Journal of Big Data

CASE STUDY Open Access

Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France

Fatima Zahra El Mazouri*, Mohammed Chaouki Abounaima and Khalid Zenkouar

*Correspondence:
Fatimazahra.elmazouri@usmba.ac.ma
Laboratory of Intelligent Systems and Applications, Faculty of Sciences and Technologies, Sidi Mohammed Ben Abdellah University, Fez, Morocco

Abstract

Introduction: The problem studied in this paper is the road insecurity, which is manifested by the big number of injuries and deaths recorded annually around the world. These victims of road accidents worry the whole community, hence the duty and the need to find solutions for reducing the number of victims and material damage. The overall purpose of this case study is to treat the problem of corporeal accidents known in France.

Case description: The case study presented in this paper is intended to help decision-makers to find and understand the all significant relationships and correlations that exist between the conditions that led to these corporeal accidents. In fact, the two French ministries of interior and transport have jointly created a unique database on corporeal accidents, called BAAC "Accident Analysis Bulletin Corporal", with the aim of allowing different exploitations of this database by different concerned administrations and research organizations. Our intervention consists to adopt a hybrid approach based on data mining techniques combined to the multicriteria decision methods. This approach allows to extract the most relevant association rules. The results thus obtained can be easily exploited by the decision-makers to choose the appropriate policies in the perspective of improving road safety.

Discussion and evaluation: The proposed approach ranks all association rules in order of importance for several quality measures of association rules. The alternatives at the top of the ranking are the results to retaining for the analysis. The approach is applied to the BAAC database of 2016, which led to the selection of three association rules. These association rules reveal that there are narrow correlations between the following elements: Driver with Pedestrian, Normal Surface of Road with Normal Atmospheric Condition and the Pavement with Pedestrian. These correlations can be justified by excess speed and carelessness of the drivers.

Conclusion: The improvement of the road safety needs mainly to work more intensively on the behavioral side of the road users. For future work, it is planned to apply the proposed approach to the French traffic accident database, containing all the data on road accidents collected over several years. In addition, other measures will be used in the ranking of the association rules.

Keywords: Data mining, Multicriteria decision aid, Association rules, Apriori algorithm, ELECTRE II method, Road safety

Introduction

Data mining "DM", as a part of the general KDD "Knowledge Discovery in Databases" process [1] (see Fig. 2), is an attractive set of methods and techniques that has known a big popularity in recent years in various fields such as the fields of finance [2], education [3] and healthcare [4]. The remarkable interest in data mining refers to the huge volume of the collected data [5], which involves many problems, DM groups all the methods, that can analyze the information from a large database in order to find useful and relevant information and all significant correlations existing between the data. This relevant information can be used by Decision Makers. Besides, the extraction of association rules "AR" was initially introduced by R. Agrawal [6], and which is a popular technique of DM that is widely used in analysis marketing, catalogue design, and other decision-making processes. Extracting AR from a set of data consists to discover the interesting relationship between variables stored in an important database. This technique proceeds in two major steps. The first one consists to the extraction of all frequent itemset, and the second step consists to generating the association rules by taking into consideration some thresholds like minimum support, noted by "minsup", and the minimum confidence, noted by "minconf".

In this context, many algorithms have been proposed to generate association rules from large data such as the algorithms presented in "Knowledge Discovery in Database Process" section. The difference between these algorithms exists in terms of response time and memory space, these methods are distinguishable so naturally by their algorithmic complexities [7–10].

DM techniques are used in different research domains and areas and have provided useful results to extract the hidden knowledge for the decision makers. For instance, we can find in [11–14] several researches and studies on the application of the association rules for various problems, and in particular the problem of safety of the road traffic [15–21]. Besides, according to the national observatory of road safety [22–24], at least 350,000 people died since 1960 as a result of a road accident in France: in 1972, there were 18,034 official deaths. And since 1945, the numbers are increasing because we speak of at least 500,000 deaths, the equivalent of a major conflict or the human losses of France during the second world war.

In 2016 alone, the French national authorities recorded 57,522 personal injuries and more than 3477 deaths within 30 days of the accident, in addition to 72,645 injuries, of which 27,187 were hospitalized, and not counting non-bodily accidents. Furthermore, road accidents are very costly for the national economy and represent a heavy burden on the state budget, in addition to the social problems that they result. Indeed, according to the published report by ONISR on accidents occurred on the roads of France in 2017 [25], the cost of the corporeal accidents in metropolitan France would be 39.7 billion euros (B€) distributed as follows:

- 11.3 B€ in mortality;
- 23.1 B€ for hospitalizations;
- 4.0 B€ n for light victims;
- 1.2 B€ for the material damage of these bodily injuries.

Material damage costs are also added to the cost of corporeal accidents. In fact, the cost of non-bodily injuries corresponding to material damage alone is estimated at 11.1 B€. The total cost of road insecurity is therefore 50.8 B€, which is the equivalent of 2.2% of the GDP "Gross Domestic Product".

All these statistics show the danger of traffic accidents which remain disturbing for everyone, and we are seriously trying to find effective solutions to fight these losses of human lives. In this context subscribe our study, which provides support and aid for the stakeholders and decision-makers to understand and find fundamental correlations between human damage and the various conditions that characterize all the recorded accidents.

The principle of this work is to treat the case of road accidents in France by using the BAAC database, with the aim of informing decision-makers about the most common conditions of accidents. To do this, we propose to extract the association rules. But given the large number of association rules generated which makes their exploitation difficult, we use the method ELECTRE II [26] to rank these different association rules from the best to the worst. The ranking result is obtained by considering different measures of association rules. The top rank represents the most relevant and interesting association rules.

The rest of the paper is organized as follows: the first section "Related work" cites some studies have been developed to analyze road accidents. "Case description" section gives a description of the case study representing the subject of this paper. "Knowledge discovery in database process" and "Overview of the ELECTRE II method" sections provide an overview on the mining of the association rules and an overview on the ELECTRE II method. "The proposed approach" section describes the proposed methodology for extracting association rules and the utility for integration of multi-criteria decision analysis approach to obtain the appropriate knowledge process. The results and discussion are presented at last section "Discussion and evaluation". Finally, we concluded by summarizing our work and giving some perspectives.

Related work

With regard to the awareness of the problem of road insecurity around the world, several research projects have been developed in the literature to analyze road accidents and determine the main causes. In this context, we find the work of Chong et al. [16] who deploy a decision tree to analyze the severity of road accidents. They found that fatal injuries were caused by many factors, including seatbelt, alcohol, and lighting conditions. In addition, other studies by Kuhnert et al. [17] CART and MARS were used to analyze a case–control epidemiological study of injuries resulting from road accidents. They also identified potential risk areas, largely caused by the driver's situation. There is also the work of Kumar et al. [13, 27], who deployed a framework to analyze road accident data. In another work, they have done a comparative analysis of heterogeneity in road accidents using data mining techniques. Moreover, Ait-Mlouk et al. [15] proposes an approach to association rule mining-based MCDA "Multi Criteria Decision Analysis" for analyzing road accident data.

We cannot forget also the interesting work of Sohn et al. [18] who 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. Another important work by Chang L, Wang H [19] analyzed the severity of road accidents by applying non-parametric classification tree techniques. Moreover, Wong and Chang [20] 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, for example, these factors included the drivers being young, male, or less experienced and their behaviors of drinking, wandering on roads around midnight, and overestimating their own driving capabilities and underestimating the possible dangers hidden in the environment. Finally, Anderson [21] studied the pattern of road accidents and used the resulting pattern to create a classification system for road accident hotspots.

This list studies is obviously far from exhaustive, road accidents have become a global disaster, and everyone contributes to find the effective solutions to save lives and reduce the huge economic losses.

On the other hand, other association rules mining algorithms have been widely used in the literature to extract frequent itemsets and build association rules. These algorithms are based primarily on minimum support and the minimum confidence. However, these algorithms generate several association rules. There is a need to make the most significant association rules by taking into account several measures, other than the support measure and the confidence measure. In this paper we propose a hybrid approach based on the Apriori algorithm and the method MCDA ELECTRE II to analyze the database of road accidents in France.

Case description

Motivation

The improvement of the road network, the implementation of measures of prevention and repression, the improvement of the behavior of the vehicles' structures and their equipment and the efficiency of the reliefs combine provided the remarkable results on road safety in France.

However, the increase in the number of means of transportation inevitably leads to an increase in the number of collisions endangering human lives. Research conducted since the 1970s has seen tremendous progress, but the current situation remains a matter of concern [24, 25].

This case study concerns the analysis of traffic accidents known in France. Its purpose is to help decision-makers to find all possible correlations and relationships between victims and the various recorded traffic accidents, with the aim of providing useful information that can help decision-makers to improve road safety even further. Before presenting our proposed model for the extraction of useful information, a description of the collected and available data on all corporeal accidents will be provided.

Analysis bulletin of corporal accident

With the aim of making it easy for the concerned administrations and researchers to have a global vision of accidents in France and having the ability to contribute to the process of road traffic safety improvement, the French Ministry of Interior and the Ministry of Transport made a unique database on corporeal accidents called BAAC available to all concerned individuals [23, 24].

The data in this database are collected with the contribution of all the competent authorities, in fact, whenever a police station or a gendarmerie brigade is informed of the occurrence of a traffic accident causing an injury, it draws up a report describing the accident which constitutes the basis of the procedure possibly leading to penal sanctions and the compensation of the victims.

In parallel, law enforcement agencies "Police or Gendarmerie" also inform the BAAC, which describes:

- The general characteristics of the accident (date, hour, place, atmospheric conditions, type of collision);
- Places (type of track, profile, condition, amenities, signage, environment);
- The vehicles (category, loading, condition, point of impact, obstacle struck, maneuvers before the accident);
- Users (place, gender, age, socio-professional category, injuries, possible alcohol, driving license characteristics, nature of the path, use of security systems).

