Abellán, J.; López-Maldonado, G.; Garach, L.; Castellano, JG. (2017). Extraction of decision rules via imprecise probabilities. International Journal of General Systems. 46(4):313-331. https://doi.org/10.1080/03081079.2017.1312359

Extraction of decision rules via imprecise probabilities

Joaquín Abellán[1], Griselda López[2], Laura Garach[3] and Javier G. Castellano[1]

[1] Department of Computer Science and Artificial Intelligence, University of Granada, Spain
{jabellan,fjgc}@decsai.ugr.es

[2]Universidad Politécnica de Valencia
grilomal@tra.upv.es

[3]Department of Civil Engineering, University of Granada, Spain
lgarach@ugr.es

Abstract. Data analysis techniques can be applied to discover important relations among features. This is the main objective of the Information Root Node Variation (IRNV) technique, a new method to extract knowledge from data via decision trees. The decision trees used by the original method were built using classic split criteria. The performance of new split criteria based on imprecise probabilities and uncertainty measures, called credal split criteria, differs significantly from the performance obtained using the classic criteria. This paper extends the IRNV method using two credal split criteria: one based on a mathematical parametric model, and other one based on a non-parametric model. The performance of the method is analyzed using a case study of traffic accident data to identify patterns related to the severity of an accident. We found that a larger number of rules is generated, significantly supplementing the information obtained using the classic split criteria.

Keywords: imprecise probabilities; Imprecise Dirichlet model; Non-parametric Predictive Inference model; uncertainty measures; decision rules; traffic accident severity

1 Overview

Data mining research is concerned with the development of techniques for knowledge extraction from data. The extraction can be done using association rules (ARs) or decision rules (DRs). DRs are interesting and useful because they are easily understood. Since their appearance in Aggraval et al. (1993), many approaches have emerged for DRs extraction in the context of several real-life applications with the aim of using the rules to support decision making processes.

In order to extract all the knowledge available in a particular dataset via a set of DRs, the Information Root Node Variation (IRNV) was presented in Abellán et al. (2013). The main feature of this method is that it varies the root node to build different decision trees (DTs). The set of DRs obtained depends heavily on the set of split criteria used to build the trees. It is important to use as many different split criteria as possible to build different tree structures, in order to find more (and different) DRs. The useful rules extracted by different DTs could be used by road safety analysts to establish specific performance indicators.

It has been shown that the performance of the new split criterion based on imprecise probabilities and uncertainty measures, called *Imprecise Info-Gain* and denoted by IIG (Abellán and Moral, 2003), differs from the classic split criteria (Matas and Abellán, 2014). In particular, this method uses the *Imprecise Dirichlet model* (IDM) (Walley, 1991) to represent information from data. With this model, we obtain a convex and closed set of probability distributions (called credal set) from the data. At this point it is necessary to quantify the information or uncertainty contained in that type of set. The uncertainty can arise from different concepts (see Li and Yu (2016)), but in our case, where credal sets are obtained, the maximum entropy function emerges as the best measure (Abellán et al. (2006), Abellán and Bossé (2017)). This measure is applied in the Imprecise Info-Gain criterion.

More recently, the *Non-parametric Predictive Inference model* (NPIM) has been presented as a similar model based on imprecise probabilities. It has been proposed by Coolen and Augustin (2005). It is important to remark that it represents a non-parametric alternative to the IDM. It is shown in Abellán et al. (2014) that these two mathematical models (IDM and NPIM) can be used in a split criterion procedure to

build DTs. They have a similar performance for classification tasks, but since their origin (sense) is different, the generated trees are different. Hence, these methods generally return different sets of DRs extracted from DTs built from the same dataset.

Imprecision handling is a key part of the difference between the methods mentioned above and the classic split criteria. This study incorporates, as a logical consequence, several different split criteria based on imprecise probabilities into the IRNV method. The main aim of this extension is to increase the amount of information extracted from a dataset in order to find all the possible DRs.

Road safety is currently one of the top priorities for the government. Identifying and detecting accident patterns can provide new insights about the causes of road accidents. This valuable information can help governments in the implementation of road safety actions. A system that could help to extract knowledge from the information available would be very useful.

Accident patterns can be determined using data mining techniques such as Association Rules (ARs). In fact, ARs have been used in the road safety field to identify accident circumstances that frequently occur together (Geurts et al., 2003; Pande and Abdel-Aty., 2009; Montella et al., 2011; Montella et al., 2012). Decision Rules (DRs) extracted to Decision Trees (DTs) are another way to identify patterns. De Oña et al. (2013a) and Abellán et al. (2013) used them to study the severity of accidents.

Currently, DTs are widely applied in road safety research. One of the reasons for using DTs to analyze the severity of traffic accidents is that the structure of a DT facilitates the extraction of DRs. The IRNV method, based on a set of DTs, was applied on accident datasets in López et al. (2014), obtaining an interesting set of informative DRs. In this paper we will show that the information obtained using the IRNV method on accident datasets can improve significantly with new split criteria based on imprecise probabilities and uncertainty measures.

This paper is organized as follows: Section 2 is devoted to the required background knowledge and the data used; Section 3 describes the extended IRNV method, considering all the split criteria used; Section 4 explains and discusses a practical case study on data about traffic accidents, and comments the results obtained with the extended method. Finally, the last section sets out the conclusions.

2 Background Knowledge

2.1 SOME MODELS BASED ON IMPRECISE PROBABILITIES: CREDAL SETS AND REACHABLE PROBABILITY INTERVALS

All the theories based on imprecise probabilities (IP) share some common characteristics (see Klir (2006)). One of them is that the information is fully described by a *lower probability function* P_* on a finite set X (with $P(X)$ its power set), or alternatively, by an *upper probability function* P^*. These functions are always regular monotone measures (Wang and Klir, 1992), and satisfy

$$\sum_{x\in X} P_*(\{x\}) \leq 1, \quad \sum_{x\in X} P^*(\{x\}) \geq 1.$$

In the various special theories of IP they have additional special properties.
One of the most general models based on IP is the model of closed and convex sets K of probability distribution functions p on a finite set X (also known as *credal sets*). Here, functions P_* and P^* associated with K are determined for each $A \in P(X)$ by the formulas

$$P_*(A) = \inf_{p\in K} \sum_{x\in A} p(x), \; P^*(A) = \sup_{p\in K} \sum_{x\in A} p(x).$$

Abellán and Klir (2005) shows a relation of generality among diverse mathematical models, based on imprecise probabilities (belief functions, probability intervals, capacities of diverse orders, etc), being the

model based on credal sets one of the most general models. All of these models generalize the classic probability theory. One of them, based on reachable probability intervals which express a special type of credal set, is related to the models used in this work.

