Accident Analysis and Prevention 90 (2016) 95–107

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier

Cong Chen [a], Guohui Zhang [a,*], Jinfu Yang [b], John C. Milton [c], Adélamar "Dely" Alcántara [d]

[a] Department of Civil Engineering, University of New Mexico, Albuquerque, NM 87131, USA
[b] School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
[c] Quality Assurance and Transportation System Safety, Washington State Department of Transportation, Seattle, WA 98101, USA
[d] Geospatial and Population Studies Traffic Research Unit, University of New Mexico, Albuquerque, NM 87106, USA

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form
26 December 2015
Accepted 1 February 2016
Available online 27 February 2016

Keywords:
Driver injury severity
Rear-end crash
Decision table/Naïve Bayes (DTNB) classifier
ROC curve
Decision rules
Traffic safety

ABSTRACT

Rear-end crashes are a major type of traffic crashes in the U.S. Of practical necessity is a comprehensive examination of its mechanism that results in injuries and fatalities. Decision table (DT) and Naïve Bayes (NB) methods have both been used widely but separately for solving classification problems in multiple areas except for traffic safety research. Based on a two-year rear-end crash dataset, this paper applies a decision table/Naïve Bayes (DTNB) hybrid classifier to select the deterministic attributes and predict driver injury outcomes in rear-end crashes. The test results show that the hybrid classifier performs reasonably well, which was indicated by several performance evaluation measurements, such as accuracy, F-measure, ROC, and AUC. Fifteen significant attributes were found to be significant in predicting driver injury severities, including weather, lighting conditions, road geometry characteristics, driver behavior information, etc. The extracted decision rules demonstrate that heavy vehicle involvement, a comfortable traffic environment, inferior lighting conditions, two-lane rural roadways, vehicle disabled damage, and two-vehicle crashes would increase the likelihood of drivers sustaining fatal injuries. The research limitations on data size, data structure, and result presentation are also summarized. The applied methodology and estimation results provide insights for developing effective countermeasures to alleviate rear-end crash injury severities and improve traffic system safety performance.

1. Introduction

Traffic crashes induce tremendous costs every year in terms of human casualties and economic losses. According to the World Health Organization (WHO) (2013), approximately 1.24 million people are killed and about 50 million are injured in traffic crashes worldwide annually. In 2009, 33,808 fatalities resulted from traffic crashes (WHO, 2013), and each fatality costs about $1.4 million on average (National Health Council, 2013). Rear-end crashes are defined as a type of crash in which the rear side of a vehicle is hit by the front side of a following vehicle (Singh, 2003). Rear-end crashes have been one of the most common types of traffic crashes in the U.S., resulting in significant casualties. The National Safety Council (2011) reported that 3.54 million rear-end crashes occurred on U.S. roadways, which accounted for 33% of total reported crashes in 2009 and resulted in 1.078 million injuries and 2100 fatalities. Significant research efforts have been made to better understand the characteristics of rear-end crashes, explore the contributing factors regarding weather and environment, roadway geometric, vehicle information, etc. for crash frequency and severity, and develop effective countermeasures to reduce crashes risks and severities. For instance, Yan et al. (2005) investigated the features of rear-end crashes happening at signalized intersections through a multiple logistic regression model. Meng and Qu (2012) applied an inverse Gaussian regression model to estimate rear-end crash frequency in urban tunnels with a proposed exposure index. Meng and Weng (2011) examined the significant factors and their respective influences on the potential of rear-end crashes in work-zone areas. Harb et al. (2007) found that light truck vehicles are a significant contributor in rear-end crashes at non-signalized intersections due to the visibility limitation of truck drivers. Li and Bai (2008) identified that rear-end crashes are the most frequent accident type of injury-related accidents in highway construction zones. Numerous studies were also conducted to seek optimized car-following strategies in order to reduce rear-end crash frequency. Duan et al. (2013) evaluated the impact of leading vehicles on headway and the minimum safe headway for following cars and proposed distinctive car-following strategies under different weather and traffic scenarios.

* Corresponding author.
E-mail address: guohui@unm.edu (G. Zhang).

http://dx.doi.org/10.1016/j.aap.2016.02.002
0001-4575/

Broughton et al. (2007) investigated different car-following methods and their corresponding contributory factors under different visibility and velocity-combined conditions. At the microscopic level, research related to rear-end crashes usually focuses on ergonomics, especially regarding the impact on human cephalic and cervical regions. For example, Farmer et al. (1999) investigated the importance of proper head restraint positioning in reducing driver neck injury severity in rear-end crashes. Boström et al. (2000) developed a Neck Injury Criterion (NIC) to assess the impact of rear-end crashes at low velocities and predict neck injury risks with the maximum NIC.

In most circumstances, traffic safety analyses are analyzing-deciding procedures to explore the characteristics of traffic crashes and help to make optimal countermeasures. Mathematical modeling methods are one of the most effective types of modeling for these procedures. Bayesian method, as a burgeoning mathematical approach in the transportation field, has been widely used to address traffic safety issues. Yanmaz-Tuzel and Ozbay (2010) applied full Bayes (FB) models to explore effective countermeasures in reducing crash frequencies. Abdalla (2005) proposed a hierarchical Bayesian model to assess the protective effects of seatbelts in traffic crashes. Riviere et al. (2006) employed a Bayesian neural network to estimate the deformation energy a vehicle receives in traffic crashes using Energy Equivalent Speed (EES). El-Basyouny and Sayed (2013) implemented an FB approach into a multivariate analysis to identify traffic crash hotspots. Yu and Abdel-Aty (2013) proposed different methods for proper informative prior assignment in hierarchical Bayesian traffic safety models. Haque et al. (2010) examined the contributing factors of motorcycle-related crashes at intersections through hierarchical Bayesian Poisson models. Chen et al. (2015b) developed a hierarchical Bayesian random intercept model with cross-level interaction settings to investigate the interactive effects between crash and vehicle variables on driver injury severity outcomes. Strauss et al. (2013) applied a Bayesian model to evaluate the characteristics and injury risks of cyclists at signalized intersections. The Naïve Bayes (NB) approach is based on the "naïve" assumption that the existence or absence of an attribute is independent from that of the others, given the class attribute. NB approach performs well in different areas for pattern recognition. Youn and Jeong (2009) presented an NB-based method for factor importance ranking. Renooij and van der Gaag (2008) conducted a sensitivity analysis on NB classifiers based on the parameter inaccuracies resulting from evidence and scenarios. Wang et al. (2011) presented a hierarchical NB model to improve search effectiveness of identity matching techniques and, therefore, facilitate identity information management. Soria et al. (2011) proposed a revised NB classifier for breast cancer data analysis. Researchers are also interested in modifying the internal configurations to improve NB model performance. Wong and Chang (2011) proposed two approaches to explore the preferable prior settings for NB classifiers considering individual attribute impacts. Lee (2007) discovered that the involvement of unlabeled training data could improve the performance of the NB learning procedure. However, NB models have not been used in traffic safety analysis, which motivates the authors to fill this gap.

The major aim of this analysis is to identify the contributing factors of driver injury severity in rear-end crashes and discover the decision rules for driver injury prediction based on these factors. Besides mathematical modeling approaches, another important approach for decision-making and pattern discovery is a decision table (DT) based on scheme-specific attribute selection. Scheme-specific attribute selection is a procedure of selecting the best subset of attributes by evaluating the performance of learning schemes using different attribute subsets (Witten et al., 2011). DTs are a type of classifier with scheme-specific attribute selection, which discover and present sophisticated logic in a concise but accurate way and have been increasingly utilized in diverse areas. Lew (1991) discussed the application of fuzzy decision tables in expert systems for representing both knowledge and working procedures. Zhang et al. (2008) implemented a DT algorithm to classify celestial objects based on astronomical databases. Han et al. (1989) discussed the applicability of a DT method in computer code design. Seagle and Duchessi (1995) introduced a computer-based approach supported by a DT analyzer to extract expert rules in heuristic classification problems. Huysmans et al. (2011) compared the comprehensibility of DTs and other rule-based models, concluding that decision tables are significantly superior to other rule-based models in terms of accuracy, response time, and answer reliability. Witlox et al. (2009) discussed the application of functional classification theory in land use planning through DT models. Despite these multi-disciplinary applications, DT method has not been used in traffic safety studies, which also inspires the authors to do this research.

The practical importance of in-depth investigations on rear-end crashes is justified by the significant loss in these crashes. These aforementioned studies provide a comprehensive and insightful understanding of Bayesian models and DT algorithms. A Bayesian modeling approach provides a considerable advantages in model fit and result interpretation, which have been proved by previous studies (Huang and Abdel-Aty, 2010; Huang et al., 2008; Washington et al., 2005). However, no Bayesian concept has been incorporated in knowledge-based non-parametric machine-learning models. NB classifiers are able to infer conditional probabilities without any presumed interdependence among the explanatory variables. DTs are capable of selecting decisive factors and presenting decision rules comprehensively and concisely. Hall and Frank (2008) developed a hybrid model by incorporating the NB classifier with DT for classification problems, but it has never been used in traffic crash analysis. Based on rear-end crash data collected in New Mexico from 2010 to 2011, this paper utilizes this decision table/Naïve Bayes (DTNB) hybrid classifier as a new knowledge-based Bayesian non-parametric machine-learning model to identify the deterministic attribute set that best predicts driver injury severities in rear-end crashes and investigates the corresponding decision rules based on these attributes. Driver injury severities were classified into three categories: no injury, injury, and fatality. The deterministic attribute set and decision rules were extracted from a comprehensive set of explanatory attributes regarding weather information, crash characteristics, vehicle information, driver demographic and behavior information, etc. The research results demonstrate that the trained classifier performs reasonably well in driver injury severity prediction. The rest of this paper is organized as follows: the data description and preprocessing procedure are provided in Section 2. Section 3 introduces model structure and specifications of the DTNB hybrid classifier, and modeling results and discussions are presented in Section 4. The research limitations are summarized in Section 5, and this research is concluded in Section 6.

