Eur. Transp. Res. Rev. (2017) 9:40
DOI 10.1007/s12544-017-0257-5

ORIGINAL PAPER

An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety

Addi Ait-Mlouk[1] · **Fatima Gharnati**[1] · **Tarik Agouti**[1]

Received: 2 December 2016 / Accepted: 18 July 2017

Abstract
Purpose Road accidents have come to be considered a major public health problem worldwide. The aim of many studies is therefore to identify the main factors contributing to the severity of crashes.
Methods This paper examines a large-scale data mining technique known as association rule mining, which can predict future accidents in advance and allow drivers to avoid the dangers. However, this technique produces a very large number of decision rules, preventing decision makers from making their own selection of the most relevant rules. In this context, the integration of a multi-criteria decision analysis approach would be particularly useful for decision makers affected by the redundancy of the extracted rules.
Conclusion An analysis of road accidents in the province of Marrakech (Morocco) between 2004 and 2014 shows that the proposed approach serves this purpose; it may provide meaningful information that could help in developing suitable prevention policies to improve road safety.

Keywords Data mining · Association rules · Road accident · Quality measurements · Multi-criteria decision analysis

✉ Addi Ait-Mlouk
aitmlouk@gmail.com

Fatima Gharnati
gharnati@uca.ma

Tarik Agouti
t.agouti@uca.ma

[1] Laboratory of Intelligent Energy Management and Information Systems, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco

1 Introduction

Data mining is defined as a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1]. Indeed, it is a vital part of business analytics and the most important trends in information technology. It involves many common classes of tasks (clustering, classification, association rules [2] etc.) which are designed for knowledge discovery in databases (KDD).

Data mining techniques are widely used in several research domains and have provided useful results to guide decision makers. Many researchers [3–7] have studied the application of data mining techniques in the domain of road accidents through association rules mining. The association rule is a powerful data mining technique for discovering a correlation between variables in the database. It is based on statistical analysis and artificial intelligence. This technique is particularly appropriate for studying road accident data by considering conditional interactions between input datasets, extracting frequent itemsets and then generating the association rules by satisfying certain parameters such as the minimum support and the minimum confidence. In this paper, the goal of the proposed approach is not to optimize road safety, but to generate insights and sufficient knowledge to enable decision makers to make the right optimization decision to avoid dangerous routes and improve road safety. This approach consists of two major steps; a rules generator using the Apriori algorithm to extract association rules, and multi-criteria decision analysis to evaluate and select the interesting rules from the large set extracted.

The rest of the paper is organized as follows: Section 2 describes the related work of data mining and machine learning techniques for accident analysis, while Section 3 describes the proposed methodology for extracting association rules and the integration of multi-criteria decision analysis approach

Published online: 27 July 2017

within the KDD process. Section 4 presents the results and a discussion of these. In the last section, we conclude by summarizing the work done in the study and describe the contributions of this work.

2 Related work

According to the World Health Organization (WHO) [8], 1.24 million people die each year on the world's roads, and as many as 50 million are injured. In addition, the Centers for Disease Control and Prevention (CDCP) have announced that road accidents cost 100 billion in medical care every year. Furthermore, the Ministry of Equipment, Transport and Logistics of Morocco [9] gives the statistics of road accidents between 2004 and 2014, as shown in Table 1. Road accidents involve not only loss of human life but also property damage.

As a review of the literature shows, many data mining techniques have been proposed to analyze road accidents. In this context, Kuhnert et al. used CART and MARS to analyze an epidemiological case-control study of injuries resulting from motor vehicle accidents. They also identified potential areas of risk, largely caused by the driver situation [10]. Ossenbruggen et al. [3] used logistic regression models to analyze the factors involved in accidents, and found that shopping areas were more dangerous than village sites. Sohn et al. [11] used the three data mining techniques of decision trees, neural networks and logistic regression to discover significant factors affecting the severity of Korean road traffic. Subsequently, Mio et al. [12] used a decision tree to analyze the severity of traffic accidents. They found that fatal injury was caused by many factors, among them seat belts, alcohol, and lighting conditions.

Chang and Wong [13] developed a CART model to analyze the relationship between drivers, severity of injury and the highway environment. Sze and Wong [14] used binary logistic regression and logistic regression diagnostics to control for the influences of demographics and the road environment. In addition, Abugessaisa [15] used clustering and classification trees to carry out interactive explorations based on brushing and linking methods in order to detect and recognize interesting patterns. Moreover, Wong and Chang [16] used several methodologies to discover factors involved in the severity of accidents, and found that a dangerous accident was caused by a combination of different factors. Anderson [17] studied the spatial patterns of road accident injury and used the resultant patterns to create a classification system for road accident hotspots. Zelalem [18] studied driver responsibility using the ID3, J48, and multilayer perceptron (MLP) algorithms to discover the related factors, and found that many factors have a direct impact on the severity of accidents, such as license grades and the driver's age and experience. Pakgohar et al. [19] used CART and multinomial logistic regression (MLR) to explore the roles played by the characteristics of drivers, and found that the CART method provided relatively precise results. Demirel et al. [20] used remote sensing for regional scale analysis and effective management of environmental factors. They concluded that this technology could be useful in the prevention of some type of accidents. Wu et al. [21] used the global positioning system (GPS) in the prevention of collision accidents. Zhang et al. [22] concluded that the lack of use of seat belts and inadequate training were also two important factors. Sanmiquel [5] analyzed the main causes of accidents using Bayesian classifiers and a decision tree.

Other association rule mining algorithms have been widely used in the literature to extract frequent itemsets and build decision rules. These algorithms are based primarily on minimum support and the minimum confidence. However, most of them produce a large number of results, which prevents decision makers from making their own selection of the most relevant ones. It is therefore important to propose an approach that can help decision makers to make their choice. Multi-criteria decision analysis (MCDA) offers a powerful solution; its advantages include taking into account the decision makers' preferences and a diversity of criteria. This paper proposes an approach to association rule mining-based MCDA for analyzing road accident data.

3 Proposed methodology

In this section, we discuss the various steps used in the construction of our proposed methodology. We start by developing the association rule mining, as described below.

Table 1 Road accident statistics in Morocco between 2004 and 2014

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Death	3894	3617	3754	3838	4162	4042	3778	4222	1351	2632	2214
Injuries	80,150	77,264	82,651	89,264	98,907	102,743	98,472	102,011	102,011	61,207	28,150

Table 2 Example of dataset with five transactions

ID	Milk	Bread	Diapers	Beer	Cola	Eggs
1	1	1	0	0	0	0
2	0	1	1	1	1	0
3	1	0	1	1	1	0
4	1	1	1	1	0	0
5	1	1	1	0	1	0

3.1 Association rule mining

The association rules technique is a powerful data mining method for discovering the relationship between variables in large databases. It was proposed by Agrawal [2] for analyzing transactional databases. It is defined as follows: let $I = \{i_1, i_2 \ldots i_n\}$ denote the set of $\boldsymbol{n}$ binary items, and let $D = \{t_1, t_2 \ldots t_m\}$ denote the set of transactions. Each transaction in $\boldsymbol{D}$ has a unique $\boldsymbol{Id}$ and contains a subset of items in $\boldsymbol{I}$. The details are given in Table 2.

