Accident Analysis and Prevention 43 (2011) 402–411

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks

Juan de Oña*,[1], Randa Oqab Mujalli[1], Francisco J. Calvo

TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 18 June 2010
Received in revised form
16 September 2010
Accepted 23 September 2010

Keywords:
Bayesian networks
Injury severity
Traffic accidents
Classification

ABSTRACT

Several different factors contribute to injury severity in traffic accidents, such as driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics, and atmospheric factors. This paper shows the possibility of using Bayesian Networks (BNs) to classify traffic accidents according to their injury severity. BNs are capable of making predictions without the need for pre assumptions and are used to make graphic representations of complex systems with interrelated components. This paper presents an analysis of 1536 accidents on rural highways in Spain, where 18 variables representing the aforementioned contributing factors were used to build 3 different BNs that classified the severity of accidents into slightly injured and killed or severely injured. The variables that best identify the factors that are associated with a killed or seriously injured accident (accident type, driver age, lighting and number of injuries) were identified by inference.

1. Introduction

The number of traffic accidents and their effects, mainly human fatalities and injuries, justify the importance of analyzing the factors that contribute to their occurrence. Identifying the factors that significantly influence the injury severity of traffic accidents was the main objective of many previous studies. Factors affecting injury severity of a traffic accident are usually caused by one or more of the following factors: driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics and atmospheric factors (Kopelias et al., 2007; Chang and Wang, 2006).

Regression analysis has been widely used to determine the contributing factors that cause a specific injury severity. The most commonly used regression models in traffic injury analysis are the logistic regression model and the ordered Probit model (Al-Ghamdi, 2002; Milton et al., 2008; Bédard et al., 2002; Yau et al., 2006; Yamamoto and Shankar, 2004; Kockelman and Kweon, 2002). However, most of the regression models that are used to model traffic injury severity have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (i.e. linear relations between the variables) (Chang and Wang, 2006). If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of severe injury.

Gregoriades (2007) highlighted the interest of using Bayesian Networks (BNs) to model traffic accidents and discussed the need to not consider traffic accidents as a deterministic assessment problem. Instead, researchers should model the uncertainties involved in the factors that can lead to road accidents. He listed a number of candidate approaches for modeling uncertainty, such as, Bayesian probability.

BNs make it easy to describe accidents that involve many interdependent variables. The relationship and structure of the variables can be studied and trained from accident data. They do not need to know any pre-defined relationships between dependent and independent variables.

The three main advantages of BNs are bi-directional induction, incorporation of missing variables and probabilistic inference. By using BNs, it is relatively easy to discover the underlying patterns of data, to investigate the relationships between variables and to make predictions using these relationships. Incident data used in a study by Ozbay and Noyan (2006) were collected from incident clearance survey forms to understand incident clearance characteristics and then used to develop incident duration prediction models. The researchers modeled the incidents' clearance durations using BNs and were able to represent the stochastic nature of incidents.

Using BNs to analyze traffic accident injury severity is scarce. A two car accident injury severity model was constructed using BNs (Simoncic, 2004). A BN was built using several variables, and the Most Probable Explanation (MPE) was calculated for the most probable configuration of values for all the variables in the BN, in order to serve as an indication of the quality of the estimated BN. The results pointed out that BNs could be applied in road accident

* Corresponding author. Tel.: +34 958 24 99 79.
E-mail address: jdona@ugr.es (J. de Oña).
[1] These authors have contributed equally to this work.

0001-4575/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2010.09.010

modeling, and some improvements, such as using more variables and larger datasets, were recommended. Although this study highlighted the possibility of using BNs to model traffic accidents, the results were based on building only one possible network, without measuring the performance of the Bayesian classifier.

The scope of this paper is to validate the possibility of using BNs to classify traffic accidents according to their injury severity, and to find out the best BN classification performance along with the best graphical representation, in order to be capable of identifying the relevant variables that affect the injury severity of traffic accidents.