2.1.1 Reachable set of probability intervals: Cases of the IDM and the A-NPIM

In the theory of probability intervals, we have lower and upper probabilities P_* and P^* that are determined for all sets $A \in P(X)$ by intervals $[l(x), u(x)]$ of probabilities on singletons ($x \in X$). Here, $l(x) = P_*\{(x\}) \in [0, 1]$ and $u(x) = P^*(\{x\}) \in [0, 1]$. Each given set of probability intervals $I = \{[l(x), u(x)] | x \in X\}$ associated with a credal set $K(I)$ of probability distribution functions p is defined as follows:

$$K(I) = \{p(x) \mid x \in X, p(x) \in [l(x), u(x)], \sum_{x \in X} p(x) = 1\}.$$

A given set I of probability intervals may be such that some combinations of values taken from the intervals do not correspond to any probability distribution function. This indicates that the intervals are unnecessarily broad. To avoid this deficiency, the concept of reachability was introduced in the theory (De Campos et al., 1994).

A given set I is called *reachable* (or *feasible*) if and only if for each $x \in X$ and every value $v(x) \in [l(x), u(y)]$ there is a probability distribution function p for which $p(x) = v(x)$. The reachability of any given set I can be easily checked: the set is reachable if and only if it passes the following tests:

(a) $\sum_{x \in X} l(x) + u(y) - l(y) \leq 1$ for all $y \in X$, $y \neq x$;

(b) $\sum_{x \in X} u(x) + l(y) - u(y) \geq 1$ for all $y \in X$, $y \neq x$.

If I is not reachable, it can be easily converted to the set $I' = \{[l'(x), u'(y)] \mid x \in X\}$ of reachable intervals using the following formulas:

$$l'(x) = \max\{l(x), 1 - \sum_{y \neq x} u(y)\},$$

$$u'(\{x\}) = \min\{u(x), 1 - \sum_{y \neq x} l(y)\}$$

for all $x \in X$.

Given a reachable set I of probability intervals, the lower and upper probabilities are determined for each $A \in P(X)$ by the formulas

$$P_*(A) = \max\{\sum_{x \in A} l(x), 1 - \sum_{x \notin A} u(x)\},$$

$$P^*(A) = \min\{\sum_{x \in A} u(x), 1 - \sum_{x \notin A} l(x)\}.$$

A specific model based on reachable probability intervals is the **Imprecise Dirichlet mode**l (IDM) of Walley (1992). The IDM was introduced to make inferences about the probability distribution of a categorical variable. Assume that Z is a finite variable and that we have a sample of size N of independent and identically distributed outcomes of Z. If we want to make inferences about the probabilities, $h_z = p(z)$, from which Z takes its values, a common Bayesian procedure consists in assuming an *a priori* Dirichlet distribution for the parameter vector (h_x), and then taking the *a posteriori* expectation of the parameters given the sample. The final expression for the inference of the probabilities has the following expression:

$$h_z = \frac{n_z + t_z s}{N + s}, t_z \in [0,1], \sum_z t_z = 1$$

where $\{t_z\}$ is a set of parameters, n_z is the frequency of the set of values (Z=z) in the dataset, N the sample size and *s* a given hyperparameter. Walley does not give a definitive recommendation for *s*, but he advocates values between *s=1* and *s=2*. Now, to obtain minimum and maximum values of probability, the values 0 and 1 are considered for t_z. Therefore, the credal set associated via the IDM for a variable Z (with values belonging to $\{z_1,\ldots,z_k\}$) obtained from a dataset, can be expressed as the following K set of probability distributions p on Z:

$$\mathrm{K(Z)} = \left\{ p \mid p(z_j) \in \left[\frac{n_{z_j}}{N+s}, \frac{n_{z_j}+s}{N+s} \right], j = 1,\ldots,k \right\}$$

As we can see, the IDM for multinomial data is a parametric model which depends on a hyperparameter representing prior knowledge. This model has been applied to various statistical problems (see Bernard, 2005). As an alternative to the IDM, Coolen and Augustin (2005) presented the **Nonparametric Predictive Inference model** (NPIM), which differs from the IDM mainly in the sense that the NPIM learns from data in the absence of prior knowledge, and no unjustified assumptions are required.

Like the IDM, the NPIM can be used in a similar way to represent information from data. It requires considering a set of strong constraints and returns a set of probabilities that is not an actual credal set (see Abellán et al. (2011)). This characteristic of the model can be a serious drawback if we want to apply information measures on this type of set. To avoid this issue, Abellán et al. (2011) presented an approximate model called A-NPIM that considers the convex hull of the set of probabilities obtained using the NPIM and returns a credal set, that it is also a specific model of reachable probability intervals. It has been proved that the exact NPIM and A-NPIM have a similar performance (see Abellán et al. (2013)).

The above considerations suggest that A-NPIM is more appropriate for our aims. This model provides a set of probabilities that is a reachable set of probability intervals, similar to the set obtained using the IDM. Using A-NPIM, the set of probabilities obtained can be expressed in a way that it is similar to the set obtained via the IDM. It is denoted by K' in this case, and has the following expression:

$$\mathrm{K'(Z)} = \left\{ p \mid p(z_j) \in \left[\max\{0, \frac{n_{z_j}-1}{N}\}, \min\{\frac{n_{z_j}+1}{N}, 1\} \right], j = 1,\ldots,k \right\}$$

2.2 DECISION TREES AND SPLIT CRITERIA

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 values, the task is termed classification; otherwise, it is termed regression. In the present case the class variable is the severity of the accident, so classification trees are developed.

Within a DT, each node represents an attribute variable or feature (a characteristic of each item in the dataset) and each branch represents one of the values or states of this variable. A tree leaf specifies the expected value of the class variable. A split criterion (the criterion for branching) associates to each node the most informative variable which has not been selected already in the path from the root to this node. If the information about the class variable does not improve or there are no features left, a leaf node is added with the most probable class value for the partition of the dataset associated to that node. Finally, a DT can be interpreted as a compact set of rules about the class variable.

