Noname manuscript No.
(will be inserted by the editor)

Evidential Data mining: precise support and confidence

Ahmed Samet · Eric Lefèvre · Sadok Ben Yahia

Received: date / Accepted: date

Abstract Associative classification has been shown to provide interesting results whenever of use to classify data. With the increasing complexity of new databases, retrieving valuable information and classifying incoming data is becoming a thriving and compelling issue. The evidential database is a new type of database that represents imprecision and uncertainty. In this respect, extracting pertinent information such as frequent patterns and association rules is of paramount importance task. In this work, we tackle the problem of pertinent information extraction from an evidential database. A new data mining approach, denoted EDMA, is introduced that extracts frequent patterns overcoming the limits of pioneering works of the literature. A new classifier based on evidential association rules is thus introduced. The obtained association rules, as well as their respective confidence values, are studied and weighted with respect to their relevance. The proposed methods are thoroughly experimented on several synthetic evidential databases and showed performance improvement.

Keywords Evidential database · Evidential confidence · Evidential support · Associative classification

A. Samet
Univ. Lille Nord de France UArtois, EA 3926 LGI2A
F-62400, Béthune, France E-mail: ahmed.samet@univ-artois.fr

E. Lefèvre
Univ. Lille Nord de France UArtois, EA 3926 LGI2A
F-62400, Béthune, France E-mail: eric.lefevre@univ-artois.fr

S. Ben Yahia
University of Tunis El Manar, LIPAH Laboratory, Faculty of Sciences of Tunis, Tunisia
E-mail: sadok.benyahia@fst.rnu.tn

1 Introduction

The extraction of hidden information from large databases held a great importance. Within the encountered masses of data in databases lie hidden knowledge nuggets of strategic importance. One of the newest answer for this problematic is the knowledge discovery domain. Indeed, the latter proposes a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. For that reason, the application fields of data mining are so various that it assimilates many fields, e.g., image mining [25], web mining [22], medical domain [24] and classification of multiple databases [35], to cite but a few.

Chronologically, the Apriori algorithm [3] was the first one that aimed to extract frequent patterns and then derive interesting association rules thanks to a level-wise sweeping of the search space. Later, several improvements were brought giving rise to new and more efficient extraction algorithms such that [6, 26, 33]. Provided algorithms were applied on precise and certain data constituting boolean databases.

Nevertheless, in real world, gathering such types of data is hard achievable since almost all acquired data might suffer from imperfection. Therefore, uncertain data mining has become a hot topic in data mining community [1, 2, 34, 36]. Uncertainty is generally represented with probabilities. Recently, this discipline has flourished with new mining algorithms such as UApriori [7], UFP-Growth [20] and UH-Mine [2]. However, the uncertainty is not the only origin of imperfection. In [9], Dubois and Prade highlighted two possible origins for imperfection that are imprecision and uncertainty. In [18], Lee detailed both sides of imperfection that could manifest in data and proposed a new database handling imperfection. This database uses the belief function theory, also called the evidence theory, as a formalism to represent information [19]. This data structure was denoted as the *evidential database*. In [5] an imprecise and uncertain answer tuples of a query with evidence theory is presented. This permits the partial values to be overlapping sets, rather than disjoint sets forming a partition. Correspondingly, mass function values are attached to the partial values in order to represent degrees of uncertainty in the attribute values.

With the growing interest for those databases, studying them from a data mining view has never been more challenging. The redefinition of data mining tools gave rise to a particular interest as it was also the case for fuzzy data mining [15]. In this context, Hewawasam et al. [13] proposed a methodology to assess patterns' support and model them through a tree-based representation. Interestingly enough, the authors [13] paid attention to associative classification where the authors introduced *association rules*. The pertinence of association rule is assessed through a conditional belief.

