Expert Systems with Applications 40 (2013) 6047–6054

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Analysis of traffic accident severity using Decision Rules via Decision Trees

Joaquín Abellán [a,*], Griselda López [b], Juan de Oña [b]

[a] Department of Computer Science & Artificial Intelligence, University of Granada, ETSI Informática, c/Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
[b] 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

Keywords:
Traffic accident
Severity
Road safety
Decision Trees
Decision Rules

ABSTRACT

A Decision Tree (DT) is a potential method for studying traffic accident severity. One of its main advantages is that Decision Rules (DRs) can be extracted from its structure. And these DRs can be used to identify safety problems and establish certain measures of performance. However, when only one DT is used, rule extraction is limited to the structure of that DT and some important relationships between variables cannot be extracted. This paper presents a more effective method for extracting rules from DTs. The method's effectiveness when applied to a particular traffic accident dataset is shown. Specifically, our study focuses on traffic accident data from rural roads in Granada (Spain) from 2003 to 2009 (both included). The results show that we can obtain more than 70 relevant rules from our data using the new method, whereas with only one DT we would have extracted only five relevant rules from the same dataset.

1. Introduction

The current large number of road accidents implies an unacceptable burden on the community in terms of human injury and economic cost. Therefore, one of the main tasks of safety analysts is to make a comprehensive assessment of traffic accidents to determine what caused them, so measures can be taken to mitigate the severity of their consequences.

Usually, an accident severity analysis is carried out to study a particular dataset of traffic accidents with the aim of obtaining useful knowledge to tackle this problem. In most countries, traffic accidents are recorded in accident reports by police officers, and subsequently the information is stored in a dataset. A huge amount of information can be obtained from such datasets. It could be said that their true potential consists in the knowledge that can be extracted from them.

Traditionally, regression techniques such as Logit and Porbit have been used to analyze traffic accident severity (Kashani & Mohaymany, 2011; Mujalli & de Oña, 2013; Savolainen, Mannering, Lord, & Quddus, 2011). However, these techniques establish their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If the assumptions are violated, the model can lead to erroneous estimations of injury likelihood (Chang & Wang, 2006).

Data Mining (DM) techniques are one of the solutions used to analyze huge amounts of data and turn it into useful information and knowledge (Han & Kamber, 2006). DM has been widely used in crash severity analysis with satisfactory results. Abdel Wahab and Abdel-Aty (2001) investigated the use of Artificial Neural Network models for predicting injury severity in two-vehicle crashes at signalized intersections. Recently, Bayesian Networks have been used to analyze traffic accident severity (De Oña, López, Mujalli, & Calvo, 2013b, 2011; Mujalli & de Oña, 2011). Decision Trees (DT) is another DM technique used to study crash severity (Chang & Chien, 2013; Chang & Wang, 2006; De Oña, López, & Abellán, 2013a; Montella, Aria, D'Ambrosio, & Mauriello, 2011, 2012).

DTs, in particular, represent a set of useful methods for analyzing traffic accident severity because, normally, they are non-parametric methods that do not depend on any functional form and require no prior probabilistic knowledge on the phenomena under study. Moreover, the structure of a DT permits the extraction of Decision Rules (DR) that can be used to discover behaviors that occur within a specific dataset. Safety analysts could use these rules to understand the events leading up to a crash and identify the variables that determine how serious an accident will be (De Oña et al., 2013a).

DTs have been largely reported in road safety literature. Specifically, the most widely used method in the literature on traffic accident severity is the CART method (Chang & Chien, 2013; Chang & Wang, 2006; De Oña et al., 2013a; Kashani & Mohaymany, 2011; Kashani, Mohaymany, & Ranjbari, 2011; Kuhnert, Do, & McClure, 2000; Montella et al., 2011, 2012; Pakgohar, Tabrizi, Khalili, & Esmaeili, 2010). However, CART always yields binary trees, which sometimes cannot be summarized as efficiently for interpretation and/or presentation (Breiman, Friedman, Olshen, & Stone, 1984).

[*] Corresponding author. Tel.: +34 958242376; fax: +34 958243371.
E-mail address: jabellan@decsai.ugr.es (J. Abellán).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.05.027

In the case of road accidents, they may not be very practical when it comes to analyzing the impact of a specific category of variable on crash severity. The C4.5 algorithm (Quinlan, 1993) is another method that is frequently used in several fields because it does not present the binary restriction when tree building. It has been used before to analyze traffic accident severity (De Oña et al., 2013a). An important difference between the two methods (CART vs. C4.5) is the split criterion: the CART method uses the Gini Index, based on a measure of diversity; and the C4.5 algorithm uses the Info Gain Ratio (IGR), based on the entropy measure on probabilities (Shannon, 1948).

However, using DRs from DTs to extract knowledge from a specific dataset also poses certain limitations. The extraction of knowledge is constrained by the tree's structure, for instance, and the DRs are dependent on a DT's structure. The DRs are extracted from each tree branch from the root node to the terminal node, and therefore knowledge is extracted only in that direction. However, there could be other important rules that depend on the root node from which the tree is built, and that are not detected by the tree's structure.

