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, 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)

Abstract

A Decision Tree (DT) is a potential method for studying traffic accident severity. One of its main advantages is that Decision Rules can be extracted from its structure and used to identify safety problems and establish certain measures of performance. However, when it used only one DT, the rule extraction is limited to the structure of that DT and some important relationships between variables cannot be extracted. This paper presents a method for extracting rules from a DT more effectively. The method's effectiveness when applied to a particular traffic accidents 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 had extracted only 5 rules from the same dataset.

Keywords: traffic accident; severity; decision trees; decision rules; road safety

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. 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 analyse traffic accident severity (Kashani and Mohaymany, 2011; Mujalli and de Oña, in press; Savolainen et al., 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 and 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 and 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

* Corresponding author. Phone: +34 958 24 23 76; email: jabellan@decsai.ugr.es

analyze traffic accident severity (De Oña et al., 2011; De Oña et al., 2013b; Mujalli and de Oña, 2011). Decision Trees (DT) is another DM technique used to study crash severity (Chang and Wang, 2006; Chang and Chien, 2013; De Oña et al., 2013a; Montella et al., 2011; Montella et al., 2012).

DTs, in particular, represent a very useful method for analysing traffic accidents severity because, normally, they are a non-parametric method that does not depend on any functional form and require no prior probabilistic knowledge on the phenomena under study. Moreover, their structure permits the extraction of Decision Rules (DR) that can be used to discover behaviours 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 and Chien, 2013; Chang and Wang, 2006; De Oña et al., 2013a; Kashani et al., 2011; Kashani and Mohaymany, 2011; Kuhnert et al., 2000; Montella et al., 2011; Montella et al., 2012; Pakgohar et al., 2010;). However, CART always yields binary trees, which sometimes cannot be summarized as efficiently for interpretation and/or presentation (Breiman et al., 1984). 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 analyse 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 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 DT's structure. The DRs are extracted from each tree branch from the root node to the terminal node. Therefore, knowledge is extracted only in the direction dictated from the root to the terminal node. However, there could be other important rules that depend on the root node from which the tree is built 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 useful rules could be used by road safety analyst to establish specific measures of performance.

Moreover, to conduct a full analysis of the dataset using our method for extracting DRs, we build DTs using two different split criteria, both each with a different meaning. In fact, the two criteria complement each other and a previous study recommends using both 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 and 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 1,801.

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 and Wang, 2006; De Oña et al., 2011; Kashani and 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 in order 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 this node (as long as this variable provides more information than if it had not been included). In the latter case, a leaf node is added 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.