In this work, we tackle the problem of the extraction of hidden and pertinent information from an evidential database. To do so, we shed the light on evidential support measure limits and we introduce a new alternative that is denoted *precise support*. The latter not only brings coherence with binary and probabilistic support measures but also flags out interesting running time. In addition, we address the problem of association rules' extraction. A new confidence measure is provided. The gathered rules are used for classification purposes. An Evidential Data Mining Apriori algorithm (EDMA) is then introduced for the information extraction. The retrieved association rules are then used for classification with a rule fusion system thanks to the Evidential Associative Classifier algorithm (EvAC).

This paper is organized as follows: Section 2 recalls the main concepts of belief function theory and the evidential database. The pioneering state-of-the-art works on confidence measure are scrutinized and we highlight their limits. In Section 3, a ramification for their method that improves the performance is presented. In addition, we introduce a new method for evidential itemsets' support computing providing more precision in its estimation. In Section 4, a new method for association rule generation is detailed. The provided rules are screen out and combined through a fusion system with the EvAC algorithm. The performance of this algorithm is studied in Section 6. Finally, we conclude and we sketch issues of future work.

2 Evidential Database: fundamental concepts

In this section, we briefly review evidence theory, also known as *belief functions theory* or *Dempster-Shafer theory*, and extend it to introduce the basic concepts of evidential databases [19].

2.1 Belief Function Theory

The belief function theory presents a large framework for imperfection handling. As highlighted by [9], the belief function theory not only models imprecision but also uncertainty. Several interpretations for this theory exist such that [8,12,32]. In our work, we rely on the Transferable Belief Model (TBM) interpretation that was originally introduced by Smets in [32]. The TBM model is a non-probabilistic interpretation of the belief function theory that represents quantified beliefs following two distinct levels: (i) a credal level where beliefs are entertained and quantified by belief functions; (ii) a pignistic level where beliefs can be used to make decisions and are quantified by probability functions. The evidence theory is based on several fundamentals such as the Basic Belief Assignment (BBA). A BBA m is the mapping from elements of the power set 2^{Θ} onto [0, 1], i.e.,

$$m : 2^{\Theta} \longrightarrow [0,1]$$

where Θ is the *frame of discernment*. It is the set of possible answers for a treated problem and is composed of N exhaustive and exclusive hypotheses:

$$\Theta = \{H_1, H_2, ..., H_N\}.$$

A BBA m also fulfills some constraints such that:

$$\begin{cases} \sum\limits_{A \subseteq \Theta} m(A) = 1 \\ m(\emptyset) \geq 0. \end{cases} \tag{1}$$

Each subset X of 2^{Θ} fulfilling $m(X) > 0$ is called a focal element. A categorical BBA is a BBA with only one focal element A and is defined as follows:

$$m(A) = 1 \quad \forall A \subset \Theta \quad \text{and} \quad m(B) = 0 \quad \forall B \subseteq \Theta,\ B \neq A. \tag{2}$$

Constraining $m(\emptyset) = 0$ is the normalized form of a BBA and this corresponds to a closed-world assumption [30], while allowing $m(\emptyset) \geq 0$ corresponds to an open world assumption [32].

From a BBA, other functions are commonly defined from 2^{Θ} into $[0,1]$: the first one, $Bel(A)$, called as the *belief function*, is interpreted as the degree of justified support given to the proposition A by the available evidence and is defined as:

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B). \tag{3}$$

On the other hand, $Pl(.)$ is the plausibility function and is defined as follows:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{4}$$

The plausibility denotes the maximum potential support that could be given to a hypothesis, if further evidence becomes available.
One of the asset of belief function theory is information fusion. From a multisource context, a combination operator can be applied in order to extract the veracious proposition. In the case of two sources S_1 and S_2, both defined in Θ, we define the conjunctive combination rule that was initially introduced in Smets' work [32]:

$$m_{\bigcirc\!\!\!\!\cap}(A) = \sum_{B \cap C = A} m_1(B) \times m_2(C) \qquad \forall A \subseteq \Theta. \tag{5}$$