2. Data description and preprocessing

This study is conducted based on two-year rear-end crash data collected in New Mexico from 2010 to 2011, provided by New Mexico Department of Transportation (NMDOT) Traffic Safety Division and the Geospatial and Population Studies Transportation Research Unit (GPS-TRU) at the University of New Mexico (UNM). The studied dataset is composed of three major sub-datasets: crash data, vehicle data, and driver data, including all the information regarding crash types, locations, occurrence time, driver and occupant injury severities, roadway geometric characteristics, weather conditions, vehicle characteristics, and driver demographic and behavior. Careful examination of the data occurred to eliminate

incomplete and outlier information and improve data quality. The response variable *SEV* is defined to categorize driver injury severity into three levels: no injury (property damage only), injury, and fatality. Compared to discrete variables, processing continuous and numeric variables with an NB classifier is significantly more computational-intensive due to the estimation efficiency, and the produced results are less accurate (Dougherty et al., 1995; Kohavi and Sahami, 1996). Therefore, continuous and numeric variables were categorized as a countable number of exclusive nominal states to enhance model performance. In this study, continuous and numeric variables were categorized following previous related studies (de Oña et al., 2011; Huang et al., 2008) or based on engineering experience. Additional effort was made to ensure an approximately comparable amount of data records in each exclusive state category for the same variable. For instance, *DTINC*, a categorized variable representing the distance (in miles) from the crash location to the nearest intersection, was coded as three nominal alternative states: NEAR, MID, and FAR. NEAR denotes the distance was less than 0.1 mile, MID indicates it was between 0.1 and 1.0 mile, and FAR means it was larger than 1.0 mile. Correlation analyses were also conducted to avoid significant correlations among the explanatory variables, according to the "naïve" assumption of inter-independence for NB models. For variables with significant correlations, variables most related to driver injury severities were kept for model estimation and less significant ones were removed, based on traffic engineering experience. For example, crash occurrence time, which is collinear with environment lighting conditions, was removed since environment lighting conditions are more closely associated with drivers' visibility and, therefore, driver injury severities. In total, a dataset including 23,433 driver injury records from 11,383 rear-end crashes were retained for model development and decision rule learning. The descriptive statistics after preprocessing are illustrated in Table 1.

3. Model description and specification

3.1. DT formation

A decision table (DT) is a scheme-specific learning algorithm modeling that presents complicated logics (Witten et al., 2011). It is defined as a table representing a complete set of decision rules under all mutually exclusive conditional scenarios in a pre-defined problem (Witlox et al., 2009). A standard DT consists of four parts. In a DT the upper left part is a list of all the conditions, denoted as C_i for $i = 1, \ldots, c$, where c is the number of conditions in the problem. A condition-state set CS_i contains all the possible alternative states that C_i is able to attain within a particular pre-defined problem:

$$CS_i = \{S_{i1},\ S_{i2},\ \ldots,\ S_{it_i}\} \tag{1}$$

where t_i is the number of alternative states for the ith condition C_i in the pre-defined problem.

The upper right part of a decision table is its condition space, which is a Cartesian product of all the condition-state sets CS_i ($i = 1, \ldots, c$), as shown below:

$$\begin{aligned} SP(C) &= CS_1 \times CS_2 \times \ldots \times CS_c \quad \text{for } c > 1 \\ &= CS_1 \quad \text{for } c = 1 \end{aligned} \tag{2}$$

Each element in the condition space is a condition entry (CE) with ordered c dimensions (also known as an ordered c-tuple) (Witlox et al., 2009), and the whole set of these condition entries in the decision table is defined as the domain of a DT, denoted as DOM(DT).

The lower left part in a DT consists of all the possible action subjects used to express the decisions, represented as A_j for $j = 1, \ldots, a$, where a is the number of all possible actions. Similar to CS_i, an action-state set AT_j includes all the attainable states for action A_t within a particular pre-defined problem, defined as:

$$AT_j = \{T_{j1},\ T_{j2},\ \ldots,\ T_{jm_j}\} \tag{3}$$

where m_j is the number of alternatives for A_j in the pre-defined problem.

The lower right part of a decision table is its action space, which is also a Cartesian product of the all the action sets AT_j for $j = 1, \ldots, a$,

$$\begin{aligned} SP(A) &= AT_1 \times AT_2 \times \ldots \times AT_a \quad \text{for } a > 1 \\ &= AT_1 \quad \text{for } a = 1 \end{aligned} \tag{4}$$

Similar to the condition space, each element in the action space is an a-dimensional action entry (AE).

The presentation of a complete DT is a matrix and could be written as follows: Let n be the number of decision rules (columns) and c be the number of conditions (rows). The condition part of a DT is then expressed as,

$$D = (d_{ir}), \quad i = 1,\ \ldots,\ c \text{ and } r = 1,\ \ldots,\ n$$

where $d_{ir} \in CS_i$.

The action part could be expressed as:

$$E = (e_{jr}), \quad j = 1,\ \ldots,\ a \text{ and } r = 1,\ \ldots,\ n$$

Where a is the number of actions (rows) and $e_{jr} \in AT_j$. Therefore, A DT specifies the relations between condition space and action space as,

$$DT = (dt_{qr}) = \begin{pmatrix} D \\ E \end{pmatrix}$$

$$\begin{aligned} \text{Where } dt_{qr} &= d_{qr}, \quad \text{for } q = 1,\ \ldots,\ c \text{ and } r = 1,\ \ldots,\ n \\ &= e_{(q-c)r}, \quad \text{for } q = c+1,\ \ldots,\ c+a \text{ and } r = 1,\ \ldots,\ n \end{aligned}$$

In application a DT is used as a lookup table based on the selected attributes. Each entry in the DT is associated with the class probability estimation based on the observed frequencies in the original dataset. The critical procedure of learning a decision table is the selection of highly discriminative attributes, given the class variable, and it is normally conducted by maximizing cross-validated performance (Hall and Frank, 2008). Cross-validation is efficient for DT learning since the learned structure would not change with the addition or deletion of instances, and only the class counts vary according to the entries. A detailed explanation of the cross-validation procedure is provided in Section 3.3.

3.2. NB classifier

In a classification task, assume Y is the class variable and $X = (X_1, X_2, \ldots, X_n)$ is the set of attribute variables. A Bayes classifier prediction of the value y of class variable Y is a process to find y that $P(Y = y_i)$ has the highest posterior conditional probability given $x = (x_1, x_2, \ldots, x_n)$, shown in Eq. (5). Using Bayes' Theorem (Chen et al., 2015c; Kruschke, 2015), it can be expressed as Eq. (6).

$$\begin{aligned} &P(Y = y_i | X = x = (x_1, x_2,\ \ldots x_n)) \\ &> P(Y = y_j | X = x = (x_1, x_2,\ \ldots x_n)), \quad \forall j, j \neq i \end{aligned} \tag{5}$$

$$\begin{aligned} &P(Y = y_i | X = x = (x_1, x_2,\ \ldots,\ x_n)) \\ &= \frac{P(X = x = (x_1, x_2,\ \ldots,\ x_n) | Y = y_i) P(Y = y_i)}{P(X = x = (x_1, x_2,\ \ldots,\ x_n))} \end{aligned} \tag{6}$$

An NB classifier is a Bayes model with the conditional independence assumption that the presence and absence of an attribute is independent from the presence and absence of other attributes in

Table 1
Variable definition and descriptive statistics.