NUM	VARIABLES	DESCRIPTION: CODE	COUNT	SEVERITY %SI	SEVERITY %KSI
1	ACT: Accident type	Fixed objects collision: **CO**	19	76.47	23.53
		Collision with pedestrian: **CP**	152	33.33	66.67
		Other (collision with animals, etc.): **OT**	32	68.57	31.43
		Rollover (in carriage without any collision): **RO**	118	61.86	38.14
		Run off road (with or without collision): **ROR**	1480	51.77	48.23
2	AGE: Age	≤ 20: **≤ 20**	219	52.73	47.27
		[21-27]: **[21-27]**	492	50	50
		[28-60]: **[28-60]**	948	51.76	48.24
		≥ 61: **≥ 61**	110	59.68	40.32
		Unknown: **UN**	32	27.59	72.41
3	ATF: Atmospheric factors	Good weather: **GW**	1540	50.58	49.42
		Heavy rain: **HR**	43	63.16	36.84
		Light rain: **LR**	161	58.75	41.25
		Other: **O**	57	51.06	48.94
4	BAR: Safety barriers	No: **N**	1740	48.3	54.7
		Yes: **Y**	61	53.6	46.4
5	CAU: Cause	Driver characteristics: **DC**	1471	48.99	51.01
		Combination of factors: **CO**	262	61.16	38.84
		Other: **OT**	29	72.73	27.27
		Road characteristics: **RC**	24	84	16
		Vehicle characteristics: **VC**	15	63.64	36.36
6	DAY: Day	Working day after the weekend or public holiday: **APH**	131	57.62	42.38
		Working day before the weekend or public holiday: **BPH**	286	52.26	47.74
		On a weekend or public holiday: **PH**	532	50.36	49.64
		Regular working day: **WD**	852	51.05	48.95
7	LAW: Lane width	< 3,25 m: **THI**	503	46.87	53.13
		[3,25-3,75] m: **MED**	1264	53.2	46.8
		> 3,75 m: **WID**	34	58.54	41.46
8	LIG: Lighting	Daylight: **DAY**	958	55.49	44.51
		Dusk: **DU**	103	54.29	45.71
		Insufficient (night-time): **IL**	131	51.15	48.85
		Sufficient (night-time): **SL**	66	59.72	48.28
		Without lighting (night-time): **WL**	543	43.1	56.9
9	MON: Month	Autumn: **AUT**	412	53.07	46.93
		Spring: **SPR**	440	53.64	46.36
		Summer: **SUM**	479	51.63	48.37
		Winter: **WIN**	470	47.92	52.08
10	NOI: Number of injuries	1 injury: **[1]**	1233	53.43	46.57
		> 1 injury: **[>1]**	568	47.35	52.65
11	OI: Occupants involved	1 occupant: **[1]**	1171	51.2	48.8
		2 occupants: **[2]**	374	51.48	48.52
		> 2 occupants: **[>2]**	256	53.71	46.29
12	SHT: shoulder type	No: **N**	309	49.35	50.65
		Non existent or impassable: **NE**	580	50.89	49.11
		Yes: **Y**	912	52.74	47.26
13	PAW: Pavement width	[6-7] m: **MED**	530	53.19	46.81
		< 6 m: **THI**	282	45.56	54.44
		> 7 m: **WID**	989	52.27	47.73
14	ROM: Pavement markings	Does not exist or was deleted: **DME**	168	52.35	47.65
		Separate margins of roadway: **DMR**	180	48.31	51.69
		Separate lanes and define road margins: **SLD**	1368	52.23	47.77
		Separate lanes only: **SLO**	85	46.59	53.41
15	SEX: Gender	Female: **F**	286	62.18	37.82
		Male: **M**	1513	49.61	50.39
		Unknown: **UN**	2	75	25
16	SHW: Shoulder width	< 1.5 m: **THI**	699	52.54	47.46
		[1.5-2.5] m: **MED**	898	50.28	49.72
		Non existent or impassable: **NE**	204	50.57	49.43
17	SID: Sight distance	Atmospheric: **ATM**	30	67.5	32.5
		Building: **BU**	6	36.36	63.64
		Other: **OT**	12	50	50
		Topography: **TOP**	420	49.39	50.61
		Vegetation: **VEG**	13	50	50
		Without restriction: **WR**	1320	51.94	48.06
18	TIM: Time	[00:00-05:59]: **[0-6)**	340	48.06	51.94
		[06:00-11:59]: **[6-12)**	380	58.73	41.27
		[12:00-17:59]: **[12-18)**	591	52.77	47.23
		[18:00-23:59]: **[18-24)**	490	47.22	52.78
19	VEH: Vehicle type	Cars: **CAR**	1287	47.1	52.9
		Trucks: **TRU**	78	53.8	46.2
		Motorbikes and motorcycles: **MOT**	385	35.6	64.4
		Other: **OT**	51	50.6	49.4
20	SEV: Severity	Accident with slightly injured: **SI**	929	-	-
		Accidents with killed or seriously injured: **KSI**	872		

Table 1. Variable description.

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 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 variables); 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.

Figure 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 shows 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*≤25 years AND *speed*≤80 km/h) THEN (*severity* = slight accident).