An association rule is defined as an implication of the form $A \rightarrow B$ such that $A, B \subset I$ and $A \cap B = \phi$. Each rule is composed of two different sets of items, A and B, where A is called the antecedent and B the consequent. To extract association rules, two measures are required: the support and the confidence. The support is defined as the proportion of transactions in the database which contain the items A. The formal definition is (1):

$$Supp(A \rightarrow B) = Supp(A \cup B) = \frac{|t(A \cup B)|}{t(A)} \quad (1)$$

The confidence determines how frequently items in B appear in a transaction that contains A. The formal definition is (2):

$$Confidence(A \rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)} \quad (2)$$

An initial step towards improving association rules algorithms is to decompose the problem into two main steps. The first is to find all itemsets that satisfy the minimum support; this step is generally expensive, due to the requirement for multiple passes over the database (see Fig. 1).

The second step is the generation of association rules. This step is responsible for extracting all high-confidence rules from the frequent itemsets found in the previous step. The association rules technique has led to significant gains in other areas and can also be used to improve the transportation sector.

3.2 Multi-criterion decision analysis

Keeney and Raiffa's [23] seminal book on MCDA defines this as "an extension of decision theory that covers any decision with multiple objectives. A methodology for appraising alternatives on the individual, often conflicting, criteria, and combining them into one overall appraisal". Roy [24] distinguishes three types of problematic: choice, sorting and ranking (see Fig. 2). Due to the large number of extracted association rules, we are interested in the multi-

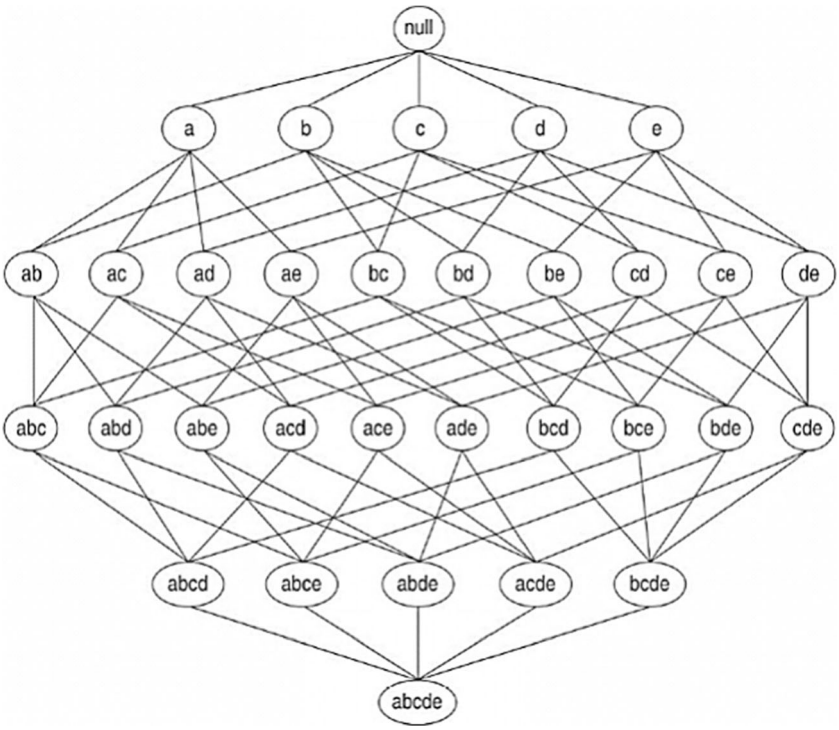

Fig. 1 An itemset lattice

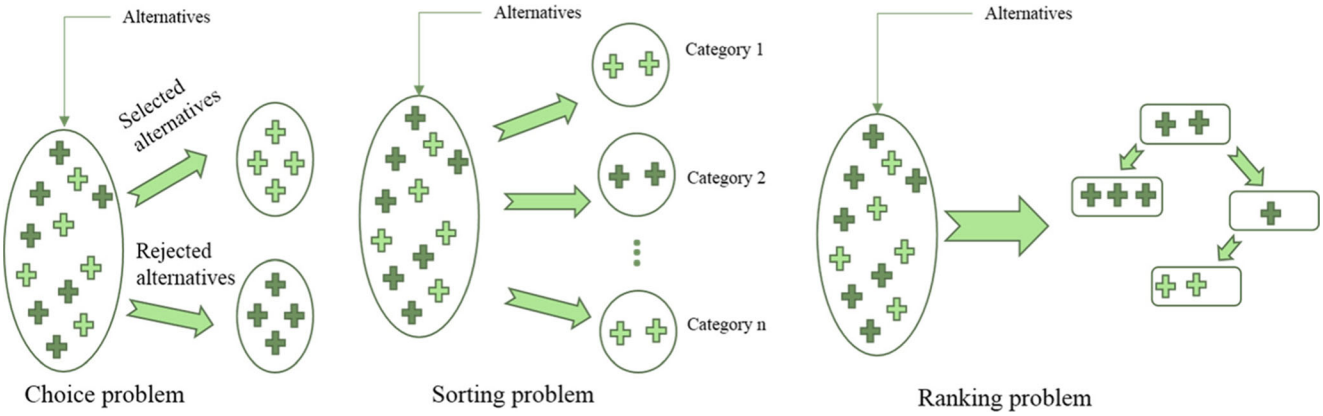

Fig. 2 MCDA problematic

criteria sorting problematic, using an existing method called ELECTRE TRI [25].

3.2.1 ELECTRE TRI

ELECTRE TRI is a multi-criteria sorting method that assigns alternatives to pre-defined categories. Each category must be characterized by a lower and upper profile. The details are given in Fig. 3.

In the data mining field, the association rule algorithms produce Pγ,a large number of extracted rules that do not allow an expert to make their own selection of the most interesting. To deal with this problem, the integration of MCDA, and particularly the existing method known as ELECTRE TRI, offers the ability to sort the results [26–29].

Let A= A = {a_1, a_2, a_3, … , a_m}{a_1, a_2, a_3,…,a_m} denote the set of alternatives, C= A = {a_1, a_2, a_3, … , a_m}{C_1, C_2, C_3,…,C_h} the set of categories, and B= A = {a_1, a_2, a_3, … , a_m}{b_1, b_2, b_3,…,b_h} the set of profiles. The alternatives are compared, not with each other, but with thresholds reflecting the boundary between h categories. ELECTRE TRI assigns alternatives to categories using two consecutive steps:

Step 1: Construct an outranking relation S by validating the assertion aSb_h, whose meaning is "a is at least as good as b_h", and build the degree of credibility $\sigma(a, b_h)$. The assertion aSb_h is considered to be valid if $\sigma(a, b_h) \succ \lambda$, λ being a "cutting level" such that $\lambda \in [0.5, 1]$.