The split criterion is a key part of a procedure to build a DT. In the literature there are many works focused on the use of classic split criteria. The most used ones are the Information Gain (IG), the Information Gain Ratio (IGR), and the Gini Index (GInf). IG and IGR were presented in Quinlan (1986) and Quinlan (1993), respectively; and GInf in Breiman et al. (1984).

The Imprecise Info Gain (IIG), presented in Abellán and Moral (2003), has a different performance when compared to the classic criteria (see Matas and Abellán, 2014). It is based on the use of imprecise

probabilities and uncertainty measures. This criterion can be defined as follows: In a classification problem, let C be the class variable (the variable under study), $\{X_1,\ldots,X_m\}$ the set of features, and X a feature; then

$$IIG(C,X) = H^*(K(C)) - \sum_i P(X = x_i)\, H^*(K(C \mid X = x_i)),$$

where K(C) and $K(C|X=x_i)$ are the convex sets of probability distributions obtained via the Imprecise Dirichlet Model for the C and (C|X=xi) variables respectively (for more information, see Abellán and Moral (2003)), and function H*(K(Z)) is the maximum Shannon's entropy function of all the probability distributions that belong to set K. This measure is a well stablished measure of uncertainty on credal sets (Abellán et al., 2006). The H* value can be obtained using the algorithm described in Mantas and Abellán (2014).

Also, using the maximum entropy as a measure of uncertainty, we can express a split criterion similar to the criterion used by the IDM as follows:

$$IIG'(C,X) = H^*(K'(C)) - \sum_i P(X = x_i)\, H^*(K'(C \mid X = x_i))$$

where K'(C) and $K'(C|X=x_i)$ are the convex sets of probability distributions obtained via the A-NPIM for the C and (C|X=xi) variables respectively. Here, the H* value can be obtained using the algorithm described in Abellán et al. (2011).

For simplicity, and to avoid confusion, we will refer to this new credal split criterion (IIG') as NPIM. Since both criteria are obtained via credal sets, we can consider them as *credal split criteria*.

2.3 DATA DESCRIPTION

The data used in the case study shown here comes from the Directorate-General for Traffic (DGT). The study only consider accidents that occurred on two-lane rural highways in the province of Granada (Spain), as most accidents with injured people occur on these roads (74% of injury crashes (DGT, 2011a)).

The study period was seven years (2003-2009). A first check was performed to filter out unrealistic data. The accidents analyzed involved one vehicle and did not occur on intersections. The resulting database contains 1,801 accidents.

The class variable is the severity of the accidents (SEV). It was defined according to the level of injury for the worst injured occupant (following previous studies such as Chang and Wang (2006); De Oña et al. (2013); Kashani and Mohaymany (2011)). With the original classification of accidents by severity, there are 149 fatal, 723 serious and 929 minor. Since the different categories of the SEV variable are not balanced and this issue affects the overall accuracy of the model (Kashani and Mohaymany (2011)), the class variable was re-coded in two levels: SI - accidents with slightly injured people (929); and KSI - accidents with killed or seriously injured people (872).

Nineteen variables (see Table 1) were used for the class variable SEV in an attempt to identify the important patterns of an accident regarding its severity. The dataset includes variables describing the conditions that contributed to the accident, such as roadway information (safety barriers, pavement width, lane width, shoulder type, paved shoulder, road markings, and sight distance), environmental information (atmospheric factors and lighting) and accident information (causes, day, hour, month, occupants). It also includes variables describing the injury severity of the accident (number of injuries and severity level of injuries).

Table 1. Description of the variables in the dataset

NUM.	VARIABLE/CODE	VALUES/CODE	TOTAL	SEVERITY %SI	SEVERITY %KSI
1	Accident type: ACT	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 (carriage without 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	Atmospheric factors: ATF	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	Safety barriers: BAR	No: **N**	1740	48.3	54.7
		Yes: **Y**	61	53.6	46.4
5	Cause: CAU	Driver characteristics: **DC**	1471	48.99	51.01
		Combination of factors: **COF**	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 weekend or public holiday: **APH**	131	57.62	42.38
		Working day before 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	Lane width: LAW	< 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	Lighting: LIG	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
		No lighting (night-time): **WL**	543	43.1	56.9
9	Month: MON	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	Number of injuries: NOI	1 injury: **[1]**	1233	53.43	46.57
		> 1 injury: **[>1]**	568	47.35	52.65
11	Occupants involved: OI	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	Paved shoulder: SHT	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	Pavement width: PAW	[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	Pavement markings: ROM	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	Gender: SEX	Female: **F**	286	62.18	37.82
		Male: **M**	1513	49.61	50.39
		Unknown: **UN**	2	75	25
16	Shoulder type: SHW	< 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	Sight distance: SID	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	Time: TIM	[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	Vehicle type: VEH	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

The choice of the variables and their categorization were mainly guided by previous studies (see Chang and Wang, 2006; De Oña et al., 2011; Montella et al., 2011). Eleven variables (BAR, CAU, DAY, LAW, LIG, PAS, PAW, ROM, SEX, SHT, SID) were taken directly from the original dataset (provided by DGT). Seven variables (ATF, TIM, DAY, OI, NOI, ACT, VEH) were re-coded in a reduced number

of categories for ease of use. And the continuous variable AGE was transformed into a categorical variable in order to identify the relevant age group of the driver.

2.4 DECISION RULES AND PARAMETERS FOR THE ACCIDENTS DATASET

A Decision Rule (DR) can be seen as the result of the data mining process. An analyst could use DRs to make decisions or as an aid for decision-making. These rules can be obtained through different approaches (Sikora and Wróbel, 2014); such as those derived from rough set theory (Guo and Chankong, 2002); relational concept analysis (Dolques et al., 2016) or decision trees (Alkhalid et al., 2013), since they can be linearized automatically into DRs.

A DR has a logical-conditional structure of the type "IF (A) → THEN (B)", where A is the antecedent of the rule (in our case, a set of states of several attribute variables); and B is the consequent (in our case, a single state of the class variable).

The extraction of DRs from a DT is an easy procedure. Each rule starts at the root node, and each variable that intervenes in tree division is a part of the rule antecedent, which ends in leaf nodes as the consequent (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 being analyzed.