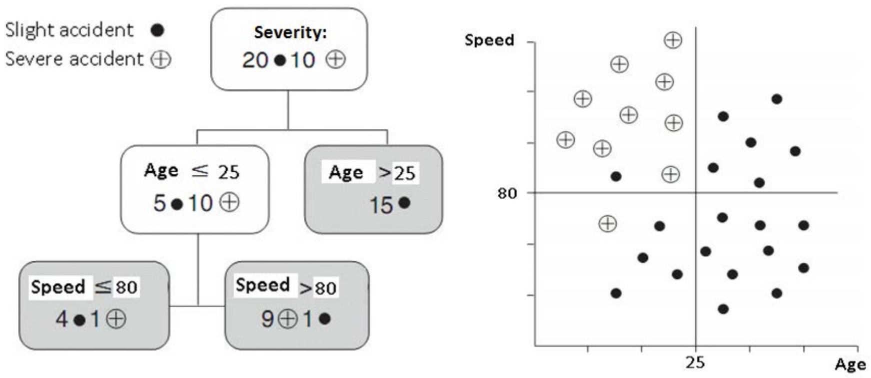

Figure 1. Example of a DT´s structure and classification.

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

- The criteria used for selecting the attribute to be placed in a node and branching. This criterion is known as the split criterion.
- The criteria used to stop branching the tree.
- The method for assigning a class label or a probability distribution at the leaf nodes.
- The pruning process (pre or post building process), which simplifies the structure of the tree and prevents over-fitting (i.e. the dependence of the data used to build the model).

DTs started to play an important role in machine learning following publication of the CART method (Breiman et al., 1984) and Quinlan's ID3 algorithm (Quinlan, 1987). The former uses a split criterion based on the Gini Index. The Quinlan method uses a split criterion, called Information Gain (IG), based on the entropy measure on probabilities (Shannon, 1948). Subsequently, Quinlan (1993) also presented the algorithm C4.5, which is an advanced version of ID3 with a split criterion, called Information Gain Ratio (IGR), which is similar to the one used in the ID3 procedure penalizing the variables with many states. Since then, C4.5 has been considered as a standard model in supervised classification. It has also been widely applied to very different fields as a data analysis tool.

The Gini Index is a measure of diversity and for a variable C (for example, the class variable in a classification problem), it can be expressed as follows:

$$gini(C) = 1 - \sum_j p^2(C = c_j). \quad (1)$$

In the same line, Shannon's entropy is a measure of information based on uncertainty that can be expressed as:

$$H(C) = -\sum_j p(C = c_j) log(p(C = c_j)). \quad (2)$$

The split criterion used in CART, that we call GInf, is based on the Gini Index, as follows:

$$GInf(C, X) = gini(C|X) - gini(X), \quad (3)$$

where $gini\,(C|X) = \sum_t p(x_t) gini\,(C|X = x_t)$ and X is another known variable (for example, a feature variable in a classification problem).

In the C4.5 procedure, the split criterion is called the Info Gain Ratio and it is a measure based on Shannon's entropy. It is defined as:

$$IGR(C, X) = \frac{IG(C,X)}{H(X)}, \quad (4)$$

where $\mathrm{IG}(C, X) = H(C) - H(C|X)$, IG is the Info Gain measure defined by Quinlan (1986) and $H(C)$ is the entropy of C. The probability of each value of the class variable is estimated in the training dataset. In the same way, $H(C|X) = -\sum_t \sum_j p(c_j|x_t) \log(p(c_j|x_t))$, where x_t, t=1,…,|X|, is each possible state of X; and c_j, j=1,…,k, each possible state of C.

2.3. Procedure for building Decision Trees

In this section, how to build a simple DT using the above mentioned split criteria, is explained. The procedure of Abellán and Moral (2003) to build DTs using imprecise probabilities and uncertainty measures is used. The method can easily be adapted to be used with precise probabilities; for example, via the GInf or IGR split criteria.

Each node N in a DT produces a partition D of the dataset (for the root node the entire dataset is considered). Also, each node N has associated a list "Γ" of labels of features (features that are not in the path from the root node to N). The recursive procedure of Abellán and Moral (2003) for building DTs can be expressed in the algorithm shown in Figure 2.

Procedure BuildTree (N, Γ)

1. *If Γ =Φ, then* ***Exit***
2. *Let D be the partition associated with node N*
3. *Compute the value of the maximum gain of information for a feature on D (using a split criterion: SC)*
   δ= max SC(C,X)
4. *If δ is lower than or equal to 0 then* ***Exit***
5. *Else*
   6. *Let X_t be the variable for which the maximum δ is attained*
   7. *Remove X_t from Γ*
   8. *Assing X_t to the node N*
   9. *For each possible value x_t of X_t*
   10. *Add a node N_t*
   11. *Make N_t a child of N*
   12. *Call BuilTree (N_t , Γ)*

Figure 2. Algorithm for building DTs.