The BAAC data are centralized by the General Directorate of the National Gendarmerie "DGGN" and the General Directorate of the National Police "DGPN" and transmitted to the Technical Studies Department of Roads and Highways "SETRA" which merges this data into a single file "The national file of corporeal accidents". The data is then monitored, validated and disseminated by the National Inter-Ministerial Observatory of Road Safety "ONISR". Figure 1 describes the BAAC management scheme [23, 24].

Configuration of the BAAC

The BAAC, which must be filled and completed by the authorities, is divided into four parts called banners. Each of these parts contains a category of information that is used to describe an accident. The list of banners is as follows:

- The characteristics of the accident;

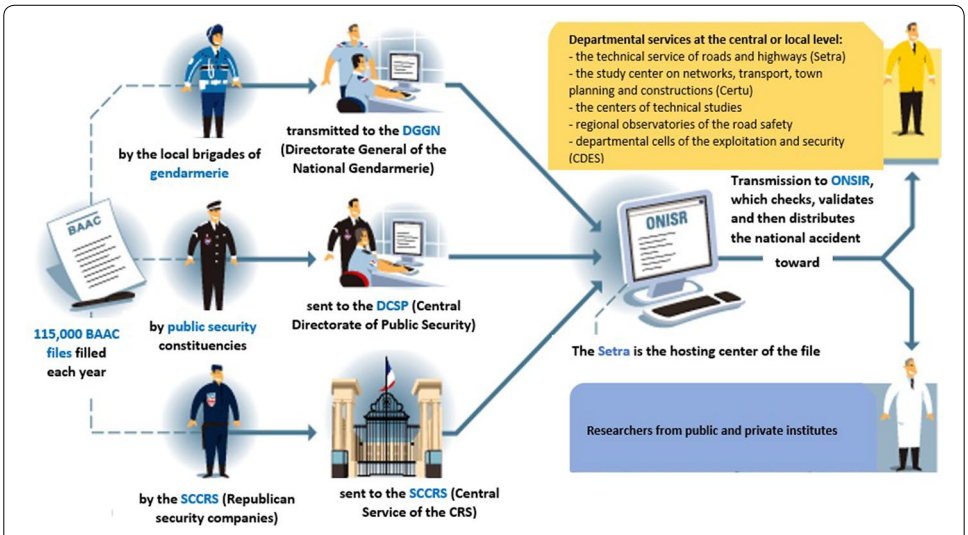

Fig. 1 BAAC management scheme

- The place where the accident occurred;
- The vehicles involved in the accident;
- The users involved in this accident.

These four banners are completed by an "identifier: Num_Acc" located at the top of the BAAC and designed to uniquely identify the accident (Fig. 1).

All data collected for the BAAC database are described at "The Proposed approach" section.

Knowledge discovery in database process

Data mining is a tool for exploring decision data. It includes a whole family of methods and techniques that facilitate the exploration and analysis of data contained within a Data Warehouse [28] or DataMart decision base or even a Big Data [29]. The DataMart is a subset of the Data Warehouse, consisting of tables at the level of detail and at more aggregated levels, to render the full spectrum of a particular business activity. The set of "DataMarts" of the company constitutes the Data Warehouse.

The Big Data represents the large data sets which exceed the capabilities of traditional database management tools. Currently, the Big data is associated with four key concepts: volume, variety, velocity and veracity.

These data mining tools are particularly effective in extracting significant information and extracting exactly the useful knowledge, from large amounts of data [30].

Steps of the knowledge extraction process

The KDD is a semi-automatic and iterative process consisting of several steps, going from the selection and preparation with transformation of data to the interpretation of results and passing by the knowledge extraction: data mining step. The different steps of this process are presented in Fig. 2.

The data mining phase designates the application, to prepared data, of the methods and the techniques that provide a set of information about the data: such as association rules and frequent patterns.

Data mining programs

Currently, there are different development platforms and technologies that offer a variety of machine learning and data mining techniques. These programs are very efficient and

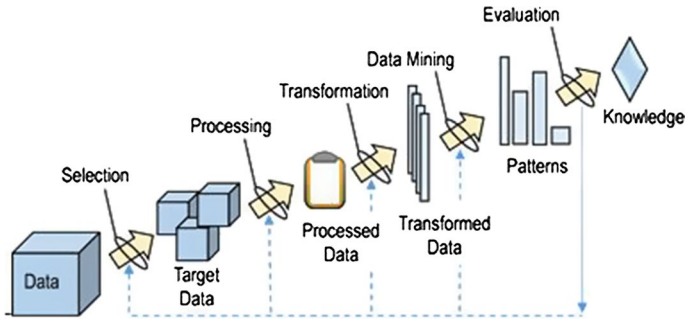

Fig. 2 Knowledge discovery process

able to process a large amount of data. For example, we cite, the most famous tools in Big Data applications such as the technology of distributed processing Apache Hadoop [http://hadoop.apache.org/] and as a second example of technology, the cluster computing Apache Spark [https://spark.apache.org/]. In this work, we opted for the IDE RStudio "Integrate Development Environment RStudio" [https://www.rstudio.com/]. This environment uses the R language which integrates several statistics functions and very advanced graphical visualization tools.

Definition of an association rule

Agrawal et al. [6] initiated the concept of association rules, for the analysis of large database of transactions in order to detect relationships and correlations among a set of items (books, products, etc.).

For the first time, Agrawal et al. [6] introduced association rules to uncover product regularities in transaction data recorded by point-of-sale systems in supermarkets. For example, the rule found in supermarket sales data would indicate that if a customer buys "Onions" and "Potatoes" together, he will probably buy "Meat". This rule is noted by: Onions, Potatoes → Meat. "Onions" and "Potatoes" represent the premise of the rule and "Meat" represents the conclusion of the rule.

The search for association rules, from a transactions database, is mainly aimed at constructing a model on conditional rules of type: A → B, that means, if we select the choice A we will, very probably, choose B. The principle of an association rule is similar to the algorithmic instruction of choice IF… THEN.

An association rule can be formally defined las the following:

- Let $I = \{i_1, i_2, \ldots, i_m\}$ a set of items and $T = \{t_1, t_2, \ldots, t_n\}$ a set of transaction, called data base of transactions, such as t_i a subset of I ($t_i \subset I$).
- m is the total number of elements and n is the considered number of transactions.
  For the above supermarket example, each transaction t_i of the database T represents the market basket relating to a single customer where we find all the products purchased. All the products offered for sale, such as the "Onions", "Potatoes" and "Meat" values, represent the set I of the items.
- An association rule is defined as an implication of the form given by the formula (1).

$$X \rightarrow Y, \quad where\, X \subseteq I,\ Y \subseteq T\ and\ X \cap Y = \emptyset \tag{1}$$

  - The itemset X is called "premise" or Left-Hand-Side "LHS"
  - The second Y is called "conclusion" or Right-Hand-Side "RHS"

In general, a rule is defined only between a sub set of items (X) and a single item (Y) as shown in the formula (2).

$$X \rightarrow i_j, \quad for\ i_j \in I \tag{2}$$

Extraction of the most relevant association rules

Some measures are very useful to extract the most relevant rules like support, confidence and lift. This list is not exhaustive we can find in the literature, like in this thesis [41], more other interest measures for association rules.

The measure of support

The support is an indication of how frequently the itemset appears in the dataset. The support of a sub set of items X, called pattern, is precisely the proportion of transactions t_i, from the base T, containing X. The support ranges from 0 to 1 and is defined by the formula (3).

$$supp(X) = \frac{card(\{t_i \in T / X \subseteq t_i\})}{card(T)} \tag{3}$$

The support of a single pattern X measures the frequency of X in the set of transactions T.

For an association rule $X \rightarrow Y$, the support measures the frequency of the pattern XUY *where* X and Y appear together.

The measure of confidence

The confidence value of a rule, $X \rightarrow Y$, represents the proportion of transactions containing X that also contains Y, ranges from 0 to 1. The confidence is computed by the formula (4).

$$conf\,(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{4}$$

The measure of lift

The lift of a rule $X \rightarrow Y$ measures the improvement made by the association rule with respect to a set of random transactions, where X and Y are independent. The lift is computed by the formula (5).

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)} \tag{5}$$

The lift ranges within $[0, +\infty]$. The possible interpretations of the measure are:

- A lift greater than 1 reflects a positive correlation between the premise X and the conclusion Y.
- A lift equal to 1 indicates a zero correlation.
- A lower lift of 1 indicates a negative correlation, also called anti-correlation.

For a better correlation, the lift must be much greater than unity.

Apriori algorithm for extract the association rules

In the literature, many algorithms have been proposed to extract frequent pattern with the association rules from a transactional database, these algorithms can be classified into 3 categories:

- Algorithms for extraction the frequent itemset: Apriori [9], AprioriTID [31], FP-Growth [32], Partition [33], Dic [34], Eclat [35], Apriori-Hybrid [36] and Relim [37].
- Algorithms for mining closed itemset: Close [38], Pascal [39].
- Algorithms for extraction the maximal itemset: MaxMiner [40].

For a review of the literature on all these algorithms, we refer readers to the work of Pasquier [7]. In this context, a comparative study has been done [8] which consists to choose the most appropriate algorithm according to decision maker's preferences by application of the famous MCDA method ELECTRE I, that permit to isolate a subset of relevant solution (algorithms of association rules extraction), the result of [8] show that we can consider Apriori and FP-Growth as the suitable algorithms for extraction association rules. Indeed, in this work the most important algorithms of association rules mining are compared by using multi-criteria approach, in fact the application of this approach has confirmed that Apriori and FP-Growth are the best selected algorithms according three criteria: time response, memory space and algorithm performance.

The Apriori algorithm have two major steps:

- Step 1: Extraction of frequent itemset based on the anti-monotonicity property: «All subsets of a frequent itemset are frequent». In order to find the itemset (pattern) with support greater than or equal to the minimum support threshold "minsup" (support $\geq$ minsup). At this step we use the Algorithm 1.
- Step 2: Apriori exploit the frequent itemset found in the previous step 1 to generate association rules. At this step we use the Algorithm 3.