Attribute		Value		SEV						Total
	NO INJURY		Percentage	INJURY	Percentage	FATALITY	Percentage			
DAY	Day	MON	Monday	2286	64.09%	1275	35.74%	6	0.17%	3567
		TUE	Tuesday	2516	61.76%	1550	38.05%	8	0.20%	4074
		WED	Wednesday	2610	63.88%	1470	35.98%	6	0.15%	4086
		THU	Thursday	2525	62.45%	1512	37.40%	6	0.15%	4043
		FRI	Friday	2712	61.04%	1710	38.49%	21	0.47%	4443
		SAT	Saturday	1259	61.96%	760	37.40%	13	0.64%	2032
		SUN	Sunday	711	59.85%	471	39.65%	6	0.51%	1188
RDREL	First harmful event location	ONWAY	On roadway	14,567	62.38%	8719	37.34%	66	0.28%	23,352
		OFFWAY	Off roadway	52	64.20%	29	35.80%	0	0.00%	81
LIGHT	Lighting condition	DAYLIGHT	Daylight	12,600	62.84%	7420	37.01%	31	0.15%	20,051
		DARK	Dark	1547	58.58%	1061	40.17%	33	1.25%	2641
		DAWN/DUSK	Dawn or dusk	472	63.70%	267	36.03%	2	0.27%	741
CURVE	Curvature	CURVE	Curve road	616	67.62%	295	32.38%	0	0.00%	911
		STAIGHT	Straight road	14,003	62.17%	8453	37.53%	66	0.29%	22,522
RDGRD	Road grade	LEVEL	Level	12,755	62.84%	7584	37.36%	59	0.29%	20,298
		HCRST	Hillcrest	365	57.21%	273	42.79%	0	0.00%	638
		ONGRADE	On grade	1434	59.50%	969	40.21%	7	0.29%	2410
		DIP	Dip	45	73.77%	16	26.23%	0	0.00%	61
		OTHER	Other road grade	20	76.92%	6	23.08%	0	0.00%	26
DRESD	Driver residency	ST	NM residency	12,679	63.01%	7423	36.89%	21	0.10%	20,123
		NST	Other state residency	1940	58.61%	1325	40.03%	45	1.36%	3310
NVEH	Number of vehicles involved	TWO	Two vehicles	11,872	67.51%	5671	32.25%	43	0.24%	17,586
		THREE	Three vehicles	2192	49.58%	2215	50.10%	14	0.32%	4421
		MORE	More than three vehicles	555	38.92%	862	60.45%	9	0.63%	1426
RDFUNC	Road function	URBN	Urban road	13,306	63.21%	7719	36.67%	26	0.12%	21,051
		RINT	Rural interstate	460	52.27%	404	45.91%	16	1.82%	880
		RNINT	Rural non-interstate	853	56.79%	625	41.61%	24	1.60%	1502
PEDINV	Pedestrian involvement	Y	Involved	5	38.46%	8	61.54%	0	0.00%	13
		N	Not involved	14,614	62.40%	8740	37.32%	66	0.28%	23,420
MCINV	Motorcycle involvement	Y	Involved	116	29.74%	265	67.95%	9	2.31%	390
		N	Not involved	14,503	62.94%	8483	36.81%	57	0.25%	23,043
HEVINV	Heavy Vehicle Involvement (including bus, pickup, semi-truck and lorries)	Y	Involved	388	52.29%	311	41.91%	43	5.80%	742
		N	Not involved	14,231	62.72%	8437	37.18%	23	0.10%	22,691
HZINV	Hazard material involvement	Y	Involved	6	46.15%	5	38.46%	2	15.38%	13
		N	Not involved	14,613	62.40%	8743	37.33%	64	0.27%	23,420
DTINC	Distance from accident location to intersection	NEAR	<0.1mile	4624	58.87%	3191	40.62%	40	0.51%	7855
		MID	0.1–1.0 mile	596	52.37%	536	47.10%	6	0.53%	1138
		FAR	>1.0 mile	9399	65.09%	5021	34.77%	20	0.14%	14,440
DLRST	Driver license restriction	RST	With restriction	3625	62.52%	2153	37.13%	20	0.34%	5798
		NORST	No restriction	10,994	62.34%	6595	37.40%	46	0.26%	17,635
RDPV	Road paving condition	PAVED	Paved surface	14,552	62.34%	8724	37.37%	66	0.28%	23,342
		UNPAVED	Unpaved surface	67	73.63%	24	26.37%	0	0.00%	91
TRFCTL	Traffic control	NCTL	No control	11,132	61.59%	6899	38.17%	42	0.23%	18,073
		SYSIGN	Stop/yield sign control	124	73.37%	45	26.63%	0	0.00%	169
		SGCTL	Signal control	507	64.18%	279	35.32%	4	0.51%	790
		RRGATE	Railroad gate	6	85.71%	1	14.29%	0	0.00%	7
		OTHER	Other control measures, such as passing zones, detours, etc.	2850	64.86%	1524	34.68%	20	0.46%	4394
NLANE	Number of lanes with same direction at accident location	ONE	One lane	3363	64.40%	1846	35.35%	13	0.25%	5222
		TWO	Two lanes	6024	62.65%	3550	36.92%	41	0.43%	9615
		MORE	More than two lanes	5232	60.87%	3352	38.99%	12	0.14%	8596
VACT	Vehicle action	STRT	Straight	9176	62.69%	5416	37.00%	46	0.31%	14,638
		BACK	Backup	35	89.74%	4	10.26%	0	0.00%	39
		SLOW	Slow	1411	62.24%	854	37.67%	2	0.09%	2267
		LTURN	Left turn	484	65.58%	250	33.88%	4	0.54%	738
		RTURN	Right turn	432	70.70%	176	28.81%	3	0.49%	611
		UTURN	U-turn	20	66.67%	10	33.33%	0	0.00%	30
		OTK	Overtaking	130	64.68%	71	35.32%	0	0.00%	201
		OTHER	Other action	2931	59.71%	1967	40.07%	11	0.22%	4909
VTYPE	Vehicle type	LVEH	Light vehicle, including passenger car or van	10,747	62.35%	6463	37.49%	27	0.16%	17,237
		HVEH	Heavy vehicle, including bus, pickup, semi-truck and lorries	3454	63.53%	1950	35.87%	33	0.61%	5437
		MC	Motorcycle	59	29.65%	136	68.34%	4	2.01%	199
		OTHER	Other	359	64.11%	199	35.54%	2	0.36%	560
DBELT	Driver seatbelt use	Y	Seatbelt used	13,698	62.03%	8332	37.73%	52	0.24%	22,082
		N	Seatbelt not used	921	68.17%	416	30.79%	14	1.04%	1351
DAGE	driver age	YOUNG	16–25	4744	65.28%	2510	34.54%	13	0.18%	7267
		MID	26–63	8814	60.91%	5608	38.75%	49	0.34%	14,471
		OLD	64 or older	1061	62.60%	630	37.17%	4	0.24%	1695

Table 1 (*Continued*)

Attribute		Value		SEV						Total
	NO INJURY		Percentage	INJURY	Percentage	FATALITY	Percentage			
DALC	Driver alcohol involvement	Y	Involved	115	44.40%	139	53.67%	5	1.93%	259
		N	Not involved	14,504	62.59%	8609	37.15%	61	0.26%	23,174
DSEX	Driver sex	M	Male	7967	63.89%	4454	35.72%	49	0.39%	12,470
		F	Female	6652	60.68%	4294	39.17%	17	0.16%	10,963
MAXDAM	Most serious vehicle damage	NSLT	No damage or slight damage	6147	72.65%	2312	27.33%	2	0.02%	8461
		FUNC	Functional damage that affects operations of vehicle	4284	68.61%	1960	31.39%	0	0.00%	6244
		DSABL	Disabled damage that vehicles cannot be driven	4188	47.98%	4476	51.28%	64	0.73%	8728

the attribute set, given the class variable value. Therefore, the predicting probability of class variable $Y = y_i$ conditioned on $X = x = (x_1, x_2, \ldots x_n)$ is as follows (Domingos and Plazzani, 1997):

$$P(Y = y_i|X = x = (x_1, x_2, \ldots, x_n)) = \frac{P(X = x = (x_1, x_2, \ldots, x_n)|Y = y_i)P(Y = y_i)}{P(X = x = (x_1, x_2, \ldots, x_n))}$$
$$\propto P(X = |x = (x_1, x_2, \ldots, x_n)|Y = y_i)P(Y = y_i) = P(Y =_{y_i})\Pi_{j=1}^{n} P(x_j|Y = y_i) \quad (7)$$

Although the NB classifier is based on the "naïve" conditional independent assumption, compared to the other classification algorithms, it still demonstrates preferable performance in analyzing many real datasets that do not strictly follow the conditional independent assumption. Specifically, the impact of the "naïve" assumption on the classification performance of an NB classifier would be insignificant if the classification tool was evaluated by zero-one loss or accuracy (Domingos and Plazzani, 1997). For attributes X_j with discrete values, the probability $p(x_j/y_i)$ was estimated by the proportion of the training instances with both $X_j = x_j$ and the class variable $Y = y_i$ over the number of all instances with the class variable $Y = y_i$ in the training dataset. Continuous or numeric attributes X_j are usually categorized with discretization techniques to enhance model performance, which was also conducted in this research, as discussed in Section 2. The probability inference method for discrete variables is also applicable for the categorized continuous and numeric variables, such as *DAGE*, *DTINC* and *NVEH* in this study.

Similar to the DT learning procedure, the NB model learning procedure with cross-validated performance measurement is also very efficient since the frequency of each class could be updated in constant time (Hall and Frank, 2008). Therefore, this paper also applies cross-validated model performance measurement into the NB component learning of the hybrid model.

3.3. DTNB hybrid model

As a hybrid classification model, a DTNB is an incorporation of a DT and an NB classifier (Hall and Frank, 2008). The learning algorithm for a DTNB is similar to learning stand-alone DTs. At each point of attribute search, the learning algorithm assesses the merit of splitting the entire attribute set into two disjointed attribute subsets, with one modeled with the DT model and the other by a NB classifier. As discussed in Section 3.1, the standard method to choose an optimal attribute set for a DT is to maximize cross-validated performance. In a typical cross-validation procedure, the entire dataset is divided into two segments: one for model learning and the other for model validation. The training and validation sets must cross-over successively so that each data in the entire dataset is validated (Refaeilzadeh et al., 2009). A commonly used cross-validation method is leave-one-out cross-validation (LOO-CV) (Witten et al., 2011), which is also applied in this study. LOO-CV is a special type of n-fold cross validation in which n is equal to the number of instances in the dataset. In each step of the cross-validation, a single instance in the dataset would be put aside and the rest of the dataset would be used for the training procedure. The trained classifier is tested by its prediction on the left instance, with 1 for success and 0 for failure. This procedure repeats n times and ends when each instance in the dataset is used at least once for validation (Kohavi, 1995). Numerous evaluation measurements are generally used for cross-validation, including root mean-squared error (RMSE) for numeric classes, accuracy for discrete classes, and the area under a receiver operating characteristic (ROC) curve (AUC). Starting with all attributes modeled by the DT, a greedy search algorithm with forward selection approach was used for the attribute splitting procedure in this study, where the selected attributes were modeled with the NB classifier and the remaining ones were modeled by the DT model in each step. LOO-CV accuracy was applied as an evaluation measurement to assess the quality of attribute split based on probability estimation produced by the hybrid model.

The classification results and probability estimations of response classes from the DT and NB classifier are combined to generate overall modeling results (Hall and Frank, 2008). Let X^{DT} be the attribute set in the decision table and X^{NB} be the one in the NB model, where X^{DT} and X^{NB} are complementary with each other. The overall class probability is calculated as follows,

$$P(y|X) = \alpha \times P_{DT}(y|X^{DT}) \times \frac{P_{NB}(y|X^{NB})}{P(y)} \quad (8)$$

where $P_{DT}(y|X^{DT})$ and $P_{NB}(y|X^{NB})$ are the class probabilities estimated by the decision table and Naïve Bayes model respectively, α is a normalization constant, and $P(y)$ is the prior probability of the class. The Laplace-corrected observed counts are used in the estimation of all probabilities.

4. Result analysis and discussion

4.1. Statistical data analysis

In 2010 and 2011 in New Mexico, there were 23,343 vehicles involved in 11,383 rear-end crashes, in which 14,619 units were with no injury (property damage only), 8478 with driver injuries, and 66 with driver deaths. The driver injury severity distributions across other variables are illustrated in Table 1. It is illustrated that there are more vehicles involved in rear-end crashes on weekdays than on weekends. This is probably due to the higher traffic volume on weekdays, but the proportions of driver injuries for each severity on each day were not significantly different. Environment lighting condition is a key factor for rear-end crash injury severity, especially for serious casualties. Although only 2641 vehicles (11.27%) get involved in rear-accidents in dark conditions, more

Table 2
DTNB classification overall performance.