Each Exit state in the above procedure corresponds to a leaf node. Here, the most probable value of the class variable, associated with the corresponding partition, is selected.

2.4. Method to obtain Decision Rules: Information Root Node Variation method.

When rules are obtained from a single classification tree, they are determined by the variable that is used as a root. In other words, the information we select from our dataset depends on the direction indicated by the root variable. This one is the most informative variable about the class variable via a split criterion.

The method that we will propose for obtaining rules, which we call the Information Root Node Variation (IRNV) method, is based on using trees obtained by varying the root node. If there are m features, and RX_i is the feature that occupies position i in importance with regards to the split criterion, RX_i will be used as the root to build DT_i (i=1,…,m). We use the simple method for building trees explained in section 2.2, however now the root node is selected directly for each tree (the rest of the building procedure remains the same). Thus, we obtain m trees and m rule sets, DT_i and RS_i (i=1,…,m), respectively. Each RS_i is checked in the test set to obtain the final rule set. The entire procedure is carried out using GInf and IGR criteria.

The following chart gives a more systematic explanation of the entire process:

1. Select Ginf as the split criterion SC for building trees.
2. Build DT_i using RX_i as the root node and SC, for i=1,…,m.
3. Extract RS_i from each DT_i.
4. Check RS_i in the TEST set → Selection of rules from RS_i.
5. Extract the final rule set obtained by using the SC.
6. Use the IGR as SC and go back to step 2. Skip if IGR was used before.
7. Join the final rule sets obtained using GInf and IGR.

Figure 3 gives a graphic explanation of the procedure for each split criterion. In other words, the method shown in Figure 3 must be applied as many times as the split criteria used.

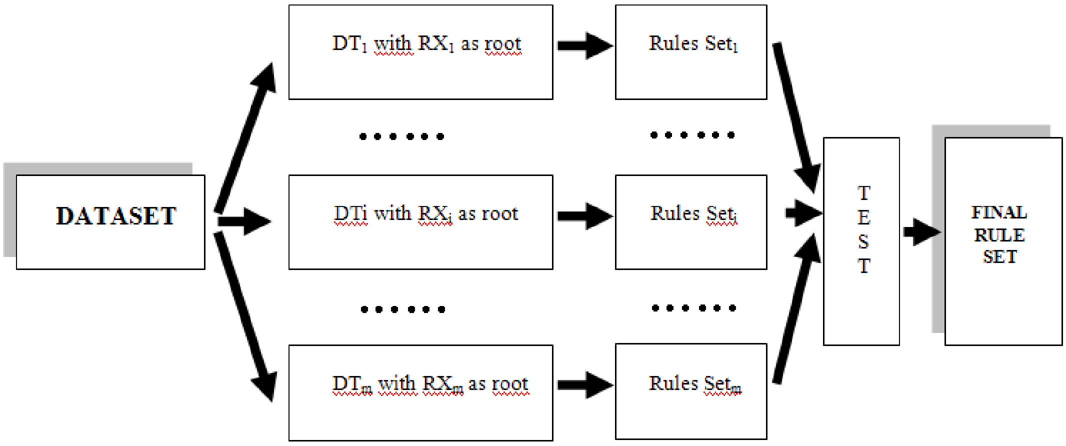

Figure 3. Information Root Node Variation method for each split criterion.

The IRNV method allows 19 possible DTs (i.e. one DT for each one of the features) for each one of the split criteria (GInf and IGR) (see Table 1) to be built from our dataset. All the DRs are extracted for each of the DTs built. Finally, each RS_i obtained from each DT_i (with each split criterion) is verified on the corresponding test set.

It is important to point out that we use two very different split criteria that can be used to build different trees, despite the fact that they begin with the same root node.