Algorithm 1: Apriori algorithm for extracting frequent patterns

Input :
T : database of transactions
minsup : threshold for the measurement of support
Output :
$\bigcup_k L_k$: set of all itemset frequents of the T
Begin
$L_1 \leftarrow$ *{1-itemset frequents}*
For (*k*=2; $L_k \neq \phi$; k++) **Do**
-- Apriori_Gen is a function that calculates the itemsets of size *k* from those of size *k-1*
$C_k \leftarrow$ *Apriori_Gen* (L_{k-1})
For (*all* $t \in T$) **Do**
$C_t \leftarrow$ *Subset*(C_k,*t*) --This function select the itemsets of C_k contained in the transaction *t*
For (*all* $c \in C_t$) **Do**
support(c) $\leftarrow$ *support(c)*+*1*
End For
End For
-- Selection of the itemset verifying the support constraint
$L_k \leftarrow$ { $c \in C_k$ /*support(c)* $\geq$ *minsup*}
End For
Return $\bigcup_k L_k$
End

Algorithm 2: Apriori_Gen

Input:
- F: set of the itemsets of size k

Output:
- C: set of the itemsets of size $k+1$ deducted from the set F

Begin
$C \leftarrow \{c=f_1 \cup f_2 \,/\, (f_1, f_2) \in F \times F,\ size(c)=k+1\}$
For *(all* $c \in C$*)* **Do**
For *(all* $s \subset c$ *and size(s)=k)* **Do**
If $s \notin F$ **Then**
$C \leftarrow C-\{c\}$
End For
End For
Return C
End

Algorithm 3: Algorithm for extracting association rules from a set of frequent patterns

Input:
- F: set of frequent patterns
- *minconf:* threshold for confidence measure

Output:
- R: set of the association rules verifying the constraint of the confidence measure

Begin
$R \leftarrow \varnothing$
For *(all itemset* $FI \in F$*)* **Do**
For *(all* $SI \subset FI$ *)* **Do**
If confidence$(SI \rightarrow FI/SI) \geq minconf$ **Then**
$R \leftarrow R \cup \{SI \rightarrow FI/SI\}$
End For
End For
Return R
End

In this work, we use the Apriori algorithm to extract the association rules linking the conditions recorded during accidents and the victims of these accidents. These association rules will allow the responsible to have a very clear vision on the main causes of the accidents in France. They will help thereafter to adopt the most appropriate policies to at least reduce the number of accidents and the number of victims.

The problem encountered at this level is that the application of the Apriori algorithm, as is the case for the other algorithms, leads to a huge number of association rules, which does not allow the decision maker to choose unambiguously the most relevant rules. Therefore, we propose to use the method ELECTRE II to rank in order of importance all the association rules.

Overview of the ELECTRE II method

In this paper, we use the multicriteria aggregation ELECTRE II method [26]. Based on a family F of criteria and a set of alternatives A, this method will allow us to rank the set of association rules from the best to the worst association rule.

Motivation of the ELECTRE II method

The choice of the ELECTRE II method is justified for several reasons. Indeed, it is a multicriteria aggregation procedure belonging to the outranking approach. In addition, it is a method that's not directly compensatory, in the sense, a weak evaluation on a criterion can be compensated by a strong evaluation on another criterion. The method ELECTRE II proceeds by pairwise comparisons between alternatives. Moreover, the results are not definitive to take a decision, but they are analyzed following a step named robustness analysis. Also, the ELECTRE II method doesn't suppose any constraint on the scales of measurement of all the criteria and can be used without any problem for many cases where the criteria are measured differently on heterogeneous scales of measurement.

Finally, this method has been used successfully to solve several concrete problems of multicriteria decision [41, 42].

Basic principle the ELECTRE II method

The method ELECTRE II proceeds in two steps, in the first step, two nested outranking relations S^1 and S^2 are built, and at the second step, the two relations will be exploited to elaborate the final ranking.

In the procedure of building the relations S^1 and S^2, and in general for any outranking relation S to be built on the set A, we must make pairwise comparisons of all pairs *(a,b)* of alternatives. We confirm that "the alternative a outrank the alternative b", denoted by $a\ S\ b$, if two conditions are verified:

- a is at least as good as b for the majority of criteria,
- Without being too much worse for the other criteria.

These two conditions are called respectively condition of concordance and condition of non-discordance for the criteria with the proposition "$a\ S\ b$".

Building outranking relations

For the construction of the two outranking relations, two thresholds of concordance requirement $c_1 > c_2$ are used. These concordance thresholds express the minimum required majority of the criteria that are for the affirmation of the outranking relationship. For example, the threshold equal to 0.5 then imposes a majority of criteria that exceeds 50%.

Furthermore, we will also need, for each criterion g_j, two other thresholds of discordance $d_{1j} < d_{2j}$. If exist at least one criterion g_j, such that the difference $g_j(a)–g_j(b)$, between two alternatives a and b, exceeds the required threshold of discordance, then the relation aSb will be refused.

If we take $d_{1j} = d_{2j} = 0$, then this means that the effect of discordance will not be taken into account under criterion g_j.

The construction of relations S^1 and S^2 is given by the formula (6).

$$aS^i b \Leftrightarrow \begin{cases} \text{Condition of the concordance:} \\ C(a,b) = \dfrac{\sum_{gj(a)\geq gj(b)} wj}{\sum_{1\leq j\leq m} wj} \geq ci \\ \text{Condition of the non-discordance:} \\ \forall gj \in F / gj(b) > gj(a) : gj(b) - gj(a) \leq dji \end{cases} \tag{6}$$

With:

- $i \in \{1,2\}$
- F is the family of criteria to be considered for ranking problem
- a and b are two alternatives of the set A of the all alternatives considered
- w_j represents the weight of importance relative to criterion g_j. In case of an equal weighting, we take $w_j = 1$
- m is the number of criteria considered ($m \geq 2$)
- $g_j(a)$ and $g_j(b)$ are the respective performances of the alternative a and b on criterion g_j.

Remarks

- In the formula (6), we assume the case of the maximization of all criteria g_j. However, for the case of the minimization of a criterion g_j, it just to replace all the evaluations $g_j(a)$ by the opposite evaluations: $-g_j(a)$, then we apply the same calculations of the case the maximization.
- The first relation S^1 is called the strong outranking relation and the second S^2 is called the weak outranking relation. We prove that: $S^1 \subset S^2$. Indeed, the thresholds of requirement of S^1 are tighter than those of S^2, and one has: for every pair of alternatives a and b: if aS^1b then aS^2b. Indeed:
  - We suppose that: aS^1b, for any two alternatives a and b, then that's mean by formulas (6): $C(a,b) \geq c1$ and $\forall gj \in F / gj(b) > gj(a) : gj(b) - gj(a) \leq dj1$
  - Moreover, we know by hypothesis that: $c1 > c2$ and $d1 < d2$
  - That implies: $C(a,b) \geq c2$ and $\forall gj \in F / gj(b) > gj(a) : gj(b) - gj(a) \leq dj2$
  - Which gives by formulas (6): aS^2b
  - That finally justifies: $S^1 \subset S^2$

Exploitation of outranking relations

Once the two strong and weak outranking relations are constructed, two antagonistic pre-orders will be established, the first, named V_1, is obtained by traversing the graph, associated to the strong outranking relation S^1, from the top to the bottom. This crossing is called descendant exploration. A second pre-order, named V_2, is obtained by traversing this time the graph in the opposite direction. This crossing is called the ascendant exploration. Before the last step, a median pre-order V is deduced from the two pre-orders V_1 and V_2: $V = (V_1 + V_2)/2$.

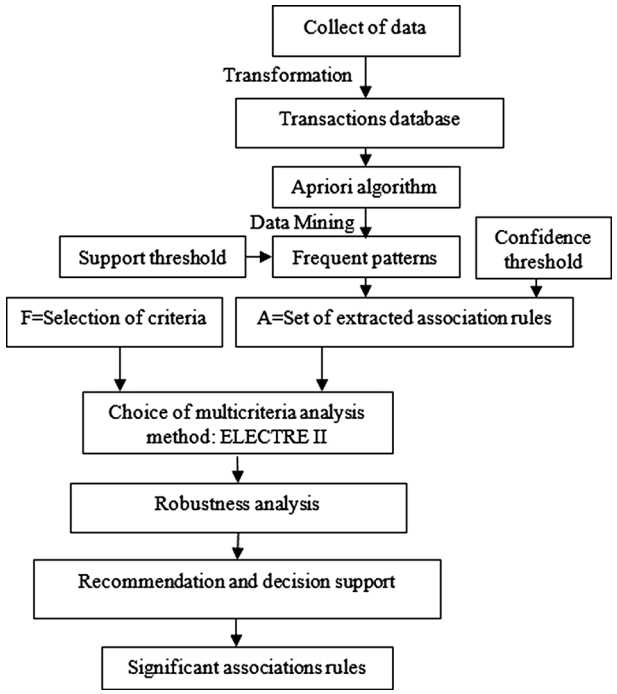

Fig. 3 Flowchart of the proposed approach

Finally, a final ranking is obtained by separating, by the relation of outranking S^2, the all equally placed of the median ranking V. For more details you can see [26].

The proposed approach

To have an adequate model to extract knowledge about the most relevant conditions that causes the majority of road accident in France, we propose an approach summarized in 5 steps, the flowchart given in Fig. 3 shows the full process. This proposed approach is described by the following steps.

- Step 1: Collection of data

The data used in our study are collected from database BAAC administered by the ONISR. This file is made from the fact sheets and contains information describing each accident. These fact sheets are elaborated by the competent authorities, such as the French National Police and French National Gendarmerie, who intervened directly at the scene of the accident.