Determination of the outranking relation consists of the following steps:

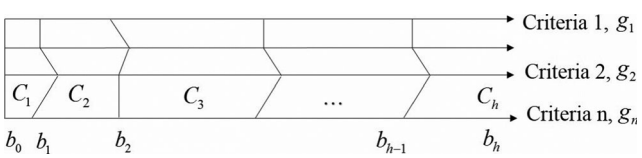

Fig. 3 Definition of categories using limit profiles

Computation of the partial concordance indices $c_j(a,b_h)$:

$$c_j(a,b_h)\begin{cases} 0 \text{ if } g_j(b_h)-g_j(a)\geq p_j(b_h) \\ 1 \text{ if } g_j(b_h)-g_j(a)\leq q_j(b_h) \\ \dfrac{p_j(b_h)+g_j(a)-g_j(b_h)}{p_j(b_h)-q_j(b_h)} \text{otherwise} \end{cases} \quad (3)$$

Computation of the concordance index $c(a,b_h)$:

$$C(a,b_h)=\frac{\sum_{j\in F}K_jC_j(a,b_h)}{\sum_{j\in F}K_j} \quad (4)$$

Computation of the discordance indices $d_j(a,b_k)$:

$$d_j(a,b_h)\begin{cases} 0 \text{ if } g_j(b_h)-g_j(a)\leq p_j(b_h) \\ 1 \text{ if } g_j(b_h)-g_j(a)\succ q_j(b_h) \\ \dfrac{p_j(b_h)+g_j(a)-p_j(b_h)}{v_j(b_h)-p_j(b_h)} \text{otherwise} \end{cases} \quad (5)$$

Computation of the credibility index $\sigma(a, b_h)$:

$$\sigma(a,b_h)=C(a,b_h)\prod_{j\in\overline{F}}\frac{1-d_j(a,b_h)}{1-C(a,b_h)} \quad (6)$$

where:

K_j is the weight of criteria j
$C_j(a,b_h)$ is the partial concordance index of criteria j
$F = \{j\in F: d_j(a,b_h) > C(a,b_h)\}$

The outranking relation is defined based on the index of credibility σ(a,bh) and λ-cut indices as follows:

σ(a,bh) ≥ λandσ(bh,a) ≥ λ ⇒ aSbhand $\sigma(b_h, a)\geq\lambda \rightarrow aSb_h$ and bhSa ⇒ aIbh, a is indifferent to b_h.
σ(a,bh) ≥ λandσ(bh,a) < λ ⇒ aSbh $\sigma(a, b_h)<\lambda$ and $\sigma(b_h, a)\geq\lambda \rightarrow a$ does not outrank bh and b_hSa →bh outranks a.
σ(a,bh) < λandσ(bh,a) < λ ⇒ a does not outrank bhandbhand b_h does not outrank aa; in this case, a and b are incomparable.

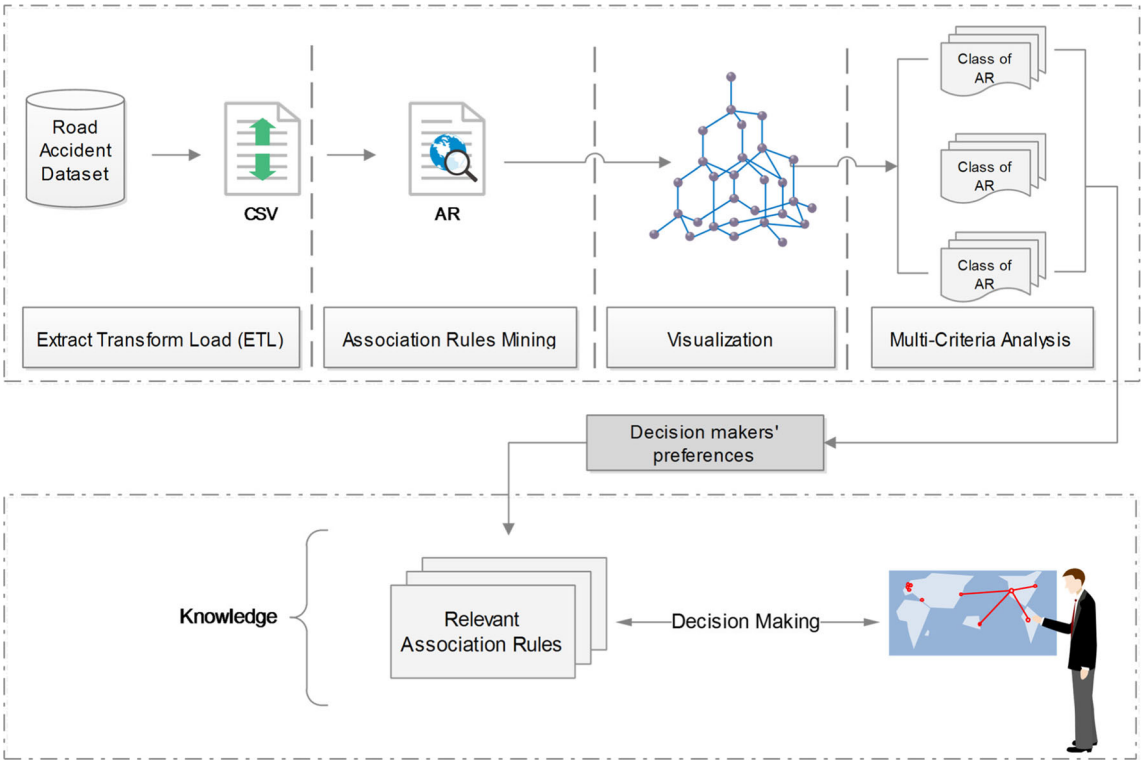

Fig. 4 The proposed approach

The values of $\sigma(a, b_h)$ and λ determine the preference between the alternative a and the profile b_h. The alternatives are not compared with each other, but with thresholds reflecting boundaries between h categories. Three situations are then possible: aIb_h indifferent, aRb_h incomparable, and aSb_h outranking.

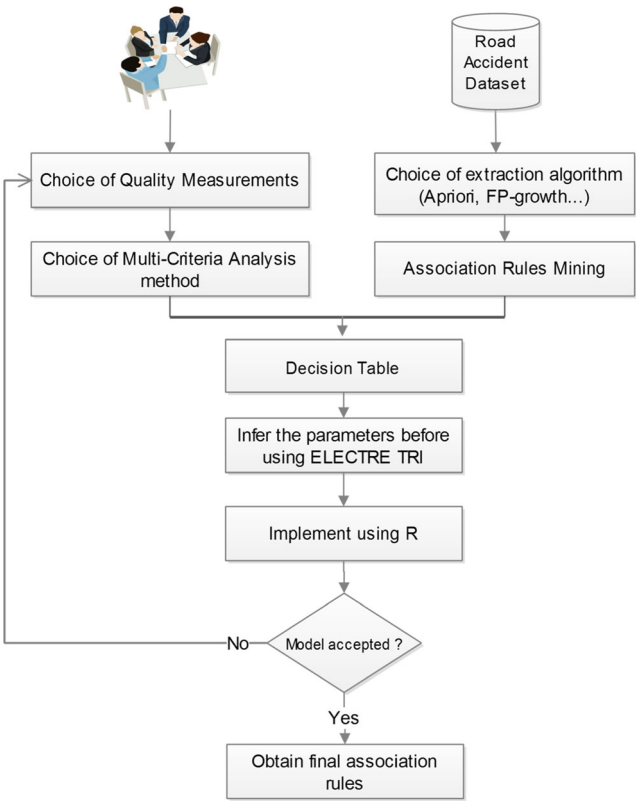

Fig. 5 The overall model

Step 2: ***Assignment Procedures***

Two assignment procedures, pessimistic and optimistic are then available.

