show that it performs better when the sample size is small and the sample mean value is low (Geedipally et al., submitted for publication).

The coefficients of NB models can be estimated using different statistical tools (see reference Lord et al., 2008b about the four-step process for developing crash prediction models). The most common tools are the GLM methods (known as the Frequentist approach in the statistical literature) and Bayesian methods. The GLMs are currently the most popular tools, since its theoretical framework is well developed (McCullagh and Nelder, 1989; Myers et al., 2002) and the method is already integrated in commercial statistical software programs, such as SAS (SAS Institute Inc., 2006) and Genstat (Payne, 2000). Over the last few years, Bayesian methods (Carlin and Louis, 2000; Gelman et al., 2003) have become more popular in highway safety research, mainly due to the increasing computing power and the availability of Bayesian software programs, such as WinBUGS (Spiegelhalter et al., 2003) and the Bayesian function in MatLAB (The MathWorks, Inc., 2007). Bayesian methods are often used for estimating the coefficients of complex modeling structures, such as multivariate models described above, and when the transportation safety analyst is interested in finding the distribution for each coefficient of the model rather than a point estimate and associated uncertainties (e.g., see Miaou and Song, 2005).

In recent years, a few researchers have proposed new and innovative statistical models for the sole purpose of predicting motor vehicle crashes. For instance, researchers in various fields of research, including highway safety, have proposed the use of neural network models, such as BPNN, for modeling and predicting crash data (Xie et al., 2007; Mussone et al., 1999; Abdelwahab and Abdel-Aty, 2002; Riviere et al., 2006). Unfortunately, these models have been criticized to work as a black-box and can sometimes over-fit the data, especially when the sample size is small (Vogt and Bared, 1998). To circumvent this problem, Bayesian neural networks (BNN) have been proposed by a few researchers (Liang, 2005; Mackay, 1992; Neal, 1995). According to Marzban and Witt (2001), BNN models can effectively reduce the over-fitting phenomenon, while still keep the strong nonlinear approximation ability of neural networks. In highway safety, Xie et al. (2007) examined the application of BNN models for predicting motor vehicle crashes. They found that BNN were much more efficient than NB models for predicting crashes.

SVMs are a set of related supervised learning methods used for classification and regression, and possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy (Kecman, 2005). The statistical learning theory and structural risk minimization are the theoretical foundations for the learning algorithms of SVMs (Kecman, 2005). It has been found that SVMs show better or comparable results than the outcomes estimated by neural networks and other statistical models (Kecman, 2005). Zhang and Xie (2007) found that v-SVM (v is a new parameter introduced to the SVMs) has better modeling performance than multilayer feed-forward neural network (MFFNN) in short-term freeway traffic volume forecasting. So far, no applications of SVMs for predicting highway crash frequencies have been identified. Theoretically, SVMs have less over-fitting problem and better generalization ability than traditional neural networks because SVMs are based on structural risk minimization (Suykens et al., 2002), whereas traditional neural networks are based on empirical risk minimization (Zhang and Xie, 2007). While neural networks may have local minimization problems, with input parameters (C, v, γ) (to be explained below) the learning algorithm of SVMs is to solve a quadratic programming, which has the global optimal solution.

Similar to any statistical and mathematical models, SVMs have also some disadvantages. First, SVMs work as black-boxes, similar to traditional neural networks. There are no formal functional form between crashes and the covariates. Second, three parameters (C, v, γ) need to be determined before the training phase. Algorithms such as the grid searching algorithm (Chang and Lin, 2001) and hybrid genetic algorithm (Huang and Wang, 2006) have been developed to select the parameter values automatically based on the data. But this process may not provide a global optimum. Third, when the number of data points or observations is large (usually greater than 2000), solving the quadratic programming (QP) during the training procedure becomes difficult with standard QP solvers (Kecman, 2005). This, however, should not be a problem in highway safety studies because the sample size commonly found in highway crash databases is usually not very large (Lord and Bonneson, 2005). Fourth, while traditional statistical methods have the advantages of being able to ascertain whether the model's coefficients are significant, the SVMs cannot be used for such purpose. The inclusion of insignificant variables may not improve prediction accuracy and may face the risk of over-fitting. However, once a list of candidate variables are identified from prior knowledge or traditional statistical models, SVMs can identify better relationships and improve the model's fit as well as the predictive performance (see Zhang, 2006, for a discussion about a method to include relevant variables in SVMs).