For a first exploitation and illustration, we propose in this paper, to use the data collected on all the accidents that were occurred during the year 2016, we have 59 432 recorded accidents. The database is available online at [https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/]. It is composed with the 4 tables in csv format: Characteristics—Places—Vehicles—Users. Our first task is to create a database under the ORACLE DBMS "Data Base Management System" for a better exploitation of the data of the 4 tables, especially at their joins for the generation of transactions database.

Table 1 Attributes retained for the description of accidents

Attribute name	Possible values: items	Description
Attributes retained from the table characteristics		
num_Acc	Alphanumeric code	Accident identification number
Month	January, February, …, December	Month of the accident
hour_of_the_accident	Morning, night	Day part of the accident
Light	Full day, dawn, night with public lighting lights, night without public lighting, Night with public lighting not lit	lighting conditions in which the accident occurred
Urbain	In-urban, Extra-urban	Location
Intersection	Out of intersection, gyratory, intersection in T, intersection in X, intersection in Y, intersection with more than 4 branches, place, level crossing	Intersection
atmospheric_condition	Normal, light rain, heavy rain, snow–hail, fog–smoke, strong wind–storm, dazzling time, overcast	Atmospheric conditions
Type_of_collision	Two vehicles—frontal, two vehicles—from the rear, two vehicles—by the side, three vehicles and more—in chain, three or more vehicles—multiple collisions, other collision, without collision	The type of collision
indicator_of_source	Metropole, antilles, guyane, reunion, mayotte	Indicator of source
Attributes retained from the table places		
num_Acc	Alphanumeric code	Accident identification number
category_of_the_road	Highway, national road, departmental road, communal road, off public network, parking lot open to public traffic or other road	Category of the road
traffic_regime	One-way, bidirectional, separate pavements, with tracks assignment variable	Traffic regime
reserved_lane	Bike path, cycle bank, reserved lan, reserved_lane	Indicates the existence of a reserved lane
declivity_of_the_road	Dish, slope, peak of coast, down of coast	declivity of the road at the place of the accident
drawing_in_plan	Straight part, curved on the left, curved right, in S	Drawing in plan of the road
state_of_surface	Normal surface, wet, puddles, flooded, snow, mud, icy, oily material, other surface state	State of surface
equipment	Underground, bridge, exchanger, railway, carrefour amenaged, pedestrian area, toll area	Equipment and infrastructure
Accident_Situation	On the pavement, on emergency stop band, on the verge, on the sidewalk, On bike path	Accident situation
Attributes retained from the table vehicles		
num_Acc	Alphanumeric code	Identifier of the accident
num_veh	Alphanumeric code	Identification of the vehicle

Table 1 (continued)

Attribute name	Possible values: items	Description
category_of_vehicle	Bicycle, Moped < 50 cm^3, VL only, VL + caravan, VL + trailer, VU only 1.5 T, PL only ≤ 7.5 T, PL only > 7.5 T, PL > 3.5 T + trailer, Tractoronly, Tractor + semi-trailer, Specialmachinery, FarmTractor, Scooter < 50 cm^3, Motorcycle ≤ 125 cm^3, Scooter > 50 cm^3, Motorcycle > 125 cm^3, Scooter > 125 cm^3, LightweightQuad ≤ 50 cm^3, HeavyQuad > 50 cm^3, Bus, Coach, Train, Tramway, Othervehicle	Category of vehicle
fixed_Obstacl_struck	Vehicle parked, tree, metal crash barrier, concrete crash barrier, other crash barrier, building, vertical sign, pole, urban equipment, parapet, ilot, sidewalk border, ditch, another obstacle on the road, another obstacle on sidewalk, unobstructed causeway exit	Fixed obstacle eventually hit by vehicle
mobile_obstacle_struck	Pedestrian, vehicle, rail vehicle, domestic animal, wild animal, other mobile	Mobile obstacle eventually hit by vehicle
initial_shock_point	Before, front right, left front, rear, right back, left rear, right side, left side, multiple shocks	Initial shock point
manoeuvre	No change of direction, same direction and same file, between 2 files, In reverse, reverse direction, crossing the median, in the bus corridor in the same direction, in the bus lane in the opposite direction, by inserting, turning back on the roadway, changing lane left, changing lane right, deported left, deported right, turning left, turning right, exceeding left, exceeding right, crossing the roadway Parking maneuver, avoidance maneuver, opening of door, stopped off parking	Principal maneuver before the accident
nb_occupants_pub_trans	Integer	Number of occupants of the vehicle
Attributes retained from the table users		
num_Acc	Alphanumeric code	Identifier of the accident
num_veh	Alphanumeric code	Identification of the vehicle
place_of_user	An Integer between 1 and 9	Allows to locate the place occupied in the vehicle by the user at the time of the accident
category_of_user	Driver, passenger, pedestrian, pedestrian in rollerblade or scooter	Category of user
Severity_accident_user	Unharmed, Killed, Hospitalized wounded, Wounded light	Severity of the accident for each user
user_gender	M, F	Gender of user
age_of_user	< 20, [20, 30], [31, 45], [46, 60], > 60	Age of user
reason_displacement	Home-work, home-school, shopping, professional use, promenade, other reason for displacement	Reason for displacement the user at the time of the accident

Table 1 (continued)

Attribute name	Possible values: items	Description
security_equipment	Belt, helmet, child device, belt used, belt unused, belt indefinite use, helmet used, helmet unused, helmet indefinite use, child device used, child device unused, child device indefinite use, reflective equipment used, reflective equipment unused, reflective equipment indefinite use, other security equipment used, other security equipment unused, other security equipment indefinite use	the existence and use of security equipment
location_pedestrian	On pavement A + 50 m from the pedestrian crossing, on pavement A − 50 m from the pedestrian crossing, on pedestrian crossing without light signalling, on pedestrian crossing with light signalling, on the sidewalk, on the accoutrement, on refuge or BAU, on against aisle	Location of the pedestrian at time the accident
pedestrian_action	Moving unspecified, moving in the direction striking vehicle, moving opposite direction of the striking vehicle, crossing, hidden, playing-current, with animal, other pedestrian action, on against aisle	Pedestrian action at time the accident
accompanying_pedestrian	Alone, accompanied, in a group	This variable indicates if the injured pedestrian was alone or not

- Step 2: Transformation of the data for construction the set of transactions: T

In the second step, under the ORACLE DBMS, we proceeded to the joining of the 4 tables to have a global vision on all the accidents. For that, we obtained a descriptive file with various information on the characteristics of the accident, accident places, information on vehicles crash in an accident and users (drivers and pedestrians). Then, we transformed the result thus obtained into a transactions table in csv format, which will be used as input for the Apriori algorithm. Each transaction contains all the information on the accident concerned, and 38 attributes were selected for the description of each accident. All attributes with their possible values and descriptions are provided in Table 1. All the 38 retained attributes are divided into 4 main tables:

- Characteristics: This table describes the general circumstances of the accident. There are 9 attributes that are retained with 59 432 records.
- Places: This table describes the location where the accident occurred. There are also 9 attributes that are retained with 59 432 places described.
- Vehicles: The table describes the vehicles involved in the accident. There are also 8 attributes that are used for this table and we have 101 924 vehicles recorded.
- Users: This table gives all the information on the users involved in the accidents, a user can be a driver or a pedestrian. This table contains 12 attributes with 133 422 users registered.

Table 2 Experimentation of several thresholds of confidence and support

Confidence threshold	Support threshold	Number of rules extracted	Execution time (in seconds)
0.5	0.5	324	0.54
0.5	0.6	66	0.51
0.5	0.7	10	0.45
0.5	0.8	0	0.36
0.5	0.9	0	0.32
0.6	0.5	321	0.5
0.6	0.6	66	0.57
0.6	0.7	10	0.45
0.6	0.8	0	0.37
0.6	0.9	0	0.33
0.7	0.5	290	0.56
0.7	0.6	63	0.49
0.7	0.7	10	0.43
0.7	0.8	0	0.35
0.7	0.9	0	0.36
0.8	0.5	185	0.69
0.8	0.6	46	0.51
0.8	0.7	9	0.51
0.8	0.8	0	0.37
0.8	0.9	0	0.28
0.9	0.5	77	0.5
0.9	0.6	20	0.49
0.9	0.7	5	0.44
0.9	0.8	0	0.39
0.9	0.9	0	0.31
1	0.5	0	0.53
1	0.6	0	0.5
1	0.7	0	0.49
1	0.8	0	0.36
1	0.9	0	0.31

The possible values taken by each attribute will represent the items for the set I. There are 220 of items in total.

- Step 3: Application of Apriori algorithm for exaction of all association rules

In this step, we use the Apriori algorithm implemented under the IDE RStudio for the extraction of all possible association rules from the transactions table. We have retained of course only the association rules respecting the thresholds of confidence and support. For our case, we did the experiment with several thresholds (50%, 60%, 70%, 90% and 100%) (see Table 2).

According to Table 2, the choice of thresholds minconf = 0.9 and minsup = 0.7 leads to the extraction of a minimum of association rules, a total of 5 association rules, which seems a more logical choice of the thresholds. But these 5 rules are extracted only at the base of the two measures support and confidence. In this first experiment, the thresholds minconf = 0.8 and minsup = 0.6 are chosen. Indeed, these choices

allow the Aprirori algorithm to extract an average of 46 association rules. The ELECTRE III method will be used to choose the best rules among this set of 46 association rules, which seems richer than selection in a set of 5 association rules. This choice must in no way alter the final result, because the 5 association rules, extracted for the choices minconf = 0.9 and minsup = 0.7, are implicitly included in the set of association rules extracted for the choices minconf = 0.8 and minsup = 0.6. What makes these 5 rules will also be examined by the ELECTRE III method with the rest of the other association rules in order to find the best association rules.

- Step 4: Application of ELECTRE II for the selection of the most relevant association rules

For decision making, the two measures of confidence and support are not sufficient to extract the most significant association rules from the set of transactions. Indeed, it is also necessary to consider the measure of independence for each association rule. A third criterion was added which expresses this independence, it is the *lift* quality measure, previously introduced at the formula (5).