	Training		Test	
Correctly classified instances	Number	Percentage	Number	Percentage
	8506	74.06%	7494	62.73%
Incorrectly classified instances	Number	Percentage	Number	Percentage
	2980	25.94%	4453	37.27%
Total number of instances	11,486		11,947	

Table 3
DTNB estimation accuracies classified by driver injury severities in rear-end crashes.

Driver injury severity	TP rate		FP rate		Precision		F-measure		ROC area: AUC	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
NO INJURY	0.825	0.788	0.325	0.621	0.807	0.68	0.816	0.73	0.804	0.631
INJURY	0.6	0.36	0.13	0.206	0.735	0.508	0.66	0.421	0.798	0.621
FATALITY	0.879	0.121	0.055	0.011	0.044	0.03	0.083	0.048	0.975	0.736
Weighted average	0.741	0.627	0.251	0.465	0.778	0.614	0.755	0.613	0.802	0.627

than 40% of them result in driver injuries or death. This is the highest percentage of fatalities occurring across all lighting conditions. This can be seen as 33 of the 2641 cases in dark conditions had fatal driver injuries, which was equal to the sum of driver fatalities under the other two lighting conditions. This is primarily due to the fact that speed differential is more significant under nighttime free flow conditions and tends to result in slower driver perception-reaction (e.g. driving under the influence) in inferior lighting conditions or in some enclosed structures, such as tunnels or other dark environments. The involvement of motorcycles, heavy vehicles, or hazardous material increases the probability of drivers being killed in rear-end crashes, indicated by the percentage rates for fatalities (2.31%, 5.80%, and 15.38%, respectively). Therefore, special attention should be paid to motorcycles, trucks, and lorries with hazardous materials. It is also revealed in *VTYPE* that motorcyclists have the highest injury (68.34%) and fatality rates (2.01%) compared to other vehicle drivers, while the injury (35.87%) and fatality rates (0.61%) of heavy vehicle drivers are not significantly higher than that of other vehicle types. This is likely because heavy vehicles, most of which are trucks in this research, have significant sizes and weights increasing drivers' capability to withstand the impact in traffic crashes, according to Levine et al. (1999). However, the involvement of trucks increases the probability of other drivers being severely injured or killed in rear-end crashes. Motorcycle drivers are more exposed to open traffic environments compared to other vehicle drivers and are, therefore, more vulnerable in traffic crashes, which is verified by Kockelman and Kweon (2002) and Chiang et al. (2014). As shown in Table 1, 53.67% of alcohol-involved drivers were injured and 1.93% were killed in rear-end crashes. These rates are both higher than those for non-alcohol-involved drivers (37.15% for injures and 0.26% for deaths), verifying the necessity of drunk driving prohibitions. The protective effect of seatbelts (*DBELT*) in reducing fatalities in rear-end crashes is shown in illustrated Table 1. The Table shows that 1.04% of drivers who did not wear seatbelts sustained fatal injuries, which is more than three times higher than that for drivers wearing seatbelts (0.24%). The most serious vehicle damage (*MAXDAM*) in an accident was positively associated with driver injury and death rates, as revealed by the fact that the driver injury distributions were 27.33% for no/slight damage (NSLT), 31.39% for functional damage (FUNC), and 51.28% for disabled vehicle damage (DSABL). A similar pattern occurred for death cases as well. A reasonable explanation is that the most serious vehicle damage in a crash could be considered an indicator of the impact from vehicle collisions in a rear-end crash, and the impact is transferrable from vehicles to drivers inside the vehicles, resulting in injuries or deaths.

The dataset reveals some deterministic features of these explanatory variables regarding driver injury patterns in rear-end crashes. It was also implemented in a DNTB classifier for attribute selection and decision rule extraction in order to predict driver injury severities in rear-end crashes.

4.2. DTNB model performance discussion

The dataset was modeled with a DTNB classifier embedded in WEKA (Waikato Environment for Knowledge Analysis) software (Bouckaert et al., 2013) developed by the University of Waikato, New Zealand for contributing attribute selection and decision rule extraction. The whole dataset was divided into two approximately equal-sized datasets according to the year of crash occurrences. The 2010 crash dataset was employed for DTNB attribute selection and decision rule extraction, and the 2011 crash dataset was used for model validation and performance evaluation. All the variables were nominalized before modeling. The overall estimation results for the training and test datasets are shown in Table 2. Tables 3 and 4 summarize the estimation results classified by each driver injury severity and the corresponding confusion table for the test dataset.

As is shown in Table 2, the overall classification accuracies for the training and test datasets are 74.06% and 62.73%, respectively. Compared to the model accuracies ranging from 60% to 67% in previous traffic crash severity studies with different machine-learning techniques (Abdelwahab and Abdel-Aty, 2001; Chen et al., 2015a; de Oña et al., 2011), the results obtained by the DTNB classifier are preferable or equally acceptable. The relatively large variance (11.33%) between the classification accuracies for training and testing datasets indicates that the learned classifier was more specific

Table 4
Classification confusion matrix for the test dataset.

		Predicted Instances Classified by Severity		
		NO INJURY (8670)	INJURY (3144)	FATALITY (133)
Observed instances classified by severity	NO INJURY (7477)	5894	1531	52
	INJURY (4437)	2764	1596	77
	FATALITY (33)	12	17	4

to the training dataset, and a more comprehensive training dataset including sufficient records for each injury severity is desirable to produce a more compatible classifier.

Several alternative performance measurements are listed in Table 3 quantifying model estimation for each injury severity. True positive (TP), true negative (TN), false positive (FP) and false negative (FN) are four fundamental statistical measurements in classification problems. As a multiclass classification problem in this study, TP measures the number of actual positives that were correctly identified as such (e.g. the number of driver fatalities that were correctly identified as FATALITY). TN measures the number of negatives that were correctly identified as such (e.g. the number of non-fatalities which were correctly identified as NO INJURY or INJURY). On the other hand, FP defines the number of the estimated instances incorrectly classified as positive when they were actually negative (e.g. the number of non-fatalities, including no injuries and injuries that were incorrectly identified as FATALITY). FN defines the number of the estimated instances incorrectly classified as negative when they are actually positive (e.g. the number of fatalities that were incorrectly identified as NO INJURY or INJURY). TP rate and FP rate are two statistics derived from these measurements. TP rate is used to measure the proportion of instances correctly predicted as positive in all actual positive instances and expressed as:

$$TP\ rate = \frac{TP}{TP + FN} \tag{9}$$

Similarly, FP rate is used to measure the proportion of instances incorrectly classified as positive in all actual negative instances as follows:

$$FP\ rate = \frac{FP}{FP + TN} \tag{10}$$

The TP rates range from 0.121 for FATALTY to 0.788 for NO INJURY with a weighted average of 0.627 for the testing dataset, as illustrated in Table 3. These results demonstrate that the DTNB classifier is able to classify 78.8% of instances with no injuries correctly, while its capability of classifying injury and fatal instances is relatively inferior. This implies that this classifier performs better on no injuries and injuries than fatal cases since the majority of the training dataset are no injury and injury records, with which more representative decision rules could be extracted for injury severity prediction. However, a high TP rate is not necessarily associated with a low FP rate, as is demonstrated in Table 3. As suggested by Eqs. (9) and (10), the TP rate only indicates the portion of correct classified instances in the entire positive domain without considering the correctly classified instances in the negative domain (TN). Similarly, the FP rate only shows the portion of falsely classified records in the entire actual negative domain and ignores the falsely classified records in the actual positive domain (FN). So TP and FP rates are not ideal measurements to some extent, and other performance measurements need to be developed, as discussed below.

Precision and Recall are two measurements defined based TP, FP, TN and FN as follows (Davis and Goadrich, 2006):

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = TP\ rate = \frac{TP}{TP + FN} \tag{12}$$

Precision indicates the proportion of correctly classified instances in all instances that are predicted as positive, and Recall has the same definition as TP rate in explaining the model performance. They are proposed to define F-measure, a weighted average of Precision and Recall (van Rijsbergen, 1979), as follows:

$$F = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{13}$$

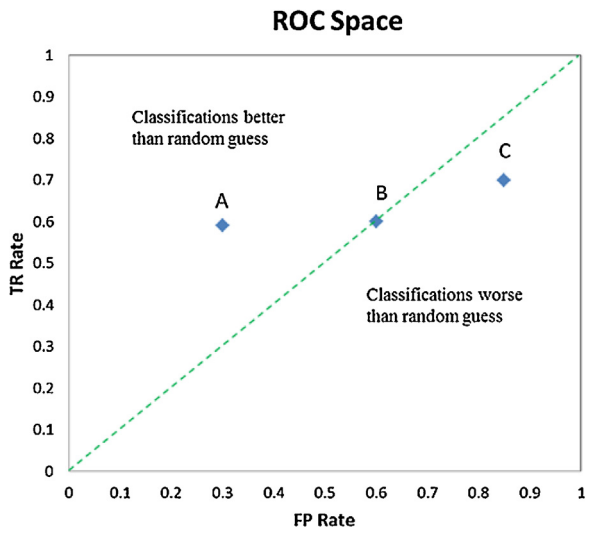

Fig. 1. An example of ROC space.

F-measure is a harmonic mean of precision and recall and averages the proportion of instances correctly classified as positive in actual positive instances and that in predicted positive instances. It ranges from 0 to 1 with 0 representing an extremely poor result and 1 indicating a perfect test. As shown in Table 3, the DTNB hybrid model has the best performance in predicting no injury instances in the testing dataset, and its F-measure is equal to 0.73. For instances with fatalities, the trained classifier performs inferiorly due to the limited sample size, with its F-measure equal to 0.048. Overall, the average F-Measure is 0.613 for the entire test dataset, implying an acceptable model performance of the trained classifier.

