Accident Analysis and Prevention 90 (2016) 128–139

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Investigating driver injury severity patterns in rollover crashes using support vector machine models

Cong Chen[a], Guohui Zhang[a,*], Zhen Qian[b], Rafiqul A. Tarefder[a], Zong Tian[c]

[a] Department of Civil Engineering, University of New Mexico, Albuquerque, NM 87131, USA
[b] Department of Civil and Environmental Engineering and Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[c] Department of Civil and Environmental Engineering, University of Nevada, Reno, NV 89557, USA

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form 20 January 2016
Accepted 17 February 2016

Keywords:
Driver injury severity
Rollover crash
Support vector machine model
Kernel function
Traffic safety

ABSTRACT

Rollover crash is one of the major types of traffic crashes that induce fatal injuries. It is important to investigate the factors that affect rollover crashes and their influence on driver injury severity outcomes. This study employs support vector machine (SVM) models to investigate driver injury severity patterns in rollover crashes based on two-year crash data gathered in New Mexico. The impacts of various explanatory variables are examined in terms of crash and environmental information, vehicle features, and driver demographics and behavior characteristics. A classification and regression tree (CART) model is utilized to identify significant variables and SVM models with polynomial and Gaussian radius basis function (RBF) kernels are used for model performance evaluation. It is shown that the SVM models produce reasonable prediction performance and the polynomial kernel outperforms the Gaussian RBF kernel. Variable impact analysis reveals that factors including comfortable driving environment conditions, driver alcohol or drug involvement, seatbelt use, number of travel lanes, driver demographic features, maximum vehicle damages in crashes, crash time, and crash location are significantly associated with driver incapacitating injuries and fatalities. These findings provide insights for better understanding rollover crash causes and the impacts of various explanatory factors on driver injury severity patterns.

1. Introduction

As reported by the National Highway Traffic Safety Administration (NHTSA), there were 3009 rollover crashes in the U.S. in 2012, which accounted for 10% of all fatal crashes in the country (National Highway Traffic Safety Adminstriation, 2013). These numbers were even higher for New Mexico. According to the New Mexico Department of Transportation (NMDOT) (New Mexico Department of Transportation, 2012), rollover crashes accounted for 5.2% of total statewide reported crashes, but resulted in 34.6% of total fatal crashes and 36.2% of occupant fatalities. Statistics also revealed that rollover crashes were mostly single-vehicle involved and accounted for 35% of all fatalities in single-vehicle crashes (Fréchède et al., 2011). The significant loss of life resulting from rollover crashes indicates the emergent need of comprehensive and in-depth investigation of rollover crash mechanisms. Numerous studies have been conducted to examine rollover crashes and their contributing factors, injury outcome patterns, and effective countermeasures. Due to the significant weight and size, rollover crashes are most likely to occur when heavy vehicles, such as pickup trucks, semi-trailers, and farming tractors are used. For instance, Farmer and Lund (2002) concluded that light trucks experience a higher potential of rollover crashes than passenger cars. Significant studies were also performed to investigate the injury patterns in rollover crashes. For example, Huelke and Compton (1983) discovered that ejected occupants have a 17 times higher risk of more serious and fatal injuries than restrained occupants. Head, neck, and spine injuries are the primary injuries in rollover crashes because of the roof deformation and its crushing impact on human cephalic and vertebral parts (Conroy et al., 2006; Funk et al., 2012; Mandell et al., 2010). To reduce crash risk and injury severities in rollover crashes, various preventive countermeasures have been proposed, tested, and implemented in many peer studies (Chen et al., 2012; Harris et al., 2011; Liu and Koc, 2013; Mangado et al., 2007; Reynolds and Groves, 2000; Wu et al., 2014). For example, Liu and Koc (2013) devised a mobile application based on an IOS system for monitoring tractor running stability and reporting rollover incidents with detailed spatial, temporal, and other relative information.

* Corresponding author. Fax: +1 505 277 1988.
E-mail address: guohui@unm.edu (G. Zhang).

http://dx.doi.org/10.1016/j.aap.2016.02.011
0001-4575/

Statistical models are the primary method used in traffic crash analyses (Chen et al., 2015a,b; Liu et al., 2015b; Lord and Mannering, 2010; Wu et al., 2015). However, these models are based on certain assumptions regarding data and model structures, which inevitably pose limitations in these studies. For instance, ordered logit models have been widely used in crash severity analyses to model the ordinal nature of injury severity outcomes under the proportional odds assumption that the impacts of a contributing factor are identical across different ordinal levels, which in most circumstances will not hold. When these assumptions are violated, statistical bias or erroneous results are induced (Lord and Mannering, 2010; Savolainen et al., 2011). To overcome the limitations of statistical models in traffic crash analyses, non-parametric models, such as neural network, classification and regression tree (CART), decision tables, have been introduced and widely utilized (Abdelwahab and Abdel-Aty, 2001; Chang and Chien, 2013; Chang and Wang, 2006; Chen et al., 2016). Among these non-parametric models, support vector machine (SVM) is a modeling technique developed to address classification and regression problems. SVM models have been increasingly implemented in transportation research to address traffic flow forecasting (Cheu et al., 2006; Huang, 2015; Wei and Liu, 2013; Yu et al., 2013; Zhang et al., 2013), crash frequency assessment (Li et al., 2008; Ren and Zhou, 2011; Suárez Sánchez et al., 2011; Yu and Abdel-Aty, 2013a,b), and travel mode and travel pattern prediction (Allahviranloo and Recker, 2013; Sheng and Xiao, 2015). Recently, several peer studies have also applied SVM to examine injury severities at crash level, i.e. the most severe occupant injury outcome in a crash (Guo et al., 2012; Li et al., 2012; Yu and Abdel-Aty, 2014). For example, Guo et al. (2012) proposed a pedestrian recognition model applicable to intelligent transportation systems based on AdaBoost algorithms and SVM to reduce pedestrian suffering in traffic crashes. However, SVM models have never been used to examine injury severity patterns in rollover crashes. Besides, in previous crash severity analyses, SVM models considered each crash as a unit and examined the most severe injury outcome in each crash, which may not enough reveal the detail injury severity patterns in the crash. For instance, assuming that there are equal numbers of people involved in two different crashes where there is only one fatality in one crash but ten fatalities in the other, in traditional crash severity analysis, these two crashes are both considered as fatal crashes, but in fact the second one is much more severe than the first in terms of number of fatalities. Therefore, the authors are motivated to conduct this research to investigate the individual injury patterns taking each driver/vehicle record as a research unit.

The primary objectives of this study are to investigate the applicability of SVM models in driver injury severity analyses and use it to examine driver injury severity patterns in rollover crashes. A disadvantage of the SVM models is that it lacks the capability of automatically selecting significant factors contributing to the target variable. Therefore, a CART analysis is conducted to rank variable relative importance and identify significant variables for driver injury prediction. The rest of this paper is organized as follows: Section 2 provides a comprehensive and in-depth literature review regarding rollover crashes and SVM applications. The data description and processing procedure are introduced in Section 3, followed by the methodology design and model specifications in Section 4. Research results are explicitly discussed in Section 5, and the research limitations and overall research effort are summarized in Section 6.

2. State of the art

Rollover crashes have been widely investigated from multiple perspectives with different methods. Abundant studies have been performed to examine the crash mechanisms and injury patterns in rollover crashes. Hu and Donnell (2011) applied a multinomial logit model to examine the significant factors in predicting rollover crash severity in rural divided highways. Through logistic regression analyses, Bambach et al. (2013a,b) examined risk factors among human, vehicle, and environmental features on occupant thoracic injuries in rollover crashes. Chang et al. (2006) investigated passenger injury outcomes and the associated risk factors in motor coach rollover crashes and discovered that upside passengers thrown from seats and their downside neighbors were most likely to suffer major injuries. Rollover crashes primarily occurr among heavy vehicles, and considerable research has been conducted to explore the mechanisms in terms of contributing factors and kinetic features. Whitfield and Jones (1995) revealed that overhead weight on top of SUVs and pickups increases vehicle rollover risk. Franceschetti et al. (2014) proposed a model to simulate the kinetic features of tractor lateral rollovers considering the geometric features, tractor inertia nature, and environmental conditions at crash occurrence. Albertsson et al. (2006) investigated the injury outcome patterns, injury risk and mechanism, and the protective effect of seatbelts in rollover coach crashes. Rollover crashes are intimately associated with vehicle roof and side-structure deformation, resulting in severe head or neck injuries of vehicle occupants. Yoganandan et al. (1990) developed a kinetic 3-dimensional model using articulated body structure to examine the impacts of vehicle roof deformation and head clearance on head and neck injury risk in rollover crashes and identified the optimal roof and head clearance configuration to minimize injury risk. Freeman et al. (2012) discovered that a significant correlation exists between vehicle roof deformation and occupant head and neck injury and developed a head and neck injury score (HNIS) to predict roof crash potential in rollover crashes. Dobbertin et al. (2013) also verified the cause-effect association between vehicle roof crash deformation and occupant head, neck, and spine injury severity in rollover crashes. Bambach et al. (2013b) explored the mechanisms of spine injuries of passengers with seatbelt restraints in rollover crashes, concluding that a roof intrusion of 20 centimeters is preferable in order to avoid spine injuries.