This paper is organized as follows. Section 2 presents the data used and briefly reviews the concept of BNs and Bayesian learning. The methods used for preprocessing and evaluating the data are also presented; finally a brief description of inference is presented. In Section 3, the results and their discussion are presented. In Section 4, summary and conclusions are given.

2. Materials and methods

2.1. Accident data

Accident data were obtained from the Spanish General Traffic Directorate (DGT) for rural highways for the province of Granada (South of Spain) for three years (2003–2005). The total number of accidents obtained for this period was 3302. The data were first checked out for questionable data, and those which 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 1536 records. Table 1 provides information on the data used for this study.

Eighteen variables were used with the class variable of injury severity (SEV) in an attempt to identify the important patterns of an accident that usually require an explanation.

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., fatal, severe, slight). Following previous studies (Chang and 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., grade, pavement width, lane width, shoulder type, pavement markings, sight distance, if the shoulder was paved or not, etc.).
- 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.

2.2. BN definition

Over the last decade, BNs have become a popular representation for encoding uncertain expert knowledge in expert systems. The field of BNs has grown enormously, with theoretical and computational developments in many areas (Mittal et al., 2007) such as: modeling knowledge in bioinformatics, medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming, and law.

Let $U = \{x_1, \ldots, x_n\}$, $n \geq 1$ be a set of variables. A BN over a set of variables U is a network structure, which is a Directed Acyclic Graph (DAG) over U and a set of probability tables $B_p = \{p(x_i|pa(x_i), x_i \in U)\}$ where $pa(x_i)$ is the set of parents or antecedents of x_i in BN and $i = 1, 2, 3, \ldots, n$. A BN represents joint probability distributions $P(U) = \prod_{x_i \in U} p(x_i|pa(x_i))$.

The classification task consists in classifying a variable $y = x_0$ called the class variable, given a set of variables $U = x_1, \ldots, x_n$, called attribute variables. A classifier h: $U \to y$ is a function that maps an instance of U to a value of y. The classifier is learned from a dataset D consisting of samples over (U, y). The learning task consists of finding an appropriate BN given a data set D over U.

BNs are graphical models of interactions among a set of variables, where the variables are represented as nodes (also known as vertices) of a graph and the interactions (direct dependences) as directed links (also known as arcs and edges) between the nodes. Any pair of unconnected/nonadjacent nodes of such a graph indicates (conditional) independence between the variables represented by these nodes under particular circumstances that can easily be read from the graph. Each node contains the states of the random variable and it represents a conditional probability table. The conditional probability table of a node contains the probabilities of the node being in a specific state, given the states of its parents.

Fig. 1 shows that the dependencies and independencies among the factors that affect the time of journey (the class variable) are represented in the form of direct edges (arrows) between factors that are represented as nodes. For example, the variable (vehicle type) is a parent (antecedent) of the two variables (cost and velocity) called children or descendents. Any knowledge (evidence) about the parent variable affects the probabilities of occurrence of the children or descendent variables.

It should be noticed that the edges in a BN are not necessarily causal. That is, a BN can satisfy the probability distribution of the variables in the BN without the edges being causal (Neapolitan, 2009). Thus, the edges between variables in a non-causal BN could imply a sort of interrelationship(s) among these variables.

2.3. BN learning and the scoring metrics used

When there are masses of data available and it is necessary to interpret them and to provide a model for predicting the behavior of unobserved cases, the learning of both structure and parameters is used (Cooper and Herskovits, 1992). There are two main approaches to structure learning in BNs:

- **Constraint based**: Perform tests of conditional independence on the data, and search for a network that is consistent with the observed dependencies and independencies.
- **Score based**: Define a score that evaluates how well the dependencies or independencies in a structure match the data and search for a structure that maximizes the score.

The advantage of score-based methods over the constraint-based methods is that they are less sensitive to errors in individual tests; compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge. Because of the aforementioned advantages, the score based method is followed in this study.

Weka software (Witten and Frank, 2005) was used in this study to build the BN. This software is freely available, it is implemented in Java language, it contains a collection of data processing and modeling techniques and it contains a graphical user interface. The BNs built here used all the nineteen variables of the 1536 records.