3. Characteristics of models

As discussed above, NB regression models are the most widely used statistical models in highway safety (Xie et al., 2007). The SVM model based on the structural risk minimization has the advantages of better generalization ability than traditional neural network models (Zhang and Xie, 2007). Therefore, the fitting and prediction capabilities of these two models should be thoroughly investigated. Below, the fundamental characteristics of NB regression models and SVM are briefly described.

3.1. Negative Binomial regression models

The probability mass function (PMF) of the NB regression model is usually given by the following (Miaou, 1994):

$$\mathrm{Prob}(Y_i = y_i) = \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi} \quad (1)$$

$$\text{Expectation of } Y_i \text{ is } \mu_i = g(x_i) \quad (2)$$

$$\text{Variance of } Y_i \text{ is } \mathrm{Var}(Y_i) = \mu_i + \frac{{\mu_i}^2}{\phi} \quad (3)$$

where y_i is the crash number collected at site i; Y_i is the dependent random variable following a NB distribution with the inverse dispersion parameter ϕ (note: $\alpha = 1/\phi$ is referred to as the dispersion parameter, see Table 2); x_i is a vector representing the crash-related variables at site i; and $g(x_i)$ is the functional form of the NB regression model.

3.2. Support Vector Machine

For this study, the v-SVM, which is based on the ε-SVM, was employed. The principles of the v-SVM are briefly described below. The details of the ε-SVM and v-SVM can be found in references Suykens et al. (2002) and Schölkopf et al. (2000), respectively.

Assume the training input is defined as vectors $x(i) \in R^{\mathrm{In}}$ for $i = 1, \ldots, N$, which are independent and identically distributed data with sample size N. The training output is defined as $y(i) \in R^1$ for $i = 1, \ldots, N$. The v-SVM maps $x(i)$ into a feature space $R^h (h > \mathrm{In})$ with higher dimension using a function $\Phi(x(i))$ to linearize the nonlinear relationship between $x(i)$ and $y(i)$. The estimation function of $y(i)$ is

$$\widehat{y} = f(x) = w^{\mathrm{T}}\Phi(x) + b \quad (4)$$

where $w \in R^h$ and $b \in R^1$ are coefficients. Schölkopf et al. (2000) showed that the coefficients are derived by solving the following optimization problem.

$$\mathrm{Min}\, Z(w, \varepsilon, \xi_i, \xi_i^*) = \frac{1}{2}w^{\mathrm{T}}w + C\left\{v\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*)\right\}$$

subject to

$$w^{\mathrm{T}}\Phi(x(i)) + b - y(i) \le \varepsilon + \xi_i \quad \forall i = 1, \ldots, N \quad (5)$$

$$y(i) - w^{\mathrm{T}}\Phi(x(i)) - b \le \varepsilon + \xi_i^* \quad \forall i = 1, \ldots, N \quad (6)$$

$$\xi_i, \xi_i^* \ge 0 \quad \forall i = 1, \ldots, N$$

$$\varepsilon \ge 0$$

where ξ_i, ξ_i^* are slack variables; C is a regularization parameter, and v is a second parameter. For each $x(i)$ the allowable error is ε. Slack variables ξ_i, ξ_i^* capture errors above ε and are penalized in the objective function via a regularization constant C (Schölkopf et al., 2000).

In reference (Schölkopf et al., 2000), the estimated function of $y(i)$ becomes

$$\hat{y} = f(x) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\Phi(x(i))^{\mathrm{T}}\Phi(x) + b$$
$$= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\cdot K(x(i), x) + b \quad (7)$$

where $K(x(i), x(j)) = \Phi(x(i))^{\mathrm{T}}\Phi(x(j))$ is the kernel function, α_i and α_i^* are Lagrange multipliers for the constraints in Eqs. (5) and (6), respectively. A radial basis kernel function was used for this study. It is defined as

$$K(x(i), x(j)) = \exp\left\{-\gamma\left\|x(i) - x(j)\right\|^2\right\} \quad (8)$$

where γ is a parameter. With the radial basis function as the kernel function, the v-SVM has three parameters (C, v, γ) that need to be determined. There is always a globally optimal solution to w and b with the input of three parameters (C, v, γ) (Burges, 2007).

4. Model implementation

Two software programs were utilized for estimating the NB and SVM models. The coefficients of the NB models were estimated using the function GENMOD in SAS (SAS Institute Inc., 2006). On the other hand, the MATLAB software (The MathWorks, Inc., 2007) was utilized to estimate the SVM models with the LIBSVM tool developed by Chang and Lin (2001, 2007). The tool LIBSVM also provides a grid searching algorithm for determining the three parameters of SVM models. Recently, Huang and Wang (2006) combined a genetic algorithm with the LIBSVM tool to estimate these parameters. They found this hybrid genetic algorithm to provide a better model prediction accuracy than the grid searching algorithm. Zhang and Xie (2007) implemented this hybrid genetic algorithm to determine the three parameters (C, v, γ), and this algorithm was consequently used in this research. The genetic algorithm toolbox in MATLAB software (The MathWorks, Inc., 2007) was used to calculate the three parameters (C, v, γ) with the input of crash frequency y_i and exploratory variables $\mathbf{x}_i$.