In this paper, a particular method for extracting DRs from DTs is used to extract all the knowledge from a particular dataset. The main characteristic of this method is that different DTs are built by varying the root node. Thus, every possible set of DRs is obtained from each tree. The resulting useful rules could be used by road safety analysts to establish specific measures of performance.

To conduct a full analysis of the dataset, in our method for extracting DRs, we use different DTs built using two different split criteria, both each with a different meaning. In fact, the two criteria complement each other, and even a previous study recommends using the both criteria for a full analysis (De Oña et al., 2013a). By doing so, a broader range of rules can be obtained from a single dataset.

The paper is structured as follows: Section 2 shows the main features of the traffic accident data used to validate the methodology. The necessary prior knowledge on decisions trees and the procedure to build them is presented. It also describes the method used to obtain Decision Rules, and how to obtain the importance of each of the variables considered in the model. Section 3 presents the main results obtained and the discussion. Finally, the last section presents the conclusions.

2. Materials and methods

2.1. Traffic accident data

Traffic accidents where only 1 vehicle was involved, for two-lane rural highways in Granada (Spain), were collected from the Spanish General Traffic Accident Directorate (DGT). The study period was 7 years (2003–2009) and accidents at intersections were not considered. Thus, the total number of accidents was 1801.

In order to identify the main factors that had an impact on accident severity and taking into account the available variables in the original dataset, 19 variables were used (see Table 1). The variables described characteristics related to the driver (age and gender); accident (month, time, day, number of injuries, occupants involved, accident type and cause); road (safety barriers, pavement width, lane width, shoulder width, shoulder type, road markings and sight distance); vehicle (vehicle type); and environment (atmospherics factors and lighting conditions).

The class variable was accident severity (SEV in Table 1). Following previous studies (Chang & Wang, 2006; De Oña et al., 2011; Kashani & Mohaymany, 2011), accident severity was defined according to the worst injured occupant, and two levels of severity were identified: accident with slightly injured (SI) and accidents with killed or seriously injured (KSI).

2.2. Classification and Decisions Trees

In the general domain of DM, a supervised classification problem is normally defined as follows: given a dataset of observations, called a *training set*, we want to obtain a set of rules that can be used to assign a value of the variable to be predicted to each new observation. To verify the quality of this set of rules, a different set of observations is used; this set is called the *test set*. The variable to be predicted (classified) is called *class variable* and the rest of variables in the dataset are called *predictive attributes* or *features*. There are important applications of classification in fields such as medicine, bioinformatics, physics, pattern recognition, economics, civil engineering, etc.

A DT is a structure that can be used in classification and regression tasks. If the class variable (i.e., the variable under study) has a finite set of possible states or values, the task is called a classification; otherwise, it is called a regression.

Within a DT, each node represents a feature and each branch represents one of the states of this variable. A tree leaf (or terminal node) specifies the expected value of the class variable depending on the information contained in the training dataset. Associated to each node is the most informative variable which has not already been selected in the path from the root to the node (as long as this variable provides more information than if it had not been included). In the latter case, a leaf node is created with the most probable class value for the partition of the dataset defined with the configuration given by the path from the root node to that leaf node.

When a new sample or instance of the test dataset is obtained, a decision or prediction about the state of the class variable can be made by following the path in the tree from the root to a leaf, using the sample values and the tree's structure.

A DT allows us to extract DRs directly. A DR is a logic conditional structure of the type "IF A THEN B". Where A is the antecedent of the rules (in our case, a set of statuses of several attribute variable); and B is the consequent (in our case, it is only one state of the class variable). Thus, each rule starts at the root node, and each variable that intervenes in tree division makes an IF of the rule, which ends in leaf nodes with a value of THEN (which is associated with the state resulting from the leaf node). The resulting state is the status of the class variable that shows the highest number of cases in the leaf node analyzed. Thus, a priori, the number of rules can be identified with the number of terminal nodes in the tree.

Fig. 1 shows an example of a DT built using a dataset of accidents. The DT is formed by two attribute variables, and the class variable is the *severity* (two states) of the accidents. This example shows how accidents are classified by each status of the class variable (slight accidents vs. severe accidents). In addition, the chart gives the number of cases shown in each leaf or terminal node (shaded nodes in the tree), distinguishing the cases that are predicted correctly in each terminal node. One example of DRs is the following: IF (*age* $\leqslant$25 yrs AND *speed* $\leqslant$80 km/h) THEN (*severity* = slight accident).

There is a wealth of information in the literature about different procedures to build DT, but normally they have the following characteristics in common:

- The criterion used for selecting the attribute to be placed in a node and branching. This criterion is known as the split criterion.
- The criterion used to stop the branching of the tree.
- The method for assigning a class label or a probability distribution at the leaf nodes.