A priori, one rule is extracted from each terminal node of the tree. Later, in order to extract significant rules, specific parameters and thresholds are used (Abellán et al., 2013; De Oña et al., 2013a):

- *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%.

The Po, P, and S thresholds for the parameters are usually selected depending on the nature of the data (balanced or unbalanced), significant interest in fatal crashes (rare events), and sample size (small or large datasets). The thresholds have been established following previous studies (De Oña et al., 2013; López et al., 2014; Montella et al., 2012).

In the literature about association rules we can find that the Lift (L) parameter is extensively used. Lift relates the frequency of concurrence of the antecedent and the consequent to the expected frequency of concurrence under the assumption of conditional independence. If L<1, this indicates negative interdependence between antecedent and consequent; L=1 indicates independence; and L>1 indicates positive interdependence (i.e., the number of times that the sets of items occur together is greater than it would be if the antecedent and consequent were independent of each other). The higher the lift, the greater the strength of the association rule. We found that this parameter is very interesting for our purposes. Hence, in this paper Lift is also used, taking into account previous studies (López et al., 2014; Montella et al., 2012), and the minimum threshold established is L≥1.5.

3 The IRNV extended method

The procedure described in Abellán and Moral (2003) to build DTs can be explained as follows: Each node No in a DT produces a partition D of the dataset (for the root node, the entire dataset is considered). Also, each node No has an associated list "I" of feature labels (features which are not present in the path from the root node to No). A recursive and simple procedure to build a DT can be expressed by the algorithm shown in Figure 1.

Procedure Build Tree (No, Γ)

1. *If Γ =Φ, then* ***Exit***
2. *Let D be the partition associated with node No*
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. *Assign X_t to node No*
   9. *For each possible value x_t of X_t*
   10. *Add a node N_t*
   11. *Make N_t a child of No*
   12. *Call BuildTree (N_t , Γ)*

Fig. 1. Algorithm to build a DT.

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.

The Information Root Node Variation (IRNV) method used to extract DRs is based on the use of the different trees obtained through root node variation. In this method, if there are m features, and RX_i is the feature that occupies position i by an importance order (gain of information via a split criterion), then RX_i is used as the root to build DT_i (i=1,…,m). We use the simple method for building trees as explained above. However, the root node is now selected directly for each tree (the rest of the build 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 using the test set to obtain the final rule set. The entire procedure is carried out using the GInf and IGR criteria.

The process followed by this method can be explained via the following scheme:

1. Select a split criterion (SC) to build trees.
2. Build DTi using RXi as the root node, and SC (i=1,…,m).
3. Extract RSi from each DTi.
4. Check RSi using the corresponding TEST set → Selection of rules from RSi.
5. Extract the final rule set obtained using the SC.
6. Change the SC and go back to step 2.
7. Join the final rule sets using all the SCs.

In the original IRNV method, the classic split criteria GInf and IGR were used as SCs (Abellán et al., 2013). In this work, the imprecise split criteria IIG and NPIM have been incorporated into the IRNV method.

In Figure 2 we can see a scheme about the procedure used by the IRNV method to the extraction of DRs via a split criterion. For each feature it is built a DT using it as root node. The rest of process to build the tree is similar to the one presented in Figure 1. Hence, for each feature we have a tree, which give us a Rule Set (RS). All the RSs obtained from all the DTs are validated in the test set. Only the DRs that surpass the values of the thresholds (in the training set and in the test set) are considered in the Final Rule Set for a split criterion. If we repeat this procedure for each split criterion, we have the total set of DRs obtained by the IRNV method.

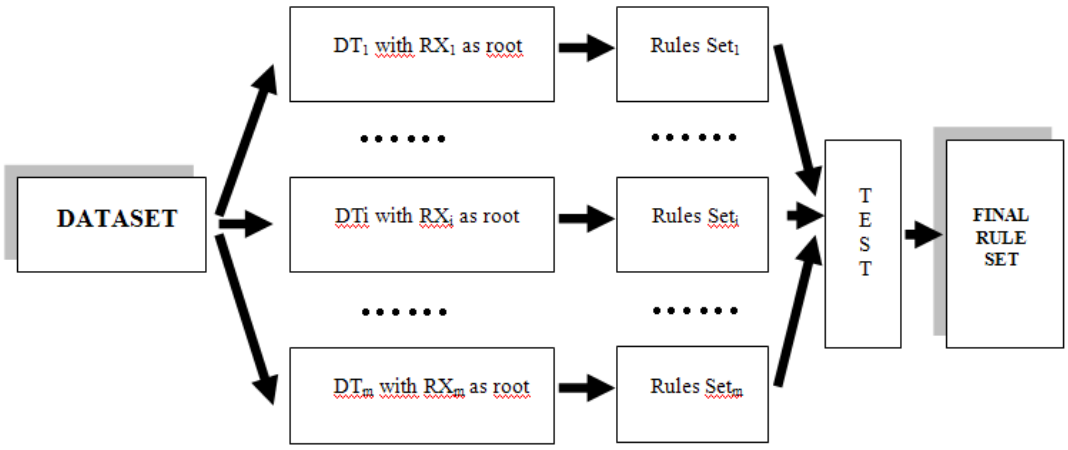

Fig. 2. Procedure to obtain validated DRs for each split criterion

4 Experiments and Results

Due to the large number of patterns considered, DTs have a very high risk of Type-1 errors, 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 Chien, 2013; De Oña et al., 2013a; Kashani and Mohaymany, 2011; Montella et al., 2011), the rules extracted on the training set with the minimum threshold values of the parameters are validated using the test set. Following previous studies (Chang and Wang, 2006; Kashani and Mohaymany, 2011), the dataset was randomly split into two different sets: A training set used to build the models, and a test set used to evaluate the rules obtained. The training set contains 70% of the data (1,260 accidents, with the following severity distribution: 51.3%-KSI and 48.7%-SI) and the test set contains the remaining 30%. Figure 2 we can also see that the final set of rules obtained must be validated to be considered in the Final set of rules.

Next, the IRNV method is applied using the training set. The different DTs obtained by root node variation were built using the Weka platform (Witten and Frank, 2005). The procedures for building the DTs based on Imprecise Info-Gain and the root node variation procedure were implemented using the method proposed in Abellán and Masegosa (2010). DTs were built with four levels of proof; previous studies such as Montella et al. (2012), and Abellán et al. (2013), use the same number of levels. This number of levels allows safety analysts to find useful and understandable rules.

