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