Because of these kinetic and injury characteristics of rollover crashes, multiple protective countermeasures have been proposed to reduce driver injury risk and severity in rollover crashes. van der Westhuizen and Els (2013) verified that slow lateral maneuverers could significantly reduce the potential of vehicle rollovers. Roll-over protective structures (ROPS) are popular devices installed on vehicles to minimize rollover crash casualties. Mangado et al. (2007) developed an affordable ROPS for agricultural tractors and established a performance model to examine the effectiveness of the proposed ROPS based on their beam physical features. Chen et al. (2012) investigated the association between the lateral stiffness coefficients (LSC) of a ROPS and human injury levels in rollover crashes, concluding that appropriate LSC varies across different ROPSs. Reynolds and Groves (2000) concluded that law enforcement and driver educational programs are recommended to improve the effectiveness of ROPS in roll over crash prevention. Harris et al. (2011) assessed the performance of existing cost-effective ROPS at reducing rollover vehicle fatalities according to national standards, concluding that a re-design of these devices is necessary before implementation.

The SVM model is a relatively new method in classification problems, and has been increasingly utilized in traffic safety research, including traffic incident detection, crash frequency prediction, and crash severity investigation. Li et al. (2008) evaluated the efficiency of SVM models in predicting motor vehicle crash occurrences, concluding that SVM models outperform negative binomial models and back-propagation neural networks in crash data prediction. Suárez Sánchez et al. (2011) applied a SVM model to forecast work-related accidents and also discovered that it is superior to back-propagation neural network models in accident classification. Ren and Zhou (2011) proposed a hybrid method by incorporating particle swarm optimization and SVM for traffic safety forecasting. Yu and Abdel-Aty (2013) applied a SVM model to predict real-time crash potential considering actual traffic data 5–10 min before crash occurrence. SVM model has been utilized in crash injury severity in a few studies as well. Li et al. (2012) applied the SVM model in crash injury severity analyses, concluding that SVM models outperform the popular ordered probit model in injury severity prediction and factor impact assessment. Yu and Abdel-Aty (2014) compared the performance of the SVM model, random parameter models, and fixed parameter models in predicting crash injury severity, concluding SVM and random parameter models outperform fix parameter models.

3. Data descriptions

A two-year crash dataset containing all rollover crashes in New Mexico from 2010 to 2011 was utilized in this research. The crash data were obtained from the Traffic Safety Division at NMDOT and the Geospatial and Population Studies Transportation Research Unit (GPS-TRU) at the University of New Mexico (UNM), and were extracted from standard crash police reports. The entire dataset consists of three subsets: crash data including explicit crash-level information regarding crash time, crash location, crash severity geometry, and environmental information at crash occurrence; vehicle data containing information about vehicle type and actions, occupant injury number and severity, travel lane features, and traffic control measures; driver data explaining driver injury severity, driver demographic features, and driver behavior characteristics. The response variable, driver injury severity, is defined by NMDOT with five injury levels: no apparent injury (coded as O), complaint of injury (coded as C), visible injury (coded as B), incapacitating injury (coded as A), and fatality (coded as K). Preliminary SVM analyses were conducted using the original five injury severity levels, and it was found that the trained model performs poorly on higher injury severity levels due to insufficient sample size. In order to obtain relatively satisfactory classification performance while minimize loss of information on injury severity, three categorical driver injury severity levels were defined in this research as follows: no injury (original Category O, coded as **N**), non-incapacitating injury (original Categories B and C, coded as **I**), and incapacitating injury and fatality (original Categories A and K, coded as **F**). In this dataset, each record indicates a vehicle/driver unit that is involved in a rollover crash along with its corresponding crash characteristics. Before data pre-processing, the studied dataset was scrutinized to eliminate incomplete and erroneous records, such as records where driver gender was “unknown”. Finally, a rollover crash dataset containing 3158 vehicle/driver records from 3106 rollover crashes was used for SVM modeling in this analysis. Continuous variables, such as crash location (in terms of the distance to the nearest intersection) and numeric variables with continuous integers, such as driver age and number of vehicles in a crash, were categorized based on authentic traffic crash studies or engineering research experience (Ding et al., 2015; Liu et al., 2015a; Wu and Zhang, 2016; Wu et al., 2014; Zou et al., 2013). Variables with similar impacts but with limited records of presence, such as drivers under the influence of alcohol or drugs, were combined as a single variable for model simplicity. Multiple categorical values with similar patterns within a variable, such as left turn and right turn in the variable “Vehicle Actions,” were also combined as “Turn” action. The descriptive statistics of these variables are listed in Table 1.

4. Methodology

4.1. Research design

In this study, SVM models are utilized for driver injury severity prediction. A SVM model treats driver injury prediction as a classification problem given the heterogeneous conditions present at the crash occurrence. Driver injury severity is considered as categorical with multiple exclusive nominal categories. Compared with other non-parametric models (Mathworks Inc., 2015a), such as decision trees, nearest neighbor classifiers, etc., SVM models produce higher predictive accuracy, and therefore have been gaining increasing popularity in traffic safety studies. However, similar to Bayesian network (BN) model (Chen et al., 2015a,c), a disadvantage of SVM models is that it lacks the capability of automatically selecting the relevant factors regarding the response variable and removes the insignificant ones based on certain criteria. Therefore, a variable selection procedure is indispensable to achieve feasible and efficient SVM modeling.

Variable importance ranking method is one of the common ways to evaluate variable importance when predicting the target variable. Additionally, there are several popular methods of evaluating variable relative importance, such as discrete choice models, CART, random forest (RF), etc. The CART model has proven to be an effective method in traffic crash analysis (Chang and Chen, 2005; Kashani and Mohaymany, 2011; Yu and Abdel-Aty, 2013b). In this study, the CART method was applied to assess variable importance with respect to driver injury severity outcomes, and significant variables were selected as input for SVM model training. After the optimal classifier was trained, sensitivity analysis was conducted using data perturbation and before-after result comparison techniques to evaluate the influences of explanatory variables on driver injury severity patterns.

4.2. SVM and kernel function

SVM model is a non-parametric method solving classification problems based on statistical learning theory, and it is a kernel-based classifier. Since the SVM model has been well documented in many previous studies (Chen et al., 2009; Li et al., 2012; Yu and Abdel-Aty, 2013b), the development procedure is not duplicated but summarized below. For training data of N records that are linearly separable,

$$(x_1, y_1), \ldots, (x_i, y_i), i = 1, 2, .., N \quad (1)$$

where y_i is the class variable and $y_i = \pm 1$, and $x_i \in R^k$ represents the vector composed of k explanatory variables. Learning an SVM model is a procedure to find the best hyperplane so that training records with $y_i = \pm 1$ are separated on each side of the hyperplane, and the distance of the closest records to this hyperplane on each side is maximized. This maximization problem could be solved by introducing Lagrange multiplier and the trained SVM classifier has the basic form as follows:

$$f(x) = \text{sign}\left[\sum_{\forall i, \alpha_{i>0}} y_i \alpha_i (x_i \times x) + b\right] \quad (2)$$

where α_i are the Lagrange multipliers, x is the support vector of the hyperplane which classifies records, b is a real number used to define the basic function of the hyperplane $\omega \times x + b = 0$, in which ω is a normal vector that is perpendicular to the hyperplane, and $\omega \times x$ is the dot product of ω and x.

For data which are not able to be separated by a linear hyperplane, non-linear transformation function Φ is needed to map data into higher dimensional space. There is a kernel function applied for this non-linear transformation and is defined as follows:

$$k\left(x_i \times x_j\right) = \Phi\left(x_i\right) \times \Phi\left(x_j\right) \quad (3)$$

There are two major types of kernel functions that have been developed and applied into SVM modeling: the inhomogeneous

Table 1
Variable definition and data description.