Based on the three criteria thus retained (confidence, support and lift), the method MCDA ELECTRE II is chosen to rank, in order of importance, the association rules result of step 3.

The assumptions of the ELECTRE II method are:

- A consistent F family of criteria, which are in our application: $F = \{g_1$*:support,* g_2*: confidence,* g_3*:lift*$\}$
- A set of alternatives, which are A = Set of association rules obtained by the Apriori algorithm.
- The concordance thresholds $c_1 > c_2 > 0.5$
- The discordance thresholds for each criterion g_j: $0 < d_{1j} < d_{2j}$.
- The weight w_j of each criterion g_j, for: $1 \leq j \leq 3$.
- Step 5: Recommendation of the final solution

At the last step, according to the classification of the association rules obtained by the ELECTRE II method, we recommend to the decision-maker the most relevant associations in order to define, for example, the security measures to be considered in the short-term, median-term or in the long-term.

In practice, for a better recommendation with a great satisfaction of the decision-maker, a robustness analysis of the results is always desirable. This analysis consists to check the stability of the results, by varying substantially the parameters of the method such as the concordance and discordance thresholds.

Discussion and evaluation

In this section, we present the result of the construction of the transactions table, then we show the extraction result of the association rules obtained by the Apriori algorithm, and we give the ranking of association rules by application of the ELECTRE II method. Finally, a robustness analysis and discussions of the results will be provided. As algorithm

Table 3 The table of experiment environment

Software	Node environment
The database management system Database Oracle XE IDE RStudio MCDA- Ulaval version 0.6.1	ProBook i5-6200U Single station PC

	A	B	C	D	E	F	G	H	I	J
1	***Month***	***Hour***	***Light***	***Age_user***	***User_gender***	***Severity_accident_user***	***Urbain***	***Intersection***	***Atmospheric_conditior***	***Type_of_collision***
2	February	Afternoon	Fullday	[31–45]	F	Unharmed	In-urban	Out of intersec	Covered time	Two vehicles_byside
3	February	Afternoon	Fullday	<20	M	Hospitalized wounded	In-urban	Out of intersec	Covered time	Two vehicles_byside
4	March	Evening	Fullday	[46–60]	M	Hospitalized wounded	In-urban	Gyratory	Normal ATM	Other collision
5	March	Evening	Fullday	<20	M	Hospitalized wounded	In-urban	Gyratory	Normal ATM	Other collision
6	March	Evening	Fullday	[46–60]	F	Hospitalized wounded	In-urban	Gyratory	Normal ATM	Other collision
7	July	Evening	Fullday	[20–30]	M	Unharmed	Extra-urban	Out of intersec	Normal ATM	Other collision
8	July	Evening	Fullday	[46–60]	M	Hospitalized wounded	Extra-urban	Out of intersec	Normal ATM	Other collision
9	August	Evening	Dawn	<20	M	Hospitalized wounded	In-urban	Out of intersec	Dazzling time	Two vehicles_byside
10	August	Evening	Dawn	<20	M	Wounded light	In-urban	Out of intersec	Dazzling time	Two vehicles_byside
11	August	Evening	Dawn	[31–45]	F	Unharmed	In-urban	Out of intersec	Dazzling time	Two vehicles_byside
12	December	Morning	Fullday	>60	F	Hospitalized wounded	In-urban	Intersection in	Normal ATM	Two vehicles_byside

Fig. 4 Table of transactions *T*

of extraction of the association rules we opted for the algorithm Apriori of the IDE RStudio. Our entire experimental environment is given in Table 3.

We experiment our proposed approach using road accident database of France.

Building the transactions table T

The transactions table is obtained by merging the four tables characteristics, places, vehicles and users, with an SQL join. The SQL integral code used to create the table of transactions *T* is attached to this paper (Additional file 1).

The table of transactions *T* is shown in Fig. 4. After joining the 4 tables, we get exactly 133,422 transactions, and each transaction is described in maximum by 220 items. Due to the size of the table of transactions, we present only an excerpt of this table. A copy of this table is attached to this paper (Additional file 2).

Application of the Apriori algorithm

In the second step, we apply the Apriori algorithm to generate all association rules. A first selection was made to retain only the important association rules, setting a confidence threshold at 80% and a support threshold at 60%. Finally, Table 4 gives the 46 rules resulting of this step.

Remarks:

- In this work, it's a first application of our proposed approach, we intend to apply it to a database on the road circulation spread over several years. In this case the data will have a very varied distribution. This will lead to measures of lift for a better distribution.
- In general, the circular and partially symmetrical rules, do not present any inconvenience in the operation of extraction of knowledge, in fact, this type of rules shows that there is a mutual correlation between the premises and the conclusions of these rules, that leads to a richer relation like the equivalence relation.

Table 4 The extracted association rules

	Premise (LHS)	Conclusion (RHS)	Support	Confidence	Lift
R1	{Mobile_obstacle_struck = Vehicle}	{Pedestrian_action = Moving unspecified}	0.62	1.00	1.09
R2	{Category_of_vehicle = VLonly}	{Pedestrian_action = Moving unspecified}	0.60	0.91	0.99
R3	{Light = Fullday}	{Pedestrian_action = Moving unspecified}	0.63	0.91	1.00
R4	{Intersection = Out of intersection}	{Pedestrian_action = Moving unspecified}	0.64	0.91	1.00
R5	{User_gender = M}	{Pedestrian_action = Moving unspecified}	0.65	0.93	1.02
R6	{Category_of_user = Driver}	{Pedestrian_action = Moving unspecified}	0.74	1.00	1.09
R7	{Pedestrian_action = Moving unspecified}	{Category_of_user = Driver}	0.74	0.81	1.09
R8	{declivity_of_the_road = Dish}	{Drawing_in_plan = Straight part}	0.62	0.84	1.11
R9	{Drawing_in_plan = Straight part}	{declivity_of_the_road = Dish}	0.62	0.82	1.11
R10	{declivity_of_the_road = Dish}	{Atmospheric_condition = Normal ATM}	0.60	0.81	1.01
R11	{declivity_of_the_road = Dish}	{Accident_Situation = On the pavement}	0.65	0.87	1.02
R12	{declivity_of_the_road = Dish}	{Pedestrian_action = Moving unspecified}	0.67	0.91	0.99
R13	{State_of_surface = Normal surface}	{Drawing_in_plan = Straight part}	0.61	0.80	1.06
R14	{State_of_surface = Normal surface}	{Atmospheric_condition = Normal ATM}	0.72	0.95	1.18
R15	{Atmospheric_condition = Normal ATM}	{State_of_surface = Normal surface}	0.72	0.89	1.18
R16	{State_of_surface = Normal surface}	{Accident_Situation = On the pavement}	0.65	0.87	1.02
R17	{State_of_surface = Normal surface}	{Pedestrian_action = Moving unspecified}	0.69	0.91	1.00
R18	{Drawing_in_plan = Straight part}	{Atmospheric_condition = Normal ATM}	0.62	0.82	1.01
R19	{Drawing_in_plan = Straight part}	{Accident_Situation = On the pavement}	0.67	0.88	1.04
R20	{Drawing_in_plan = Straight part}	{Pedestrian_action = Moving unspecified}	0.69	0.91	0.99
R21	{indicator_of_origin = Metropole}	{Atmospheric_condition = Normal ATM}	0.62	0.80	1.00
R22	{indicator_of_origin = Metropole}	{Accident_Situation = On the pavement}	0.65	0.84	0.99
R23	{indicator_of_origin = Metropole}	{Pedestrian_action = Moving unspecified}	0.71	0.92	1.00
R24	{Atmospheric_condition = Normal ATM}	{Accident_Situation = On the pavement}	0.69	0.85	1.00
R25	{Accident_Situation = On the pavement}	{Atmospheric_condition = Normal ATM}	0.69	0.81	1.00
R26	{Atmospheric_condition = Normal ATM}	{Pedestrian_action = Moving unspecified}	0.74	0.91	1.00
R27	{Pedestrian_action = Moving unspecified}	{Atmospheric_condition = Normal ATM}	0.74	0.81	1.00
R28	{Accident_Situation = On the pavement}	{Pedestrian_action = Moving unspecified}	0.78	0.91	1.00
R29	{Pedestrian_action = Moving unspecified}	{Accident_Situation = On the pavement}	0.78	0.85	1.00

Table 4 (continued)

	Premise (LHS)	Conclusion (RHS)	Support	Confidence	Lift
R30	{Atmospheric_condition = Normal ATM;Category_of_user = Driver}	{Pedestrian_action = Moving unspecified}	0.60	1.00	1.09
R31	{Category_of_ user = Driver;Pedestrian_ action = Moving unspecified}	{Atmospheric_condition = Normal ATM}	0.60	0.81	1.01
R32	{Atmospheric_condition = Normal ATM;Pedestrian_action = Moving unspecified}	{Category_of_user = Driver}	0.60	0.82	1.10
R33	{Accident_Situation = On the pavement;Category_of_ user = Driver}	{Pedestrian_action = Moving unspecified}	0.64	1.00	1.09
R34	{Category_of_ user = Driver;Pedestrian_ action = Moving unspecified}	{Accident_Situation = On the pavement}	0.64	0.86	1.01
R35	{Accident_Situation = On the pavement;Pedestrian_ action = Moving unspecified}	{Category_of_user = Driver}	0.64	0.82	1.10
R36	{Atmospheric_condition = Normal ATM;State_of_surface = Normal surface}	{Accident_Situation = On the pavement}	0.62	0.87	1.02
R37	{State_of_surface = Normal surface;Accident_Situation = On the pavement}	{Atmospheric_condition = Normal ATM}	0.62	0.95	1.18
R38	{Atmospheric_condition = Normal ATM;Accident_Situation = On the pavement}	{State_of_surface = Normal surface}	0.62	0.90	1.20
R39	{Atmospheric_condition = Normal ATM;State_of_surface = Normal surface}	{Pedestrian_action = Moving unspecified}	0.65	0.91	1.00
R40	{State_of_surface = Normal surface;Pedestrian_action = Moving unspecified}	{Atmospheric_condition = Normal ATM}	0.65	0.95	1.18
R41	{Atmospheric_condition = Normal ATM;Pedestrian_action = Moving unspecified}	{State_of_surface = Normal surface}	0.65	0.89	1.18
R42	{Drawing_in_plan = Straight part;Accident_Situation = On the pavement}	{Pedestrian_action = Moving unspecified}	0.61	0.91	0.99
R43	{Drawing_in_plan = Straight part;Pedestrian_action = Moving unspecified}	{Accident_Situation = On the pavement}	0.61	0.88	1.04
R44	{Atmospheric_condition = Normal ATM;Accident_Situation = On the pavement}	{Pedestrian_action = Moving unspecified}	0.63	0.92	1.00
R45	{Atmospheric_condition = Normal ATM;Pedestrian_action = Moving unspecified}	{Accident_Situation = On the pavement}	0.63	0.85	1.00
R46	{Accident_Situation = On the pavement;Pedestrian_ action = Moving unspecified}	{Atmospheric_condition = Normal ATM}	0.63	0.81	1.00

