Accident Analysis and Prevention 108 (2017) 27–36

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Comparison of four statistical and machine learning methods for crash severity prediction

Amirfarrokh Iranitalab[a,*], Aemal Khattak[b]

[a] Department of Civil Engineering and Nebraska Transportation Center, University of Nebraska-Lincoln, 330P Prem S. Paul Research Center at Whittier School, Lincoln, NE, 68583-0851, United States

[b] Department of Civil Engineering and Nebraska Transportation Center, University of Nebraska-Lincoln, 330E Prem S. Paul Research Center at Whittier School, Lincoln, NE, 68583-0851, United States

ARTICLE INFO

Keywords:
Traffic crash severity prediction
Multinomial logit
Nearest neighbor classification
Support vector machines
Random forests
Crash costs

ABSTRACT

Crash severity prediction models enable different agencies to predict the severity of a reported crash with unknown severity or the severity of crashes that may be expected to occur sometime in the future. This paper had three main objectives: comparison of the performance of four statistical and machine learning methods including Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF), in predicting traffic crash severity; developing a crash costs-based approach for comparison of crash severity prediction methods; and investigating the effects of data clustering methods comprising K-means Clustering (KC) and Latent Class Clustering (LCC), on the performance of crash severity prediction models. The 2012–2015 reported crash data from Nebraska, United States was obtained and two-vehicle crashes were extracted as the analysis data. The dataset was split into training/estimation (2012–2014) and validation (2015) subsets. The four prediction methods were trained/estimated using the training/estimation dataset and the correct prediction rates for each crash severity level, overall correct prediction rate and a proposed crash costs-based accuracy measure were obtained for the validation dataset. The correct prediction rates and the proposed approach showed NNC had the best prediction performance in overall and in more severe crashes. RF and SVM had the next two sufficient performances and MNL was the weakest method. Data clustering did not affect the prediction results of SVM, but KC improved the prediction performance of MNL, NNC and RF, while LCC caused improvement in MNL and RF but weakened the performance of NNC. Overall correct prediction rate had almost the exact opposite results compared to the proposed approach, showing that neglecting the crash costs can lead to misjudgment in choosing the right prediction method.

1. Introduction

Motor vehicle crash severity modeling has been an important topic of traffic safety research for the past many years. It involves the use of statistical techniques (and recently data mining/machine learning methods) to gain insights into factors that affect or are associated with crash severity, along with predicting the severity outcome of crashes with unknown severity levels. This study is focused on crash severity prediction and its application using four statistical and machine learning methods.

Different agencies may benefit from the ability to predict the severity of a reported crash with unknown severity or the severity of crashes that may be expected to occur sometime in the future. In this study, three types of these stakeholder agencies are mainly considered: transportation safety planners who are interested in predicting the crash costs imposed upon a community in a future year; the hospitals and agencies that provide emergency care that need to predict the injury severity of people involved in traffic crashes using out-of-hospital variables to be able to provide them with proper medical care as fast as possible; and insurance companies that determine their costumers premiums and perform their economic analysis based on a lot of factors, including the costs of the possible crashes they might be involved in, which depends on crash severity.

This paper has three main objectives: 1) comparison of the performance of four statistical and machine learning methods including Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF), in predicting traffic crash severity; 2) developing a crash costs-based approach for comparison of crash severity prediction methods; and 3) investigating the effects of data clustering methods comprising K-means

* Corresponding author.
E-mail addresses: airanitalab2@unl.edu (A. Iranitalab), khattak@unl.edu (A. Khattak).

http://dx.doi.org/10.1016/j.aap.2017.08.008
Received 24 February 2017; Received in revised form 3 August 2017; Accepted 6 August 2017

Clustering (KC) and Latent Class Clustering (LCC), on the performance of crash severity prediction. A literature review is presented in the second section of this paper while the methods mentioned above along with the proposed comparison method are introduced in the third section. The fourth section of this paper presents a crash dataset that the modeling was based on. The fifth section presents a comparison of the prediction results using the proposed approach. Conclusions and discussion about the performance of these prediction methods, the effects of clustering on prediction and the application of the proposed comparison approach complete the paper.

2. Literature review

This literature review is focused on crash severity modeling and prediction and use of prediction and clustering methods in traffic safety research. A significant body of safety research is focused on crash severity modeling. A common approach in these studies is using a statistical modeling tool with crash severity as the dependent variable and characteristics of crash, driver, roadway, weather, etc. as independent variables. Some research papers reported using binary discrete outcome models (e.g., binary logit or probit) with two levels of crash severity (Fan et al., 2016; Khattak et al., 1998; Shibata and Fukuda, 1994) or multinomial models (e.g. Multinomial Logit (MNL) and Probit) with multiple levels of crash severity (Malyshkina and Mannering, 2010; Ye and Lord, 2014). A number of studies used generalized extreme value models, such as nested logit, to address the correlation of unobserved portion of utility of crash severity levels (Shankar et al., 1996; Yasmin and Eluru, 2013), while ordered probit and logit models were used to account for the ordinal nature of crash severity variable in many other studies (Khattak et al., 2002; Kockelman and Kweon, 2002). The prediction performance of crash severity models is mostly used as a validation tool for the estimated/trained models, rather than being the major focus of papers. Data mining/machine learning algorithms are more frequently observed as prediction tools in the literature, rather than statistical methods. For example, Artificial Neural Networks (ANN) (Abdel-aty and Abdelwahab, 2004; Abdelwahab and Abdel-Aty, 2001; Delen et al., 2006), and Decision Trees (Abellán et al., 2013; Chong et al., 2005) were used in a number of traffic safety-related papers. (Abdelwahab and Abdel-Aty, 2001) compared the prediction performance of two well-known ANN paradigms, the multilayer perceptron (MLP) and fuzzy adaptive resonance theory (ART) neural networks on crash severity data with ordered logit models and found out that MLP had the best overall correct prediction.