Variable description		Driver injury severity						Total
Severity **N**	Percentage	Severity **I**	Percentage	Severity **F**	Percentage			
Driver injury severity		1478	46.80%	1286	40.72%	394	12.48%	3158
Crash-level variables								
First harmful event location								
	On road	1014	46.47%	887	40.65%	281	12.88%	2182
	Off road	464	47.54%	399	40.88%	113	11.58%	976
Lighting condition								
	Dark	494	44.54%	457	41.21%	158	14.25%	1109
	Dawn/dusk	89	54.94%	53	32.72%	20	12.35%	162
	Daylight	895	47.43%	776	41.12%	216	11.45%	1887
Weather								
	Sunny	1010	41.07%	1096	44.57%	353	14.36%	2459
	Adverse	468	66.95%	190	27.18%	41	5.87%	699
Road curvature								
	Curve road	372	44.93%	342	41.30%	114	13.77%	828
	Straight road	1106	47.47%	944	40.52%	280	12.02%	2330
Road grade								
	Road with grade	413	50.80%	309	38.01%	91	11.19%	813
	Level road	1065	45.42%	977	41.66%	303	12.92%	2345
Number of vehicles in crash								
	Single vehicle	1391	46.11%	1240	41.10%	386	12.79%	3017
	Two or more vehicles	87	61.70%	46	32.62%	8	5.67%	141
Road function								
	Urban	386	48.43%	333	41.78%	78	9.79%	797
	Rural non-interstate	713	45.44%	655	41.75%	201	12.81%	1569
	Rural interstate	379	47.85%	298	37.63%	115	14.52%	792
Maximum vehicle damage in crash								
	Slight damage	229	71.12%	76	23.60%	17	5.28%	322
	Functional damage	185	64.01%	90	31.14%	14	4.84%	289
	Disabled damage	1064	41.77%	1120	43.97%	363	14.25%	2547
Crash location (Distance to nearest intersection)								
	Within 0.1 mile	935	47.66%	786	40.06%	241	12.28%	1962
	0.1–1.0 mile	125	38.82%	149	46.27%	48	14.91%	322
	Further than 1.0 mile	418	47.83%	351	40.16%	105	12.01%	874
Crash time								
	Morning (6:00am–12:00pm)	451	49.24%	377	41.16%	88	9.61%	916
	Afternoon (12:00–pm-6:00pm)	460	47.72%	386	40.04%	118	12.24%	964
	Evening (6:00pm–0:00am)	335	46.66%	276	38.44%	107	14.90%	718
	Night (0:00am–6:00am)	232	41.43%	247	44.11%	81	14.46%	560
Vehicle-level variables								
Driver residency								
	Non New Mexico driver	436	49.49%	338	38.37%	107	12.15%	881
	New Mexico driver	1042	45.76%	948	41.63%	287	12.60%	2277
Driver license restriction								
	No restriction	1170	47.72%	975	39.76%	307	12.52%	2452
	With restriction	308	43.63%	311	44.05%	87	12.32%	706
Road pavement								
	Road paved	1337	46.20%	1189	41.09%	368	12.72%	2894
	Road not paved	141	53.41%	97	36.74%	26	9.85%	264
Road surface								
	Adverse road	601	66.12%	260	28.60%	48	5.28%	909
	Dry road	877	39.00%	1026	45.62%	346	15.38%	2249
Traffic control								
	Traffic control	393	44.61%	380	43.13%	108	12.26%	881
	No traffic control	1085	47.65%	906	39.79%	286	12.56%	2277
Number of lanes available for that car's travel								
	One lane	751	46.16%	676	41.55%	200	12.29%	1627
	Two lanes	612	47.85%	503	39.33%	164	12.82%	1279
	Three or more	115	45.63%	107	42.46%	30	11.90%	252
Vehicle type								
	Light vehicle	456	42.11%	484	44.69%	143	13.20%	1083
	Heavy vehicle	1022	49.25%	802	38.65%	251	12.10%	2075
Vehicle action								
	Go straight	1334	46.59%	1163	40.62%	366	12.78%	2863
	Acceleration or deceleration	50	46.73%	42	39.25%	15	14.02%	107
	Turn	94	50.00%	81	43.09%	13	6.91%	188
Driver seatbelt use								
	Seatbelt is used	1462	48.91%	1233	41.25%	294	9.84%	2989

Table 1 (*Continued*)

Variable description		Driver injury severity						Total
Severity **N**	Percentage	Severity **I**	Percentage	Severity **F**	Percentage			
	Seatbelt not used	16	9.47%	53	31.36%	100	59.17%	169
Driver age								
	Young: 24 or younger	451	44.88%	424	42.19%	130	12.94%	1005
	Mid-aged: between 25 to 63	937	48.08%	787	40.38%	225	11.54%	1949
	Senior: 64 or older	90	44.12%	75	36.76%	39	19.12%	204
Driver under influence								
	Driver under influence	112	26.73%	194	46.30%	113	26.97%	419
	Driver not under influence	1366	49.87%	1092	39.87%	281	10.26%	2739
Driver gender								
	Male	1082	50.73%	793	37.18%	258	12.10%	2133
	Female	396	38.63%	493	48.10%	136	13.27%	1025

polynomial function and the Gaussian radial basis function (RBF), as defined below:

$$K_{\text{poly}}\left(x_i \times x_j\right) = \left[\left(x_i \times x_j\right) + 1\right]^P \quad (4)$$

$$K_{\text{Gaussian}}\left(x_i \times x_j\right) = \exp\left[-\gamma x_i - {x_j}^2\right] \quad (5)$$

where P is the exponential parameter defining the polynomial function and γ is the kernel parameter that controls the width of Gaussian. In this study, three polynomial kernels are considered: linear kernel ($P=1$), quadratic kernel ($P=2$) and cubic kernel ($P=3$); while for Gaussian RBF kernel, three specific kernel settings are used: fine Gaussian Kernel ($\gamma = 0.1$), medium Gaussian kernel ($\gamma = 0.5$), and coarse Gaussian kernel ($\gamma = 1$).

The box constraint level, parameter used to "keep the allowable values of the Lagrange multipliers in a 'box', a bounded region" (Mathworks Inc., 2015b), is used for model calibration. The SVM classifiers are tuned by increasing the box constraint level. An increase of the box constraint level leads to a decrease of the number of support vectors, but will increase training time. In this research, to better compare the performance of these SVM models with different classifiers, the box constraint level is set as 1.

As indicated before, the SVM model was originally designed for binary classification problems, but it is applicable to multi-categorical problems after some modifications regarding variable definition and model structure. Lingras and Butz (2007) developed one-versus-one and one-versus-all approaches to address multi-categorical classifications in SVM. For a classification problem with Q classes in the predicted variable, a one-versus-all strategy is used to define Q binary SVM classifiers and each classifier are trained to identify one class from all other (Q-1) classes; a one-versus-one strategy is used to train $Q(Q\text{-}1)/2$ binary SVM classifiers for all possible pairs of classes and each classifier is used to examine each pair of interested classes. In this study, the one-versus-one strategy is utilized to fully examine the discrepancy among three injury severity levels.

4.3. *Variable importance ranking and predicting variable selection*

Variable selection is an indispensable procedure in classification problems in order to reduce the noise introduced by insignificant factors and improve model estimation accuracy and efficiency, especially for classification models lacking inborn variable selection procedures, such as BN, artificial neural network (ANN), SVM, etc. Several machine learning techniques, including decision tree, CART, RF, etc., have been developed to evaluate variable relative importance with respect to predicted variables based on certain criteria to assist variable selection. The CART technique has been utilized to address multiple aspects of traffic safety problems and has proven to be effective in variable selection and crash outcome prediction (Chang and Chen, 2005; Hossain and Muromachi, 2013; Kashani and Mohaymany, 2011; Montella et al., 2012). Therefore, in this study, the CART technique was used to evaluate the relative importance of predictor variables and the most important variables were selected for SVM model prediction. Since the CART model specifications have been described explicitly in previous authentic studies (Chang and Chen, 2005; Kashani and Mohaymany, 2011; Kuhnert et al., 2000), the development procedure of the CART model is not duplicated here.

5. Result discussion

5.1. *Variable selection result*

The variable selection procedure based on three injury severity levels was conducted on Salford Predictive Modeler, a data mining and predictive analytic platform developed by Salford Systems Company. The entire dataset with a total of 22 predictor variables was imported for variable importance analysis, and the relative variable importance through CART modeling is listed in Table 2. According to Banerjee et al. (2008), the variable importance score learned from the CART model measures and sums the contribution that a variable makes as a primary splitter or a surrogate to the primary splitter in the CART structure in improving response prediction. The variable with the largest overall improvement is scored 100, and all other variables have their scores relatively scaled to the best performing variable and ranged downwards toward zero. It was shown that among all these variables, driver seatbelt usage is most related to driver injury severity outcomes in rollover crashes with a ranking score of 100. Lighting condition and road grade were surprisingly found not to be associated with driver injury severity suggested by their ranking scores equal to 0. It was also found that road curvature and first harmful event location have little importance to driver injury outcomes, which was suggested by their score values of less than 1.0. Therefore, in this study, the last four variables with the lowest ranking scores were removed: road grade, lighting condition, road curvature and first harmful event location, and the rest 18 variables were used as the input for SVM classifier training and injury severity prediction. It should be noted that these scores are specific to a certain trained model, and may be totally different if they are learned from another CART structure. Therefore, in this study, these variable importance scores were only used for variable selection, but not for quantitative interpretations of variable influence.