Pessimistic assignment: compare the alternative a successively to b_i for $i = h,h\text{-}1,..,0,$ then assign a to the category $c_{h+1}(a \rightarrow c_{h+1})$.

Optimistic assignment: compare the alternative a successively to b_i for $i = 1...h.$ then assign a to the category $c_h(a \rightarrow c_h)$.

3.3 Proposed approach

Road accident analysis can be conducted using three different categories of methods: analytical methods, statistical methods, and simulation. Each method has certain strengths and weaknesses. Generally, simulation methods require sophisticated resources, making them time-consuming. Analytical methods are fast to apply but cannot be used in complex problems. Due to the weakness of these methods, statistical methods are best suited to our goal of understanding complex road accidents. However, traditional statistical methods do not offer a high level of automation when it comes to analyzing large data.

Table 3 Road accident data attributes

Attribute name	Attribute values	Description
Accident_ID	Integer	Identification of accident
Accident_Type	Fatal, Injury, Property Damage	Accident type
Driver_Age	< 20, [21–27], [28–60] > 61	Driver's age
Driver_Sex	M, F	Driver's sex
Driver_Experience	<1, [2–4], >5	Driver's experience
Vehicle_Age	[1–2], [3–4], [5–6] > 7	Service year of the vehicle
Vehicle_Type	Car, Truck, Motorcycle, Other	Type of vehicle
Light_Condition	Daylight, Twilight, Public Lighting, Night	Light conditions
Weather_Condition	Normal Weather, Rain, Fog, Wind, Snow	Weather conditions
Road_Condition	Highway, Icy Road, Collapsed Road, Unpaved Road	Road conditions
Road_Geometry	Horizontal, Alignment, Bridge, Tunnel	Road geometry
Road_Age	[1–2], [3–5], [6–10], [11–20] > 20	The age of road
Time	[00–6], [6–12], [12–18], [18–00]	Accident time
City	Marrakesh, Casablanca, Rabat...	Name of the city where the accident occurred.
Particular_Area	School, Market, Shop...	Where the accident occurred: in a school or market area.
Season	Autumn, Spring, Summer, Winter	Season of the year
Day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday	Days of week
Accident_Causes	Effects of Alcohol, Fatigue, Loss of Control, Speed, Pushed by Another Vehicle, Brake Failure	Causes of accident
Number_of_Injuries	1, [2–5], [6–10],> 10	Number of injuries
Number_of_Deaths	1, [2–5], [6–10],> 10	Number of deaths
Victim_Age	< 1, [1–2], [3–5] > 5	Victim Age

Data mining is often used as an approach which integrates concepts from statistics and artificial intelligence. Hence, it is a powerful tool that can discover complex and hidden relationships in large datasets. It has a clear advantage over other traditional statistical methods, particularly in the case of complex systems; this is certainly the case in the current study of road safety optimization.

To construct an adequate model for discovering interesting rules from an accidents database, it is important to integrate decision-making methods into the association rule mining process, in order to improve the quality of the extracted rules and build a performance model for road accident analysis.

The proposed approach is divided into two modules. The first is the association rules generator for extracting rules using the Apriori algorithm. The second is the decision support module for measuring the accuracy and relevance of results, as well as helping the expert to make the right decision concerning road network planning and new policies for road safety etc. The details of the proposed approach are shown in Fig. 4.

The global process of the proposed approach is presented in Fig. 5, wherein three steps are required. Firstly, preprocessing of the data is carried out, for which we use an extract transform load (ETL) tool to prepare and cleanse the data. Secondly, the correlations between variables in the data are extracted using the association rules technique, and the

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Accident_Type	Drive_Age	Drive_Sex	Drive_Exp	Vehicle_Age	vehicle_Type	Light_Condition	Weather_Condition	Road_Condition	Road_Geometry	Time	Season	Day	Causes
2	Fatal	<20	M	<1	<2	Car	Day	Clear	Collapse road	Horizontal	[6-12]	Spring	Md	Loss of Control
3	Injury	[21-27]	F	>6	<5	Car	Day	Run	Highway	Crossing	[12-18]	Summer	S	Alcohol effects
4	Injury	[28-60]	F	>7	<10	Car	Night	Clear	Collapse road	Alignment	[18-00]	Autumn	W	Speed
5	Injury	>60	F	<1	<15	Car	Day	Run	Highway	Horizontal	[12-18]	Summer	Sa	Speed
6	Injury	<21	F	<2	<10	Truck	Day	Clear	Unpaved road	Alignment	[12-18]	Summer	T	Brake Failure
7	Injury	[21-27]	F	<3	<5	Car	Day	Wind	Highway	Alignment	[6-12]	Winter	Md	Speed
8	Property damage	[28-60]	M	[2-6]	<15	Car	Day	wind	Collapse road	Horizontal	[12-18]	Summer	T	Loss of Control
9	Injury	<21	F	[2-6]	<10	Truck	Day	wind	Unpaved road	Alignment	[12-18]	Autumn	S	Speed
10	Injury	[21-27]	F	[2-6]	<5	Truck	Day	Clear	Highway	Alignment	[12-18]	Summer	W	Pushed by another vehicle
11	Injury	[28-60]	F	[2-6]	<15	Pedestrian	Day	Clear	Collapse road	Crossing	[6-12]	Autumn	Md	Alcohol effects
12	Injury	>61	F	>6	<5	Truck	Day	Clear	Unpaved road	Alignment	[6-12]	Summer	S	Speed

Fig. 6 Data model

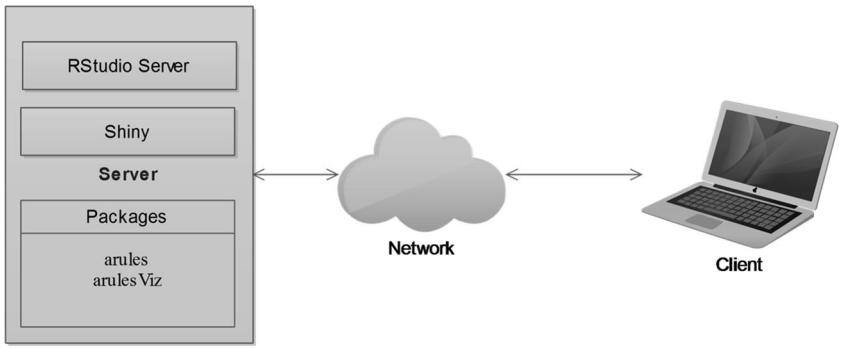

Fig. 7 Technical architecture

results are sorted according to the decision makers' preferences using the ELECTRE TRI method. Finally, the results are visualized using the arulesViz [30] package in R.

3.3.1 Variables setup

The accident data were obtained from the Ministry of Equipment, Transport and Logistics [9] in the province of Marrakech (Morocco) for the period 2003–2014. Each road accident has a record in the police database; this consists of various important attributes of the road accident. We select a set of records as the input for the algorithm. In order to identify the main factors that affect road accidents, 21 variables were used (see Table 3) [31]. These variables describe characteristics of the accident (type of collision, road users, injuries etc.), traffic conditions (maximum speed, priority regulations etc.), environmental conditions (weather, light conditions etc.), road conditions (road surface, obstacles etc.), human conditions (fatigue, alcohol etc.), and geographical conditions (location, physical characteristics etc.). The data model used is given in Fig. 6; this contains the data records related to the road accidents. In the first step, the algorithm takes as input the accident dataset, the minimum support and the minimum confidence for mining the association rules.