The receiver operating characteristic (ROC) curve is another important measurement to evaluate the overall model performance of the DTNB classifier. The ROC curve demonstrates the tradeoffs between TP rate and FP rate in a two-dimensional ROC space, where its vertical axis represents the TP rate and the horizontal axis represents the FP rate, as shown in Fig. 1. In Fig. 1, points representing model classification better than random guess are located above the diagonal line, such as Point A; points indicating model classification worse than random guess lie below the diagonal line, such as Point C; and points on the diagonal line suggest a totally random guess, such as Point B. ROC curves are normally employed to measure the performance of binary classifications. For multi-class classification problems, ROC curve analysis could be modified by producing a ROC curve for each class correspondingly rather than extending the two-dimensional space into a polytope with $(n^2 - n)$ surfaces (Fawcett, 2003). In this study, the ROC curves for each of the three severity levels are depicted in Figs. 2–4 for the testing dataset. It is shown that all the ROC curves are located

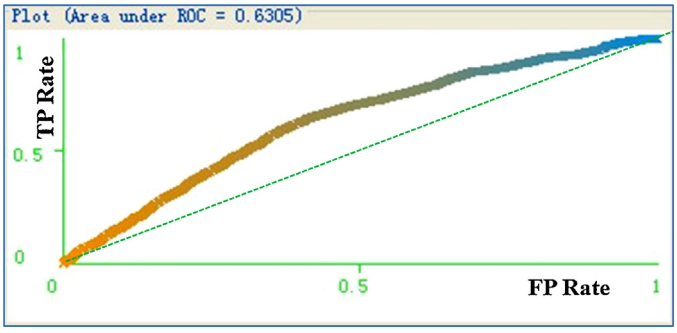

Fig. 2. ROC curve for the category of NO INJURY.

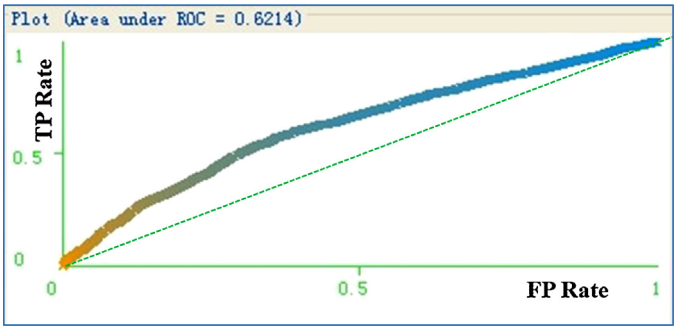

Fig. 3. ROC curve for the category of INJURY.

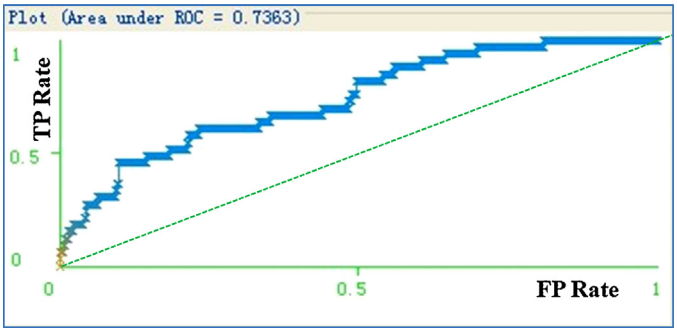

Fig. 4. ROC curve for the category of FATALITY.

above the diagonal line, indicating that the DTNB classifier performed acceptably well on three injury severity predictions for the testing dataset. The Area Under an ROC Curve (AUC) is defined to quantitatively evaluate the overall performance of a classifier. The AUC is the area enclosed by the ROC curve, the horizontal axis and the right boundary of the ROC space, with a maximum value of 1 indicating a perfect classification result. A value of 0.5 indicates that an absolutely random classification result is produced by the classifier. In this research, a corresponding AUC based on the testing dataset was calculated for each driver injury severity level (NO INJURY, INJURY, and FATALITY) and is shown in Figs. 2–4 and Table 3. It is revealed that the DTNB model achieves the best performance for fatal records, with an AUC of 0.736. This is followed by that for no injury instances and injury instances, with AUCs of 0.631 and 0.621, respectively. The overall AUC indicated that model performance of the entire dataset should be calculated as a weighted sum of the AUCs for all injury outcome levels as follows (Provost and Domingos, 2001),

$$AUC_{overall} = \sum_{i=1}^{n} AUC_{c_i} p(c_i) \quad (14)$$

where AUC_{c_i} is the AUC for injury severity outcome c_i, $p(c_i)$ is the probability of occurrence for injury severity outcome c_i in the dataset, and n is the number of injury severities, which is equal to 3 in this research. In this study, the overall AUC for the test dataset is 0.627, suggesting that the learned DTNB classifier is able to effectively extract the injury severity patterns and produce an acceptable performance based on the criteria proposed by Tape (2001). This result is also supported by our previous study (Chen et al., 2015a), where $AUC_{overall} = 0.658$ is obtained in the analysis of crash injury severity on Spanish rural highways.

Table 4 is the confusion matrix signifying the discrepancy between the predicted results and actual observations in the test dataset, where each row in this matrix represents the actual number of instances for each injury severity, and each column indicates the predicted number of instances for each injury severity category. There are considerable misclassifications distributed in the off-diagonal cells, resulting in overestimation or underestimation for each severity level. As is illustrated in Table 4, 1531 instances of no injuries are misclassified as injury cases, 52 no injuries are misclassified as fatal instances, 2764 injuries are misclassified as no injuries, 77 injuries are misclassified as fatalities, 12 fatalities are misclassified as no injury cases, and 17 fatalities are misclassified as injury cases. The overall match rate (accuracy) is 62.73%, illustrating an acceptable model performance was produced by the DTNB hybrid model.

4.3. Model performance comparison with other methods

In this study, the performance of this DTNB classifier is compared with other popular methods. First, the two individual classifiers that consists of the DTNB classifier, the DT model and the NB classifier, are used as the control models and applied on the same training and testing datasets. The results produced from the DT model are shown in Tables 5 and 6, and the results from the NB classifier are shown in Tables 7 and 8.

As are shown in Tables 2, 5 and 7, the DTNB hybrid classifier performs best on the training dataset, with classification accuracy equal to 74.06%, but its prediction accuracy on the test dataset is the lowest among these models, equal to 62.73%. Meanwhile, the difference of prediction accuracy for training and testing datasets from the DTNB classifier is the largest among all the three models, indicating that the machine-learning method is more specific

Table 5
DT classification overall performance.

	Training		Test	
Correctly classified instances	Number	Percentage	Number	Percentage
	7861	68.44%	7670	64.20%
Incorrectly classified instances	Number	Percentage	Number	Percentage
	3625	31.56%	4277	35.80%
Total number of instances	11,486		11,947	

Table 6
DT classification confusion matrix for the test dataset.

		Predicted instances classified by severity		
		NO INJURY (9361)	INJURY (2584)	FATALITY (2)
Observed instances classified by severity	NO INJURY (7477)	6296	1181	0
	INJURY (4437)	3061	1374	2
	FATALITY (33)	4	29	0

Table 7
NB classification overall performance.

	Training		Test	
Correctly classified instances	Number	Percentage	Number	Percentage
	7515	65.43%	7721	64.63%
Incorrectly classified instances	Number	Percentage	Number	Percentage
	3971	34.57%	4226	35.37%
Total number of instances	11,486		11,947	

Table 8
NB classification confusion matrix for the test dataset.

		Predicted instances classified by severity		
		NO INJURY (9027)	INJURY (2677)	FATALITY (243)
Observed instances classified by severity	NO INJURY (7477)	6207	1163	107
	INJURY (4437)	2817	1499	121
	FATALITY (33)	3	15	15

Table 9
MNL-BN estimation results.

	Training		Test	
Correctly classified instances	Number	Percentage	Number	Percentage
	7677	66.84%	7856	65.76%
Incorrectly classified instances	Number	Percentage	Number	Percentage
	3809	33.16%	4091	34.24%
Total number of instances	11,486		11,947	

to learning scheme and training dataset. A detailed examination of Tables 4, 6 and 8 reveals that the DT model performs best on records of no injury and is able to correctly classify 6296 no injury records, but performs worst on injury and fatality levels with the least number of correctly classified records, which are 1374 for injury records and 0 for fatality records, respectively. The NB classifier has the highest prediction accuracy on fatality records and is able to correctly classify 15 of the 33 fatality instances. The DTNB classifier produced the worst performance on driver no injury records among the three models, and works inferior on fatality record prediction than NB classifier, but it is able to predict 1596 injury records correctly, which significantly improves the performance of the DT and NB classifiers on injury records.

This DTNB classifier is considered as a knowledge-based non-parametric machine-learning model for qualitative crash data analysis, and its performance based on classification accuracy is also compared with other regression models and non-regression models, and the traditional multinomial logit (MNL) model and a multinomial logit-Bayesian network (MNL-BN) model are used as the representatives of these two model types, respectively. The model results are shown below in Tables 9 and 10 (Chen et al., 2015a).

As shown in Tables 2 and 9, in terms of prediction accuracy, the DTNB classifier outperforms the proposed MNL-BN model on the training dataset, but performs inferior on the testing dataset. Besides, the variance of estimation accuracies for the proposed MNL-BN model on training and testing datasets is around 1%, and that for the DTNB is around 12%, indicating that the DTNB method is more specific to learning scheme and training dataset and applicable to exploratory analysis, but the proposed MNL-BN model produces more reliable and less biased results for independent datasets once the model structure is trained. For the modeling results from the MNL model, as shown in Table 10, the prediction accuracy is 6664 (6664 = 4881 + 1782 + 1), and the overall correction rate of this MNL prediction is 55.78%, which is significantly lower than the accuracy from DTNB approach (62.73%), indicating that the proposed DTNB approach outperforms the traditional MNL model.

The results provided by this DTNB classifier is a trained hybrid classifier consisting of extracted decision rules, which could be interpreted by the "if-then" rule similar to decision tree models (shown in Section 4.4) and is easy to understand. Comparing to other non-parametric methods, such as support vector machines, near neighbor classifiers, and ensemble classifiers, although decision tree models may not be able produce a competitive prediction accuracy, they faster in prediction speed and fitting speed, and requires less memory storage, and therefore is a promising model for classification tasks (MathWorks.Inc, 2015), and these features were also found during the modeling procedure for the DTNB classifier in this study.