Applying the IRNV method, different DTs are built by varying the root node. The first one (DT1) is the DT obtained when the IRNV method is not applied. This DT is built using the variable accident type (ACT) as the root node. Because the number of variables in the dataset is 19 (including ACT variable), 18 new trees are constructed. DTs are created using each of the remaining variables as the root node.

Table 2 provides the different DTs built using the different root nodes, as well as the number of rules obtained in each DT. The IRNV method was applied using the imprecise split criteria IIG and NPIM.

The main results are the following:

- The root node in DT1 is the ACT variable (using the IIG or NPIM criterion). This variable is also the root node when the classic split criteria (GInf and IIG) are used (see Abellán et al., 2014). In this tree (DT1), 12 rules can be extracted from the training set (see Table 2).

- 278 and 281 rules are obtained using IIG and NPIM respectively. In addition, for both algorithms TIM is the variable that generates the highest number of rules when used as the root node (24 and 28 rules respectively in Table 2).

- When the rules are validated using the test set, the number of rules decreases. When the IIG criterion is used, valid decision rules (DRs) can be obtained from all the generated trees. In contrast, when using the NPIM criterion, DT19 (root node: the paved shoulder variable) does not generate any valid DRs (in other words, rules that verify the minimum thresholds set for the S, Po, P, and Lift parameters).

- Using the IIG criterion, DT13 (root node: LAW variable) generates the highest number of valid rules (5). With the IGR criterion, DT11 (root node: DAY variable) generates the highest number of valid rules (6). In both cases, the number of validated rules exceeds the number of validated rules in DT1 (DT built when IRNV method is not applied), which only validates 3 rules using IIG and 4 rules using NPIM.

Table 2. Number of rules obtained with IRNV using the imprecise split criteria

	IRNV: IIG			**IRNV: NPIM**		
DTS	**R.N**	**R.T.**	**V.R.**	**R.N.**	**R.T.**	**V.R.**
DT_1	ACT	12	3	ACT	12	4
DT_2	LIG	19	1	SEX	15	3
DT_3	SEX	13	2	LIG	18	1
DT_4	CAU	15	1	CAU	17	2
DT_5	VEH	6	3	VEH	6	-
DT_6	ATF	9	1	ATF	14	3
DT_7	PAW	14	1	PAW	15	2
DT_8	TIM	24	4	TIM	28	4
DT_9	AGE	16	3	NOI	16	3
DT_{10}	NOI	13	3	AGE	19	3
DT_{11}	DAY	17	4	DAY	18	6
DT_{12}	SID	20	3	SID	18	3
DT_{13}	LAW	14	5	LAW	10	4
DT_{14}	MON	18	2	MON	19	3
DT_{15}	OI	19	2	OI	18	2
DT_{16}	ROM	13	3	SHW	11	3
DT_{17}	SHW	16	3	ROM	16	3
DT_{18}	BAR	9	2	BAR	11	2
DT_{19}	SHT	11	4	SHT	-	-
TOTAL		**278**	**50**		**281**	**51**

R.N. Root node; R.T. Rules from the Training set; V. R. Validated Rules

In Abellán et al. (2013), the IRNV method was applied with two different classic split criteria: GInf (based on the Gini Index) and IGR (Info Gain Ratio). Comparing the use of classic split criteria (Abellán et al., 2013) and imprecise split criteria (this paper) in the IRNV method, the following results are obtained regarding the number of generated rules:

- The IRNV method produced 227 rules using the GInf criterion and 174 rules using the IGR criterion (for the training set in both cases). Using imprecise criteria, a larger set of rules is obtained: 278 rules using IIG and 281 rules using NPIM (for the training set; see Table 2).

- Analyzing the results for the test set with classic criteria, 78 rules were validated rules using Ginf, and 81 rules using IGR, as can be seen in Abellán et al., 2013. In this paper the Lift parameter was not used for the validation procedure. If this parameter is used, the number of validated rules decreases to 30 (GInf) and 23 (IGR). Nevertheless, the imprecise criteria generate a larger number of validated rules in the test set, 50 using IIG and 51 using NPIM.

The key results for parameters in the significant rules are the following:

- Rules extracted using the classic and imprecise split criteria present probability ranges between 60% and 100%.

- The rules with the highest support (4.92%) value are obtained for the imprecise split criteria. Regarding the population parameter, the larger values are also obtained for the imprecise split criteria (with a value of 6.35% for IIG and NPIM). Finally, the rules with the highest Lift (2.1) value are obtained using the imprecise split criterion IIG and the classic split criterion GInf.

- Regarding the mean values for each parameter (Table 3), the imprecise split criteria achieve the highest values for all parameters. In particular, IIG obtains the larger population value (2.04%, very similar to the value obtained using NPIM, 2.02%) and the larger support value (1.73%, again, very similar to the value obtained with NPIM, 1.72%). The NPIM split criterion generates the larger probability value (a mean of 87.2%) and the larger Lift (1.75).

Table 3. Mean parameter values for the different split criteria

Validated Rules	Po(%)	S(%)	P(%)	Lift
IIG: 51	**2.04**	**1.73**	86.37	1.72
NPIM: 50	2.02	1.72	**87.16**	**1.75**
Ginf: 30	1.81	1.53	86.09	1.71
IGR: 23	1.97	1.63	84.38	1.70

Table 4 shows the number of rules shared between the different split criteria. The number of rules according to the severity of the accident (SI or KSI) is also indicated. NPIM and IIG are the split criteria that share the greatest number of rules (36 rules in Table 4). The number of rules identified using IIG that are also identified using GInf is lower (13 rules), and with IGR only 6 rules are shared. The number of rules identified using NPIM that are also identified with the classic split criteria is 12, and 1 for GInf and IGR respectively.