5. Application of the models

This section describes the application of the models using the Texas data. The first part covers the characteristics of the data. The second part provides details about the different sample sizes used for developing the models. The third part describes the data normalization method employed for feeding the data into the SVM models. The last part presents the performance indexes needed for comparing both types of model.

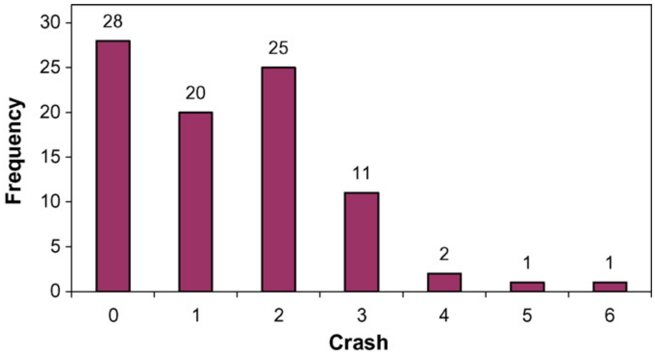

Fig. 1. Histogram of crash data.

5.1. Data

The data used for evaluating the models were collected at 88 segments located on rural frontage roads in Texas. The data were assembled for a project related to the development of accident modification factors (AMFs) for the Texas Department of Transportation (TxDOT) (Lord and Bonneson, 2007). In total, 122 crashes occurred during the 5-year period. The histogram illustrating the crash data distribution is shown in Fig. 1. The summary statistics are presented in Table 1. The ratio between variance and mean was found to be about 1.2, which indicates that the NB model is suitable for the selected dataset. It should be pointed out that Washington et al. (2003) suggest a more rigorous approach to justify the choice of Poisson regression or NB regression. However, since the authors were interested in comparing the results with those of Xie et al. (2007), the same dataset used by Xie et al. (2007) was employed in this study.

In this study, the same functional form as the one used by Xie et al. (2007) was adopted for $g(x_i)$ in Eq. (2), which is described in Eq. (9):

$$\mu_i = \beta_0 \left(\frac{365 \times F_i \times L_i \times \mathrm{Off}_i}{1,000,000} \right) \exp(\beta_1 \mathrm{LW}_i + \beta_2 \mathrm{RS}_i) \tag{9}$$

where F_i is the average daily traffic (ADT) for segment i; L_i is the length of segment in miles; Off_i is the number of years during which the crash data y_i are collected. In this research, Off_i is equal to 5; LW_i is the lane width of segment i in feet; RS_i is the right-shoulder width of segment i in feet; and, β_0, β_1, β_2 are regression coefficients.

5.2. Fitting and predicting samples

The original plan was to separate the dataset into three subsets: one for fitting, one for stop-learning to calculate the parameters (C, v, γ), and one for calculating the model prediction errors. This approach, recommended by Zhang and Xie (2007), can reduce the likelihood of over-fitting. However, given the fact that the original sample used in this study is already very small, it was determined that this approach would not provide adequate results, especially for the NB regression models (see Lord, 2006 and Lord and Miranda-Moreno, 2008). In addition, in order to simplify the comparison with the results documented in Xie et al. (2007), the dataset was randomly separated into two new subsets. This random separation process was conducted for fifteen times (5 separations for each subset size × 3 subset sizes). For each subset size, one separation was used only to determine the parameters, and the remaining four separations were used to calculate the prediction errors. In the end, this random separation process was deemed adequate for the purpose of this study.

Different sample sizes were utilized for examining the fitting and prediction capabilities of the models. The fitting sample sizes were equal to 60, 70, and 80, respectively. Besides the fitting samples, the remaining data were used as predicting samples. Consequently, the predicting sample sizes were 28, 18 and 8, respectively.

5.3. Data normalization

Because the data have dissimilar units and magnitudes, the data for each variable had to be normalized. Data normalization can improve the data fitting as well as prediction performances and is required for input into SVM models. The normalization was accomplished using the following equation:

$$xn_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{10}$$

where x_i is a vector representing the variables of F_i, L_i, LW_i, and RS_i.