5.2. *Model performance*

The selected 18 variables in Section 5.1 were used as inputs for SVM classifier learning. In order to comprehensively investigate the applicability and performance of SVM modeling on driver injury severity prediction, both quadratic and cubic kernel functions were employed for SVM classifier training. The whole dataset

Table 2
CART variable importance ranking result.

Variable	Score	Variable	Score
Driver seatbelt use	100.000	Road pavement	4.747
Road surface	**49.169**	**Driver age**	**4.726**
Weather	**41.760**	**Road function**	**4.668**
Maximum vehicle damage	**24.033**	**Driver residency**	**3.202**
Driver under influence	**12.148**	**Vehicle type**	**2.896**
Crash time	**9.908**	**Driver gender**	**1.688**
Traffic control	**9.622**	**Driver license restriction**	**1.145**
Vehicle action	**6.786**	First harmful event location	0.553
Number of vehicle in crash	**5.743**	Road curvature	0.212
Crash location	**5.628**	Lighting condition	0.000
Number of lanes available for that car's travel	**5.382**	Road grade	0.000

Table 3
SVM medium Gaussian RBF kernel classifier performance.

		Three injury severity levels				Two injury severity levels			
Training		Test		Training		Test			
Training dataset 1(60% of the whole dataset)	Correctly classified instances	Number 1785	Percentage 94.20%	Number 578	Percentage 45.76%	Number 1804	Percentage 95.20%	Number 681	Percentage 53.92%
	Incorrectly classified instances	Number 110	Percentage 5.80%	Number 685	Percentage 54.24%	Number 91	Percentage 4.80%	Number 577	Percentage 46.08%
	Total number of instances	1895		1263		1895		1263	
Training dataset 2(70% of the whole dataset)	Correctly classified instances	Number 2068	Percentage 93.57%	Number 429	Percentage 45.25%	Number 2096	Percentage 94.84%	Number 510	Percentage 53.80%
	Incorrectly classified instances	Number 142	Percentage 6.43%	Number 519	Percentage 54.75%	Number 114	Percentage 5.16%	Number 438	Percentage 46.20%
	Total number of instances	2210		948		2210		948	
Training dataset 3(80% of the whole dataset)	Correctly classified instances	Number 2359	Percentage 93.39%	Number 277	Percentage 43.83%	Number 2390	Percentage 94.62%	Number 344	Percentage 54.43%
	Incorrectly classified instances	Number 167	Percentage 6.61%	Number 355	Percentage 56.17%	Number 136	Percentage 5.38%	Number 288	Percentage 45.57
	Total number of instances	2526		632		2526		632	

Table 4
SVM cubic kernel classifier performance.

		Three injury severity levels				Two injury severity levels			
Training		Test		Training		Test			
Training dataset 1(60% of the whole dataset)	Correctly classified instances	Number 1556	Percentage 82.11%	Number 643	Percentage 50.91%	Number 1586	Percentage 83.69%	Number 791	Percentage 62.63%
	Incorrectly classified instances	Number 339	Percentage 17.89%	Number 620	Percentage 49.09%	Number 309	Percentage 16.31%	Number 502	Percentage 37.37%
	Total number of instances	1895		1263		1895		1263	
Training dataset 2(70% of the whole dataset)	Correctly classified instances	Number 1765	Percentage 79.86%	Number 481	Percentage 50.74%	Number 1812	Percentage 81.99%	Number 591	Percentage 62.34%
	Incorrectly classified instances	Number 445	Percentage 20.14%	Number 467	Percentage 49.26%	Number 398	Percentage 18.01%	Number 357	Percentage 37.66%
	Total number of instances	2210		948		2210		948	
Training dataset 3(80% of the whole dataset)	Correctly classified instances	Number 1989	Percentage 78.74%	Number 311	Percentage 49.21%	Number 2055	Percentage 81.35%	Number 379	Percentage 59.97%
	Incorrectly classified instances	Number 537	Percentage 22.26%	Number 321	Percentage 50.79%	Number 471	Percentage 18.65%	Number 253	Percentage 40.03%
	Total number of instances	2526		632		2526		632	

was divided into two sub-datasets, with one for model training and the other for model performance testing, based on three different splitting ratios: 6:4 (60% of the whole dataset as training dataset), 7:3 (70% of the whole dataset as training dataset) and 8:2 (80% of the whole dataset as training dataset). It was also revealed that converting a multi-categorical response variable into a binary response variable was a solution to improve SVM classification performance (Li et al., 2012). Therefore, the driver injury severity was converted as a binary outcome by aggregating non-incapacitating injury (Severity **I**) and incapacitating injury/fatality (Severity **F**) into a single category, and a new SVM classifier was trained for model performance comparison purposes. Preliminary performance tests indicate that medium Gaussian RBF kernels and cubic kernels perform best within each kernel family respectively, and therefore they are selected as the candidate kernel functions for this analysis. The performance of training SVM classifiers based on these two kernel functions is shown in Table 3 and 4, respectively.

Numerous performance patterns could be revealed by separate and comparative examinations of Tables 3 and 4. By comparing the general performance of these two kernel functions, it was revealed that the SVM classifier with medium Gaussian RBF kernel performed better than the SVM classifier with cubic kernel function on the training datasets, regardless of the number of injury severity levels. For instance, with the three training and testing dataset settings, the medium Gaussian SVM classifier produced training accuracies ranging from 93.39% to 94.20% on three injury severity levels and from 94.62% to 95.20% on two injury severity levels; the cubic SVM classifier produced relative inferior classification results, showing training accuracies from 78.74% to 82.11% and from 81.35% to 83.69% for three injury severity and two injury severity definitions, respectively. Although the trained medium Gaussian SVM classifier performed better than the cubic SVM classifier, it produced lower prediction accuracies on testing datasets, indicated by the cross-comparison of prediction accuracy on the same testing dataset. For example, the trained medium Gaussian SVM classifier produced a training accuracy of 94.20% on Training Dataset 1 for three injury level classifications. This is better than the performance of the SVM cubic kernel classifier on the same dataset (82.11%), but it is only able to correctly predict 45.76% of the corresponding testing dataset, which is significantly lower than that from the cubic SVM kernel classifier (50.91%). Consistent results were also found for all the other training and testing dataset pairs. These results reveal that there is an overfitting issue using medium Gaussian SVM classifiers in this study, which may generate biased performance results. Therefore, the cubic SVM classifier was preferred and selected as the optimal classifier for performance discussion in the following paragraphs and variable impact analysis in Section 5.2.

An examination of Table 4 reveals that the trained cubic SVM classifier outputs comparative performances for Testing Dataset 1 (40% of the whole dataset) and Testing Dataset 2 (30% of the whole dataset) with the prediction accuracies of 50.91% and 50.74% for three injury level classifications. It works relatively poorly on Testing Dataset 3 (20% of the whole dataset), outputting a correct prediction rate of 49.21%. Identical patterns were also illustrated for binary classifications. Therefore, it is suggested that a sufficient testing data size is necessary to generate the reasonable model classification performance and avoid biased estimation. Lateral performance comparisons in Table 4 indicate that it is an effective way to improve model performance by converting a multi-classification problem into a binary classification problem. For example, the cubic SVM classifier is able to correctly predict 50.91% of Testing Dataset 1 (40% of the whole dataset) with three driver injury severity levels, while it is significantly improved to 62.63% on the same testing data size after it was aggregated into a binary injury prediction problem. This pattern is also verified by the medium Gaussian SVM performance shown in Table 3. This finding is consistent with many authentic studies and it proves that converting a response variable with multiple values into a binary outcome variable is a popular approach to improve model prediction performance (Delen et al., 2006; Li et al., 2012; Tax and Duin, 2002).