In the second step, MCDA is used to evaluate the extracted rules according to the decision makers' preferences in order to reduce the large number of rules, and shows only the most relevant. An analysis of this information can produce good results that can help decision makers to understand the factors behind road accidents; hence, appropriate preventive efforts can be undertaken.

3.4 Implementation

The new contribution of this work is the application of these techniques to general business problems using computerized approaches with graphical interfaces, meaning that the tools are easy to use and available to business experts. The technical architecture of the proposed approach is given in Fig. 7. The implementation is based on R [32] and Shiny [33], the open-source programming language and software environment for statistical computing and graphics. The server is composed of two components: the Rstudio Server and R packages for association rule mining and visualization. Shiny is an R package that

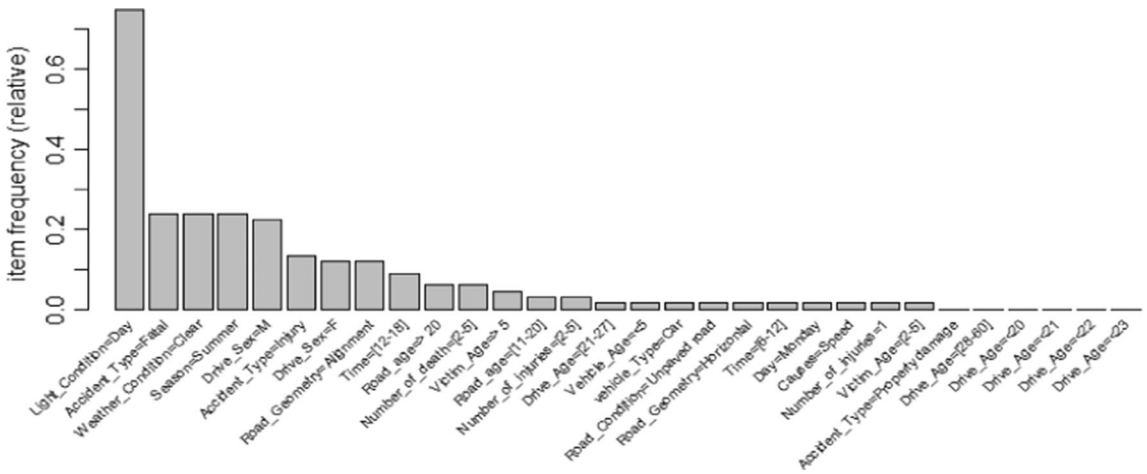

Fig. 8 Frequent itemsets

Table 4 Extracted association rules

N	Antecedent	Consequent	Support	Confidence	Lift
1	{}	= > {Light_Condition = Day}	0.850	0.850	1.000
2	{Road_Geometry = Horizontal}	= > {Light_Condition = Day}	0.300	1.000	1.176
3	{Drive_Age= [21–27]}	= > {Light_Condition = Day}	0.300	1.000	1.176
4	{Day = Monday}	= > {Light_Condition = Day}	0.300	1.000	1.176
5	{Road_Condition = Unpaved Road}	= > {Light_Condition = Day}	0.300	0.857	1.008
6	{Causes = Speed}	= > {Road_age= [11–20]}	0.300	0.857	1.905
7	{Victim_Age= [2–5]}	= > {Light_Condition = Day}	0.300	0.857	1.008
8	{Number_of_injuries = 1}	= > {Light_Condition = Day}	0.350	1.000	1.176
9	{Vehicle_Age = <5}	= > {Light_Condition = Day}	0.300	0.857	1.008
10	{Time= [6–12]}	= > {Light_Condition = Day}	0.350	1.000	1.176
11	{Road_age= > 20}	= > {Season = Summer}	0.300	0.750	1.364
12	{Road_age= > 20}	= > {Light_Condition = Day}	0.350	0.875	1.029
13	{Accident_Type = Fatal}	= > {Weather_Condition = Clear}	0.300	0.750	1.364
14	{Accident_Type = Fatal}	= > {Drive_Sex = M}	0.400	1.000	1.818
15	{Drive_Sex = M}	= > {Accident_Type = Fatal}	0.400	0.727	1.818
16	{Accident_Type = Fatal}	= > {Light_Condition = Day}	0.350	0.875	1.029
17	{Vehicle_Type = Car}	= > {Light_Condition = Day}	0.350	0.778	0.915
18	{Road_age= [11–20]}	= > {Light_Condition = Day}	0.350	0.778	0.915
19	{Drive_Sex = F}	= > {Accident_Type = Injury}	0.450	1.000	2.000
20	{Accident_Type = Injury}	= > {Drive_Sex = F}	0.450	0.900	2.000
21	{Drive_Sex = F}	= > {Light_Condition = Day}	0.400	0.889	1.046
22	{Victim_Age= > 5}	= > {Light_Condition = Day}	0.400	0.889	1.046
23	{Time= [12–18]}	= > {Season = Summer}	0.450	0.900	1.636
24	{Season = Summer}	= > {Time= [12–18]}	0.450	0.818	1.636
25	{Time= [12–18]}	= > {Light_Condition = Day}	0.500	1.000	1.176
26	{Number_of_Injuries= [2–5]}	= > {Road_ Geometry = Alignment}	0.350	0.700	1.273
27	{Number_of_Injuries= [2–5]}	= > {Light_Condition = Day}	0.350	0.700	0.824
...	…	…	…	…	…
53	{Time= [12–18] Season = Summer}	= > {Light_Condition = Day}	0.450	1.000	1.176
54	{Light_Condition = Day Time= [12–18]}	= > {Season = Summer}	0.450	0.900	1.636
55	{Light_Condition = Day Season = Summer}	= > {Time= [12–18]}	0.450	0.818	1.636
56	{Season = Summer Number_of_Deaths= [2–5]}	= > {Light_Condition = Day}	0.300	1.000	1.176
57	{Light_Condition = Day Number_of_Deaths= [25]}	= > {Season = Summer}	0.300	0.750	1.364
58	{Weather_Condition = Clear Road_Geometry = Alignment}	= > {Light_Condition = Day}	0.300	0.857	1.008
59	{Light_Condition = Day Road_Geometry = Alignment}	= > {Weather_Condition = Clear}	0.300	0.750	1.364
60	{Weather_Condition = Clear Season = Summer}	= > {Light_Condition = Day}	0.350	1.000	1.176
61	{Light_Condition = Day Weather_Condition = Clear}	= > {Season = Summer}	0.350	0.700	1.273
62	{Drive_Sex = M Season = Summer}	= > {Light_Condition = Day}	0.300	1.000	1.176
63	{Drive_Sex = M Weather_Condition = Clear}	= > {Light_Condition = Day}	0.300	1.000	1.176
64	{Accident_Type = Fatal Driver_Sex = M Weather_Condition = Clear}	= > {Light_Condition = Day}	0.300	1.000	1.176
65	{Accident_Type = Fatal Light_Condition = Day Weather_ Condition = Clear}	= > {Drive_Sex = M}	0.300	1.000	1.818
66	{Accident_Type = Fatal Driver_Sex = M Light_Condition = Day}	= > {Weather_Condition = Clear}	0.300	0.857	1.558
67	{Driver_Sex = M Light_Condition = Day Weather_Condition = Clear}	= > {Accident_Type = Fatal}	0.300	1.000	2.500

makes it easy to build interactive web applications directly using R. The individual components are clients; these are connected to a network and send a request to the server, and the server responds accordingly. The web application is interactive, scalable and suitable for road accident analysis.