The crash data for y_i were also normalized. The fitting and predicting values yn_i generated from the SVM models are denormalized to $\hat{y}_i$ using the following equation:

$$\hat{y}_i = yn_i(\max(y_i) - \min(y_i)) + \min(y_i) \tag{11}$$

5.4. Performance index

Two evaluation criteria proposed by Oh et al. (2003) were adopted for this study to compare the model performance. The measures of effectiveness (MOEs) are described in Eqs. (12) and (13) (Oh et al., 2003).

$$\text{Mean absolute deviation (MAD)} = \frac{1}{n}\sum_{i=1}^{n} \left| \hat{y}_i - y_i \right| \tag{12}$$

$$\text{Mean squared predictor error (MSPE)} = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2 \tag{13}$$

where n is the size of fitting or predicting sample; $\hat{y}_i$ is the estimated number of crashes at site i; and y_i is the observed number of crashes. The model performance is better if the values of MAD and MSPE are smaller.

6. Analysis results

This section describes the results of the analysis. The first part summarizes the modeling results of the NB model. Recall that SVM models only provide predicted values. Consequently, no model

Table 1
Summary statistics of the explanatory variables

	Length (mile)	ADT (vpd)	Right-shoulder length (ft)	Lane width (ft)	Number of crashes in 5 years
Mean	2.16	939	1.38	10.67	1.39
S.D.	0.99	1186	2.12	0.84	1.28
Minimum	0.69	110	0	9	0
Maximum	5.34	6400	9	13	6

Table 2
Results of NB regression models

Fitting set size	Parameter	Estimate	Standard error	Wald 95% confidence limits		$P > \chi^2$
60	Intercept ($\ln \beta_0$)	2.6195	1.6740	−0.6615	5.9005	0.1176
	Lane width (β_1)	−0.2880	0.1600	−0.6015	0.0256	0.0719
	Right shoulder (β_2)	−0.1943	0.0704	−0.3324	−0.0563	0.0058
	Dispersion parameter	0.1557	0.1361	−0.1111	0.4225	
70	Intercept ($\ln \beta_0$)	1.3495	1.6394	−1.8637	4.5626	0.4104
	Lane width (β_1)	−0.1726	0.1581	−0.4826	0.1373	0.2751
	Right shoulder (β_2)	−0.1447	0.0683	−0.2786	−0.0108	0.0341
	Dispersion parameter	0.1756	0.1399	−0.0985	0.4498	
80	Intercept ($\ln \beta_0$)	1.6448	1.3689	−1.0382	4.3278	0.2295
	Lane width (β_1)	−0.2029	0.1315	−0.4606	0.0549	0.1230
	Right shoulder (β_2)	−0.0967	0.0606	−0.2155	0.0220	0.1103
	Dispersion parameter	0.1033	0.1188	−0.1296	0.3362	

characteristics can be described. The second part describes the results of the comparison analysis.

6.1. NB regression model output

The NB regression results are shown in Table 2. The functional forms of the models developed from this dataset are summarized below:

(1) Fitting sample size of 60:

$$\mu_i = \mathrm{e}^{(2.6195)} \times \left(\frac{365 \times F_i \times L_i \times \mathrm{Off}_i}{1,000,000} \right) \times \mathrm{e}^{(-0.288 \times \mathrm{LW}_i - 0.1943 \times \mathrm{RS}_i)} \qquad (14)$$

(2) Fitting sample size of 70:

$$\mu_i = \mathrm{e}^{(1.3495)} \times \left(\frac{365 \times F_i \times L_i \times \mathrm{Off}_i}{1,000,000} \right) \times \mathrm{e}^{(-0.1726 \times \mathrm{LW}_i - 0.1447 \times \mathrm{RS}_i)} \qquad (15)$$

(3) Fitting sample size of 80:

$$\mu_i = \mathrm{e}^{(1.6448)} \times \left(\frac{365 \times F_i \times L_i \times \mathrm{Off}_i}{1,000,000} \right) \times \mathrm{e}^{(-0.2029 \times \mathrm{LW}_i - 0.0967 \times \mathrm{RS}_i)} \qquad (16)$$

Table 2 shows that, while the signs for all coefficients seem appropriate, many are marginally or not significant. The coefficients become more significant and their magnitude becomes smaller as the sample size increases. This characteristic is not unusual and was observed in Xie et al. (2007) and Lord and Bonneson (2007). This is mainly explained by the bias introduced by the small sample size and low sample mean values. Data characterized by these two conditions can seriously affect NB regression models (see Lord, 2006; Lord and Miranda-Moreno, 2008).