Table 5 is the classification matrix produced by the cubic SVM model on Testing Dataset 1 (40% of the whole dataset), and it illustrates the overestimation and underestimation between each pair of injury severity levels. In this table, each row represents the actual number of observed instances for each injury severity level, and each column shows the number of predicted instances for each injury severity level. The diagonal cells display the number of correction predictions, and non-diagonal cells are the amounts of misclassifications. As it is shown, the trained SVM classifier performs best on the no injury category with a prediction accuracy equal to 58.77%, which is followed by a prediction accuracy of 50.46% for the non-incapacitating injury category; the trained model performs most poorly with the incapacitating injury and fatality category and is only able to correctly predict 22.67% of all fatal records. This finding is consistent with Li et al. (2012), where the trained model performs inferior for higher injury severities due to the insufficient number of observations in these injury severity levels.

With the development of mathematical and computational techniques, more advanced statistical models, such as mixed-logit models, have been proposed to model traffic crash data by capturing crash injury severity patterns, evaluating factor influence on injury outcome and predicting injury severity for new records. In this study, a mixed logit model was utilized the same training and testing datasets used in Table 5 (6:4 data splitting ratio) for model comparison analysis, and the produced confusion matrix is shown in Table 6. It is found in Table 6 that the overall prediction accuracy is 55.58% (702 correct predictions out of 1263 records), which is higher than that from the cubic SVM model (45.76%). However, a deeper examination of the confusion matrix reveals that the mixed logit model performs inferior on fatal record prediction, with 0 fatal record correctly classified. This is because that the mixed logit model is developed based on certain statistical assumptions regarding model development and data structure, which may not always hold for a typical crash dataset. While the SVM model provides a non-parametric alternative to predict injury severity outcomes in traffic crashes by releasing these statistical restrictions and working as a “black-box”, which is more universally applicable to different crash datasets.

5.3. Variable Impact Analysis

As a non-parametric classification model, SVM has been criticized for its veiled performance, where the impacts of contributing factors on the response variable are not accessed. Sensitivity analysis is an effective method of measuring variable impact from a statistical perspective. A two-stage sensitivity analysis method has been utilized in existing SVM crash studies (Li et al., 2012; Yu and Abdel-Aty, 2013b), and it is also used in this research as follows: first, each explanatory variable was changed by a user-defined amount while other variables remain unchanged; then the probabilities of each injury severity level before and after this perturbation were simulated in the cubic SVM model and recorded in Table 7. Similar to the pseudo-elasticity analysis in (Kim et al., 2007) and taking one of the values for each variable as the base category, the probability percentage change is calculated and shown in Table 8 to illustrate variable influence on driver injury outcomes. The model training procedure was disabled to ensure that the trained model structure was not altered by new testing datasets. For binary indicator variables, i.e. driver seatbelt use, the probabilities with the presence and absence of the condition were evaluated and compared. For variables with multi-categorical values, such as vehicle type, driver action, etc., the impact of each categorical value on driver injury severity was assessed and recorded.

Evident injury outcome patterns could be found in Tables 7 and 8. Weather condition at crash occurrence is an important factor contributing to driver injury outcomes. It was found that drivers are more likely to suffer incapacitating and fatal injuries in rollover crashes happening under adverse weather conditions, with a probability of 0.113, which is 31.40% higher than that in sunny weather which is equal to 0.086. This is reasonable since adverse weather conditions, such as rainy, snowy, windy, etc., may reduce drivers’ visibility and vehicles’ maneuverability, which requires drivers to make more effort to maintain normal vehicle operations. Similar conclusions could also be drawn on road pavement conditions, as it was shown that unpaved roads (0.108) tend to induce more driver incapacitating/fatal injuries, with a probability 28.57% higher than that for its counterpart. Road surface condition is also a significant factor to driver injury severity, and it is a factor similar to but not necessarily related to weather conditions. It is found in this study, on the contrary, that adverse road surface tends to induce less driver injuries and fatalities than inferior road surface conditions, i.e. slush, icy, snow, etc., with the corresponding probabilities decreased by

Table 5
Cubic SVM classification confusion matrix (6:4 data splitting ratio).

		Predicted instances classified by severity			
		Severity **N** (596)	Severity **I** (544)	Severity **F** (123)	True positive rate
Observed instances classified by severity	Severity **N** (570)	335	200	35	58.77%
	Severity **I** (543)	215	274	54	50.46%
	Severity **F** (150)	46	70	34	22.67%

Table 6
Mixed logit model classification confusion matrix (6:4 data splitting ratio).

		Predicted instances classified by severity			
		Severity **N** (551)	Severity **I** (712)	Severity **F** (0)	True positive rate
Observed instances classified by severity	Severity **N** (570)	343	227	0	60.18%
	Severity **I** (543)	184	359	0	66.11%
	Severity **F** (150)	24	126	0	0%

Table 7
Variable impact analysis results.

Variable	Value	Severity		
		Severity **N**	Severity **I**	Severity **F**
Weather	Sunny	0.471	0.443	0.086
	Adverse	0.63	0.256	0.113
Road pavement	Unpaved	0.537	0.356	0.108
	Paved	0.496	0.42	0.084
Road surface	Dry	0.444	0.463	0.092
	Adverse	0.67	0.239	0.091
Maximum vehicle damage	Slight	0.662	0.245	0.093
	Functional	0.583	0.334	0.083
	Disable	0.47	0.446	0.084
Road function	Urban	0.484	0.423	0.093
	Rural non-interstate	0.499	0.428	0.073
	Rural interstate	0.523	0.351	0.126
Traffic control	No control	0.52	0.395	0.085
	Traffic control	0.457	0.454	0.089
Crash location	Within 0.1 mile	0.513	0.396	0.091
	0.1–1.0 mile	0.439	0.478	0.083
	Further than 1.0 mile	0.54	0.374	0.087
Number of vehicles in crash	Single	0.471	0.442	0.087
	Two or more	0.607	0.285	0.108
Number of lanes for car's travel	One lane	0.51	0.413	0.077
	Two lanes	0.524	0.388	0.088
	Three or more	0.411	0.488	0.101
Crash time	Morning	0.47	0.426	0.103
	Afternoon	0.485	0.422	0.093
	Evening	0.49	0.42	0.09
	Night	0.526	0.387	0.087
Vehicle action	Go straight	0.491	0.428	0.08
	Acceleration or deceleration	0.326	0.519	0.155
	Turn	0.496	0.342	0.162
Vehicle type	Light vehicle	0.517	0.391	0.092
	Heavy vehicle	0.489	0.43	0.081
Driver seatbelt use	Seatbelt used	0.508	0.431	0.061
	Seatbelt not used	0.169	0.295	0.536
Driver under influence	Under influence	0.337	0.402	0.261
	Not under influence	0.51	0.423	0.067
Driver age	Young: 24 or younger	0.534	0.373	0.093
	Mid-aged: between 25 to 63	0.489	0.443	0.068
	Senior: 64 or older	0.434	0.397	0.169
Driver gender	Female	0.365	0.564	0.071
	Male	0.567	0.344	0.089
Driver residency	New Mexico	0.495	0.411	0.094
	Non-New Mexico	0.499	0.436	0.065
Driver license restriction	No restriction	0.512	0.416	0.072
	With restriction	0.463	0.405	0.132

48.38% (from 0.463 to 0.239) and 1.09% (from 0.092 to 0.091), respectively. It is explainable that drivers tend to be more cautious when experiencing inferior road surface conditions and operate vehicles more discreetly. However, they are prone to speeding and careless driving under favorable road conditions, and these reckless behaviors compromise driving safety on roadways.

For maximum vehicle damage in a crash, it was found that with the increase of maximum vehicle damage in a rollover crash, the probability of driver non-incapacitating injury was also augmented.

Table 8
Sensitivity analysis results.