The normalized version of conjunctive combination rule, proposed by Dempster [8], integrates a conflict management approach that redistributes the generated conflictual mass. For two sources S_1 and S_2 having respectively m_1 and m_2 as BBA, the Dempster's rule of combination, aka orthogonal sum, is defined as follows:

$$m_{\oplus}(A) = \frac{1}{1 - m_{\bigcirc\!\!\!\!\cap}(\emptyset)} \sum_{B \cap C = A} m_1(B) \times m_2(C) = \frac{1}{1 - m_{\bigcirc\!\!\!\!\cap}(\emptyset)} m_{\bigcirc\!\!\!\!\cap}(A) \qquad \forall A \subseteq \Theta, A \neq \emptyset \tag{6}$$

where $m_{\bigcirc\!\!\!\!\cap}(\emptyset)$ is defined by:

$$m_{\bigcirc\!\!\!\!\cap}(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) \times m_2(C). \tag{7}$$

$m_{\bigcirc\!\!\!\!\cap}(\emptyset)$ represents the conflict mass between m_1 and m_2.

Generally in an information fusion problem, all considered sources do not share the same domain (frame of discernment). This constraint prevents from using usual combination tools [8]. Let us consider two belief functions m_1 and m_2 defined respectively in Θ_1 and Θ_2, the conjunctive combination rule can be extended to this special case. A unique and larger frame $\Theta = \Theta_1 \times \Theta_2$ is used so that the combination can be expressed as follows:

$$m^{\Theta}_{1\times 2} = m_1^{\Theta_1 \uparrow \Theta} \bigcirc\!\!\!\!\cap\, m_2^{\Theta_2 \uparrow \Theta} \tag{8}$$

where ↑ is the vacuous extension that can be written as follows:

$$m^{\Theta_1 \uparrow \Theta}(A) = \begin{cases} m^{\Theta_1}(B) & if A = B \times \Theta_2 \\ 0 \quad \text{otherwise.} \end{cases} \tag{9}$$

The result of Equation (8) can be retrieved following the Cartesian product as follows:

$$m^{\Theta}_{1\times 2}(A \times B) = m_1^{\Theta_1}(A) \times m_2^{\Theta_2}(B). \tag{10}$$

After source combination which integrates the credal stage of the TBM model, taking decision is necessary. In [31], the pignistic probability is introduced allowing probabilistic decision from BBA following this formula:

$$BetP(H_n) = \sum_{A \subseteq \Theta} \frac{|H_n \cap A|}{|A|\,(1 - m(\emptyset))} \times m(A) \qquad \forall H_n \in \Theta \tag{11}$$

where $|\cdot|$ is the cardinality operator. In the following subsection, we present the basic concepts of the evidential databases that rely on the evidence theory to handle uncertainty.

2.2 Evidence database concept

An evidential database stores data that could be either perfect or imperfect [18]. Data's imperfection in such database is expressed through the belief function theory. An evidential database, denoted by $\mathcal{EDB}$, with n columns and d lines where each column i $(1 \leq i \leq n)$ has a domain Θ_i of discrete values. m_{ij} the cell, of line j and column i, contains a normalized BBA as follows:

$$m_{ij} : 2^{\Theta_i} \to [0, 1] \quad \text{with}$$

$$\begin{cases} m_{ij}(\emptyset) \geq 0 \\ \sum\limits_{A \subseteq \Theta_i} m_{ij}(A) = 1. \end{cases} \tag{12}$$

Unlike binary, probabilistic or fuzzy database, the columns of the evidential database represent attributes rather than items. Indeed, columns are assimilated to questions. Each line is an information source. The BBA m_{ij} could be seen as the answer of an information source i to a question j. Such kind of modelling makes the evidential database a generalisation of several types of databases [29]. For example, a fuzzy database can be obtained by constructing a consonant BBAs within its cells.