6.2. Model performance comparisons

The performances of NB regression models and SVM models are shown in Table 3. This table shows that the values of MAD and MSPE for the SVM models are consistently lower than those for NB regression models for all fitting sample sizes. For the fitting sample size of 80, the NB regression model has large prediction errors (the last two columns). This is probably caused by the limitations of NB regression models (fixed functional form), and possibly because the data were randomly separated into prediction and fitting samples. On the other hand, the SVM models both fitted and predicted values very well, with values for the MOEs rarely exceeding 1.20. Overall, this table shows that the SVM models perform much better than the NB regression models.

Table 3
Performance comparison of NB regression and SVM models

Fitting sample size	MOEs	Fitting		Prediction	
		NB	SVM	NB	SVM
60	MAD	1.14	0.86	1.07	0.90
	MSPE	2.62	1.19	1.96	1.12
70	MAD	1.09	0.88	1.22	0.85
	MSPE	2.46	1.26	2.24	0.98
80	MAD	1.00	0.88	2.49	0.91
	MSPE	1.87	1.15	11.89	1.18

Using the same samples analyzed in reference (Xie et al., 2007), the performances of the NB regression models and BPNN models were compared with the SVM models developed in this work; the BPNN models were not re-estimated, and the results were taken directly from Xie et al. (2007). The results of the comparison are summarized in Table 4. The MOEs show that the SVM models perform better than the NB models (developed in their work). The MOEs also show that the SVM models either perform better or have comparable performance than the BPNN models.

Table 4
Performance comparison using samples in Xie et al. (2007)

Fitting sample size	MOEs	Fitting			Prediction		
		NB	BPNN	SVM	NB	BPNN	SVM
60	MAD	0.99	0.85	0.85	1.28	1.00	0.99
	MSPE	1.76	1.17	1.04	2.96	1.42	1.38
70	MAD	0.96	0.82	0.80	1.58	1.15	0.80
	MSPE	1.63	1.10	0.96	4.24	1.75	1.29
80	MAD	1.04	0.85	0.81	1.86	1.41	1.41
	MSPE	2.04	1.12	1.07	6.53	2.52	2.53
60	MAD	1.20	0.94	0.95	0.98	0.87	0.75
	MSPE	2.89	1.34	1.28	1.82	1.00	0.85
70	MAD	1.16	0.92	0.88	0.85	0.79	0.80
	MSPE	2.68	1.33	1.32	0.94	0.83	0.84
80	MAD	1.14	0.92	0.87	0.80	0.86	0.73
	MSPE	2.56	1.28	1.24	0.87	1.02	0.75
60	MAD	1.19	0.91	0.92	1.05	0.87	0.91
	MSPE	2.90	1.23	1.17	2.00	1.20	1.18
70	MAD	1.14	0.88	0.79	1.11	0.99	0.98
	MSPE	2.70	1.17	0.77	1.81	1.49	1.40
80	MAD	1.14	0.88	0.84	1.40	1.17	1.11
	MSPE	2.73	1.17	1.13	2.57	2.05	1.94

Table 5
Data used for the sensitivity analysis

Site number	Crash count	ADT (vehicle/day)	Segment length (miles)	Lane width (feet)	Right-shoulder width (feet)
14	3	6168	2.40	11	4
88	1	428	1.15	10	0

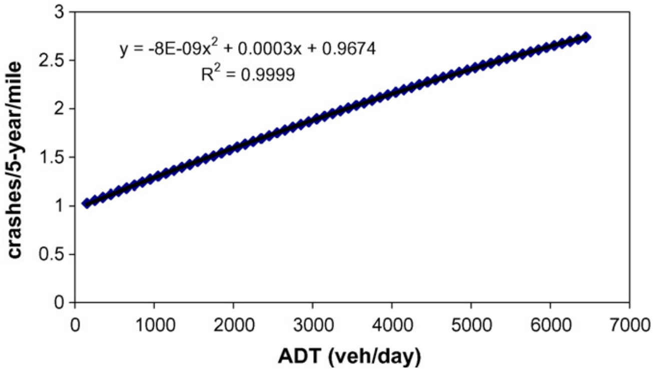

Fig. 2. Sensitivity analysis for the variable ADT for site 14.