Variable	Value	Severity		
		Severity **N**	Severity **I**	Severity **F**
Weather	**Sunny***	**0.00**	**0.00**	**0.00**
	Adverse	33.76%	−42.21%	31.40%
Road pavement	Unpaved	8.27%	−15.24%	28.57%
	Paved	**0.00**	**0.00**	**0.00**
Road surface	**Dry**	**0.00**	**0.00**	**0.00**
	Adverse	50.90%	−48.38%	−1.09%
Maximum vehicle damage	**Slight**	**0.00**	**0.00**	**0.00**
	Functional	−11.93%	36.33%	−10.75%
	Disable	−29.00%	82.04%	−9.68%
Road function	**Urban**	**0.00**	**0.00**	**0.00**
	Rural non-interstate	3.10%	1.18%	−21.51%
	Rural interstate	8.06%	−17.02%	35.48%
Traffic control	**No control**	**0.00**	**0.00**	**0.00**
	Traffic control	−12.12%	14.94%	4.71%
Crash location	Within 0.1 mile	16.86%	−17.15%	9.64%
	0.1–1.0 mile	**0.00**	**0.00**	**0.00**
	Further than 1.0 mile	23.01%	−21.76%	4.82%
Number of vehicles in crash	**Single**	**0.00**	**0.00**	**0.00**
	Two or more	28.87%	−35.52%	24.14%
Number of lanes for car's travel	**One lane**	**0.00**	**0.00**	**0.00**
	Two lanes	2.75%	−6.05%	14.29%
	Three or more	−19.41%	18.16%	31.17%
Crash time	**Morning**	**0.00**	**0.00**	**0.00**
	Afternoon	3.19%	−0.94%	−9.71%
	Evening	4.26%	−1.41%	−12.62%
	Night	11.91%	−9.15%	−15.53%
Vehicle action	**Go straight**	**0.00**	**0.00**	**0.00**
	Acceleration or deceleration	−33.60%	21.26%	93.75%
	Turn	1.02%	−20.09%	102.50%
Vehicle type	**Light vehicle**	**0.00**	**0.00**	**0.00**
	Heavy vehicle	−5.42%	9.97%	−11.96%
Driver seatbelt use	**Seatbelt used**	**0.00**	**0.00**	**0.00**
	Seatbelt not used	−66.73%	−31.55%	778.69%
Driver under influence	Under influence	−33.92%	−4.96%	289.55%
	Not under influence	**0.00**	**0.00**	**0.00**
Driver age	Young: 24 or younger	9.20%	−15.80%	36.76%
	Mid-aged: between 25 to 63	**0.00**	**0.00**	**0.00**
	Senior: 64 or older	−11.25%	−10.38%	148.53%
Driver gender	Female	−35.63%	63.95%	−20.22%
	Male	**0.00**	**0.00**	**0.00**
Driver residency	New Mexico	−0.80%	−5.73%	44.62%
	Non-New Mexico	**0.00**	**0.00**	**0.00**
Driver license restriction	**No restriction**	**0.00**	**0.00**	**0.00**
	With restriction	−9.57%	−2.64%	83.33%

*Categories in bold are selected as based category.

The corresponding probabilities of driver non-incapacitating injury related to slight, functional, and disabled vehicle damage are 0.245, 0.334, and 0.446, respectively. The probabilities of the no injury category decrease accordingly. The maximum vehicle damage could be considered as a visible reflection of the impact generated from the crash, and significant vehicle deformation was generally associated with severe casualties. Analogously, it was shown that rollover crashes on rural interstate roadways were most likely to result in driver incapacitating and fatal injuries, with a highest probability equal to 0.126 among all three road types. Additionally, rollover crashes on rural non-interstate roadways have the largest likelihood of inducing driver non-incapacitating injuries (0.428). These findings are reasonable since rural roadways usually have higher speed limits than urban roads. Drivers are also more likely to speed on rural roadways where there is less traffic volume. Both of these factors generate higher impact at crash occurrence and leave less time for drivers to respond properly, resulting in incapacitating injuries and fatalities. However, it is shown in Table 7 that the combined probability for injury and fatality is highest on urban roads among all road types, partly because of the relative denser population and complicated driving environment. Therefore, further individual investigation is desired for the urban and rural crashes regarding their distinctive injury patterns.

Traffic control measurement was identified as significant in CART variable importance ranking procedures. It is shown in Table 7 that the presence of a traffic control device, such as a no passing zone sign, stop or yield sign, signal control, etc., tends to increase the probability of driver injuries and fatalities compared with roadway sections without traffic control measurements. A similar pattern was also detected regarding crash location to the nearest intersection. It was shown that rollover crashes near intersections (less than 0.1 mile) are most likely to induce driver incapacitating injuries and fatalities, with a probability equal to 0.091, while roadways with a distance of 0.1–1.0 mile to the nearest intersection have the highest likelihood of resulting in driver non-incapacitating injuries in rollover crashes. A probable explanation is that roadways near intersections are accompanied by various types of traffic control devices and drivers need to follow them and alter vehicle operations accordingly. Any inappropriate acceleration or deceleration and insufficient driver reaction and perception time would result in crash occurrence and therefore lead to driver casualties.

The role that the number of vehicles in a rollover crash plays in deciding driver injuries severities could not be neglected. It is shown through comparisons in Table 7 that single-vehicle rollover crashes are more likely to result in a driver sustaining visible injuries, with a probability of 0.442. Multi-vehicle rollovers are more likely to result in driver incapacitating/fatal injuries and property damage only, with the probabilities equal to 0.108 and 0.607, respectively. The overall probabilities of injury and fatality for single-vehicle and multi-vehicle rollovers are 0.529 and 0.393, respectively. These statistics suggest that, although a slightly higher probability of driver incapacitating injuries and deaths exists, multi-vehicle rollover crashes still tend to produce less severe injury outcomes than single-vehicle rollover crashes. The number of available travel lanes demonstrates a monotonic effect on the driver incapacitating injury/death outcome, as illustrated by the estimated probabilistic influences for single-lane (0.077), two-lane (0.088), and multi-lane (0.101) designs. This indicates that the number of available travel lanes was positively associated with driver incapacitating and fatal injury outcomes in rollover crashes. It was also revealed that multi-lane design had the highest likelihood of inducing driver non-incapacitating injuries in rollover crashes, suggested by the estimated evidence probability (0.488).

A monotonic pattern was discovered regarding crash time impact on driver injury severity. It was found in Table 8 that as time goes on since morning time, the probabilities of drivers suffering non-incapacitating injury and incapacitating/fatal injuries decreased by 9.15% (from 0.426 to 0.387) and 15.53% (from 0.103 to 0.087), respectively. These findings imply that rollover crashes occurring in the morning are most likely to induce driver injuries and deaths, but the probability differences among different time periods with respect to the same injury severity levels are trivial.

Consistent conclusions could also be reached from the vehicle action impact analysis. It was found that, compared with running straight, vehicle speed change (acceleration and deceleration) and turning actions resulted in a higher likelihood of driver incapacitating injuries and deaths in rollover crashes. This is indicated by their corresponding probabilities of 0.155 and 0.162, which are 93.75% and 102.50% higher than that of vehicle running straight (0.080), respectively. These findings are consistent with Parenteau et al. (2003) discovering that trip-over crashes resulting from sudden slowdown or stop during vehicle lateral motion are the most prevalent type in passenger car and light truck vehicle rollover crashes. In terms of vehicle type, it has been revealed in data descriptions that heavy vehicles are more prone to suffering rollover crashes. No evident impact discrepancy has been detected on injury severity outcomes in our study as the estimated probabilities of both vehicle types for each injury category are comparable, as shown in Tables 7 and 8.

As is shown in Section 5.1, driver seatbelt use is the most important variable determining driver injury severity outcomes. It shows consistent results that drivers who do not use seatbelts suffer a significantly higher probability of incapacitating injuries and death than those wearing seatbelts in rollover crashes. The corresponding probability increased by 778.69%, from 0.061 to 0.536. The protective effect of seatbelts has been verified and evaluated in abundant studies (Carpenter and Stehr, 2008; Gross et al., 2007; Lerner et al., 2001). For instance, Carpenter and Stehr (2008) found that mandatory seatbelt use decreased severe and fatal injuries by almost 10% in fatal crashes. Therefore, seatbelt equipment utilization should be enforced at regional and national levels. Driver alcohol use or drug involvement also significantly increases driver incapacitating injury and fatality in rollover crashes. As shown in Table 7, compared with sober drivers, the probability of incapacitating and fatal injuries on drunk or drug-used drivers increased significantly by 289.55% from 0.067 to 0.261. It is understandable that drunk or drug-using drivers suffer from visibility and recognition impairment, which impairs their abilities of judging and driving properly.

Several other driver demographic characteristics were also closely related to driver injury outcomes: driver age, driver gender, driver residency, and driving license restrictions. Table 8 shows that compared with mid-age drivers, young drivers and senior drivers are more likely to suffer incapacitating and fatal injuries, with the corresponding probabilities 36.76% and 148.53% higher than that for mid-age drivers, respectively, and senior drivers are the most vulnerable group in rollover crashes overall. It is understandable that young drivers lack experience and tend to perform reckless driving more often, and senior drivers are less acute in responding and slower in operating vehicles properly than the other two groups at the occurrence of emergency. Both of these factors increase the risk of the driver suffering severe injuries and fatalities. It is also revealed that mid-age drivers are the group most associated with non-incapacitating injuries with a probability of 0.443. Driver residency was also associated with driver injury outcomes. The potential of New Mexico drivers suffering incapacitating and fatal injuries is 0.094, which is 44.62% higher than that for drivers from outside of the state. However, the two driver groups withstand comparative potential on both of the other two injury categories. This result is explainable since local drivers tend to perform more reckless driving because they are more familiar with local driving environments and traffic conditions, while non-local drivers tend to drive more carefully on strange roadways. Female drivers were found to have significantly larger potential to sustain visible injuries than male drivers in rollover crashes, with a probability equal to 0.564 which is 63.95% higher than the probability for male drivers, but in the meantime they have lower probabilities of sustaining no injuries and fatalities, which are 35.63% and 20.22% lower than the corresponding probabilities for male drivers, respectively. This discovery verifies the injury severity discrepancy between males and females, which has been widely discussed in previous studies (Islam and Mannering, 2006; Kockelman and Kweon, 2002; Massie et al., 1995). Moreover, it is suggested in Tables 7 and 8 that driver license restrictions, such as contact lenses, daytime driving restrictions, handicapped devices, etc., tend to increase the risk of the driver sustaining severe and fatal injuries with the probability increasing by 83.33% from 0.072 to 0.132. Therefore, efforts should be made by law enforcement, with the use of protective device development, to improve the safety of drivers with special driving needs.