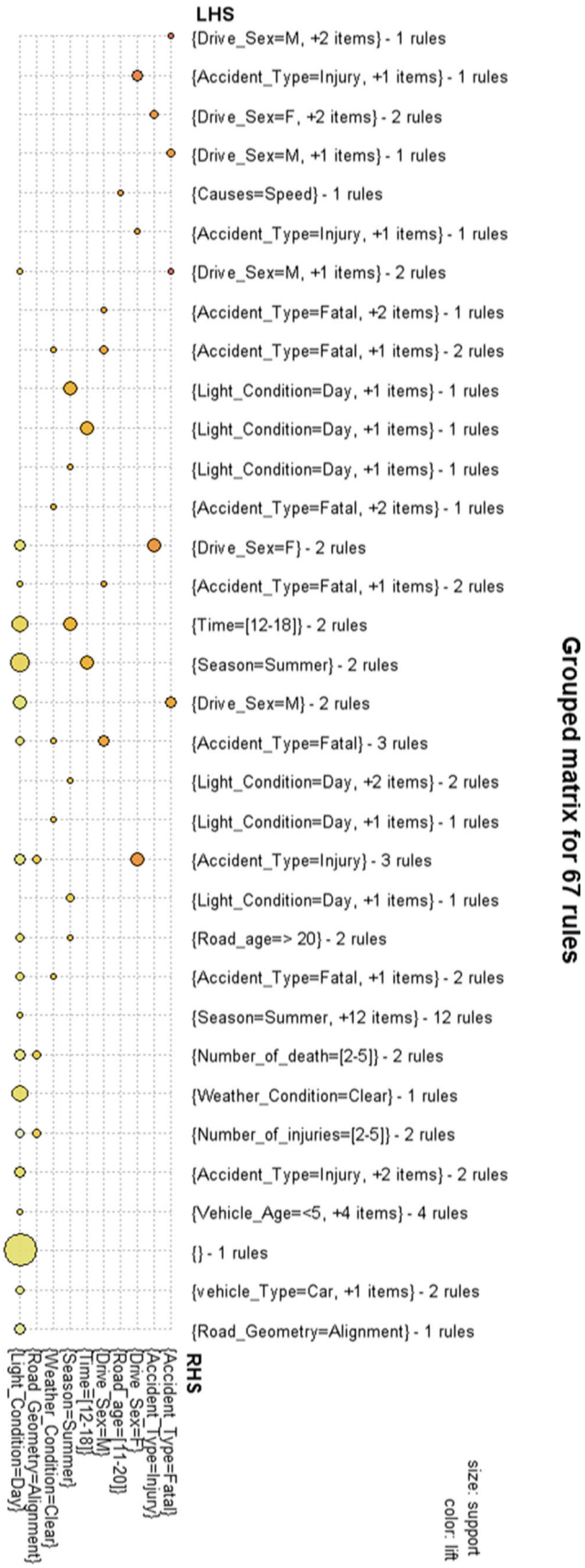

Fig. 9 Grouped matrix-based visualization

4 Results and discussion

Following data cleansing, we select a set of significant records which identify the factors related to road accidents. Then, we apply the proposed approach using two steps. The first is the extraction of association rules from datasets using the Apriori algorithm with the minimum support = 0.33 to extract frequent itemsets (see Fig. 8). This figure illustrates the itemsets by frequency. The results are sensitive to the minimum support introduced in the first step of Apriori algorithm. The second step is to generate the association rules from the frequent itemsets previously extracted. The extracted rules are given in Table 4.

To visualize the extracted rules, we use arulesViz [30] as an R package extension; this implements several known and novel visualization techniques such as matrix-, group-, and graph-based visualization. The frequent itemsets are shown in Fig. 8. The matrix-based visualization technique presents the antecedent and consequent items on the X and Y axes. This technique is enhanced using a grouped matrix, by grouping the extracted rules using clustering; an example of a grouped matrix-based visualization is given in Fig. 9. The group of the most interesting rules according to the lift (this measures how far the antecedent and consequent rules are from independence) are shown in the top left-hand corner of the plot. There is one rule which contains "*Driver_sex = M*", and two other items in the antecedent (LHS); the consequent (RHS) is "*Accident_type = Fatal*".

Graph-based visualization uses vertices and edges (see Fig. 10). The vertices typically represent items or itemsets, and edges indicate a relationship between rules. Interesting measures are typically added to the plot as labels for the edges.

The Apriori algorithm and its derivatives provide an effective solution for the extraction of association rules. However, these algorithms produce a large number of rules, preventing decision makers from making their own selection of the most interesting rules. To solve this problem, the integration of multi-criteria decision analysis approach is useful in practice for decision makers affected by redundancy in the extracted rules [29–31]. In this context, we use the ELECTRE TRI method, considering a set of extracted rules as the alternatives and support, confidence and lift as the criteria.

The support used in the first step is to count frequent itemsets using the Apriori algorithm, which satisfies the minimum support requirements defined by the user. This step is generally expensive due to the use of multiple passes over the database. For the second step, after the extraction of association rules in the form of $A \rightarrow B$, the support, confidence, and lift of each extracted rule is computed using the Apriori algorithm. We use multi-criteria decision support to prioritize the extracted rules; each method in MCDS is based on the decision matrix (evaluation table), where the values of this table are given by the decision makers (domain expert) according to their preferences. In this case, we used