Table 1: Example of an evidential database $\mathcal{EDB}$

Transaction	Attribute A	Attribute B
T_1	$m_{11}(A_1) = 0.7$ $m_{11}(\Theta_A) = 0.3$	$m_{12}(B_1) = 0.4$ $m_{12}(B_2) = 0.2$ $m_{12}(\Theta_B) = 0.4$
T_2	$m_{21}(A_2) = 0.3$ $m_{21}(\Theta_A) = 0.7$	$m_{22}(B_1) = 1$

In an evidential database, as shown in Table 1, an *item* corresponds to a focal element. An *itemset* (i.e., *pattern*) corresponds to a conjunction of focal elements having different domains. Two different itemsets can be related via the inclusion or intersection operator. Indeed, the inclusion operator [4,27] for itemsets is defined as follows, let X and Y are two itemsets:

$$X \subseteq Y \iff \forall x_i \in X, x_i \subseteq y_j.$$

where x_i and y_j are the i^{th} and the j^{th} elements of respectively X and Y. For the same itemsets X and Y, the intersection operator [27] is defined as follows:

$$X \cap Y = Z \iff \forall z_i \in Z, z_i \subseteq x_j \text{ and } z_i \subseteq y_k.$$

An *association rule* [28] R is a causal relationship between two itemsets that can be written as follows $R : X \to Y$ fulfilling $X \cap Y = \emptyset$.

Example 1 In Table 1, A_1 is an item and $\Theta_A \times B_1$ is an itemset such that $A_1 \subset \Theta_A \times B_1$ and $A_1 \cap \Theta_A \times B_1 = A_1$. In Table 1, $A_1 \to B_1$ is considered as an association rule.

In the following, we present how to extract hidden information within evidential databases. Evidential data mining is detailed through the definition of literature support and confidence measures.

2.3 Evidential Data mining

Unfortunately, only a few studies were carried out on evidential data mining. In [14], Hewawasam et al. proposed a methodology to estimate itemsets' support and model them in a tree-based representation: *Belief Itemset Tree* (BIT). The BIT representation brings easiness and rapidity for the estimation of the association rule's confidence. Their proposed support measure relies on the following inclusion operator:

$$X \subseteq Y \iff \forall x_i \in X, x_i \in Y.$$

In [4], the authors introduced a new approach for itemset support computing and applied it on a *Frequent Itemset Maintenance* (FIM) problem. All approaches [4,13] were based on Cartesian product between BBAs. In this context, we study the support of an itemset $X = \prod\limits_{i \in [1 \ldots n]} x_i$ such that x_i is an item belonging to the frame of discernment Θ_i. Since, items do not share the same frame of discernment, fusion rules cannot be applied. The belief support introduced by [4] is computed by the following equation:

$$m_j(X) = \prod_{x_i \in X} m_{ij}(x_i) \tag{13}$$

where $m_j(X)$ is the Cartesian product of all BBA in the transaction T_j. Thus, the BBA of the itemset X expressed in the $\mathcal{EDB}$ database becomes:

$$m_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} m_j(X). \tag{14}$$

Then, the support of X in the $\mathcal{EDB}$ database is deduced as follows:

$$Support^{Bel}_{\mathcal{EDB}}(X) = Bel_{\mathcal{EDB}}(X). \tag{15}$$

Remark 1 An itemset or a pattern is said to be *frequent* if and only if its allocated support is greater than or equal to a min-threshold fixed by the user otherwise it is called *infrequent.*

The Cartesian product-based support, as presented above, fulfils several mathematical properties such that the *anti-monotony* property, i.e., the supersets of an infrequent itemset are also infrequent. The opposite is true, all subsets of a frequent itemset are also frequent. Owe to this property, the construction of an Apriori-based algorithm becomes straightforward [4].