6. Conclusion

Vehicle rollover crashes are a major source of fatal traffic crashes, and in-depth investigation of the injury severity distribution and heterogeneous factor impacts on injury severities in rollover crashes are of practical importance. SVM models are a popular non-parametric classification tool that has been widely used in transportation research, but is still relatively new in traffic crash analysis. Based on a two-year crash dataset in the state of New Mexico, this paper applies SVM models to predict driver injury severity outcomes in rollover crashes and investigate the probabilistic influences of contributing factors regarding crash, vehicle, and driver information on driver injury severity patterns. The driver injury severities are aggregated as a three-level categorical variable: property damage only (**N**), non-incapacitating injury (**I**), and incapacitating injury and fatality (**F**). Two popular kernel functions, including inhomogeneous polynomial kernel and Gaussian RBF kernel, are utilized to examine the applicability and performance of SVM models on crash driver injury prediction. A CART model is utilized to identify significant variables for driver injury severity prediction based on variable relative importance ranking, and the sensitivity analysis is conducted to estimate variable impacts on driver injury severity. Compared with peer studies, the trained cubic SVM classifier produces reasonable performances. In this study, the cubic SVM classifier outperforms the medium Gaussian RBF SVM classifier, and the trained cubic SVM classifier works best on no injury instances and worst on the incapacitating/fatal injury category. It is also verified that aggregating a multi-categorical response variable into a binary response variable is an effective approach to improve classification model performance.

The sensitivity analyses are conducted through data perturbation and before-after comparison techniques to quantify the contribution of the explanatory variables on the probability distribution of driver injury severities. It is found that driver alcohol or drug involvement is the most significant cause of driver incapacitating injuries and fatalities in rollover crashes, while seatbelt equipment is the most effective way to protect drivers from sustaining incapacitating injuries or being killed. For other driver and vehicle characteristics, it is revealed that senior drivers are the most vulnerable groups in rollover crashes; local drivers are more likely to suffer incapacitating injuries and deaths than non-local drivers; female drivers have a higher likelihood to be injured; male drivers are more self-protective in crash emergencies, but do suffer a slightly higher potential of incapacitating injuries and deaths. Vehicle movement change, including speed variation and turning actions, also increased the potential of incapacitating injuries and fatalities. The increasing number of traveling lanes, traffic control devices, unpaved roadways, and dry road surfaces also tend to increase injury or fatality potential. At the crash level, the maximum vehicle damage is positively associated with driver non-incapacitating injury potential. Other crash-level factors that increase the potential of incapacitating injuries and fatalities are rural interstate ways, multi-vehicle rollovers, morning crash times (6:00am-12:00pm), and crash locations within 0.1 mile to the nearest intersection. These results enhance the understanding of the impacts of these significant variables on driver injury and fatality in rollover crashes.

There are some limitations that need to be generalized, which may affect result estimation and interpretations. First, this research is based on a two-year rollover crash dataset, where the numbers of incapacitating injuries and fatalities were limited. The driver injury severity was aggregated into three levels due to the limited sample sizes, which inevitably led to loss of information to some extent. Meanwhile, the instances with less severe injuries, such as no injury and complaint of injury crashes, may be under-reported. Therefore, more complete datasets with sufficient records for each type of injury severity outcomes are desirable. In addition to Li et al. (2012) pointing out that the SVM model performance highly depends on learning procedure and parameter selection, it is suggested that performance also depends on training and testing datasets. In this study, the Gaussian RBF kernel function illustrates overfitting issues and performed worse on testing datasets than cubic SVM classifiers, but it may not be transferrable to datasets regarding other topics. More performance comparisons of these two kernel functions should be made on different datasets to compressively examine their applicability and effectiveness. Moreover, as is shown in this study, the trained cubic SVM classifier produces inferior performance on the incapacitating/fatal injury category. Other common kernel functions, such as homogeneous polynomial kernel function and hyperbolic tangent kernel function, may be applied and tested to improve model performance in the future.

Acknowledgment

This research was funded in part by the National Natural Science Foundation of China (grant nos. 51138003 and 51329801).

References

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec. 1746, 6–13.

Albertsson, P., Falkmer, T., Kirk, A., Mayrhofer, E., Björnstig, U., 2006. Case study: 128 injured in rollover coach crashes in Sweden—injury outcome, mechanisms and possible effects of seat belts. Saf. Sci. 44, 87–109.

Allahviranloo, M., Recker, W., 2013. Daily activity pattern recognition by using support vector machines with multiple classes. Transp. Res. Part B: Methodol. 58, 16–43.

Bambach, M.R., Grzebieta, R.H., McIntosh, A.S., 2013a. Thoracic injuries to contained and restrained occupants in single-vehicle pure rollover crashes. Accid. Anal. Prev. 50, 115–121.

Bambach, M.R., Grzebieta, R.H., McIntosh, A.S., Mattos, G.A., 2013b. Cervical and thoracic spine injury from interactions with vehicle roofs in pure rollover crashes. Accid. Anal. Prev. 50, 34–43.

Banerjee, A., Arora, N., Murty, U.S., 2008. Classification and regression tree (CART) analysis for deriving variable importance of parameters influencing average flexibility of CaMK kinase family. Electron. J. Biol. 4, 27–33.

Carpenter, C.S., Stehr, M., 2008. The effects of mandatory seatbelt laws on seatbelt use, motor vehicle fatalities, and crash-related injuries among youths. J. Health Econ. 27, 642–662.

Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 36, 365–375.

Chang, L.-Y., Chien, J.-T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51, 17–22.

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38, 1019–1027.

Chang, W.-H., Guo, H.-R., Lin, H.-J., Chang, Y.-H., 2006. Association between major injuries and seat locations in a motorcoach rollover accident. Accid. Anal. Prev. 38, 949–953.

Chen, S., Wang, W., van Zuylen, H., 2009. Construct support vector machine ensemble to detect traffic incident. Expert Syst. Appl. 36, 10976–10986.

Chen, C., Wang, G., Zhang, Y., Zhang, Y., Si, J., 2012. Effect of lateral stiffness coefficient of loader ROPS on human injury in a lateral rollover incident. Biosyst. Eng. 113, 207–219.

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., Guan, H., 2015a. A multinomial logit model-bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. 80, 76–88.

Chen, C., Zhang, G., Tian, Z., Bogus, S.M., Yang, Y., 2015b. Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. Accid. Anal. Prev. 85, 186–198.

Chen, C., Zhang, G., Wang, H., Yang, J., Jin, P.J., Walton, C.M., 2015c. Bayesian network-based formulation and analysis for toll road utilization Supported by traffic information provision. Transp. Res. Part C: Emerg. Technol. 60, 339–359.

Chen, C., Zhang, G., Yang, J., Milton, J.C., Alcántara Dely, A., 2016. An Explanatory analysis of driver injury severity in rear-end crashes using a decision table/naïve bayes (DTNB) hybrid classifier. Accid. Anal. Prev., http://dx.doi.org/10.1016/j.aap.2016.02.002 (in press).

Cheu, R.L., Xu, J., Kek, A.G.H., Lim, W.P., Chen, W.L., 2006. Forecasting of shared-use vehicle trips using neural networks and support vector machines. Transp. Res. Rec. 1968, 40–46.

Conroy, C., Hoyt, D.B., Eastman, A.B., Erwin, S., Pacyna, S., Holbrook, T.L., Vaughan, T., Sise, M., Kennedy, F., Velky, T., 2006. Rollover crashes: predicting serious injury based on occupant vehicle, and crash characteristics. Accid. Anal. Prev. 38, 835–842.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38, 434–444.

Ding, C., Ma, X., Wang, Y., Wang, Y., 2015. Exploring the influential factors in incident clearance time: disentangling causation from self-selection bias. Accid. Anal. Prev. 85, 58–65.