2.5. Significant Decision Rules

In order to extract significant and useful rules (i.e. rules that could provide useful information for the implementation of road safety strategies in the future), of the type "A→B", the parameters and the minimum threshold used by Montella et al. (2011) and De Oña et al. (2013a) are used:

- *Support* (S) is the percentage of the dataset where "A & B" appear. Minimum threshold is S≥ 0.6%
- *Population* (Po) is the percentage of the dataset where "A" appears. Minimum threshold is Po≥1%
- *Probability* (P) is the percentage of cases in which the rule is accurate (i.e. P=S/Po expressed as percentage). Minimum threshold is P≥60%

Due to the large number of patterns considered, DTs can suffer from an extreme risk of Type-1 error, that is, of finding patterns that appear due to chance alone to satisfy constraints on the sample data (Webb, 2007). To reduce this error and following other authors (Chang and Chen, 2013; De Oña et al., 2013a; Kashani and Mohaymany, 2011; Montella et al., 2011) the rules extracted on the training set (with the minimum parameters) are validate using the testing set.

2.6. Importance of the variables

The importance of a variable in the model is defined following Eq. 5:

$$VIM\ (X) = \sum_{i=1}^{h} \frac{n x_i}{n} I(C, X = x_i), \qquad (5)$$

where X is the variable with possible states $\{x_i...x_n\}$, C is the class variable (SEV in our case), nxi is the number of cases that X=xi, and n is the number of total cases; and I is the GInf or the IGR split criterion, i.e. an information gain measure.

3. RESULTS

The software used to build the DTs was Weka (Witten and Frank, 2005). The procedures for building the DTs based on each split criterion and the root node variation procedure were implemented using the method proposed by Abellán and Masegosa (2010).

In order to obtain DRs that would be useful and easy to understand by the analysts, we built DTs with only four levels. Previous studies (Montella et al., 2011; Montella et al., 2012) used the same number of levels.

Using the method exposed in Section 2.2 to obtain DRs we would use only one DT (DT_1 in Table 2). Using the IRNV method, by varying the root node, 19 DTs, can be used to obtain DRs, (DT_1 to DT_{19} in Table 2) for each one of the split criteria (GInf and IGR). Thus, the total number of DTs generated is 38 (19 for GInf and 19 for IGR).

DTS	GInf			IGR		
	ROOT NODE	RULES TRAINING	VALIDATED RULES	ROOT NODE	RULES TRAINING	VALIDATED RULES
DT_1	ACT	14	5	SEX	8	6
DT_2	CAU	16	4	ACT	8	2
DT_3	SEX	8	4	CAU	12	5
DT_4	LIG	17	2	CAT	15	7
DT_5	VEH	12	4	VEH	6	1
DT_6	CAT	14	6	LIG	16	7
DT_7	PAW	10	3	NOI	5	2
DT_8	AGE	7	3	SID	10	3
DT_9	TIM	12	3	PAW	7	3
DT_{10}	SID	8	4	AGE	7	3
DT_{11}	NOI	9	3	LAW	4	3
DT_{12}	DAY	12	5	TIM	11	6
DT_{13}	LAW	14	8	DAY	11	5
DT_{14}	MON	16	3	BAR	6	4
DT_{15}	ROM	8	5	MON	10	4
DT_{16}	OI	12	7	ROM	7	3
DT_{17}	SHW	14	5	OI	5	1
DT_{18}	BAR	11	3	SHW	13	10
DT_{19}	SHT	13	1	SHT	13	6
TOTAL		227	78		174	81

Table 2. Number of rules obtained in the different steps of the IRNV method.

DT_1 presents a different root node depending on the split criteria: ACT is selected as root node when GInf is used; whereas GEN is selected when using IGR (Table 2). For this DT, 23 rules were extracted from the training set (15 with GInf and 8 with IGR) but only 11 rules (5 with GInf and 6 with IGR) were validated with the testing set.

Table 2 shows, for each root node, the number of the DRs obtained from each DT. Both criteria (GInf and IGR) generate more than 170 rules for the training set that verify the minimum threshold fixed for the parameters S, Po and P. The variable that generates the highest number of rules when they are used as root node is LIG, and depending on the criteria the number of rules is: 17 rules when GInf is used, and 16 rules when using IGR.