4.4. DTNB result discussions

The primary objective of this research is to select the significant attributes affecting driver injury severities in rear-end crashes and explore the decision rules for each severity level based on these selected attributes. In this study, 15 attributes are selected as the decisive feature set by the hybrid classifier as follows: *DAY*, *RDREL*, *LIGHT*, *WEATHER*, *RDGRD*, *NVEH*, *RDFUNC*, *MCINV*, *HEVINV*, *DTINC*, *RDPV*, *NLANE*, *DBELT*, *DALC*, and *MAXDAM*. Note that the attribute set

Table 10
MNL classification confusion matrix for the testing dataset.

		Predicted Instances Classified by Severity		
		NO INJURY (7533)	INJURY (4377)	FATALITY (37)
Observed instances classified by severity	NO INJURY (7477)	4881	2580	15
	INJURY (4437)	2635	1782	20
	FATALITY (33)	16	16	1

Table 11
Decision rules for fatal injury classifications learned by the DTNB hybrid classifier.

DAY	RDREL	LIGHT	WEATHER	RDGRD	NVEH	RDFUNC	MCINV	HEVINV	DTINC	RDPV	NLANE	DBELT	DALC	MAXDAM	SEV
SUN	ONWAY	DARK	CLEAR	ONGRADE	TWO	RINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
TUE	ONWAY	DARK	SNOW	LEVEL	THREE	RNINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
WED	ONWAY	DARK	CLEAR	LEVEL	TWO	RINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
SAT	ONWAY	DARK	CLEAR	ONGRADE	TWO	RINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
MON	ONWAY	DAYLIGHT	CLEAR	LEVEL	TWO	RNINT	N	N	NEAR	PAVED	TWO	N	N	DISABLE	FATALITY
FRI	ONWAY	DAYLIGHT	CLEAR	LEVEL	TWO	RINT	N	Y	NEAR	PAVED	TWO	N	N	DISABLE	FATALITY
TUE	ONWAY	DAYLIGHT	SNOW	LEVEL	TWO	URBAN	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
TUE	ONWAY	DAYLIGHT	CLEAR	ONGRADE	THREE	RNINT	N	N	NEAR	PAVED	ONE	Y	N	DISABLE	FATALITY
FRI	ONWAY	DARK	RAIN	LEVEL	TWO	RNINT	N	N	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
SAT	ONWAY	DAYLIGHT	CLEAR	LEVEL	MORE	URBAN	N	Y	FAR	PAVED	ONE	Y	Y	DISABLE	FATALITY
SAT	ONWAY	DAYLIGHT	CLEAR	LEVEL	MORE	URBAN	N	Y	FAR	PAVED	TWO	Y	N	DISABLE	FATALITY

is selected for formulating decision rules for all three injury severities based on the entire dataset, not only for a particular injury outcome. 2865 decision rules are trained by the DTNB classifier based on the selected attributes, in which 1366 rules are used for predicting no injury cases, 1488 for injury prediction, and 11 for fatality prediction. As shown in Table 4, there are 8670 instances in the test dataset predicted as no injuries, 3144 as injuries, and 133 as fatalities. On average, a decision rule for no injury prediction is used to classify 6.3 instances in the testing dataset, a decision rule for injury prediction is used to classify 2.1 instances, and a decision rule for fatality prediction is used to classify 12.1 instances. However, if only the correctly classified instances are considered, the average numbers of correct predictions are 4.3 instances for a no injury decision rule, 1.1 instances for an injury decision rule, and 0.4 instances for a fatality decision rule. Based on these results, the learned decision rules for no injury are the most efficient in severity outcome prediction, followed by those for injury. The learned decision rules for fatality are the least efficient in correct classification, which explains the lowest TP rate and F-measure of FATALITY for the testing dataset in Table 3.

A trained DT lists all the decision rules for predicting the most probable driver injury severity in rear-end crashes under a set of specific conditions of these selected variables. Fatality is always the most significant concern in traffic safety analyses, so this discussion focuses on decision rules generated for driver fatal injury prediction. Table 11 summarizes all learned decision rules for predicting driver fatality in rear-end crashes. The learned DT is fundamentally a matrix of if-then rules working with condition states and action states: if a specific set of conditions for the selected attributes is satisfied, a particular injury severity level that a driver is most likely to suffer in a rear-end crash would be returned. For example, the first decision rule in Table 11 could be written with if-then rule as follows:

If *DAY* = SUN, and *RDREL* = ONWAY, and *LIGHT* = DARK, and *WEATHER* = CLEAR, and *RDGRD* = ONGRADE, and *NVEH* = TWO, and *RDFUNC* = RINT, and *MCINV* = N, and *HEVINV* = Y, and *DTINC* = NEAR, and *RDPV* = PAVED, and *NLANE* = TWO, and *DBELT* = Y, and *DALC* = N, and *MAXDAM* = DSABL,

Then *SEV* = FATALITY.

The significant effects of some condition-states on driver fatal injury are detected in Table 11. *RDREL* has a unanimous condition state for all the 11 decision rules, *RDREL* = ONWAY, indicating that it is highly likely to result in fatalities if the first harmful event of a serial rear-end crash happens on the roadway. This is probably because in serial rear-end crashes with multiple vehicles, the first event on a roadway segment would block traffic and result in consecutive collisions due to limited response time for following vehicle drivers. There are 5 out of the 11 decision rules with the presence of dark lighting conditions (*LIGHT* = DARK), indicating that lower light conditions are an important factor in predicting driver fatal injuries in rear-end crashes using these model assumptions. The other 6 decision rules for driver fatality prediction are associated with daylight conditions, which seem contradictory to the generally accepted understanding, but recent research by Venkataraman et al. (2013) found that different lighting conditions had both positive and negative effects based on site conditions. Similarly seemingly contradictory findings are also concluded for *WEATHER*, *RDGRD*, and *RDPV*, where clear weather, level road grade, and paved road surface are the most frequent conditions in predicting driver fatal injuries. These contradictions are explainable because drivers tend to be more aware when driving in adverse conditions, such as extreme weather, lower environment lighting conditions, mountainous terrain, wet or icy pavement surfaces (associated with extreme weather), granular pavement, etc., while crash risk and severity might induce potential speeding or careless driving in comfortable driving environments. This finding receives support from multiple studies (Ding et al., 2015; Haque et al., 2012; Milton et al., 2008; Savolainen and Mannering, 2007; Shaheed et al., 2013; Yu and Abdel-Aty, 2014). For instance, Yu and Abdel-Aty (2014) discovered that snowy weather conditions tend to reduce the likelihood of serious crashes. Savolainen and Mannering (2007) found that crashes occurring on wet road surfaces also tend to be less severe.

The number of vehicles (*NVEH*) involved in a crash is significant in predicting driver fatal injuries in rear-end crashes, and two-vehicle rear-end crashes are the most common type resulting in fatalities, indicated in Table 11. Other studies revealed that the number of vehicles in a crash is closely related to severe crash injuries or fatalities (NHTSA, 2013; Wu et al., 2015). For instance, NHTSA (2013) reported that 1,661,000 single-vehicle crashes and 3,677,000 multi-vehicle crashes occurred in the U.S. in 2011, of which 17,991 and 11,766 were fatal crashes, respectively. Further analyses also found that the number of vehicles in a crash has significant influence on the mechanisms of inducing crash occurrences and casualties. Geedipally and Lord (2010) revealed that splitting modeling of single-vehicle and multi-vehicle crashes produces more reliable estimations in predicting crash injury severities compared to combined modeling. Venkataraman et al. (2013) discovered that the significant attribute sets affecting crash potentials vary for distinctive crash groups aggregated by the number of vehicles involved. Therefore, the intriguing effect of number of vehicles involved on driver injury outcomes should be further examined in future research with separate modeling of different crash groups by the number of vehicles involved.

Heavy vehicle involvement (*HEVINV*) is significant in predicting driver fatalities in rear-end crashes, indicated in Table 11. Heavy vehicle involvement was present in 8 of the 11 rules for fatality prediction, which is consistent with the statistical findings in Sections 2 and 4.1. The reasonableness and good transferability of this finding is verified by Yan et al. (2005), who also discovered that the rear-end crash risk is increasing with the increment of vehicle size. Besides, Harb et al. (2007) found that light truck vehicles are an important contributing factor in rear-end crashes at non-signalized intersections due to truck drivers' limited visibility.

Meng and Weng (2011) revealed that higher heavy vehicle percentage causes higher rear-end crash risk. However, in this study, heavy vehicle type (*VTYPE* = HEV) is not found to be significant in predicting driver injury severity in rear-end crashes, which is probably because heavy vehicles make up only a slight portion of all the studied vehicles and its influence is not as significant as *HEVINV*. Furthermore, Lao et al. (2014) verified that truck percentage has a parabolic influence on rear-end crash risk. The number of driving lanes (*NLANE*) is significant in predicting driver injury severities in rear-end crashes, and rear-end crash fatalities are most likely to happen on two-lane roadways, as shown in Table 11. The influence of the number of lanes on crash severity has been assessed by a previous study. Jung et al. (2014) discovered that an increase in the number of lanes tends to increase the likelihood of incapable injury and fatalities in crashes occurring in rainy weather, and this study examined its interactive effects across other crash-related factors.

Road function (*RDFUNC*) is a significant factor contributing to driver fatal injury in rear-end crashes. As is shown in Table 11, fatal rear-end crashes are more likely to happen on rural roadways, including rural interstate (RINT) and rural non-interstate (RNINT) roadways. This finding is supported by the fact that 55% of the overall fatalities in traffic accidents occur on rural roads (NHTSA, 2013). This is explainable because traffic in rural areas normally travels at high speeds, which may result in significant deformation of vehicles and, therefore, severe injuries on drivers in rear-end crashes. Also, the design criteria and access control are often less than other higher class facilities, resulting in increased roadway crash potential. The safety performance of rural roads is generally discussed jointly with lane numbers. In New Mexico, 65% of crash-related fatalities occurred on rural highways. More than 80% of rural highways are two-lane highways (NMDOT, 2012), which explains the highest frequency of two-lane condition in Table 11 among all categories of lane numbers. Significant research has been done to address rural crash severities, including rural two-lane highways (Chen and Chen, 2011; de Oña et al., 2013; Farah et al., 2009; Karlaftis and Golias, 2002; Khorashadi et al., 2005; Lord et al., 2005; Siskind et al., 2011; Wu et al., 2014). For example, Farah et al. (2009) investigated drivers' overtaking strategies on rural-two-lane highways through driving simulations. Siskind et al. (2011) discovered that speeding, alcohol involvement, and traffic rule violations are major factors of fatal crashes on rural roadways. Table 11 also indicates that the condition state of rural roadways (RNINT and RINT) is closely associated with heavy vehicle involvement (*HEVINV*). This could be because a considerable portion of traffic on rural roadways is heavy vehicles traveling at high speeds due to light traffic, which increases the potential of severe injuries and fatalities in rear-end crashes.

