Journal of Safety Research 42 (2011) 317–326

Contents lists available at SciVerse ScienceDirect

Journal of Safety Research

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

A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks

Randa Oqab Mujalli, Juan de Oña *

TRYSE Research Group, Department of Civil Engineering, University of Granada, Spain

ARTICLE INFO

Available online 28 September 2011

Keywords:
Injury severity
Variable selection
Bayesian networks
Data mining
Classification

ABSTRACT

Introduction: This study describes a method for reducing the number of variables frequently considered in modeling the severity of traffic accidents. The method's efficiency is assessed by constructing Bayesian networks (BN). *Method:* It is based on a two stage selection process. Several variable selection algorithms, commonly used in data mining, are applied in order to select subsets of variables. BNs are built using the selected subsets and their performance is compared with the original BN (with all the variables) using five indicators. The BNs that improve the indicators' values are further analyzed for identifying the most significant variables (accident type, age, atmospheric factors, gender, lighting, number of injured, and occupant involved). A new BN is built using these variables, where the results of the indicators indicate, in most of the cases, a statistically significant improvement with respect to the original BN. *Conclusions:* It is possible to reduce the number of variables used to model traffic accidents injury severity through BNs without reducing the performance of the model. *Impact on Industry:* The study provides the safety analysts a methodology that could be used to minimize the number of variables used in order to determine efficiently the injury severity of traffic accidents without reducing the performance of the model.

1. Introduction

A great deal of information on traffic accidents exists, extracted from different sources, in which many variables that are expected to affect injury severity in traffic accidents are considered. The number of variables used in research work could be enormous, and in some cases this number could be even higher than 100 variables (Delen, Sharda, & Bessonov, 2006). This might complicate the manner of dealing with a certain problem, where some of the variables considered might hide the effect of other more significant ones. Different types of studies tried to identify the most significant variables in order to only consider them in the analysis of traffic accidents (Chang & Wang, 2006; Chen & Jovanis, 2000; Kopelias, Papadimitriou, Papandreou, & Prevedouros, 2007; Xie, Zhang, & Liang, 2009). Therefore, researchers in the field of traffic accidents, specifically in the domain of traffic accident injury severity, focused their research on trying to identify the most significant variables that contribute to the occurrence of a specific injury severity in a traffic accident.

Most previous research used regression analysis techniques, such as logistic and ordered probit models (Al-Ghamdi, 2002; Bédard, Guyatt, Stones, & Hirdes, 2002; Kockelman & Kweon, 2002; Milton, Shankar, & Mannering, 2008; Yamamoto & Shankar, 2004; Yau, Lo, & Fung, 2006). These techniques have their own drawbacks. Chang and Wang (2006) indicated that these regression models use certain assumptions, and if any of these assumptions were violated, the ability of the model to predict the factors that contribute to the occurrence of a specific injury severity would be affected.

Recently researchers used data mining techniques such as artificial neural networks, regression trees, and Bayesian networks.

For instance, Abdelwahab and Abdel-Aty (2001) used artificial neural networks to model the relationship between driver injury severity and crash factors related to driver, vehicle, roadway, and environment characteristics. Thirteen variables were tested first for significance using the χ^2 test, and the results indicated that only six variables were found to be significant: driver gender, fault, vehicle type, seat belt, point of impact, and area type. They compared the classification performance of Multi-Layer Perceptron (MLP) neural networks and that of the Ordered Probit Model (OPM). Their findings indicated that classification accuracy of MLP neural networks outperformed that of the OPM, where 65.6% and 60.4% of cases were correctly classified for the training and testing phases, respectively, compared to 58.9% and 57.1% correctly classified cases for the training and testing phases, respectively, by the OPM.

Another study that used the neural networks to model injury severity in traffic accidents (Delen et al., 2006) classified the injury severity of a traffic accident into five categories (no injury, possible injury, minor non-incapacitating injury, incapacitating, and fatality) and they used certain techniques, such as χ^2 test, stepwise logistic regression, and decision tree induction to select the most significant variables. Out of 150 variables, they selected 17 variables as

* Corresponding author at: ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada, Spain. Tel.: +34 958 24 99 79.
E-mail address: jdona@ugr.es (J. de Oña).

0022-4375/$ – see front matter © 2011 National Safety Council and Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsr.2011.06.010

important in influencing the level of injury severity of drivers involved in accidents. They used the MLP neural networks to classify the injury severity level, where their data included "no injury" cases 10 times more than "fatal cases;" they faced an unbalanced dataset situation that affected their total accuracy (40.71%).

Other researchers used classification tree techniques to model injury severity in traffic accidents (Chang & Wang, 2006). In their study they developed a Classification and Regression Tree (CART) model to establish the relationship between injury severity and twenty explanatory variables that represented: driver/vehicle characteristics, highway/environmental variables and accident variables, where they aimed to model the injury severity of an individual involved in a traffic accident.