The aim of frequent itemsets is to find all interesting association rules. Originally, the retrieval of frequent itemsets and association rules were carried out on binary databases [3]. The proposed approach considers every itemset of size k from which it generates $2^k - 2$ potential interesting rules. The set of generated rules are then filtrated following their confidence measure. The confidence denotes the relevance of a rule and only valid[1] rules are retained. For a rule $R : R_a \rightarrow R_c$, such that R_c and R_a are respectively the conclusion and the antecedent (premise) part of the rule R, the confidence is expressed as follows:

$$Confidence(R) = P(R_c|R_a) = \frac{\sum\limits_{i=1}^{d} P_{T_i}(R_a \cap R_c)}{\sum\limits_{i=1}^{d} P_{T_i}(R_a)}. \tag{16}$$

The confidence is seen as an apriori probability (i.e., the probability of having R_c knowing that R_a is true). $P_{T_i}(X)$ indicates the probability of X appearance within the transaction T_i. In addition, even in fuzzy data mining association rule's confidence is built upon conditional fuzzy measures [16]. In this respect, Hewawasam et al. [13] applied the conditional reasoning in evidential data mining. The confidence is computed with the basics of conditional belief. Thus, the confidence of a rule R in the set of all rules $\mathcal{R}$, i.e., $R \in \mathcal{R}$, is computed as follows:

$$Confidence(R) = Bel(R_c|R_a) \tag{17}$$

where $Bel(.|.)$ is the conditional belief. Despite the existence of several interpretations and formulations for conditional belief, Hewawasam et al. [13] defined the confidence following Fagin et al. [10] interpretation such that:

$$Bel(R_c|R_a) = \frac{Bel(R_a \cap R_c)}{Bel(R_a \cap R_c) + Pl(R_a \cap \bar{R_c})}. \tag{18}$$

In [10], this conditional belief interpretation is considered more consistent than the original work of Dempster [8] which is written as follows:

$$Bel(R_c|R_a) = \frac{Bel(R_c \cup \overline{R_a}) - Bel(\overline{R_a})}{1 - Bel(\overline{R_a})}. \tag{19}$$

Example 2 Through the following example, we highlight the inadequacy of the conditional belief use. Let us consider the Transaction 1, of Table 1, from which we try to compute the confidence of $A_2 \rightarrow B_1$ (i.e., $Bel(B_1|A_2)$). The conditional belief, introduced in [8] , is equal to:

$$Bel(B_1|A_2) = \frac{Bel(B_1 \cup \overline{A_2}) - Bel(\overline{A_2})}{1 - Bel(\overline{A_2})} = \frac{Bel(B_1)}{1} = 0.4$$

The belief of B_1, knowing that A_2 is true, is equal to that of $Bel(B_1)$ due to the independence between A_2 and B_1. In addition, both hypothesis might be correlated so that the event B_1 does not occur knowing already the happening of A_2.

[1] An association rule is considered as valid if its confidence is greater than or equal to a threshold $minconf$.

In the following section, we study existing support measures within evidential databases. First, a simplification of the belief-based support is introduced. Then, we highlight the limits of this approach and we introduce a new measure called the *precise support*.

3 Evidential precise support

In the following, the state-of-the-art of evidential support measures is scrutinized. A simplification for the existing approach is discussed. The simplification brings easiness and computation rapidity. A new alternative is proposed that brings more precision comparatively to existing works. The introduced support estimation is denoted in the sequel *precise support*.

3.1 Support ramification

The support definition proposed by [4,13] relies on a Cartesian product. Despite being adequate in case of combining BBAs with different frames of discernment, its computational complexity is exponential. Indeed, for an evidential database, with d transactions, we have k attributes having each n focal elements, the arithmetic complexity of a Cartesian product is: $\mathcal{C} = d \times n^k = O(n^k)$. In the following, we provide a new formulation for the belief-based support.