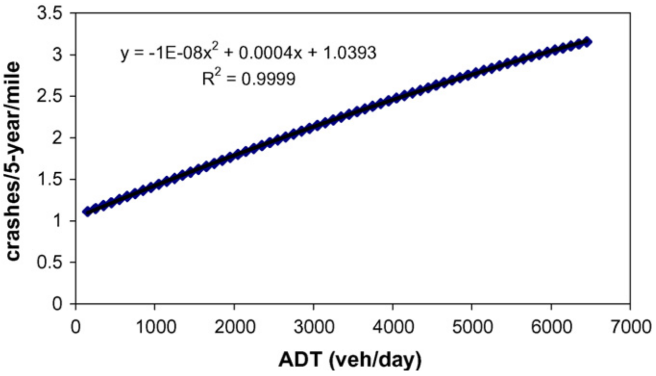

Fig. 3. Sensitivity analysis for the variable ADT for site 88.

The implementation of SVM models is faster than neural network models. With the LIBSVM tool (Chang and Lin, 2007), SVM models can be implemented conveniently in MATLAB (The MathWorks, Inc., 2007). The training of neural networks is usually computationally intensive. Because the BNN proposed by Liang (2005) and used by Xie et al. (2007) are implemented on the UNIX platform, it takes much more time to code and train the BNN models in UNIX than the SVM models in MATLAB, especially for users not familiar with UNIX.

As described by Lord (2006) and others, crash data characterized by a low sample mean value combined with a small sample size can seriously affect the estimation of the dispersion parameter of NB models as well as their predictive capabilities. The dataset used in this study is characterized by both low sample mean and small sample size. The testing results show that the SVM model provides better performance for this situation.

6.3. *Sensitivity analysis of the SVM models*

Since SVM models work as a black-box (similar to neural networks), it may be criticized for not being able to generate interpretable parameters for each explanatory variable. As discussed above, SVM models do not have a specified functional form. To minimize the black-box problem, the method by Fish and Blodgett (2003) was used to analyze the sensitivity of the explanatory variables. All explanatory variables were investigated one at a time. An example related to the sensitivity analysis with respect to the variables ADT and right-shoulder width (RS) is included in this paper. The sensitivity analysis for each of these two variables consisted of recording the changes in crash frequencies generated from a SVM model for different values of ADT or RS, within a reasonable interval (as determined by the data), while keeping all the other variables constant.

The SVM model for the fitting sample size of 60 was used for sensitivity analysis. This analysis was performed for sites #14 and #88. The collected crash data and other related characteristics are shown in Table 5. Site #14 has right shoulders and a high ADT value while site #88 has no right shoulders and a low ADT value. Figs. 2 and 3 show that, for both site #14 and site #88, the relationship between crash frequency and ADT has a quadratic functional form, and the crash frequency increases with the increase of ADT (note that the relationship is almost linear however). This relationship has been observed frequently in the safety literature (see Hauer, 1997). This has also been observed by Xie et al. (2007), although the relationship was less linear. Figs. 4 and 5 show that, for both site #14 and site #88, the relationship between crash frequency and right-shoulder width has a quadratic functional form, and the crash frequency decreases with an increase in right-shoulder width.

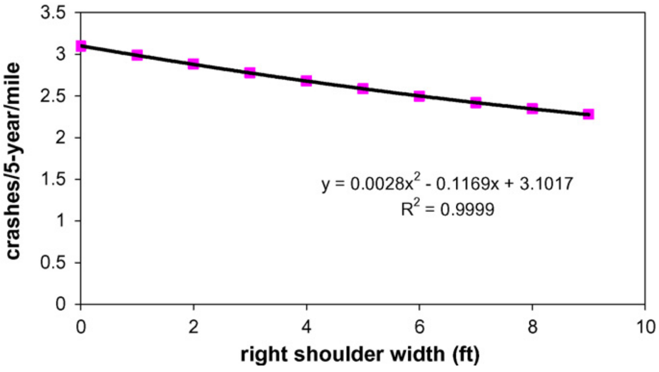

Fig. 4. Sensitivity analysis for the variable right-shoulder width for site 14.

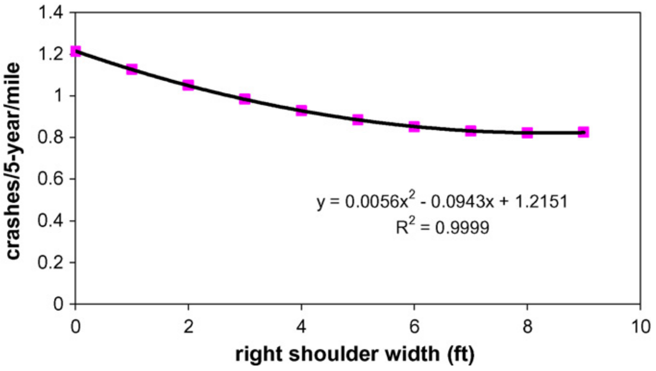

Fig. 5. Sensitivity analysis for the variable right-shoulder width for site 88.