When the rules are validated using the testing set, the number of rules decreases considerably (78 rules with GInf and 81 rules with IGR). We would like to remark that all DTs generate valid DRs. When GInf is used, the root node that generates the highest number of valid rules is LAW (8 rules). When IGR is used, the root node that generates the highest number of valid rules is SHW (10 rules). In both cases, the number of valid rules obtained from a single tree, using both criteria, is lower (5 with GInf and 6 with IGR).

VARIABLE	GInf (%)	VARIABLE	IGR (%)
ACT	100.00	SEX	100.00
CAU	77.89	ACT	94.51
LIG	69.56	CAU	81.72
SEX	69.53	ATF	69.17
VEH	67.43	VEH	51.96
ATF	61.21	LIG	36.30
PAW	44.20	NOI	33.37
TIM	41.09	SID	32.64
AGE	40.72	PAW	27.35
SID	35.47	AGE	21.81
NOI	33.78	LAW	18.29
DAY	26.60	TIM	18.25
LAW	20.91	DAY	13.59
MON	11.58	BAR	9.24
ROM	4.64	MON	5.06
OI	4.54	ROM	3.46
SHT	3.52	OI	3.15
BAR	2.02	SHT	2.23
SHW	0.77	SHW	0.46

Table 3. Normalized importance of the variables.

Table 3 shows the normalized importance of the variables in the model. Six variables were detected as having the greatest impact on accident severity with GInf, with percentages that vary from 100% to 61.21%. Five variables were detected with IGR, with percentages ranging from 100% to 51.96%. Both split criteria identify the same variables, although with different orders of importance: ACT, CAU, SEX, VEH, ATF. The variable LIG is also detected with GInf (with a percentage higher than 50%), whereas with IGR its percentage in the model is slightly lower (36.3%). However, it occupies sixth place in the importance ranking.

From the point of view of safety, these results are consistent with previous studies. Several authors (Kcoleman and Kweon, 2002; De Oña et al., 2011; De Oña et al., 2013a; 2013b) have pointed out that *accident type* is a key variable in severity. Chang and Wang (2006) stressed that the most important variable associated with crash severity was the *vehicle type*. *Causes of the accident* also match previous studies (Al-Ghamdi, 2002; Kashani and Mohyamany, 2011). Xie et al. (2009) and Mujalli and De Oña (2013) found that *atmospheric factors* have an important effect on severity. Many studies have also indicated gender differences in injury severity (Abel-Aty, 2003; Evans 2001; Obeng, 2011; Ulfarsson and Mannering 2004). *Lighting conditions* have been also identified as a variable with effects on severity. In fact, Gray et al. (2008), Abel-Aty (2003) and Helai et al. (2008) found that more severe injuries are predicted during darkness. Pande and Abel-Aty (2009) concluded that there is a significant correlation between lack of illumination and high crash severity. De Oña et al. (2011) and De Oña et al. (2013a) also pointed that KSI accidents are associated with roadways with no lighting.

In order to describe the pattern showed in the rules, only rules with the most severe consequent (accidents with killed or seriously injured, KSI) are extracted in Table 4. The IRNV method generates 4 KSI rules with GInf and 3 KSI rules with IGR (DT_1); and 36 KSI rules for GInf and 28 for IGR ($DT_{2\text{-}19}$). Due to the large number of rules obtained with each method, only rules with S>5% are extracted on Table 4. The support is a parameter that combines confidence and population. Therefore, a support higher than 5% implies that the rule is met by at least 63 accidents in the sample under study.

Num.	RULES (IF…)	THEN	S%	Po%	P%
1	NOI=[1];OCU=[1];VEH=MOT;ACT=ROR	KSI	7.62	10.79	70.59
2	**CAU=DC;VEH=MOT;ATF=GW;ACT=ROR**	**KSI**	**8.10**	**11.59**	**69.86**
3	SEX=M;ACT=ROR;CAU=DC;VEH=MOT	KSI	7.78	11.43	68.06
4	**ACT=ROR;CAU=DC;VEH=MOT;ATF=GW**	**KSI**	**8.10**	**11.59**	**69.86**
5	ATF=GW;SEX=M;ACT=ROR;VEH=MOT	KSI	8.81	12.86	68.52
6	LIG=WL;ATF=GW;SEX=M;LAW=THI	KSI	5.40	7.46	72.34
7	SID=WR;CAU=DC;VEH=MOT;BAR=N	KSI	7.06	11.67	60.54
8	TIM=[18-24];ATF=GW;LIG=WL;BAR=N	KSI	8.10	13.02	62.20
9	BAR=N;SEX=M;ACT=CP;ATF=GW	KSI	5.00	7.38	67.74
10	SHW=NE;SEX=M;ATF=GW;LIG=WL	KSI	7.06	11.35	62.24