Application of the ELECTRE II method

In the proposed approach, we have chosen the method ELECTRE II to rank the 46 association rules obtained by application of the Apriori algorithm. The ELECTRE II method starts from the construction of the decision matrix, which contains all the considered

Table 5 Matrix 46 × 3 of the decision

Alternatives	Criteria		
	Criterion1: support	Criterion2: confidence	Criterion3: lift
	MAX	MAX	MAX
	w1 = 1	w2 = 1	w3 = 1
R1	0.62	1.00	1.09
R2	0.60	0.91	0.99
R3	0.63	0.91	1.00
R4	0.64	0.91	1.00
…	…	…	…
R46	0.63	0.81	1.00

MAX: means the criterion must be maximized

criteria (support, confidence, lift) with their evaluations. An extract from the decision matrix is shown in Table 5. The first step of ELECTRE II consists to calculate the matrix of the concordance indices $C(R_i, R_k)$ for all pairs $(R_i R_k)$ of association rules (Additional file 3). An extract from this matrix is given at Table 6.

The remain preferences of the three criteria are given in Table 4.

In this work, we consider the three criteria at the same level of importance. For this, we choose for each criterion a weight equal to 1.

The second step for the ELECTRE II is now to construct the outranking relations S^1 et S^2. For concordance and discordance thresholds, we set $c_1 = 0.8$, $c_2 = 0.6$ and $c_0 = 0.7$ (average of the concordance c_1 and c_2), $d2 = 80\%$ of extent for each criterion and $d1 = 60\%$ of extent for each criterion g_j ($extent(gj) = Max(g_j(a)) – Min(g_j(a))$) (see Table 7).

Once all the thresholds are known, the ELECTRE II method proceeds to the construction of outranking S^1 and S^2 from which we deduce the three rankings: the ascending pre-order, the pre-order descending and the median pre-order, Fig. 5. Finally, the median ranking V of the target association rules is obtained by taking the average of the descendant $V1$ and ascendant $V2$ rankings (see Table 8).

Note that all calculations and verifications are done by the tool of the Multi-Criteria Decision Aid MCDA-Ulaval [http://cersvr1.fsa.ulaval.ca].

Robustness analysis

In MCDA, the robustness analysis is always recommended. This robustness analysis consists to find and validate the most stable and robust solutions. It is obtained by appropriate variations of certain parameters of the method, such as the concordance and discordance thresholds. In this study, the rate of stability in the evaluation of choices was examined by modifying the c_1, c_0, c_2, d_{j1} and d_{j2} parameters for the ELECTRE II method. All the results of the robustness analysis are given in Table 9. This table shows that the alternatives R6, R14 and R28 occupy the top of the majority of the rankings calculated for the different choices of the concordance and discordance thresholds.

Table 6 Matrix 46 × 46 of the concordance indices

[Alternative]	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20 …
R1	~	0.67	0.33	0.33	0.33	0.33	0.33	0.67	0.67	0.67	0.33	0.33	0.67	0.33	0.33	0.33	0.33	0.67	0.33	0.33
R2	0.33	~	0.67	0.67	0.33	0.33	0.67	0.67	0.67	1.00	0.67	0.67	0.67	0.33	0.67	0.67	0.67	0.67	0.67	0.67
R3	0.67	0.67	~	0.33	0.33	0.33	0.67	1.00	1.00	1.00	0.67	0.33	1.00	0.33	0.67	0.67	0.67	1.00	0.67	0.33
R4	0.67	1.00	1.00	~	0.33	0.33	0.67	1.00	1.00	1.00	0.67	0.67	1.00	0.33	0.67	0.67	0.67	1.00	0.67	0.67
R5	0.67	0.67	0.67	0.67	~	0.33	0.67	1.00	1.00	0.67	1.00	0.33	1.00	0.33	0.67	1.00	0.33	0.67	0.67	0.33
R6	0.67	0.67	0.67	0.67	0.67	~	0.67	1.00	1.00	0.67	0.67	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67	0.67
R7	0.67	0.33	0.33	0.33	0.33	0.67	~	0.67	0.67	0.67	0.33	0.33	0.67	0.67	0.67	0.33	0.33	0.33	0.33	0.33
R8	0.67	0.33	0.00	0.00	0.00	0.00	0.33	~	1.00	0.67	0.00	0.00	0.67	0.33	0.33	0.00	0.00	0.67	0.00	0.00
R9	0.67	0.33	0.00	0.00	0.00	0.00	0.33	0.67	~	0.67	0.00	0.00	0.67	0.33	0.33	0.00	0.00	0.67	0.00	0.00
R10	0.33	0.33	0.00	0.00	0.33	0.33	0.67	0.33	0.33	~	0.33	0.00	0.67	0.33	0.33	0.33	0.00	0.33	0.33	0.00
…																				

Table 7 The discordance thresholds for $d^2 = 80\%$ and $d^1 = 60\%$

	Criterion 1: support Extent = 0.18	Criterion 2: confidence Extent = 0.20	Criterion 3: support Extent = 0.21
d_1	$60\% \times 0.18 = 0.108$	$60\% \times 0.20 = 0.12$	$60\% \times 0.21 = 0.126$
d_2	$80\% \times 0.18 = 0.144$	$80\% \times 0.20 = 0.16$	$80\% \times 0.21 = 0.168$

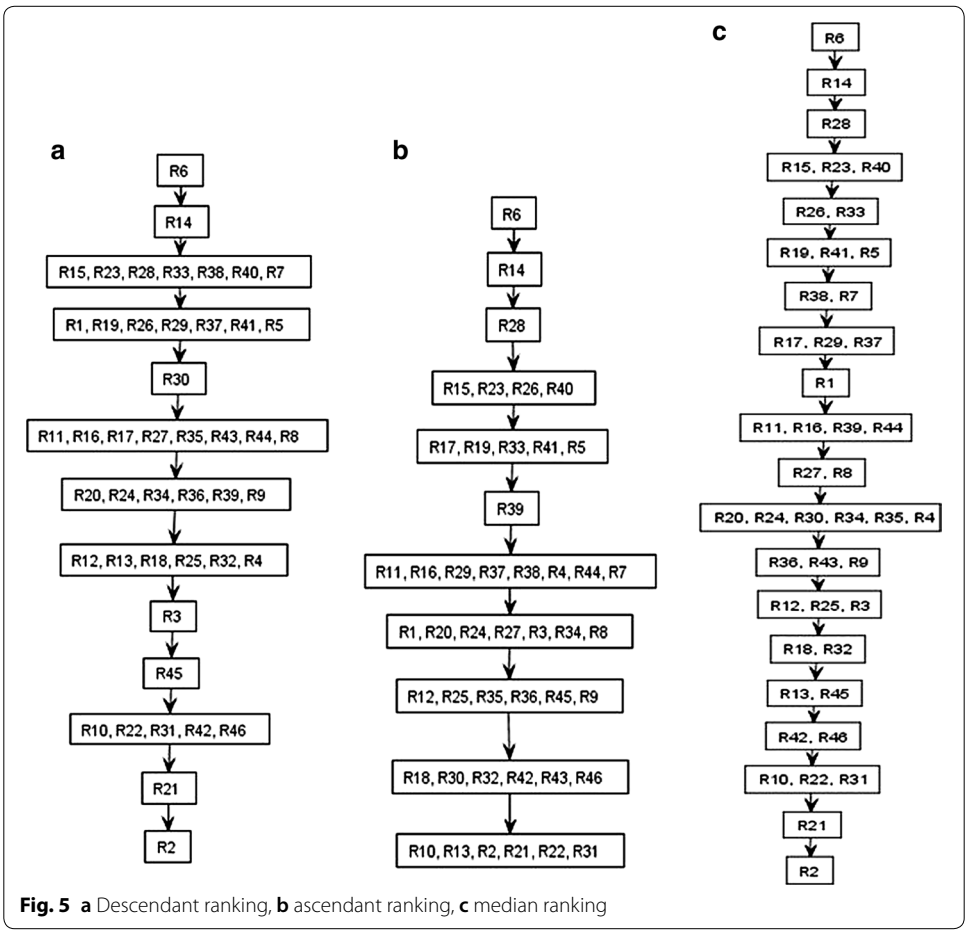

Fig. 5 **a** Descendant ranking, **b** ascendant ranking, **c** median ranking

Discussion

In this first exploitation of the BAAC database and after the robustness analysis, summarized in Table 9, we conclude that the three association rules: R6, R14 and R28 occupy the first ranks of all rankings calculated by ELECTRE II. In addition, these rules R6 (support = 0.74, confidence = 1, lift = 1.09), R14 (support = 0.72, confidence = 0.75, lift = 1.18) and R28 (support = 0.78, confidence = 0.91, lift = 1) have a significant lift, the lift is greater than 1, which shows that a positive correlation exists between the premises and the conclusions of these of association rules. These three rules deserve a special interest and will therefore be retained for further analysis. These association rules are as follows:

Table 8 ELECTRE II method, descendant, ascendant, median and final ranking

Rank	Descendant ranking: V_1	Ascendant ranking: V_2	Median ranking: V
1	R6	R6	R6
2	R14	R14	R14
3	[R7, R15, R23, R28, R33, R38, R40]	R28	R28
4	[R1, R5, R19, R26, R29, R37, R41]	[R15, R23, R26, R40]	[R15, R23, R40]
5	R30	[R5, R17, R19, R33, R41]	[R26, R33]
6	[R8, R11, R16, R17, R27, R35, R43, R44]	R39	[R5, R19, R41]
7	[R9, R20, R24, R34, R36, R39]	[R4, R7, R11, R16, R29, R37, R38, R44]	[R7, R38]
8	[R4, R12, R13, R18, R25, R32]	[R1, R3, R8, R20, R24, R27, R34]	[R17, R29, R37]
9	R3	[R9, R12, R25, R35, R36, R45]	R1
10	R45	[R18, R30, R32, R42, R43, R46]	[R11, R16, R39, R44]
11	[R10, R22, R31, R42, R46]	[R2, R10, R13, R21, R22, R31]	[R8, R27]
12	R21		[R4, R20, R24, R30, R34, R35]
13	R2		[R9, R36, R43]
14			[R3, R12, R25]
15			[R18, R32]
16			[R13, R45]
17			[R42, R46]
18			[R10, R22, R31]
19			R21
20			R2

- R6: {Category_of_user = Driver} → {Pedestrian action = Moving unspecified}
- R14: {State_of_surface = Normal surface} → {Atmospheric_condition = Normal ATM}
- R28: {Accident_Situation = On the pavement} → {Pedestrian_action = Moving unspecified}

As a first ascertainment, these association rules reveal that there are narrow correlations between the following items: "Driver" with "Pedestrian", "Normal Surface of the road" with "Normal Atmospheric Condition" and the "On the pavement with "Moving unspecified of the pedestrian". Moreover, these same items appear for the majority of the 49 rules of the Table 4, which prove, in addition to the correlations, that they are very frequent in the database of accidents: (1) "Pedestrian_action = Moving unspecified" appears in 30 rules, (2) "Atmospheric_condition = Normal ATM" appears in 21 rules, (3) "State_of_surface = Normal surface" appears in 11 rules and (4) "Accident_Situation = On the pavement" appears in 19 rules.

These correlations show that the majority of road accident victims are unfortunately pedestrians. Worse still, most of these pedestrians were hit by vehicles on the pavement. In addition, the majority of French drivers cause more accidents in normal roads and normal weather conditions; this can be justified by the excess speed in these best atmospheric and road conditions, more by the increase in driver travel, especially for leisure travel. Moreover, hitting a pedestrian on the pavement can only be justified by the imprudence and the carelessness of the drivers. This imprudence is at the origin of several reasons: the very young age of some drivers, the abuse of alcohol and drugs, using the phone while driving, etc. There is a third reason related to the quality of road

Table 9 Results of the robustness analysis

	Case1a	Case1b	Case1c	Case2a	Case2b	Case1c
c_1, c_0, c_2	0.6, 0.7, 0.8			0.7, 0.8, 0.9		
d_1, d_2	50% 70%	60% 80%	70% 90%	50% 70%	60% 80%	70% 90%
Median ranking						
1	R6	R6	R6	[R6, R14, R28]	[R6, R14, R28]	[R6, R14, R28]
2	R14	R14	R14	[R23, R26, R40]	[R23, R26, R40]	[R23, R26, R40]
3	R28	R28	R40	R15	R15	R15
4	[R15, R23, R40]	[R15, R23, R40]	R28	[R5, R17, R33, R38]	[R5, R17, R33, R38]	[R5, R17, R33, R38]
5	[R26, R33]	[R26, R33]	[R15, R23]	[R7, R19, R29, R41]	[R7, R19, R29, R41]	[R7, R19, R29, R41]
6	[R5, R19, R41]	[R5, R19, R41]	[R5, R26]	[R37, R39]	[R37, R39]	[R37, R39]
7	[R7, R38]	[R7, R38]	R33	[R1, R11, R16, R27, R44]	[R1, R11, R16, R27, R44]	[R1, R11, R16, R27, R44]
8	[R17, R29, R37]	[R17, R29, R37]	[R7, R17, R19, R38, R41]	[R4, R8, R20, R24]	[R4, R8, R20, R24]	[R4, R8, R20, R24]
9	R1	R1	[R29, R37]	[R34, R35]	[R34, R35]	[R34, R35]
10	[R11, R16, R39, R44]	[R11, R16, R39, R44]	[R1, R11, R16, R39, R44]	[R3, R9, R12, R25, R30, R36, R43]	[R3, R9, R12, R25, R30, R36, R43]	[R3, R9, R12, R25, R30, R36, R43]
11	[R8, R27]	[R8, R27]	[R8, R27]	[R18, R32, R45]	[R18, R32, R45]	[R18, R32, R45]
12	[R4, R20, R24, R30, R34, R35]	[R4, R20, R24, R30, R34, R35]	[R4, R20, R24, R34, R35]	[R13, R22, R42]	[R13, R22, R42]	[R13, R22, R42]
13	[R9, R36, R43]	[R9, R36, R43]	[R9, R30, R36, R43]	[R10, R31, R46]	[R10, R31, R46]	[R10, R31, R46]
14	[R3, R12, R25]	[R3, R12, R25]	[R3, R12, R25]	R2	R2	R2
15	[R18, R32]	[R18, R32]	[R18, R32]	R21	R21	R21
16	[R13, R45]	[R13, R45]	[R13, R45]	[R6, R14, R28]		
17	[R42, R46]	[R42, R46]	[R42, R46]			
18	[R10, R22, R31]	[R10, R22, R31]	[R10, R22, R31]			
19	R21	R21	R21			
20	R2	R2	R2			

infrastructure, but as a developed country, such as France, this problem is not posed with great intensity given the good quality of the state of the road infrastructure.

These results are very promising and fruitful. However, they need to be further analyzed, with the help of road traffic experts, in order to make the decisions and appropriate measures to improve road safety.

Concerning the performance of the proposed approach, we opted for the Apriori algorithm, implemented in the IDE RStudio, as stated earlier, which is considered as one of the most efficient algorithms [8]. Furthermore, in numerical experiments, only a few seconds of the order of the 0.42 s, in average, are necessary to extract all association rules from the database of 133,422 transactions (see Table 2). However, when we dispose a very large number of transactions, for example more than one million transactions, the FP-Growth algorithm with the cluster computing Apache Spark can be used. In addition, the question of the performance with the method ELECTRE II does not arise, because the method is used to rank only 46 association rules called the indifference threshold, can be considered without influence for the comparison of two association rules. In addition, to confirm that an association rule has a positive correlation, only if its lift is strictly exceeding the unity of a given preference threshold

In the ELECTRE methods, the indifference relation is considered, for example for a given criterion, such as lift measurement, a difference which is equal to 0.01, called the indifference threshold, can be considered without influence for the comparison of two association rules. In addition, to confirm that an association rule has a positive correlation, only if its lift is strictly exceeding the unity of a given preference threshold. However, the version II of the ELECTRE methods, used in this work, does not take into account this threshold of indifference and the preference threshold. In future work, it will be planned to use the version III of the ELECTRE methods more complete than the version II and which deploys three relations: indifference, weak preference and strict preference [43].

Conclusion

The data mining algorithms provide an important solution for extracting association rules. Nevertheless, these algorithms generate a big number of rules, which make a difficulty for the decision makers to express their own choice of the most interesting association rules. To solve this problem, the idea of multi-criteria decision analysis integration approach can be very useful for the decision makers who are suffering from many extracted associations rules.

This article examines and discusses the problem of road safety in France, which remains disturbing given the considerable number of victims recorded annually. Our contribution to this problem was to help the decision makers extract relevant knowledge in the form of association rules. Furthermore, the integration of multi-criteria decision analysis solved the problem of the big number of extracted association rules by selecting only the most interesting. At this stage, the proposed approach has adopted the ELECTRE II method.

The obtained results in this paper, show that our proposed approach easily leads to relevant associations rules, which can help decision-makers to analyze the significant relationships existing between accident characteristics in BAAC database. These analyzes

will be useful for adopting appropriate strategies and policies to improve road safety in France.

The first conclusions of the results show that the excess speed and carelessness of the drivers is the main cause of the accidents. In addition, the pedestrians were the first vulnerable victims of these accidents.

Reducing the number of serious injuries therefore requires policies that focus on the protection of vulnerable users, such as pedestrians. In general, the improvement of the road safety needs mainly to work more intensively on the behavioral side of the road users (drivers and pedestrians): the education, the prevention and the repression. Moreover, the self-driving and connected vehicles, communicating with one another (V2V technology) and with the road infrastructure (V2I technology), are a subject of extensive research nowadays and are expected to revolutionize the automotive industry in the near future and can contribute significantly to the improvement of road safety. Indeed, the sensory equipment instilled in new connected vehicles will become one of the key features of accident prevention, in addition, drivers will receive the latest reports on potentially dangerous road and weather conditions, upcoming collisions or diversions in their direct vicinity and any other need-to-know information will be sent straight to their dashboard.

The principal perspective related to the selection of association rules based on the data collected during only 1 year, is still insufficient to establish solid road safety analysis and policies. So, it is planned to use a large database on road accidents, that is spread over many years.

The next steps for this present work will be devoted to the application of the proposed approach to the Moroccan case, which turns out to be a more serious problem than the case of France.