Crash location (*DTINC*) is also intimately associated with crash injury severities. Table 11 shows that most of the fatal crashes occur within 0.1 mile of the nearest intersection (*DTINC* = NEAR). A reasonable explanation is that vehicles decelerate intensively from a high velocity and the headway between vehicles varies dramatically when approaching intersections, leading to insufficient response time and severe rear-end collisions. Therefore, fatal rear-end crashes are most likely to happen when vehicles are approaching intersections with high speeds and inadequate acceleration, insufficient deceleration, short driver perception and reaction time, etc. may dramatically contribute to severe crash occurrences. Significant studies have been conducted to examine the characteristics of intersection-related crashes, including rear-end crashes. Kim et al. (2007) modeled crash risks for different severities at rural intersections via binomial hierarchical multilevel models. Xie et al. (2013) investigated the safety performance of signalized intersections taking corridor-level correlations into account. Huang et al. (2008) studied the driver injury and vehicle damage patterns in traffic crashes in urban intersections. Wang et al. (2015) examined the temporal variation of safety performance of intersection treatment. Therefore, special attention should be paid at intersections, especially for rural intersections where vehicles approach at higher speeds.

Driver alcohol involvement (*DALC*) is also selected as a necessary factor to formulate driver injury severity prediction rules, as listed in Table 11, though driver alcohol involvement is only present in 1 of the 11 rules. This is likely because alcohol has influencing effects in impairing drivers' visibility and judgments, and the limited presence of driver alcohol involvement (*DALC* = Y) is due to the insufficient amount of fatality records. Consistent conclusions are also summarized by Yan et al. (2005), Li and Bai (2008), and Weiss et al. (2014). For instance, Yan et al. (2005) found that alcohol-driving drivers are 9.58 times more likely to get involved in rear-end crashes than non-drinking drivers even they are under legal alcohol use. Li and Bai (2008) also concluded that driver alcohol involvement induced a large portion of fatal crashes. The most serious vehicle damage in a crash (*MAXDAM*) is found to be significant in predicting driver injury severities, and vehicle disabled damage (*MAXDAM* = DSABL) appears unanimously in all decision rules for driver fatality prediction. This indicates the significant association between vehicle disabled damage and driver fatality. A rational interpretation is that vehicle damage is a reflection of the impact generated in a rear-end crash, which is transferrable from vehicle bodies to drivers, and severe vehicle damage is generally associated with high casualties. Comparing *DALC* with *MAXDAM*, it is discovered that disabled vehicle damage is shown in most of the fatality decision rules while driver alcohol or drug involvement is rarely present, which indicates that the variable *MAXDAM* has a higher weight in resulting in driver fatal injuries in rear-end crashes. However, the most serious vehicle damage is an aftermath of rear-end crashes while drivers' alcohol or drug involvement occurs before a crash happens. Therefore, the importance of drunken driving prohibition should not be understated and corresponding law enforcement should be enhanced. Similar to *DALC*, other variables, such as motorcycle involvement (*MCINV*), crash day (*DAY*), and seatbelt usage (*DBELT*), also illustrate unique patterns in predicting driver injury outcomes. Overall, the selected features and their conditions-states are consistent with the statistical analysis findings in Section 2, demonstrating the reasonableness of the results produced by the hybrid classifier.

5. Research limitation

This research produced promising results through the DTNB hybrid classifier, but also has some limitations. This study is based on a two-year rear-end crash dataset in which the crash records under adverse weather or geometric conditions were limited, as shown in Table 1. The unbalanced dataset with respect to each injury outcome may yield biased estimations. For example, there are 66 fatal instances in the entire dataset, which only account for 0.28% of the total records and may not be able to comprehensively represent the actual conditions of fatal rear-end crashes in New Mexico. More data across multiple years are desired for less biased modeling results. Additionally, some significant attributes related to driver injury severity, such as vehicle speed (Cheng et al., 2013; Liu et al., 2015; Ma et al., 2015) and traffic volume (Wang and Cheu, 2013; Zou et al., 2014, 2013), are not included in the original dataset, although the maximum vehicle damage could be a reflection of vehicle speeds to some extent. Further investigations with accident scene retrieval techniques are recommended for more comprehensive data collection. Third, as the knowledge-based non-parametric machine-learning model, this DTNB classifier is only able to produce qualitative results and lacks of quantitative measurements to evaluate variable influence on

driver injury severities. More research is needed to improve this DTNB classifier by developing qualitative analysis procedure, such as sensitivity analysis, in the future. Fourth, as discussed in Section 4.2, the attribute set for decision rule learning were selected for all three injury severities based on the entire dataset, where an attribute that is critical in predicting a specific injury level may not be significant in predicting the others. Therefore, discriminative analysis should be conducted and a unique feature set for each injury outcome should be examined in future research. Fifth, the DTNB classifier reports 2865 decision rules in total for three severity levels, which is a chaotic presentation even with a succinct and understandable tabular format. Hence, additional effort should be made to elaborate these rules in a clustered and ordered way.

6. Conclusion

Rear-end crashes are a major type of traffic accidents in the U.S., and it is necessary to examine the mechanism of driver injury severity in these crashes. DT and NB classifiers are two distinctive and effective methods for instance classification based on a certain set of attributes, which have been widely used in multiple research areas except for traffic safety analyses. Based on a two-year rear-end crash dataset in New Mexico, this paper applies a new DTNB hybrid classifier as a knowledge-based Bayesian non-parametric machine-learning model to select the attributable feature set regarding driver behavior, demographic features, vehicle factors, geometric and environmental characteristics, etc. for driver injury severities in rear-end crashes and extract the decision rules for driver injury severity prediction. The DTNB hybrid classifier produces a reasonable classification result, indicated by several performance measurements, such as F-measure, ROC curve, and AUC, as well as by the model performance comparison results with several popular models.

The DTNB hybrid classifier outputs the selected feature set for driver injury severity prediction, accompanied by a decision table with learned decision rules based on the applied dataset. 15 attributes were selected as significant in predicting driver injury fatalities, including crash day (*DAY*), first harmful event location (*RDREL*), lighting condition (*LIGHT*), weather condition (*WEATHER*), road grade (*RDGRD*), number of vehicles involved (*NVEH*), road function (*RDFUNC*), motorcycle involvement (*MCINV*), heavy vehicle involvement (*HEVINV*), distance from crash location to the nearest intersection (*DTINC*), road pavement condition (*RDPV*), number of driving lanes (*NLANE*), seatbelt use (*DBELT*), driver alcohol involvement (*DALC*), and maximum vehicle damage (*MAXDAM*). Decision rules for fatality prediction reveal that the involvement of heavy vehicles in rear-end crashes increases the probability of driver fatalities, and motorcycle involvement is also significant in predicting driver injury and fatalities. Driver fatalities are more likely to occur in a comfortable traffic environment, such as clear weather, level road grade, and paved road surface, whereas drivers would be more aware of potential risk under adverse driving conditions. Driver fatal injuries are most likely to happen on rural roads, especially on rural two-lane highways. Maximum vehicle damage in rear-end crashes is positively associated with driver injury severities, and drivers are most likely to suffer fatal injuries when vehicles involved in rear-end crashes are disabled. The number of vehicles in a rear-end crash significantly affects driver injury outcomes, and two-vehicle rear-end crash is the most frequent type resulting in driver fatalities. Fatal rear-end crashes are more likely to happen near intersections, where vehicles accelerate and decelerate dramatically, resulting in limited time for proper responses. The effectiveness of seatbelt use and drunk driving prohibition in reducing driver injury severities are verified in the extracted decision rules.

This research also has some limitations with respect to data size, data structure and attributes, and result presentation. Further research is desirable to enhance the DTNB model performance and better investigate decision rules for rear-end crash injury prediction by examining a more comprehensive dataset. Additionally, discriminative analyses should be conducted and the distinctive attribute set for each injury severity should be examined separately in future research.

Acknowledgments

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

References

Abdalla, I.M., 2005. Effectiveness of safety belts and Hierarchical Bayesian analysis of their relative use. Saf. Sci. 43, 91–103.
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.
Boström, O., Fredriksson, R.S., Håland, Y., Jakobsson, L., Krafft, M., Lövsund, P., Muser, M.H., Svensson, M.Y., 2000. Comparison of car seats in low speed rear-end impacts using the BioRID dummy and the new neck injury criterion (NIC). Accid. Anal. Prevent. 32, 321–328.
Bouckaert, R.R., Frank, E., Hall, M., Kirkby, M., Reutemann, P., Seewald, A., Scuse, D., 2013. WEKA Manual for Version 3-7-8. Hamilton, New Zealand.
Broughton, K.L.M., Switzer, F., Scott, D., 2007. Car following decisions under three visibility conditions and two speeds tested with a driving simulator. Accid. Anal. Prevent. 39, 106–116.
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. Prevent. 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. C: Emerg. Technol. 60, 339–359.
Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accid. Anal. Prevent. 43, 1677–1688.
Cheng, W., Wang, J.-H., Bryden, G., Ye, X., Jia, X., 2013. An examination of the endogeneity of speed limits and accident counts in crash models. J. Transp. Saf. Secur. 5, 314–326.
Chiang, V.X.Y., Cheng, J.Y.X., Zhang, Z.C., Teo, L.-T., 2014. Comparison of severity and pattern of injuries between motorcycle riders and their pillions: a matched study. Injury 45, 333–337.
Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, New York, pp. 233–240.
de Oña, J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accid. Anal. Prevent. 51, 1–10.
de Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid. Anal. Prevent. 43, 402–411.
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. Prevent. 85, 58–65.
Domingos, P., Plazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. 29, 103–130.
Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th International Conference on Machine Learning, San Francisco, CA, pp. 194–202.
Duan, J., Li, Z., Salvendy, G., 2013. Risk illusions in car following: is a smaller headway always perceived as more dangerous? Saf. Sci. 53, 25–33.
El-Basyouny, K., Sayed, T., 2013. Depth-based hotspot identification and multivariate ranking using the full Bayes approach. Accid. Anal. Prevent. 50, 1082–1089.
Farah, H., Bekhor, S., Polus, A., 2009. Risk evaluation by modeling of passing behavior on two-lane rural highways. Accid. Anal. Prevent. 41, 887–894.
Farmer, C.M., Wells, J.K., Werner, J.V., 1999. Relationship of head restraint positioning to driver neck injury in rear-end crashes. Accid. Anal. Prevent. 31, 719–728.
Fawcett, T., 2003. ROC graphs: notes and practical considerations for researchers. Mach. Learn., 1–38.
Geedipally, S.R., Lord, D., 2010. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models. Accid. Anal. Prevent. 42, 1273–1282.