Note: Ruels 1 and 2 have been obtained form the GInf criterion and rules 3-10 from the IGR criterion. In bold are the rules that are repeated in both methods.

Table 4. DRs from the IRNV method.

Table 4 shows the following patterns. Using the IRNV method, we identified two rules (rules 1 and 2) with GInf (and neither of them was obtained from DT_1); and seven rules (rules 3 to 10) with IGR (rule 3 was obtained from DT_1).

Rules 1 to 5 allow the identification of one of the most important concerns for road safety in Spain: run-off-road for motorcycles in two-lane rural highways (DGT, 2011). Precisely, one of the priorities of the Spanish Road Safety Strategy 2011-2020 (DGT, 2011) is to diminish this type of accidents, as well as their severity.

- Rule 1 identifies this kind of accident when only one occupant is involved (therefore, there is also only one injury). The probability of KSI in these cases is one of the highest (70.6%).
- Rules 2 and 4 are the same. Motorcyclists' run-off-road accidents under good weather conditions when the cause of the accident is due to the driver. The probability of KSI in these cases is 69.9%.
- Rule 3 identifies motorcyclists' run-off-road accidents for male drivers and due to driver characteristics. The probability of KSI is 68%.
- Rule 5 shows a similar pattern: motorcyclists' run-off-road accidents under good weather conditions when the driver is a male. The probability of KSI in these cases is 68.5%.

In this sense, the DGT is making an important effort to lower the number of accidents of this type (e.g. advertising campaigns that target motorcyclists; more stringent monitoring on two-lane rural highways; lowering the maximum speed limit on two-lane rural highways; etc.). The DGT also tries to lower the motorcycle crash severity (e.g. with projects that target improvements on the shoulders of two-lane rural highways that have no safety barriers). On the other hand, one of the priorities in the DGT's 2013-2016 Research Plan (DGT, 2011) is to identify the main factors that lead to accidents of this type (run-off-road for motorcycles on two-lane rural highways).

Table 4 shows that three rules (rules 7 to 9) identify KSI accidents on two-lane rural highways with no safety barriers:

- Rule 7 identifies motorcyclists' accidents with no-restrained sight distance due to the driver. Even if this rule does not present a very high probability (only 60.5%), it represents 11.7% of the population.
- Rule 8 identifies accidents in the evening (18-24 h) under good weather conditions on roads with no lighting. This rule presents the highest population (13.0%).
- Rule 9 identifies collision with pedestrian accidents under good weather conditions when the driver is a male.

These rules show that safety barriers play a fundamental role in crash severity on two-lane rural highways.

Finally, rules 6 and 10 share 3 variables: ACT, LIG and SEX. Thus, the pattern described for these rules refers to an accident on roads with no lighting, when atmospheric factors are good for male drivers. If the road has a lane width of < 3.25 m, rule 6 is obtained, whereas rule 10 is for roads where the shoulder non-existent or impassable. Thus, from the point of view of road safety, bad lighting conditions and bad road features increase accident severity.

4. CONCLUSIONS

If we use a single DT to extract knowledge based on a dataset, in the form of DRs, we are constrained by the DT's structure. However, the method proposed in this paper uses one DT for each variable under study (variables that describe the data), which allows us to extract much more knowledge. If we add that our model uses 2 split criteria, the extraction is even more extensive.

More than 70 significant validated rules were obtained from the practical study conducted on traffic accident data from rural roads in Granada (Spain). For the KSI rules, only one rule was repeated in both methods (rule 2 with rule 4); however some patterns were similar in both methods (rules 1 to 5). Although the criterion based on IGR detected a higher number of rules (with the set minimum parameters), it could be said that the two criteria complement each other when searching for the key factors that have an impact on accident severity, because each criterion detects different patterns within the same dataset.