Dobbertin, K.M., Freeman, M.D., Lambert, W.E., Lasarev, M.R., Kohles, S.S., 2013. The relationship between vehicle roof crush and head: neck and spine injury in rollover crashes. Accid. Anal. Prev. 58, 46–52.

Farmer, C.M., Lund, A.K., 2002. Rollover risk of cars and light trucks after accounting for driver and environmental factors. Accid. Anal. Prev. 34, 163–173.

Fréchède, B., McIntosh, A.S., Grzebieta, R., Bambach, M.R., 2011. Characteristics of single vehicle rollover fatalities in three Australian states (2000–2007). Accid. Anal. Prev. 43, 804–812.

Franceschetti, B., Lenain, R., Rondelli, V., 2014. Comparison between a rollover tractor dynamic model and actual lateral tests. Biosyst. Eng. 127, 79–91.

Freeman, M.D., Dobbertin, K., Kohles, S.S., Uhrenholt, L., Eriksson, A., 2012. Serious head and neck injury as a predictor of occupant position in fatal rollover crashes. Forensic Sci. Int. 222, 228–233.

Funk, J.R., Cormier, J.M., Manoogian, S.J., 2012. Comparison of risk factors for cervical spine head, serious, and fatal injury in rollover crashes. Accid. Anal. Prev. 45, 67–74.

Gross, E.A., Axberg, A., Mathieson, K., 2007. Predictors of seatbelt use in American Indian motor vehicle crash trauma victims on and off the reservation. Accid. Anal. Prev. 39, 1001–1005.

Guo, L., Ge, P.-S., Zhang, M.-H., Li, L.-H., Zhao, Y.-B., 2012. Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine. Expert Syst. Appl. 39, 4274–4286.

Harris, J.R., Winn, G.L., Ayers, P.D., McKenzie, E.A., 2011. Predicting the performance of cost-effective rollover protective structure designs. Saf. Sci. 49, 1252–1261.
Hossain, M., Muromachi, Y., 2013. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.
Hu, W., Donnell, E.T., 2011. Severity models of cross-median and rollover crashes on rural divided highways in Pennsylvania. J. Saf. Res. 42, 375–382.
Huang, M.-L., 2015. Intersection traffic flow forecasting based on ν-GSVR with a new hybrid evolutionary algorithm. Neurocomputing 147, 343–349.
Huelke, D.F., Comp ton, C.P., 1983. Injury frequency and severity in rollover car crashes as related to occupant ejection: contacts and roof damage. Accid. Anal. Prev. 15, 395–401.
Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. J. Saf. Res. 37, 267–276.
Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.
Kim, J.-K., Kim, S., Ulfarsson, G.F., Porrello, L.A., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. Accid. Anal. Prev. 39, 238–251.
Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34, 313–321.
Kuhnert, P.M., Do, K.-A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Comput. Stat. Data Anal. 34, 371–386.
Lerner, E.B., Jehle, D.V.K., Billittier, A.J., Moscati, R.M., Connery, C.M., Stiller, G., 2001. The influence of demographic factors on seatbelt use by adults injured in motor vehicle crashes. Accid. Anal. Prev. 33, 659–662.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using Support Vector Machine models. Accid. Anal. Prev. 40, 1611–1618.
Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.
Lingras, P., Butz, C., 2007. Rough set based 1-v-1 and 1-v-r approaches to support vector machine multi-classification. Inf. Sci. 177, 3782–3798.
Liu, B., Koc, A.B., 2013. SafeDriving: a mobile application for tractor rollover detection and emergency reporting. Computers Electron. Agric. 98, 117–120.
Liu, J., Khattak, A.J., Richards, S.H., Nambisan, S., 2015a. What are the differences in driver injury outcomes at highway-rail grade crossings? Untangling the role of pre-crash behaviors. Accid. Anal. Prev. 85, 157–169.
Liu, J., Wang, X., Khattak, A.J., Hu, J., Cui, J., Ma, J., 2015b. How big data serves for freight safety management at highway-rail grade crossings: a spatial approach fused with path analysis. Neurocomputing 181, 38–52.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transp. Res. Part A: Policy Pract. 44, 291–305.
Mandell, S.P., Kaufman, R., Mack, C.D., Bulger, E.M., 2010. Mortality and injury patterns associated with roof crush in rollover crashes. Accid. Anal. Prev. 42, 1326–1331.
Mangado, J., Arana, J.I., Jarén, C., Arazuri, S., Arnal, P., 2007. Design calculations on roll-over protective structures for agricultural tractors. Biosyst. Eng. 96, 181–191.
Massie, D.L., Campbell, K.L., Williams, A.F., 1995. Traffic accident involvement rates by driver age and gender. Accid. Anal. Prev. 27, 73–87.
Mathworks Inc., 2015. Choose a Classifier-Statistics and Machine Learning Toolbox Documentation [WWW Document].
Mathworks Inc., 2015. Classification Learner [WWW Document]. URL http://www.mathworks.com/help/stats/classificationlearner-app.html?requestedDomain=www.mathworks.com (accessed 1.17.16.).
Montella, A., Aria, M., DAmbrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72.
National Highway Traffic Safety Adminstriation, 2013. Traffic Safety Facts 2012: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. U.S. Department of Transportation, Washington, D.C http://www-nrd.nhtsa.dot.gov/Pubs/812032.pdf.
New Mexico Department of Transportation, 2012. New Mexico traffic crash annual report 2011.
Parenteau, C.S., Viano, D.C., Shah, M., Gopal, M., Davies, J., Nichols, D., Broden, J., 2003. Field relevance of a suite of rollover tests to real-world crashes and injuries. Accid. Anal. Prev. 35, 103–110.
Ren, G., Zhou, Z., 2011. Traffic safety forecasting method by particle swarm optimization and support vector machine. Expert Syst. Appl. 38, 10420–10424.
Reynolds, S.J., Groves, W., 2000. Effectiveness of roll-over protective structures in reducing farm tractor fatalities. Am. J. Prev. Med. 18, 63–69.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43, 1666–15760.
Sheng, H., Xiao, J., 2015. Electric vehicle state of charge estimation: Nonlinear correlation and fuzzy support vector machine. J. Power Sour. 281, 131–137.
Suárez Sánchez, A., Riesgo Fernández, P., Sánchez Lasheras, F., de Cos Juez, F.J., García Nieto, P.J., 2011. Prediction of work-related accidents according to working conditions using support vector machines. Appl. Math. Comput. 218, 3539–3552.
Tax, D.M.J., Duin, R.P., 2002. Using two-class classifiers for multiclass classification. Proceedings of the 16th International Conference on Pattern Recognition 2, 124–127.
van der Westhuizen, S.F., Els, P.S., 2013. Slow active suspension control for rollover prevention. J. Terramech. 50, 29–36.
Wei, D., Liu, H., 2013. An adaptive-margin support vector regression for short-term traffic flow forecast. J. Intell. Transp. Syst. 17, 317–327.
Whitfield, R.A., Jones, I.S., 1995. The Effect of passenger load on unstable vehicles in fatal, untripped rollover crashes. Am. J. Public Health 85, 1268–1271.
Wu, Q., Zhang, G., 2016. Formulating alcohol influenced driver injury severities in intersection-related crashes. Transport.
Wu, Q., Chen, F., Zhang, G., Liu, X.C., Wang, H., Bogus, S.M., 2014. Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. Accid. Anal. Prev. 72, 105–115.
Wu, Qiong, Zhang, Guohui, Ci, Yusheng, Wu, Lina, Tarefder, Rafiqul A., Alcántara, Adélamar Dely, 2015. Exploratory multinomial logit model-based driver injury severity analyses for teenage and adult drivers in intersection-related crashes. Traffic Inj. Prev., http://dx.doi.org/10.1080/15389588.2015.1100722 (in press).
Yoganandan, N., Almusallam, A., Sances, A., 1990. Head and neck dynamics in an automobile rollover. Math. Computer Model. 14, 947–952.
Yu, R., Abdel-Aty, M., 2013a. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
Yu, R., Abdel-Aty, M., 2013b. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.
Yu, R., Wang, G., Zheng, J., Wang, H., 2013. Urban Road Traffic Condition Pattern Recognition Based on Support Vector Machine. J. Transp. Syst. Eng. Inf. Technol. 13, 130–136.
Zhang, N., Zhang, Y., Wang, X., 2013. Forecasting of short-term urban rail transit passenger flow with support vector machine hybrid online model, in: Transportation Research Board 92nd Annual Meeting Compendium of Papers. p. 16p.
Zou, Y., Zhang, Y., Lord, D., 2013. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. Accid. Anal. Prev. 50, 1042–1051.