Proposition 1 *Let us consider an evidential database $\mathcal{EDB}$ and the itemset $X = x_1 \times \cdots \times x_n$ constituted by the product of items (focal elements) x_i $(1 \leq i \leq n)$ of the exclusive frames of discernment Θ_i. For a transaction T_j, we have:*

$$Support^{Bel}_{T_j}(X) = \prod_{i \in [1 \ldots n]} Support^{Bel}_{T_j}(x_i) = \prod_{i \in [1 \ldots n]} Bel(x_i) \tag{20}$$

$$Support^{Bel}_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Support^{Bel}_{T_j}(X) \tag{21}$$

Proof Let us consider two items and focal elements x_1 and x_2 belonging respectively to m_1 and m_2 BBA such that $m = m_1 \times m_2$.

$$\begin{aligned}
Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) &= \sum_{a \subseteq x_1 \times \cdots \times x_n} m_{1 \times \cdots \times n}(a) \\
Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) &= \sum_{y_1 \subseteq x_1, \ldots, y_n \subseteq x_n} m_1(y_1) \times \cdots \times m_n(y_n) \\
Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) &= \sum_{y_1 \subseteq x_1} m_1(y_1) \times \cdots \times \sum_{y_n \subseteq x_n} m_n(y_n) \\
Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) &= Bel(x_1) \times \cdots \times Bel(x_n) = \prod_{i \in [1 \ldots n]} Bel(x_i)
\end{aligned}$$

Owe to this new formulation, the initial theoretical complexity is lowered to about $\mathcal{C} = d \times |X|O(1)$.

Example 3 Let us consider the evidential database given in Table 1, the support of $A_1 \times B_1$ is computed by using Equations (14) and (15) as follows:

$$\begin{cases} m_{\mathcal{EDB}}(A_1 \times B_1) = \frac{m_{11}(A_1)\cdot m_{21}(B_1)+m_{12}(A_1)\cdot m_{22}(B_1)}{2} = 0.14 \\ m_{\mathcal{EDB}}(A_1 \times B_2) = \frac{m_{11}(A_1)\cdot m_{21}(B_2)+m_{12}(A_1)\cdot m_{22}(B_2)}{2} = 0.07 \\ m_{\mathcal{EDB}}(A_1 \times \Theta_B) = \frac{m_{11}(A_1)\cdot m_{21}(\Theta_B)+m_{12}(A_1)\cdot m_{22}(\Theta_B)}{2} = 0.14 \\ \vdots \end{cases}$$

Thus, the support of $A_1 \times B_1$ becomes:

$$Support^{Bel}_{\mathcal{EDB}}(A_1 \times B_1) = Bel(A_1 \times B_1) = 0.14$$

The same result can be found by:

$$Support^{Bel}_{\mathcal{EDB}}(A_1 \times B_1) = \frac{Bel_{T_1}(A_1) \times Bel_{T_1}(B_1) + Bel_{T_2}(A_1) \times Bel_{T_2}(B_1)}{2} = 0.14$$

The main concern of the literature approaches for the support computation rely on the belief function. As highlighted by [8], the plausibility is an upper bound whereas the belief is lower one for each hypothesis happening (occurrence probability). In fact, $Bel(.)$ assesses the belief by referring only to a small subset of the superset. The support of X is evaluated by considering only subsets included in it. In the following, we introduce the precise support definition that gets rid of this limitation.

3.2 Precise support

Let us consider an evidential database $\mathcal{EDB}$ and the itemset $X = x_1 \times \cdots \times x_n$ constituted by the product of items (focal elements) x_i $(1 \leq i \leq n)$ of the exclusive frame of discernment Θ_i. The degree of presence of an item x_i in a transaction T_j (BBA) can be measured as follows:

$$Pr : 2^{\Theta} \to [0, 1] \tag{22}$$

$$Pr(x_i) = \sum_{x \subseteq \Theta_i} \frac{|x_i \cap x|}{|x|} \times m(x) \qquad \forall x_i \in 2^{\Theta_i}. \tag{23}$$