Table 4. Number of rules shared by the different split criteria

Validated Rules	Split Criteria			
	Imprecise		Classic	
	IIG	NPIM	GInf	IGR
IIG: 50 (22 SI/28 KSI)	-	36 (17 SI/19 KSI)	12 (6 SI/ 6 KSI)	6 (5 SI/1KSI)
NPIM: 51 (27 SI/24 KSI)	-	-	13 (8 SI/ 5 KSI)	1 (1 SI /0 KSI)
GInf: 30 (12 SI/18 KSI)	-	-	-	3 (2 SI/1 KSI)
IGR: 23 15 SI/8 KSI)	-	-	-	-

So, using these new split criteria in the IRNV method, new and interesting information has been obtained from the same dataset.

Regarding the variables identified using the different split criteria, more variables were identified in the rules using NPIM and IIG than using only the GInf or IGR criteria. For example, the following new variable statuses in the rules are identified when imprecise split criteria are used: ROM=SLO (this status appears when the IIG criterion is applied); ROM=DMR (appears with IIG and NPIM).

4.1 SAFETY RULES DESCRIPTION

KSI accidents are rare events and they are the most important with regard to road safety. For this reason, only severe and mortal accidents (KSI rules) are considered in the safety analysis. Therefore, 28 KSI rules obtained using the IIG criterion (of the 50 KSI/SI rules) and 24 rules obtained using the NPIM criterion (of the 51 KSI/SI rules) are analyzed.

Table 5. KSI rules using imprecise criteria.

Num	Split criterion	Antecedent of the rules								Po(%)	S(%)	P(%)	Lift
1	IIG-NPIM	**ACT=ROR**	;	**VEH=MOT**	;	**SHT=Y**	;	**OI=[1]**		5,56	4,37	78,57	1,53
2	IIG-NPIM	ACT=ROR	;	VEH=MOT	;	SHT=Y	;	NOI=[1]		5,56	4,37	78,57	1,53
3	IIG-NPIM	ACT=ROR	;	VEH=MOT	;	SHT=Y	;	LIG=WL		2,30	2,06	89,66	1,75
4	IIG-NPIM	**ACT=ROR**	;	**VEH=MOT**	;	**LIG=WL**	;	**SHW=THI**		1,83	1,59	86,96	1,70
5	IIG-NPIM	ACT=ROR	;	VEH=MOT	;	LIG=WL	;	LAW=MED		2,30	2,06	89,66	1,75
6	IIG-NPIM	ACT=CP	;	VEH=CAR	;	LAW=MED	;	SHT=Y		3,65	3,17	86,96	1,70
7	IIG	**ACT=CP**	;	**VEH=CAR**	;	**LAW=MED**	;	**MON=WIN**		1,51	1,27	84,21	1,64
8	IIG-NPIM	ACT=CP	;	VEH=CAR	;	MON=AUT	;	NOI=[1]		1,59	1,59	100,00	1,95
9	IIG-NPIM	**ACT=CP**	;	**VEH=CAR**	;	**MON=AUT**	;	**OI=[1]**		1,59	1,59	100,00	1,95
10	IIG	ACT=CP	;	VEH=CAR	;	MON=WIN	;	ROM=SLD		1,75	1,35	77,27	1,51
11	NPIM	**ACT=CP**	;	**VEH=CAR**	;	**SEX=M**	;	**SHT=N**		**1,03**	**0,87**	**84,62**	**1,65**
12	IIG-NPIM	ACT=CP	;	VEH=CAR	;	ROM=SLD	;	SHW=THI		3,02	2,62	86,84	1,69
13	IIG	*ACT=CP*	;	*VEH=CAR*	;	*MON=AUT*	;			1,59	1,59	100,00	1,95
14	IIG-NPIM	ACT=CP	;	SHT=Y	;	ATF=GW	;	DAY=WD		2,78	2,38	85,71	1,67
15	IIG	ACT=CP	;	SHT=Y	;	AGE=(20-27]	;			1,03	0,95	92,31	1,80
16	IIG-NPIM	LIG=WL	;	ATF=GW	;	DAY=PH	;	SID=TOP		2,22	2,06	92,86	1,81
17	NPIM	LIG=WL	;	ATF=GW	;	SEX=F	;	ROM=SLD		1,51	1,27	84,21	1,64
18	IIG	LIG=WL	;	ATF=GW	;	SEX=F	;	SHT=Y		1,51	1,27	84,21	1,64
19	IIG-NPIM	LIG=WL	;	TIM=(18-24]	;	PAW=WID	;	MON=WIN		2,54	2,22	87,50	1,71
20	IIG-NPIM	LIG=WL	;	TIM=(18-24]	;	SHT=NE	;	PAW=THI		1,27	1,19	93,75	1,83
21	IIG-NPIM	LIG=WL	;	TIM=[0-6]	;	SID=WR	;	DAY=BPH		1,19	1,11	93,33	1,82
22	IIG	LIG=WL	;	LAW=THI	;	SEX=M	;	MON=WIN		3,49	2,70	77,27	1,51
23	IIG-NPIM	LIG=WL	;	LAW=THI	;	SEX=M	;	SID=WR		6,35	4,92	77,50	1,51
24	IIG-NPIM	**LIG=DAY**	;	**LAW=THI**	;	**SEX=M**	;	**AGE=<=20**		1,59	1,35	85,00	1,66
25	NPIM	SID=TOP	;	SHW=MED	;		;	ACT=ROR		1,98	1,75	88,00	1,72
26	NPIM	SID=TOP	;	LAW=MED	;	AGE=<=20	;	SEX=M		1,27	1,27	100,00	1,95
27	IIG-NPIM	SID=TOP	;	MON=WIN	;	AGE=(20-27]	;	CAU=DC		1,98	1,75	88,00	1,72
28	IIG	**SHT=NE**	;	**TIM=(18-24]**	;	**CAU=COF**	;	**PAW=MED**		1,03	0,79	76,92	1,50
29	IIG	SHT=NE	;	TIM=(12-18]	;	DAY=WD	;	SEX=M		3,10	2,54	82,05	1,60
30	NPIM	NOI=[>1]	;	ATF=GW	;	PAW=WID	;	SHT=NE		1,83	1,67	91,30	1,78
31	IIG	ACT=ROR	;	VEH=CAR	;	PAW=MED	;	ROM=SLO		2,06	1,59	76,92	1,50
32	IIG-NPIM	CAU=DC	;	DAY=WD	;	MON=SPR	;	VEH=MOT		2,78	2,14	77,14	1,50
33	IIG-NPIM	LAW=THI	;	MON=SUM	;	ROM=DMR	;	SEX=M		1,51	1,27	84,21	1,64

Note: **Rules shared with GInf**; ***Rules shared with IGR***.