7. Summary and conclusions

The objective of this study was to evaluate the application of SVM models for predicting motor vehicle crashes. To accomplish the objective of this study, NB regression and SVM models were developed and compared using data collected on rural frontage roads in Texas. Different sample sizes were used for the model comparison. The models were also compared to BPNN models documented in a previous study (Xie et al., 2007).

The study has shown that SVM models predict crash data more effectively and accurately than traditional NB models. In addition, SVM models do not over-fit the data and offer similar, if not better, performance than BPNN models. Given this characteristic and the fact that SVM models are faster to implement than BPNN models (at least in MATLAB), it is suggested to use these models if the sole purpose of the study consists of predicting motor vehicle crashes. They are particularly useful when the sample size is below 2000 observations, which happens frequently in highway safety studies.

Even though the application of SVM models offered positive results, it is suggested to evaluate this kind of model using other datasets to validate the results obtained in this research. This validation should include the assessment of their performance when traffic flow is the only covariate investigated. Flow-only models, such as baseline models or general ADT models, are the current model type that has been proposed in various chapters of the forthcoming Highway Safety Manual (Hughes et al., 2005). The authors hope that the results presented in this paper will promote new research ideas for developing innovative prediction methodologies in highway safety.

References

Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas. Transportation Research Record: Journal of the Transportation Research Board, No. 1784, TRB, National Research Council, Washington, DC, pp. 115–125.

Burges, C.J.C., 2007. A Tutorial on Support Vector Machines for Pattern Recognition. research.microsoft.com/~cburges/papers/SVMTutorial.pdf. Accessed Nov. 11, 2007.

Carlin, B., Louis, T., 2000. Bayes and Empirical Bayes Methods for Data Analysis, 2nd edition. Chapman and Hall, New York.

Chang, C.-C., Lin, C.-J., 2001. Training ν-Support Vector Classifiers: theory and algorithms. Neural Computation 13 (9), 2119–2147.

Chang, C.-C., Lin, C.-J., 2007. LIBSVM: A Library for Support Vector Machines, www.csie.ntu.edu.tw/_cjlin/libsvm. Accessed on April, 16, 2007.

Donnell, E.T., Mason Jr., J.M., 2006. Predicting the frequency of median barrier crashes on Pennsylvania Interstate Highways. Accident Analysis & Prevention 38 (3), 590–599.

Fish, K.E., Blodgett, J.G., 2003. A visual method for determining variable importance in an artificial neural network model: an empirical benchmark study. Journal of Targeting Measurement and Analysis for Marketing 11 (3), 244–254.

Geedipally, S., Guikema, S.D., Dhavala, S., Lord, D. Characterizing the performance of a Bayesian Conway-Maxwell Poisson GLM. Working paper, Zachry Department of Civil Engineering, College Station, TX, submitted for publication.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, 2nd edition. Chapman & Hall/CRC Press, New York.

Hauer, E., 1997. Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Measures on Road Safety. Elsevier Science Ltd., Oxford.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in empirical Bayes estimation. Accident Analysis & Prevention 33 (6), 799–808.

Hilbe, J.M., 2007. Negative Binomial Regression. Cambridge University Press, Boston, MA.

Huang, C.L., Wang, C.J., 2006. A GA-based feature selection and parameters optimization for support vector machines. Expert System with Applications 31 (2), 231–240.

Hughes, W., Eccles, K., Harwood, D., Potts, I., Hauer, E., 2005. Development of a Highway Safety Manual. Appendix C: Highway Safety Manual Prototype Chapter: Two-Lane Highways. NCHRP Web Document 62 (Project 17-18(4)). Washington, DC (http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w62.pdf accessed October 2007).

Kecman, V., 2005. Support vector machines—an introduction. In: Wang, L. (Ed.), Support Vector Machines: Theory and Applications. Springer-Verlag, Berlin, Heidelberg, New York, pp. 1–48.

Liang, F., 2005. Bayesian neural networks for nonlinear time series forecasting. Statistics and Computing 15 (1), 13–29.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-Gamma Models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention 38 (4), 751–766.

Lord, D., 2008. Methodology for estimating the variance and confidence intervals of the estimate of the product of baseline models and AMFs. Accident Analysis & Prevention 40 (3), 1013–1017.

Lord, D., Bonneson, J.A., 2005. Calibration of Predictive Models for Estimating the Safety of Ramp Design Configurations. Transportation Research Record: Journal of the Transportation Research Board, No. 1908, TRB, National Research Council, Washington, DC, pp. 88–95.