The proposed approach will be improved by taking into account other measures of the quality of association rules. Moreover, the BAAC database should be enriched by other relevant information that can help to make more complete analyzes, such as information on the results of alcohol and drug testing. These improvements can contribute to finding immediate and more obvious association rules for a direct exploitation by the road safety investigators.

Conclusively, in future work, the approach proposed will be expanded and applied in the context of big data using the cluster computing Apache Spark with FP-Growth parallel algorithm, to build a predictive model for road safety.

Additional files

Additional file 1. SQL code for create the table of transactions.

Additional file 2. The integral table of transactions *T*.

Additional file 3. The integral matrix of concordance indices.

Abbreviations

AR: association rule; BACC: Accident Analysis Bulletin Corporal; Conf: confidence measure; DBMS: Data Base Management System; DGGN: Directorate General of the National Gendarmerie; DGPN: General Directorate of the National Police; DM: data mining; ELECTRE: ELimination and Choice Translating Reality; GDP: gross domestic product; IDE: integrate development environment; KDD: knowledge discovery in databases; LHS: left-hand-side; MCDA: multi-criteria decision

analysis; Minconf: minimum confidence; Minsupp: minimum support; ONISR: National Inter-Ministerial Observatory of Road Safety; RHS: right-hand-side; SETRA: Technical Studies Department of Roads and Highways; Supp: support measure; SQL: structured query language.

Authors' contributions
All mentioned authors contribute in the elaboration of the paper. All authors read and approved the final manuscript.

Authors' information
Fatima Zahra El Mazouri is a Ph.D. student in Computer Science at the Faculty of Sciences and Technologies, Sidi Mohamed Ben Abdellah University, Morocco. She received his master's degree in computer science from the Sidi Mohamed Ben Abdellah University 2015. She is actively engaged in research on various aspects of information technologies ranging from multicriteria decision analysis methods and data mining algorithms to big data, fuzzy logic, and machine learning.

Mohammed Chaouki ABOUNAIMA received the PH.D. degree in Computer Science from the Mohammed V University, Rabat, Morocco, in 1997. He is currently a professor at the Department of Computer Engineering, Faculty of Science and Technology Fez Morocco. He is a member at the Laboratory of Intelligent System & Applications. His research interests lie in computer science, data mining, operational research, information systems and multicriteria decision analysis.

Khalid Zenkouar received his Ph.D. degree in Computer Science from Faculty of Sciences, University Sidi Mohamed Ben Abdellah, Fez, Morocco in 2006. Now he is a professor at the Department of Computer Engineering, Faculty of Science and Technology Fez Morocco. He is a member in the Laboratory of Intelligent System & Applications. His current research interests include image analysis, machine intelligence and pattern recognition, and he is also involved in the multi-criteria decision making.

Acknowledgements
The authors thank the anonymous reviewers for their helpful suggestions and comments.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The road accident database known in France and referred in this paper is available online at the government website: https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 30 August 2018 Accepted: 18 December 2018
Published online: 31 January 2019

References
1. Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: an overview. AI Magazine. 1992;13(3):57.
2. Dutta I, Dutta S, Raahemi B. Detecting financial restatements using data mining techniques. J Expert Syst Appl. 2017;90:374–93. https://doi.org/10.1016/j.eswa.2017.08.030.
3. Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. Comput Educ. 2017. https://doi.org/10.1016/j.compedu.2017.05.007.
4. Neesha J, Nuraini AR, Wahidah H. Data mining in healthcare—a review. Procedia Computer Science. 2015;72:306–13.
5. Feno DR. Mesures de qualité des règles d'association: normalisation et caractérisation des bases. PhD thesis of the University of Reunion. France. 2007.
6. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of SIGMOD'93; 1993. p. 207–16.
7. Pasquier N. Data Mining: Algorithmes d'Extraction et de Réduction des règles d'Association dans les Bases de Données. PhD thesis of Clermont Ferrand II University. France. 2000.
8. Ait-Mllouk A, Agouti T, Gharnati F. Comparative survey of association rule mining algorithms based on multiple criteria decision analysis approach. In: 3rd international conference on control, engineering & information technology (CEIT). 2015. p. 1–6.
9. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, _éd.par Bocca Jorge B, Jarke Matthias and Zaniolo Carlo- Morgan Kaufmann; 1994. p. 487–99.

10. Heckerman D, Mannila H, Pregibon D, Uthurusamy R, Park M. Algorithms for fast discovery of association rules. In: 3rd Intl. conf. on knowledge discovery and data mining. AAAI Press; 1997. p. 283–96.
11. Sanmiquel L, Rossell JM, Vintr C. Study of Spanish mining accidents using data mining techniques. Saf Sci. 2015;75:49–55.
12. Mirabadi A, Sharifian S. Application of association rules in Iranian railways (RAI) accident data analysis. Saf Sci. 2010;48(10):1427–35.
13. Kumar S, Toshniwal D. A data mining framework to analyze road accident data. J Big Data. 2015;2:26.
14. El Mazouri FZ, Abounaima MC, Zenkouar K. A selection of useful patterns based on multi-criteria analysis approach. In: ACM proceeding of the international conference on computing wireless & communication systems, ICCWCS'17, Larache Morocco, 14–16 Nov 2017. https://doi.org/10.1145/3167486.3167517.
15. Ait-Mlouk A, Gharnati F, Agouti T. An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety. Eur Transp Res Rev. 2017;9:40. https://doi.org/10.1007/s12544-017-0257-5.
16. Chong M, Abraham A, Paprzycki M. Traffic accident analysis using decision trees and neural networks. In: Isaias P, et al., editors. IADIS international conference on applied computing, vol. 2. Portugal: IADIS Press; 2004. p. 39–42.
17. Kuhnert PM, Do KA, McClure R. Combining nonparametric models with logistic regression: an application to motor vehicle injury data. Comput Stat Data Anal. 2000;34(3):371–86.
18. Sohn S, Hyungwon S. Pattern recognition for a road traffic accident severity in Korea. Ergonomics. 2001;44(1):101–17.
19. Chang L, Wang H. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev. 2006;38(5):1019–27.
20. Wong J, Chung Y. Comparison of methodology approach to identify causal factors of accident severity. Transp Res Rec. 2008;2083:190–8.
21. Anderson TK. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid Anal Prev. 2009;41(3):359–64.
22. https://fr.wikipedia.org/wiki/Accident_de_la_route_en_France. Accessed 2018.
23. https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation. Accessed 2018.
24. http://www.securite-routiere.gouv.fr/la-securite-routiere/l-observatoire-national-interministeriel-de-la-securite-routiere. Accessed 2018.
25. Direction de l'information légale et administrative. La sécurité routière en France: Bilan 2017. Paris: ONISR; 2018.
26. Roy B, Bertier B. La méthode ELECTRE II: Une Méthode de Classement en Présence de Critères Multiples. Note de Travail No 142, Direction Scientifique, Groupe Metra. 1971.
27. Kumar S, Toshniwal D, Parida M. A comparative analysis of heterogeneity in road accident data using data mining techniques. Evol Syst. 2017;8(2):147–55.
28. Jarke M, Lenzerini M, Vassiliou Y, Vassiliadis P. Fundamentals of data warehouses. Berlin: Springer; 2002.
29. Mayer-Schonberger V, Cukier K, Big Data: a revolution that will transform how we live, work and think. Ed. John Murray. 2013
30. Wu X, Kumar V, RossQuinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowl Inf Syst. 2007;14(1):1–37.
31. Agrawal R et Srikant R. Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB, Jarke (Matthias) et Zaniolo (Carlo), Morgan Kaufmann. 1994. p. 487–99.
32. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: ACM-SIGMOD int. conf. on management of data, Mai 2000, p. 1–12.
33. Savasere A, Omiecinsky E, Navathe S. An efficient algorithm for mining association rules in large databases. In: 21st Int'l CONF. ON VERY LARGE DATABASES (VLDB), Septembre 1995.
34. Brin S, Motwani R, Ullman J D, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Peckham, ed. Proceedings ACM SIGMOD international conference on management of data, Tucson, Arizona, USA. New York: ACM Press; 1997. p. 255–64.
35. Zaki MJ, Parthasarathy S, Ogihara M, Li W. New algorithms for fast discovery of association rules. In: Heckerman D, Mannila H, Pregibon D, Uthurusamy R, Park M, eds. 3rd intl. conf. on knowledge discovery and data mining. New York: AAAI Press; 1997. p. 283–96.
36. Yanbo W, Huiqiang W, Xuefei J, Ming Y. Research of AprioriHybird algorithm and application in network situational awareness, computer science and information technology (ICCSIT), 2010 3rd IEEE international conference, vol. 7, no, 9–11 July 2010. p. 170–172.
37. Borgelt Christian. Workshop open source data mining software (OSDM'05, Chicago, IL), 66–70. New York: ACM Press; 2005.
38. Han J, Kamber M. Data mining: concepts and techniques. New York: Morgan Kaufmann Publishers; 2000.
39. Astide Y, Taouil R, Pasquier N, Stumme G, Lakhal L. Pascal: un algorithme d'extraction des motifs fréquents". Techn Sci Inf. 2002;21(1):65–95.
40. Bayardo RJ. Efficiently mining long patterns from databases. In: Proceedings of the 1998 ACM SIGMOD international conference on management of data (SIGMOD'98). New York: ACM Press, 1998. p. 85–93.
41. Duke HC, Byeong SA, Soung HK. Prioritization of association rules in data mining: multiple criteria decision approach. J Expert Syst Appl. 2005;29:867–78.
42. Velasquez M, Hester PT. An analysis of multi-criteria decision-making methods. Int J Operat Res. 2013;10(2):56–66.
43. El Mazouri FZ, Abounaima MC, Zenkouar K. Application of the ELECTRE III method at the moroccan rural electrification program. Int J Elect Comput Eng. 2018; 8(5): 3285–3295. http://www.iaescore.com/journals/index.php/IJECE/article/view/10285/9595.