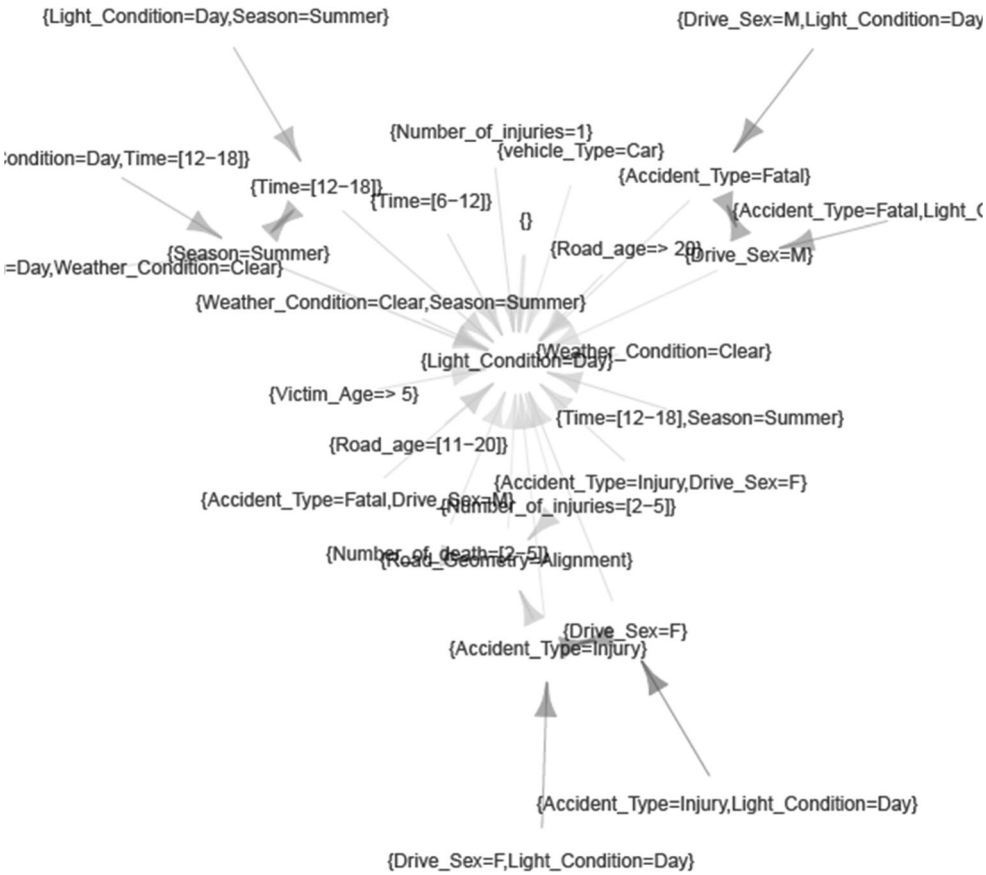

Fig. 10 Graph-based visualization with items and rules as vertices

minimum support = 0.33 to count frequent itemsets, and for the MCDS we used the values computed by the algorithm as the preference of decision makers in order to determine the performance of our approach.

Table 5 Decision matrix

Rule/Criteria	Support	Confidence	Lift
Rule1	0.85	0.85	1.00
Rule2	0.30	1.00	1.17
Rule3	0.30	1.00	1.17
Rule4	0.30	1.00	1.17
Rule5	0.30	0.85	1.00
Rule6	0.30	0.85	1.90
Rule7	0.30	0.85	1.00
Rule8	0.35	1.00	1.17
Rule9	0.30	0.85	1.00
Rule10	0.35	1.00	1.17
Rule11	0.30	0.75	1.36
Rule12	0.35	0.87	1.02
…	…	…	…
Rule63	0.30	1.00	1.17
Rule64	0.30	1.00	1.81
Rule65	0.30	0.85	1.55
Rule66	0.30	0.85	2.50
Rule67	0.30	0.85	1.55

Table 5 gives the decision matrix (evaluation table), which lists the rules as rows of the table and the criteria as columns. Then, each rule/criteria combination is scored, with a weight determined by the relative importance of the criteria, and these scores are added to give an overall score for each option. The scores for support and confidence vary between 0 and 1.

Decision matrix analysis is a useful technique for making a decision. It is particularly powerful where there are a number of good alternatives to choose from and many different factors to take into account. Decision matrix analysis helps in deciding between several options where many different criteria are involved.

The second step of ELECTRE TRI is to define a set of profiles according to the decision makers' preferences; the profiles b_1 and b_2 are the limits between categories A and B and categories B and C (see Table 6).

Table 6 Initial profiles defining the category limits

Profiles	Support	Confidence	Lift
b_1	0,5	1,0	1,2
b_2	0,4	0,9	1,0

Table 7 Parameters for the ELECTRE TRI method

Threshold	Support	Confidence	Lift
weight (K_j)	0.5	1.0	1.2
$q_j(b_1)$	0.4	0.9	1.0
$p_j(b_1)$	0.5	1.0	1.2
$v_j(b_1)$	0.4	0.9	1.0
$q_j(b_2)$	0.5	1.0	1.2
$p_j(b_2)$	0.4	0.9	1.0
$v_j(b_2)$	0.5	1.0	1.2

Each alternative is compared to the profiles; the importance of each criterion in decision making is reflected in predefined threshold scores. The preference threshold p, the indifference q, and the veto threshold v are given in Table 7. Moreover, each criterion has a weight k, reflecting its contribution to the final decision.

The third step is the computation of the concordance indexes $c_j(a,b_h)$ as in Eq. (3) and the discordance indexes $d_j(a,b_k)$ as in Eq. (5). The results are the outranking relations, which determine the relationship between the rules and profiles. The parameter that determines the preferred situation between the association rules and the profiles b_h is known as the cutting level, and its default value is $\lambda = 0.76$. The evaluation of the association rules using assignment procedures is shown in Table 8.

4.1 Discussion

Road safety is currently one of the government's highest priorities. Identifying and profiling black spots and black zones in terms of accident-related data and location characteristics needs to provide new insights into the complexity and causes of road accidents, which, in turn, provide valuable input for government actions. Data mining techniques have led to significant advances in other areas and should also be used to improve this sector. The use of inventory management systems tracking sensors generates a large amount of data; this appears to be a possible application area for data mining, and there have been prior studies of analyzing, optimizing and improving road safety in shipping and transport logistics. The existing method of optimization has long been computerized, but does not provide the type of insights that are the goal of data mining. The goal of our proposed approach is not to optimize transportation safety, but to generate insights and sufficient knowledge to enable logistics managers to make the right decision, thus enabling the optimization, the avoidance of dangerous routes and improvements in road safety.

In this study, Table 8 shows the results of assigning rules to categories (classes) C1, C2, and C3 such that the most relevant category is C1. The extracted decision rules indicate that fatal and injury-causing accidents occur mostly in the following situations.

- The first most common cause of accidents is speeding. Speed influences both the risk of a crash and its consequences;
- Females have a direct impact on the accidents;
- Most accidents occur when lighting exists.
- The number of deaths and injuries is increasing, especially in summer.
- Accidents frequently occur when the weather is clear.

Based on this study, it can be said that the integration of multi-criteria decision analysis within knowledge discovery in databases performs well and produces useful knowledge. After eliminating the non-interesting rules, 32 significant rules were obtained. The rest of the rules belong to the less interesting categories interest. The most interesting rules are given in Table 9.

The use of the Apriori algorithm and its derivatives produces a large number of association rules. It is therefore difficult to extract useful insight from this wide range of results. However, the integration of multi-criteria decision analysis approach within the association rules process selects only the most relevant rules, according to the decision makers' preferences. The results are always sensitive to the values of thresholds p_j, q_j, v_j, and the decision makers' preferences.

There is a rich literature that describes the different techniques and their outcomes in road accident analysis [4, 6, 15, 23, 34, 35]. These techniques have found an association between drivers' behaviors, weather conditions, light conditions and the severity of accidents. However, the large size of the database leads to a very high number of

Table 8 Assignment procedures

Rule	C1	C2	C3
Rule1			×
Rule2	×		
Rule3	×		
Rule4	×		
Rule5	×		
Rule6			×
Rule7	×		
Rule8	×		
Rule9	×		
Rule10			×
Rule11	×		
Rule12	×		
…	…	…	…
Rule63	×		
Rule64			×
Rule65		×	
Rule66		×	
Rule67		×	

Table 9 The final set of relevant rules

Class	Rules
C1	Rule2,Rule3,Rule4,Rule5,Rule7, Rule8 Rule9, Rule11, Rule12, Rule13,Rule16 Rule17, Rule18,Rule24,Rule26, Rule27 ,Rule28,Rule29,Rule30,Rule32Rule33,Rule34, Rule35,Rule38,Rule50, Rule52 Rule55,Rule56,Rule58, Rule59,Rule61 Rule63

extracted rules, which cannot be explored further, and which confuse decision makers. The results of our study not only confirm an association between certain variables but also show that the integration of MCDA allows decision makers to make their own selection of the most interesting rules, according to their preferences and needs, allowing the application of accident prevention efforts in the identified areas for various categories of accidents.