Table 5 shows the KSI rules obtained using only the IIG criterion (IIG code in the second column), the rules obtained using only the NPIM criterion (NPIM code in the second column), and the rules obtained simultaneously with both split criteria (IIG-NPIM code in the second column). Most of the rules identified using NPIM are also identified using IIG. The analysis of shared rules using the imprecise or classic criteria shows that only one rule is identified using the IGR split criterion (shown in italics and greyed out in Table 5) is also identified using the imprecise criteria IIG and NPIM, and 7 rules identified using the GInf criterion (in bold in Table 5) are also identified using IIG and NPIM. These results show that the NPIM and IIG split criteria are similar; however, these new split criteria implemented in the IRNV method provide new and interesting information from the same dataset.

In Table 5, the rules have been grouped in five sets to show their common patterns. In the first group the common pattern was run-off-road accidents involving motorcycles; in the second group, the common pattern was the collision with a pedestrian; in the third group, lighting condition equal to night time with no illumination; in the fourth group, accidents in which the visibility was restrained by topography; and the fifth group shows miscellaneous patterns.

According to Table 5, important road safety patterns are simultaneously identified by the classic and imprecise criteria (bold rules in Table 5). Rule 1 (shared by the imprecise criteria and the classic GInf criterion) shows that run-off-road accidents involving motorcycles usually occur on roads with paved shoulder. Rules 2 and 3 (identified only by the imprecise criteria) show a similar pattern. In this pattern, the driver perceives a wider road and increases speed, thus increasing the probability of having an accident. Another pattern of run-off-road accidents involving a motorcycle is identified by classic and imprecise criteria (rule 4 in Table 5). This pattern occurs when the luminosity is insufficient and the shoulder is nar-

row. It has been pointed out that the number of accidents involving motorcycles at night could be reduced by wearing retro-reflective material on the clothes and helmets (López et al., 2014).

Classic and imprecise criteria also identify pedestrian accidents in which the vehicle involved is a car. In particular, some patterns are related to bad weather during autumn (rules 8, 9 and 13) or winter (rules 7 and 10). In addition, rule 11 identifies pedestrian accidents involving a car driven by a man on roads without paved shoulder. These results are in agreement with López et al. (2014). Imprecise criteria have shown pedestrian accidents involving a car where the shoulder type variable has an impact. These patterns suggest that providing protection for pedestrians in rural environments would be an effective safety countermeasure.

Many of the identified patterns involve lighting conditions. They indicate that accidents occur with no light or with insufficient light, and with good weather conditions (rules 16, 17, and 18). Lighting conditions have been also identified as a variable with an impact on the severity (Gray et al., 2008; Helai et al., 2008). Rules 17 and 18 associate accidents with lighting conditions and women. This fact has already been identified in other studies (De Oña et al., 2011). This lack of light is identified in logical slots, between 18 and 24 hours and between 0 and 6 hours (rules 19, 20, and 21). Two additional patterns are related to insufficient light: narrow lane and male drivers (rules 22 and 23).

In addition, the imprecise criteria add new and interesting patterns that had not been identified using the classic criteria. These patterns are related to sight distance restrained by topography and young drivers (rules 26 and 27) or have run-off-road as accident type (rule 25). Young people aged between 20 and 27 are involved in pedestrian accidents in roads with narrow shoulder width (rule 15). Teenage drivers may be differentially vulnerable to crashes on rural roads because of their inexperience and lack of maturity compared with older and more experienced drivers (Mayhew et al., 2003; McCartt et al., 2009; Peek-Asa et al., 2010). From a safety perspective, the vertical signal (when sight distance is restrained by topography) should be improved.

The patterns identified using the new split criteria (IIG and NPIM) show the current safety problems of the road analyzed. In fact, some of them have been pointed out as a priority in the Spanish Road Safety Strategy 2011-2020 (DGT, 2011).

Conclusions

This paper presents an extension of a method called IRNV that uses Decision Trees (DTs) to obtain Decision Rules (DRs). The original IRNV method was used with classic split criteria, such as GInf (based on the Gini Index) and IGR (the Info-Gain Ratio). Two new split criteria based on imprecise probabilities have been implemented in the IRNV method: one based on a parametric mathematical model, called IIG; and another one based on a non-parametric mathematical model, called NPIM. Both use a different treatment of the data and build different DTs to represent different sets of DRs. These new split criteria represent an interesting addition to the classic criteria when all of them are applied in the IRNV method.

The extended IRNV method has been applied on data about traffic accidents. A greater number of rules are obtained using the new imprecise split criteria (IIG and NPIM) implemented in the IRNV procedure. The IIG and the NPIM criteria generate 278 and 281 rules respectively, compared to 227 rules using GInf and 174 using IGR. Also, with the imprecise criteria a greater number of rules were validated (50 and 51 respectively), compared to 30 and 23 using GInf and IGR respectively. Regarding safety, more knowledge was obtained about the road analyzed. In addition, the rules obtained using the imprecise criteria achieve the highest parameters values for population and support using IIG, and for probability and Lift using NPIM.

The imprecise criteria represent an increment in the knowledge extraction for the IRNV method. In the case study carried out about accident analysis, they add new and interesting patterns of accidents with people killed or seriously injured that had not been identified using the classic criteria. Patterns are related to sight distance restrained by topography, involving young drivers, or having run-off-road as accident type. Theses pattern remark the need for studying the conditions in the environment of two-lane rural highways (i.e. vertical or horizontal signs).

Acknowledgements

This work has been supported by the Spanish “Ministerio de Economía y Competitividad" under Project TEC2015-69496-R and FEDER funds.

This paper is an improved and notably extended version of a paper presented at the *9th International Conference on Rough Sets and Current Trends in Computing* (López et al., 2014).