Hall, M., Frank, E., 2008. Combining Naive Bayes and Decision Tables. In: 21st Florida Artificial Intelligence Research Society Conference, AAAI Press, Miami, FL, pp. 318–319.
Han, S.H., Kim, T.W., Choi, Y., Yoo, K.J., 1989. Development of a computer code AFTC for fault tree construction using decision table method and super component concept. Reliab. Eng. Syst. Saf. 25, 15–31.
Haque, M.M., Chin, H.C., Debnath, A.K., 2012. An investigation on multi-vehicle motorcycle crashes using log-linear models. Saf. Sci. 50, 352–362.
Haque, M.M., Chin, H.C., Huang, H., 2010. Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections. Accid. Anal. Prevent. 42, 203–212.
Harb, R., Radwan, E., Yan, X., Abdel-Aty, M., 2007. Light truck vehicles (LTVs) contribution to rear-end collisions. Accid. Anal. Prevent. 39, 1026–1036.
Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prevent. 42, 1556–1565.
Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accid. Anal. Prevent. 40, 45–54.
Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B., 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. 51, 141–154.
Jung, S., Jang, K., Yoon, Y., Kang, S., 2014. Contributing factors to vehicle to vehicle crash frequency and severity under rainfall. J. Saf. Res. 50, 1–10.
Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prevent. 34, 357–365.
Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accid. Anal. Prevent. 37, 910–921.
Kim, D.-G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accid. Anal. Prevent. 39, 125–134.
Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prevent. 34, 313–321.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143.
Kohavi, R., Sahami, M., 1996. Error-based and entropy-based discretization of continuous features. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portlan, Oregon, pp. 114–119.
Kruschke, J.K., 2015. Doing Bayesian data analysis. In: Doing Bayesian Data Analysis. Elsevier.
Lao, Y., Zhang, G., Wang, Y., Milton, J., 2014. Generalized nonlinear models for rear-end crash risk analysis. Accid. Anal. Prevent. 62, 9–16.
Lee, C.-H., 2007. Improving classification performance using unlabeled data: Naive Bayesian case. Knowl.-Based Syst. 20, 220–224.
Levine, E.M., Bedard, M., Molloy, D.W., Basilevsky, A., 1999. Determinants of Driver Fatality Risk in Front Impact Fixed Object Collisions. Mature Medicine, Toronto, Canada.
Lew, A., 1991. Fuzzy decision tables for expert systems. Comput. Math Appl. 21, 111–116.
Li, Y., Bai, Y., 2008. Comparison of characteristics between fatal and injury accidents in the highway construction zones. Saf. Sci. 46, 646–660.
Liu, J., Khattak, A.J., Richards, S.H., Nambisan, S., 2015. What are the differences in driver injury outcomes at highway-rail grade crossings? Untangling the role of pre-crash behaviors. Accid. Anal. Prevent. 85, 157–169.
Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accid. Anal. Prevent. 37, 185–199.
Ma, X., Chen, F., Chen, S., 2015. Modeling crash rates for a mountainous highway by using refined-scale panel data. Transp. Res. Rec.: J. Transp. Res. Board, 10–16.
MathWorks Inc., 2015. Choose a Classifier [WWW Document], http://www.mathworks.com/help/stats/choose-a-classifier.html (accessed 12.01.15).
Meng, Q., Qu, X., 2012. Estimation of rear-end vehicle crash frequencies in urban road tunnels. Accid. Anal. Prevent. 48, 254–263.
Meng, Q., Weng, J., 2011. Evaluation of rear-end crash risk at work zone using work zone traffic data. Accid. Anal. Prevent. 43, 1291–1300.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prevent. 40, 260–266.
National Health Council, 2013. Estimating the Costs of Unintentional Injuries [WWW Document], http://www.nsc.org/news_resources/injury_and_death_statistics/Pages/EstimatingtheCostsofUnintentionalInjuries.aspx.
National Highway Traffic Safety Administration (NHTSA), 2013. Fatality Analysis Reporting System Encyclopedia [WWW Document].
National Safety Council, 2011. National Safety Council Injury Facts 2011 Edition. Itasca, IL.
New Mexico Department of Transportation, 2012. New Mexico traffic crash annual report 2011.
Provost, F., Domingos, P., 2001. Well-trained PETs: Improving Probability Estimation Trees, CeDER Working Paper #IS-00-04. New York.
Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. Encycl. Database Syst.
Renooij, S., van der Gaag, L.C., 2008. Evidence and scenario sensitivities in naive Bayesian classifiers. Int. J. Approx. Reason. 49, 398–416.
Riviere, C., Lauret, P., Ramsamy, J.F.M., Page, Y., 2006. A Bayesian neural network approach to estimating the energy equivalent speed. Accid. Anal. Prevent. 38, 248–259.
Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accid. Anal. Prevent. 39, 955–963.
Seagle, J.P., Duchessi, P., 1995. Acquiring expert rules with the aid of decision tables. Eur. J. Oper. Res. 84, 150–162.
Shaheed, M.S.B., Gkritza, K., Zhang, W., Hans, Z., 2013. A mixed logit analysis of two-vehicle crash severities involving a motorcycle. Accid. Anal. Prevent. 61, 119–128.
Singh, S., 2003. Driver Attributes and Rear-end Crash Involvement Propensity. Washington, DC.
Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., Hanks, H., 2011. Risk factors for fatal crashes in rural Australia. Accid. Anal. Prevent. 43, 1082–1088.
Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., Ellis, I.O., 2011. A "non-parametric" version of the naive Bayes classifier. Knowl.-Based Syst. 24, 775–784.
Strauss, J., Miranda-Moreno, L.F., Morency, P., 2013. Cyclist activity and injury risk analysis at signalized intersections: a Bayesian modelling approach. Accid. Anal. Prevent. 59C, 9–17.
Tape, T.G., 2001. Interpretation of diagnostic tests. Ann. Interv. Med. 135, 72.
van Rijsbergen, C.J., 1979. Information Retrieval, 2nd ed. Butterworths, London,UK.
Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accid. Anal. Prevent. 59, 309–318.
Wang, G.A., Atabakhsh, H., Chen, H., 2011. A hierarchical Naïve Bayes model for approximate identity matching. Decis. Support Syst. 51, 413–423.
Wang, J.-H., Abdel-Aty, M.A., Park, J., Lee, C., Kuo, P.-F., 2015. Estimating safety performance trends over time for treatments at intersections in Florida. Accid. Anal. Prevent. 80, 37–47.
Wang, Y., Cheu, R.L., 2013. Safety impacts of auxiliary lanes at isolated freeway on-ramp junctions. J. Transp. Saf. Secur. 5, 327–343.
Washington, S.P., Congdon, P., Karlaftis, M.G., Mannering, G., 2005. Bayesian multinomial logit models: exploratory assessment of transportation applications. In: TRB 2005 Annual Meeting CD-ROM, Transportation Research Board, National Research Counci, Washington, DC.
Weiss, H.B., Kaplan, S., Prato, C.G., 2014. Analysis of factors associated with injury severity in crashes involving young New Zealand drivers. Accid. Anal. Prevent. 65, 142–155.
Witlox, F., Antrop, M., Bogaert, P., De Maeyer, P., Derudder, B., Neutens, T., Van Acker, V., Van de Weghe, N., 2009. Introducing functional classification theory to land use planning by means of decision tables. Decis. Support Syst. 46, 875–881.
Witten, I.H., Frank, F., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann Publishers, San Francisco.
Wong, T.-T., Chang, L.-H., 2011. Individual attribute prior setting methods for naïve Bayesian classifiers. Pattern Recogn. 44, 1041–1047.
World Health Organization, 2013. Global Status Report on Road Safety 2013 Supporting a Decade of Action. Geneva, Switzerland.
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. Prevent. 72, 105–115.
Wu, Q., Zhang, G., Ci, Y., Wu, L., Tarefder, R.A., Alcántara, A.D., 2015. Exploratory multinomial logit model-based driver injury severity analyses for teenage and adult drivers in intersection-related crashes. Traffic Inj. Prevent.
Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accid. Anal. Prevent. 50, 25–33.
Yan, X., Radwan, E., Abdel-Aty, M., 2005. Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. Accid. Anal. Prevent. 37, 983–995.
Yanmaz-Tuzel, O., Ozbay, K., 2010. A comparative Full Bayesian before-and-after analysis and application to urban road safety countermeasures in New Jersey. Accid. Anal. Prevent. 42, 2099–2107.
Youn, E., Jeong, M.K., 2009. Class dependent feature scaling method using naive Bayes classifier for text datamining. Pattern Recogn. Lett. 30, 477–485.
Yu, R., Abdel-Aty, M., 2013. Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. Accid. Anal. Prevent. 56, 51–58.
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.
Zhang, Y., Zhao, Y., Gao, D., 2008. Decision table for classifying point sources based on FIRST and 2MASS databases. Adv. Space Res. 41, 1949–1954.
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. Prevent. 50, 1042–1051.
Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Anal. Methods Accid. Res. 1, 39–52.