In summary, the integration of the association rules technique within multi-criteria decision analysis contributes to a better understanding of the dynamics of road accidents and can provide meaningful information to help decision makers and logistics managers to improve performance in terms of transport quality and road safety optimization. Finally, the proposed approach has the following major strengths:

- Mining and visualization of association rules
- Management of the interest level of association rules
- Reduction of the large number of extracted rules.
- Road accident analysis
- Improvements in road safety

5 Conclusion

In many countries, road transport often involves accidents, and this affects transport and shipping services. Understanding road traffic is extremely important in improving road safety. In this paper, we propose an effective method for mining strong and relevant association rules from a road accident database. With the objective of identifying the hidden relationships between the most common accidents, the road accident dataset is analyzed using the association rules technique. The proposed method uses efficient mining of association rules. Furthermore, the integration of MCDA within the association rule mining process provides a sustainable solution by selecting only the most interesting rules according to the decision makers' preferences. In particular, we study a set of rules extracted from the road accidents database, considering the criteria most commonly used in the literature. We conclude that the application of multi-criteria decision analysis to a set of extracted rules can contribute to solving the problem that arises when using traditional algorithms, in terms of redundancy and a lack of interesting rules. Furthermore, the results indicate that human and behavioral characteristics play an important role in the occurrence of all traffic accidents. Finally, the results show that the proposed approach serves its purpose and can provide meaningful information which can help in developing suitable prevention policies for improving road safety.

In further work, a new methodology combining this approach with other optimization methods will be applied in the context of big data, using VANETs, Apache Kafka for streaming, and machine learning to build a predictive model for road safety.

Acknowledgements The authors would like to thank Christine Miles for her english editing. The authors would also like to acknowledge the valuable comments from the referees on our manuscript.

References

1. Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, Menlo Park, p 1–34
2. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD), p 207–216
3. Ossenbruggen P, Pendharkar J et al (2001) Roadway safety in rural and small-urbanized areas. Accid Anal Prev 33(4):485–498
4. Oa J, Lpez G, Abelln J (2013) Extracting decision rules from police accident reports through decision trees. Accid Anal Prev 50:1151–1160
5. Sanmiquel L, Rossell JM, Vintr C (2015) Study of Spanish mining accidents using data mining techniques. Saf Sci 75:49–55
6. Mirabadi A, Sharifian S (2010) Application of association rules in Iranian railways (RAI) accident data analysis. Saf Sci 48(10):1427–1435

7. Brenac T (2009) Common before-after accident study on a road site: a low-informative Bayesian method. Eur Transp Res Rev 1(3):125–134
8. The World Health Organization. http://www.who.int/gho/road safety/en/, accessed 2016
9. The Ministry of Equipment, Transport and Logistics Morocco, http://www.equipement.gov.ma/en/Pages/home.aspx, accessed 2016
10. Kuhnert PM, Do KA, McClure R (2000) Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Comput Stat Data Anal 34(3):371–386
11. Sohn S, Hyungwon S (2001) Pattern recognition for a road traffic accident severity in Korea. Ergonomics 44(1):101–117
12. Chong M, Abraham A, Paprzycki M (2004) Traffic accident analysis using decision trees and neural networks. In: Isaias P et al (eds) IADIS International Conference on Applied Computing, vol 2. IADIS Press, Portugal, pp 39–42
13. Chang L, Wang H (2006) Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 38(5):1019–1027
14. Sze NN, Wong SC (2007) Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accid Anal Prev 39: 1267–1278
15. Abugessaisa I (2008) Knowledge discovery in road accidents database integration of visual and automatic data mining methods. Int Public Inf Syst 1:59–85
16. Wong J, Chung Y (2008) Comparison of methodology approach to identify causal factors of accident severity. Transp Res Rec 2083: 190–198
17. Anderson TK (2009) Kernel density estimation and K-means clustering to profile road accident hotspots. Accid Anal Prev 41(3):359–364
18. Zelalem R (2009) Determining the degree of drivers' responsibility for car accidents: the case of Addis Ababa traffic office. Addis Ababa University, Addis Ababa
19. Pakgohar A, Tabrizi RS, 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 Comput Sci 3:764–769
20. Demirel N, Emil MK, Duzgun HS (2011) Surface coalmine area monitoring using multi-temporal high-resolution satellite imagery. Int J Coal Geol 86:3–11
21. Wu H, Tao J, Li X, Chi X, Li H, Hua X, Yang R, Wang S, Chen N (2013) A location based service approach for collision warning systems in concrete dam construction. Saf Sci 51:338–346
22. Zhang M, Kecojevic V, Komljenovic D (2014) Investigation of haul truck-related fatal accidents in surface mining using fault tree analysis. Saf Sci 65:106–117
23. Keeney RL, Raiffa H (1993) Decisions with multiple objectives: preferences and value trade-offs. Cambridge University Press, Cambridge
24. Figueira J, Mousseau V, Roy B (2005) ELECTRE methods. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: state of the art surveys. Springer New York, New York, NY, p 133–162
25. Mousseau V, Figueira J, Naux J (2001) Using assignment examples to infer weights for ELECTRE TRI method: some experimental results. Eur J Oper Res 130(2):263–275
26. Lenca P, Meyer P, Vaillant B, Picouet P, Lallich S (2004) Évaluation et analyse multicritère des mesures de qualité des règles d'association. Revue des Nouvelles Technologies de l'Information, mesures de Qualit pour la Fouille de Donnes, RNTI-E-1, pp. 219–246
27. Ait-Mlouk A, Agouti T, Gharnati F (2015) Comparative survey of association rule mining algorithms based on multiple-criteria decision analysis approach. In: Control, engineering and information technology (CEIT), 3rd international conference on, vol., no., pp. 1-6, 25-27
28. Ait-Mlouk A, Agouti T, Gharnati F, Derbali B (2015) A choice of relevant association rules based on multi-criteria analysis approach. 2015 5th international conference on information and communication technology and accessibility (ICTA), Marrakech, pp. 1–6. doi: 10.1109/ICTA.2015.7426886
29. Ait-Mlouk A, Agouti T, Gharnati F (2016) Multi-agent-based modeling for extracting relevant association rules using a multi-criteria analysis approach. Vietnam J Comput Sci 3(4):235–245
30. Hahsler M, Chelluboina S (2011) Visualizing association rules: introduction to the R-extension package arulesViz. R project module
31. Ait-Mlouk A, Agouti T, Gharnati F (2016) An approach based on association rules mining to improve road safety in Morocco, international conference on information Technology for Organizations Development (IT4OD), 1-6, 2016, IEEE
32. https://www.r-project.org/, Accessed 2017
33. http://shiny.rstudio.com/, Accessed 2017
34. Kumar S, Toshniwal D (2017) Severity analysis of powered two wheeler traffic accidents in Uttarakhand. India Eur Transp Res Rev 9:24
35. Kumar S, Toshniwal D (2015) Analysing road accident data using association rule mining. International Conference on Computing, Communication and Security (ICCCS), Pamplemousses, pp 1–6