References

1. Abellán, J., Moral, S.: Building classification trees using the total uncertainty criterion. International Journal of Intelligent Systems, 18(12), 1215-1225 (2003)
2. J. Abellán, J., Klir, G.J: Additivity of uncertainty measures on credal sets. International Journal of General Systems, 34(6), 691–713 (2005)
3. J. Abellán, J., Klir, G.J, and Moral, S: Disaggregated total uncertainty measure for credal sets, International Journal of General Systems, 35(1), 29-44 (2006)
4. Abellán, J., Masegosa, A.: An ensemble method using credal decision trees. European Journal of Operational Research 205 (1), 218–226 (2010)
5. Abellán, J., Masegosa, A.: Requirements for total uncertainty measures in Dempster–Shafer theory of evidence, International Journal of General Systems, 37(6), 733-747 (2008)
6. Abellán, J., Baker, R.M. and Coolen, F.P.A.: Maximising entropy on the nonparametric predictive inference model for multinomial data, European Journal of Operational Research, 212(1), 112-122 (2011)
7. Abellán, J., De Oña, J., López, G.: Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Application, 40, 6047–6054 (2013)
8. Abellán, J., Baker, R.M., Coolen, F.P.A., Crossman, R. and Masegosa, A.: Classification with decision trees from a nonparametric predictive inference perspective, Computational Statistics and Data Analysis, 71, 789-802 (2014)
9. Abellán, J. and Bossé, E., Drawbacks of uncertainty measures based on the pignistic transformation, IEEE Trans. On Systems, Man and Cybernetics: Systems, in press (2017). DOI: 10.1109/TSMC.2016.2597267
10. Agrawal R, Imeilinski T and Swami A. Mining association rules between sets of items in large databases. In: *Proceeding of ACM SIGMOD Conference on Management of Data*.Washington, DC: ACM, 207-216, 1993.
11. Alkhalid, A., Amin, T., Chikalov, I., Hussain, S., Moshkov, M., Zielosko, B.: Optimization and analysis of decision trees and rules: dynamic programming approach. International Journal of General Systems 42(6), 614-634 (2013)
12. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Chapman & Hall, Belmont, CA (1984)
13. Chang L.Y., Chien, J.T.: Analysis of driver injury severity in truck involved accidents using a non-parametric classification tree model. Safety Science, 51, 17-22 (2013)
14. Chang, L.Y., Wang, H.W.: Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027 (2006)
15. F.P.A. Coolen and T. Augustin: Learning from multinomial data: a nonparametric predictive alternative to the imprecise Dirichlet model, in: F.G. Cozman, R. Nau, T. Seidenfeld (Eds.), ISIPTA'05, Proceedings of the 4th International Symposium on Imprecise Probabilities and Their Applications, Pittsburg, Pennsylvania, 125-134 (2005)
16. De Campos, L. M., Huete, J. F. and Moral, S. Probability intervals: A tool for uncertain reasoning. Intern. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2(2), 167-196 (1994)
17. De Oña, J., López, G., Abellán, J.: Extracting decision rules from police accident reports through decision tres. Accident Analysis and Prevention, 50, 1151–1160 (2013a)
18. De Oña, J., López, G., Mujalli, R.O., Calvo, F.J.: Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accident Analysis and Prevention, 51, 1-10 (2013b)
19. DGT, 2011a. *Las principales cifras de la siniestralidad vial. España 2010*. Directorate-General for Traffic, Madrid. Available at: http://www.dgt.es/was6/portal/contenidos/es/seguridad_vial/estadistica/publicaciones/princip_cifras_siniestral/cifras_siniestralidadl013.pdf
20. DGT, 2011b. Spanish Road Safety Strategy 2011-2020. Traffic General Directorate, Madrid, 222p.

21. Dolques, X., Le Ber, F., Huchard, M., Grac. C.: Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. International Journal of General Systems 45(2), 187-210 (2016)
22. Geurts, K., Wets, G., Brijs, T., Vanhoof, K. Profiling high frequency accident locations using association rules. In: Proceedings Transportation Research Board (CD-ROM), Washington, DC, USA, January 12-16 (2003)
23. Gray, R. C., Quddus, M. A., Evans, A. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research, 39, 483–495 (2008)
24. Guo J., Chankong V.: Rough set-based approach to rule generation and rule induction. International Journal of General Systems 31(6), 601-617 (2002)
25. Helai, H., Chor, C. H., Haque, M. M. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention, 40, 45–54 (2008)
26. Kashani, A., Mohaymany, A.: Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science, Vol. 49, 1314–1320 (2011)
27. Li, X. and Yu, L.: Decision making under various types of uncertainty. International Journal of General Systems 45(3), 251-252 (2016)
28. López G., Garach, L., Abellán, J., Castellan, J.G., Mantas, C. J.: Using Imprecise Probabilities to Extract Decision Rules via Decision Trees for Analysis of Traffic Accidents. 9th International Conference on Rough Sets and Current Trends in Computing. Springer International Publishing Switzerland , 288-298 (2014).
29. Mantas, C. J., Abellán, J.: Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data. Expert Systems with Applications, 41, 2514–2525 (2014)
30. Mayhew, D. R., Simpson, H. M., Pak, A. Changes in collision rates among novice drivers during the first months of driving. Accident Analysis and Prevention 25, 683−691 (2003)
31. McCartt, A. T., Mayhew, D. R., Braitman, K. A., Ferguson, S. A., Simpson, H. M.: Effects of age and experience on young driver crashes: review of recent literature. Traffic Injury Prevention 10(3), 209−219 (2009)
32. Montella, A., Aria M., D'Ambrosio A., Mauriello F.: Data Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transportation Research Record, 2237, 107-116 (2011)
33. Montella A., Aria M., D'Ambrosio A., Mauriello F.: Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, 49, 58-72 (2012)
34. Pande, A., Abdel-Aty, M. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Safety Science, 47, 145–154 (2009)
35. Peek-Asa, C., Britton, C., Young, T., Pawlovich, M., Falb, S.: Teenage driver crash incidence and factors influencing crash injury by rurality. Journal of Safety Research 41(6), 487–492 (2010)
36. Quinlan, J. R.: Induction of decision trees. Machine Learning, 1, 81–106 (1986)
37. Sikora. M., Wróbel, Ł.: Data-driven adaptive selection of rule quality measures for improving rule induction and filtration algorithms. International Journal of General Systems 42(6), 594-613 (2013)
38. Quinlan, J. R..: Programs for machine learning. Morgan Kaufmann series in Machine Learning (1993)
39. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
40. Wang, Z. and Klir, G. J. Fuzzy Measure Theory. Plenum Press, New York, 1992.
41. Webb, G.I.: Discovering significant patterns. Machine Learning 68, 1–33 (2007)
42. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA (2005)