Literature pertaining to the machine learning methods used in this paper was reviewed. The use of Support Vector Machines (SVM) was reported in some traffic safety-related studies (Chong et al., 2005; Li et al., 2012). (Li et al., 2012) compared the performance of SVM with ordered probit on modeling crash injury severity, and found it is better in terms of prediction and comparable regarding investigating the impacts of variables on crash injury severity. While not very frequent, Random Forests (RF) was used in crash severity modeling and prediction and its performance was reported satisfying (Das et al., 2009; Harb et al., 2009). Further, (Lv et al., 2009) used Nearest Neighbor Classification (NNC) for real-time highway traffic accident prediction; however, this review of traffic crash severity literature did not uncover any research that used NNC.

To account for heterogeneity in crash data (presence of unobserved factors that are correlated with observed variables), a number of crash severity studies utilized data clustering methods. While some papers preferred Latent Class Clustering (LCC) as a model-based clustering method for clustering crash data to obtain homogeneity prior to crash severity modeling (De Oña et al., 2013; Eluru et al., 2012; Kaplan and Prato, 2013; Zhao et al., 2016), others used K-means Clustering (KC) in traffic crash analysis (Anderson, 2009; Mauro et al., 2013; Zhang et al., 2007). (Mohamed et al., 2013) showed that LCC provided better results in a crash dataset compared to KC, while KC had better performance in another crash dataset, relative to LCC.

In summary, the reviewed literature on crash injury severity showed that significant attention has been on crash severity modeling but injury outcome prediction was not a common major focus. Statistical models were more frequently used in crash severity modeling compared to machine learning methods, while machine learning methods were mostly used as prediction tools. MNL, SVM, and RF were found to be used in crash severity modeling with varying popularity. Clustering methods LLC and KC were also found reported in crash analysis in general and crash severity modeling in particular, with varying levels of success.

3. Methodology

In this research, several prediction methods (also known as classification methods) were used for crash severity prediction along with two methods for clustering the crash data. This section presents the prediction and clustering methods utilized in this study. Also, an approach for comparison of the crash severity prediction accuracy, based on the monetary costs of traffic crashes, is proposed and discussed later.

3.1. Prediction methods

In statistics and machine learning, prediction (classification) methods are used for predicting the class (category or population) of an observation, based on information extracted from a dataset consisting of observations with known classes (also known as training data). This study used prediction methods Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM), and Random Forests (RF) for classifying crash injury outcomes.

3.1.1. Multinomial logit (MNL)

Multinomial logit is the most widely used discrete choice model (Train, 2002) and has a long history of use in crash severity literature. Assume that the n^{th} crash observation with severity level i has the severity utility function V_{in}:

$$V_{in} = \boldsymbol{\alpha_i} + \boldsymbol{\beta_i X_{in}} + \varepsilon_{in} \tag{1}$$

In the above equation, V_{in} is a function of independent variables that determines the severity, $\boldsymbol{\alpha_i}$ is a constant parameter, $\boldsymbol{\beta_i}$ is a vector of model parameters (coefficients), $\boldsymbol{X_{in}}$ is a vector of observable characteristics (independent variables) that influence severity, and ε_{in} is an error term that accounts for unobserved effects, which are assumed independently and identically distributed with a generalized extreme value distribution. If $P_n(i)$ is the probability of occurrence of crash severity level i for observation n, then:

$$P_n(i) = \frac{e^{\alpha_i + \beta_i X_{in}}}{\sum_i e^{\alpha_i + \beta_i X_{in}}} \tag{2}$$

So, the probability of occurrence of each crash severity level for a crash with unknown severity level can be calculated after the model is estimated and the affiliated severity level can be predicted based on the calculated probabilities (Savolainen et al., 2011; Train, 2002).

3.1.2. Nearest neighbor classification (NNC)

NNC, also known as k-nearest neighbors, is a prediction method that classifies an observation of interest by looking at the closest $k$observations. The class that the majority of the k closest observations belong to, is predicted as the class of the observation of interest. In other words, nearest neighbor decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points (Cover and Hart, 1967). In the implementation of NNC two decisions are needed: the value of k, and the distance function to use (Elkan, 2011). In this paper the value of k was determined by trying different values for this quantity and finding the best prediction accuracy and the