With regards to the special patterns detected for the KSI accidents analysed, we could highlight the high number of rules for the motorcyclists' run-off-road accidents (rules 1 to 5). These results are in line with current concerns for road safety on two-lane rural highways. The Spanish Road Safety Strategy 2011-2020 (DGT, 2011) promotes specific studies on the factors associated with the highest levels of severity in run-off-road accidents on two-lane rural highways (i.e. KSI) when motorcyclists are involved.

Our study also highlights the need for studying the conditions in the environment of two-lane rural highways (i.e. safety barriers, shoulders, visibility, lighting, etc.), because they have a substantial impact on crash severity.

Finally, it should be pointed out that the proposed method can be extrapolated for specific studies on other datasets (i.e. other infrastructure, roads and countries). This method can also provide DRs that would be useful and easy for road safety analysts and managers to use to identify problems.

References

Abdel Wahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record, 1746, 6–13.

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research, 34, 597–603.

Abellán, J., Moral, S., 2003. Building classification trees using the total uncertainty criterion. International Journal of Intelligent Systems, 18 (12), 1215–1225.

Abellán, J., Masegosa, A., 2010. An ensemble method using credal decision trees. European Journal of Operational Research, 205 (1), 218–226.

Al-Ghamdi, A., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention, 34, 729–741.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman & Hall, Belmont, CA.

Chang L.Y. and Chien, J.T., 2013. Analysis of driver injury severity in truck involved accidents using a non-parametric classification tree model. Safety Science, 51, 17-22.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

De Oña, J., López, G., Abellán, J., 2013a. Extracting decision rules from police accident reports through decision tres. Accident Analysis and Prevention, 50, 1151–1160.

De Oña, J., López, G., Mujalli, R.O., Calvo, F.J., 2013b. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks, 51, 1-10.

De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention, 43, 402–411.

DGT, 2011. Spanish Road Safety Strategy 2011-2020. Traffic General Directorate, Madrid, 222p.

Evans, L., 2001. Female compared with male fatality risk from similar physical impacts. The Journal of Trauma: Injury, Infection and Critical Care, 50, 281–288.

Gray, R.C., Quddus, M.A., Evans, A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research, 39, 483–495.

Han, J., and Kamber, M., 2006. Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.

Helai, H., Chor, C.H., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention, 40, 45–54.

Kashani, A., Mohaymany, A., 2011. Analysis of the traffic injury severity on twolane, two-way rural roads based on classification tree models. Safety Science, 49, 1314–1320.

Kashani, A., Mohaymany, A., Ranjbari, A., 2011. A data mining approach to identify key factors of traffic injury severity. Promet-Traffic & Transportation, 23 (1), 11–17.

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention, 34, 313–321.

Kuhnert, P.M., Do, K.A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis, 34, 371–386.

Montella A., Aria M., D'Ambrosio A., Mauriello F., 2011. Data Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transportation Research Record, 2237, 107-116.

Montella A., Aria M., D'Ambrosio A., Mauriello F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, 49, 58-72.

Mujalli, R.O., de Oña, J., in press. Injury severity models for motorized vehicle accidents: a review, Proceedings of the Institution of Civil Engineering – Transport, http://dx.doi.org/10.1680/tran.11.00026, in press.

Mujalli, R.O., de Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research 42, 317–326.

Obeng, K., 2011. Gender differences in injury severity risks in crashes at signalized intersections. Accident Analysis and Prevention, 43 (4), 1521–1531.

Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A., 2010. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science, 3, 764–769.

Pande, A., Abdel-Aty, M., 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Safety Science, 47, 145–154.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention, 43, 1666–1676.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423 and 623–656.

Ulfarsson, G.F., Mannering, F.L., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accident Analysis and Prevention, 36, 135–147.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA.

Xie, Y., Zhang, Y., Liang, F., 2009. Crash injury severity analysis using Bayesian ordered Probit models. Journal of Transportation Engineering ASCE, 135 (1), 18–25.