Use of Bayesian Networks (BN) as the modeling approach in analysis of crash-related injury severity has been relatively scarce. De Oña, Mujalli, and Calvo (2011) employed BN to model the relationship between injury severity and 18 variables related to driver, vehicle, roadway, and environment characteristics.

Some of these studies tend to apply the models on the datasets without selecting the most significant variables (Chang & Wang, 2006; Delen et al., 2006; Simoncic, 2004). However, Chang and Wang (2006) stated that if the model was applied on a few important variables, much more useful results could be obtained. Others like Abdelwahab and Abdel-Aty (2001) used some statistical techniques to choose the most significant variables before applying their model.

The scope of this research is to build BNs using some selected variables in order to evaluate the performance of BNs when using only the most significant variables, and to compare the results with a base model that is built using all the variables in the original dataset, in order to find out whether using only the most significant variables would affect values of the measures used to assess the built model.

This paper is organized as follows. Section 2 presents the data used. In Section 3, the method followed is presented and described, and a brief review of variable selection methods and the basic concept of BNs are presented, along with a description of the performance indicators used to assess the performance of the built BNs. In Section 4, the results and their discussion are provided. In Section 5, some conclusions are given.

2. Accident data

Accident data were obtained from the Spanish General Traffic Directorate (DGT) for rural highways in the province of Granada (southern Spain) for three years (2003–2005). The total number of accidents obtained for this period was 3,302. The data were first checked out for questionable data, and those that were found to be unrealistic were screened out. Only rural highways were considered in this study; data related to intersections were not included, since intersections have their own specific characteristics and need to be analyzed separately. Finally, the database used to conduct the study contained 1,536 records. Table 1 provides information on the data used for this study.

Eighteen (18) variables were used with the class variable of injury severity (SEV) in an attempt to identify the important variables that affect injury severity in traffic accidents.

The data contained information related to the accidents and other information related to the drivers.

The data included variables describing the conditions that contributed to the accident and injury severity.

- Injury severity variables: number of injuries (e.g., passengers, drivers and pedestrians), severity level of injuries (e.g., slight injured –SI– and killed or seriously injured –KSI). Following previous studies (Chang & Wang, 2006; Milton et al., 2008) the injury severity of an accident is determined according to the level of injury to the worst injured occupant.
- Roadway information: characteristics of the roadway on which the accidents occurred (e.g., pavement width, lane width, shoulder type, pavement markings, sight distance, if the shoulder was paved or not)
- Weather information: weather conditions when the accident occurred (e.g., good weather, rain, fog, snow, and windy)
- Accident information: contributing circumstances (e.g., type of accident, time of accident [hour, day, month and year], and vehicles involved in the accident).
- Driver data: characteristics of the driver, such as age or gender.

3. Method

The procedure used in this study has been the following:

1. The original dataset obtained from the DGT was divided into two subsets: a training set containing 2/3 of the data (1,024 records), and a testing set containing the rest of the data (512 records). The testing set was used to validate the results obtained using the training set.
2. Based on the 18 variables taken from the accident reports (see Table 1), identification of the variables that affect injury severity in traffic accidents was performed using different methods of evaluator-search algorithms.
3. For each one of the selected subsets of variables, 10 simplified BNs were built using the hill climbing search algorithm and the MDL score (De Oña et al., 2011).
4. The performance of the built BNs using the selected subsets of variables was compared with the performance of the original BN, which was built using the 18 variables (BN-18). Five performance evaluation indicators were used.
5. Of all the simplified built BNs, the selected ones are those whose results improve or maintain the results obtained by the performance indicators of BN-18 in 90% of the cases or more, and whose improvements are statistically significant.
6. For the selected BNs, the variables that repeat in more than 50% of the cases are identified and a new BN is built using these variables.
7. Finally, the results obtained by this new BN, based on a double process of variable selection procedure, are compared with those obtained by BN-18.

3.1. Variable selection methods

In machine learning, variable selection is a process that is used to select a subset of variables and to remove variables that do not contribute to the performance of the machine learning technique used.

In this study, we used 6 evaluators with 11 search methods. Weka's Select Variable Panel (Witten & Frank, 2005) was used to perform the variable selection.

A brief description of each of the evaluators used is given below:

1. Correlation-based variable selection (CfsSubsetEval): this evaluator measures the predictive ability of each variable individually and the degree of redundancy among them. It selects the sets of variables that are highly correlated with the class but have low inter-correlation with each other (Hall, 1998).
2. Consistency-based variable selection (ConsistencySubsetEval): this evaluator measures the degree of consistency of the variable sets in class values when the training values are projected onto the set. This evaluator is usually used in conjunction with a random or exhaustive search (Liu & Setiono, 1996).
3. Classifier Subset Evaluator (ClassifierSubsetEval): this evaluator uses the classifier specified in the object editor as a parameter, to evaluate sets of variables on the training data or on a separate holdout set (Witten & Frank, 2005).
4. Wrapper Subset Evaluator (WrapperSubsetEval): this evaluator uses a classifier to evaluate variable sets and it employs cross-validation to estimate the accuracy of the learning scheme for each set (Khhavi & John, 1997).