Lord, D., Bonneson, J.A., 2007. Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. Transportation Research Record 2023, pp. 20–27.

Lord, D., Guikema, S.D., Geedipally, S., 2008a. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention 40 (3), 1123–1134.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma Models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46 (5), 751–770.

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., Ivan, J.N., van Schalkwyk, I., Lyon, C., Jonsson, T., 2008b. Methodology for Estimating the Safety Performance of Multilane Rural Highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, DC. (http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w126.pdf, accessed on June 3 2008).

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further Notes on the Application of Zero Inflated Models in Highway Safety. Accident Analysis & Prevention 39 (1), 53–57.

Ma, J., Kockelman, K.M., 2006. Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. Transportation Research Record: Journal of the Transportation Research Board, No. 1950, TRB, National Research Council, Washington, DC, pp. 24–34.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis & Prevention 40 (3), 964–975.

Mackay, D.J.C., 1992. Bayesian methods for adaptive models. Ph.D. Dissertation. California Institute of Technology, Pasadena, California.

Marzban, C., Witt, A., 2001. A Bayesian neural network for severe-hail size prediction. Weather Forecast 16 (5), 600–610.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd edition. Chapman and Hall, London.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Analysis & Prevention 26 (4), 471–482.

Miaou, S.-P., Lord, D., 2003. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes. Transportation Research Record: Journal of the Transportation Research Board, No. 1840, TRB, National Research Council, Washington, DC, pp. 31–40.

Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis & Prevention 37 (4), 699–720.

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accident Analysis & Prevention 31 (6), 705–718.

Myers, R.H., Montgomery, D.C., Vining, G.G., 2002. Generalized Linear Models: With Applications in Engineering and the Sciences. Wiley Publishing Co., New York.

Neal, R.M., 1995. Bayesian learning for neural networks. Ph.D. Dissertation. University of Toronto, Toronto, Ontario.

Oh, J., Lyon, C., Washington, S., Persaud, B., Bard, J., 2003. Validation of FHWA Crash Models for Rural Intersections: Lesions Learned. Transportation Research Record: Journal of the Transportation Research Board, No. 1840, TRB, National Research Council, Washington, DC, pp. 41–49.

Park, E.S., Lord, D., 2007. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record: Journal of the Transportation Research Board, No. 2019, TRB, National Research Council, Washington, DC, pp. 1–6.

Payne, R.W. (Ed.), 2000. The Guide to Genstat. Lawes Agricultural Trust, Rothamsted Experimental Station, Oxford.

Riviere, C., Lauret, P., Ramsamy, J.F.M., Page, Y., 2006. A Bayesian neural network approach to estimating the energy equivalent speed. Accident Analysis & Prevention 38 (2), 248–259.

SAS Institute Inc., 2006. Base SAS® 9.1.3 Procedures Guide, 2nd edition. SAS Institute Inc., Cary, NC.

Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New support vector algorithms. Neural Computation 12, 1207–1245.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis & Prevention 29 (6), 829–837.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. WinBUGS Version 1.4.1 User Manual. MRC Biostatistics Unit, Cambridge (www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. Accessed Nov. 11, 2007).

Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vanderwalle, J., 2002. Least Squares Support Vector Machines. World Scientific Publishing Co. Pte. Ltd., Singapore.

The MathWorks, Inc., 2007. MATLAB Programming. Natick, MA. (Note: other useful references include: The MathWorks, Inc. (2007) Genetic Algorithm and Direct Search Toolbox 2 User's Guide. Natick, MA and The MathWorks, Inc. (2007) Neural Network Toolbox 5: User's Guide. Natick, MA).

Vogt, A., Bared, J.G., 1998. Accident Models for Two-Lane Rural Roads: Segments and Intersections. Publication FHWA-RD-98-133. FHWA, U.S. Department of Transporation.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.

Wong, S.C., Sze, N.N., Li, Y.C., 2007. Contributing factors to traffic crashes at signalized intersections in Hong Kong. Accident Analysis & Prevention 39 (6), 1107–1113.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accident Analysis & Prevention 39 (5), 922–933.

Zhang, H.H., 2006. Variable selection for vector support machines via smoothing spline ANOVA. Statistica Sinica 16, 659–674.

Zhang, Y., Xie, Y., 2007. Forecasting of Short-Term Freeway Volume with v-Support Vector Machines. Transportation Research Record: Journal of the Transportation Research Board, No. 2024, TRB, National Research Council, Washington, DC, pp. 92–99.