As illustrated above, the $Pr(.)$ is a measure that computes a probability in a single BBA. The Pr is also assimilated to the pignistic probability in case of $x_i \in \Theta_i$. The evidential support of an itemset $X = \prod\limits_{i \in [1 \ldots n]} x_i$ is then computed as follows:

$$Support^{Pr}_{T_j}(X) = \prod_{X_i \in \Theta_i, i \in [1 \ldots n]} Pr(X_i) \tag{24}$$

$$Support^{Pr}_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Support^{Pr}_{T_j}(X). \tag{25}$$

Interestingly enough, the precise support definition presents a larger element inclusion than those given in respectively [4,13]. The proposed definition allows us to overcome the limits of the belief-based support previously mentioned. Indeed, $Pr(.)$ function does not only consider all subsets of X but also those having intersection with it. The probabilistic formulation of the support sustains previous data mining support works such that [3] pioneering work on binary databases in case of certain BBAs[2]. Even fuzzy support definition [16] is consistent with the precise support in case of consonant BBAs[3]. Moreover, the formulation of the support provides an interesting performance since we avoid the pitfall of the computation of the Cartesian product.

Property 1 The precise support estimation function fulfils the anti-monotony property, i.e.,

$$Support^{Pr}_{\mathcal{EDB}}(A) \leq Support^{Pr}_{\mathcal{EDB}}(B) \quad \forall A \subseteq B. \tag{26}$$

Proof Assuming an evidential database $\mathcal{EDB}$, let us consider two evidential itemsets A and $A \times X$ where $A \subset A \times X$ such that $\forall x \in A,\ x \in A \times X$. We aim at proving this relation $Support^{Pr}_{\mathcal{EDB}}(A \times X) \leq Support^{Pr}_{\mathcal{EDB}}(A)$:

$$\begin{array}{l} Support^{Pr}_{T_j}(A \times X) = Pr(A) \times Pr(X) \\ Support^{Pr}_{T_j}(A \times X) \leq Support^{Pr}_{T_j}(A) \quad \text{Since} \quad Pr(X) \in [0,1] \quad \text{then} \\ Support^{Pr}_{\mathcal{EDB}}(A \times X) \leq Support^{Pr}_{\mathcal{EDB}}(A). \end{array}$$

The definition of the proposed support is computed with transactional precise probability (i.e., $Support^{Pr}_{T_j}(.)$). To avoid the computation of the support of a single item several times, we store all item's support in a table, which is called *Precise Table* (PT).

Example 4 Table 2 shows the precise table constructed from the evidential database $\mathcal{EDB}$ (Table 1). Each item in $\mathcal{EDB}$ rows has a Pr value.

The extraction of frequent itemsets is detailed in Algorithm 1. This Algorithm is denoted EDMA that stands for Evidential Data Mining Apriori. It generates frequent evidential itemsets in a level-wise manner as did the Apriori [3]. The use of a Apriori-based algorithm is justified for several reasons. In

[2] A BBA with only one focal element H and $H \in \Theta$ is said to be certain and is denoted $m(H) = 1$.

[3] A BBA is said consonant if focal elements are nested.

Table 2: Precise Table deduced from the evidential database $\mathcal{EDB}$ presented in Table 1

Transaction	Transactional Support
T_1	$Pr^{\Theta_A}(A_1) = 0.85$
	$Pr^{\Theta_A}(A_2) = 0.15$
	$Pr^{\Theta_A}(\Theta_A) = 1.00$
	$Pr^{\Theta_B}(B_1) = 0.60$
	$Pr^{\Theta_B}(B_2) = 0.40$
	$Pr^{\Theta_B}(\Theta_B) = 1.00$
T_2	$Pr^{\Theta_A}(A_1) = 0.35$
	$Pr^{\Theta_A}(A_2) = 0.65$
	$Pr^{\Theta_A}(\Theta_A) = 1.00$
	$Pr^{\Theta_B}(B_1) = 1.00$
	$Pr^{\Theta_B}(B_2) = 0.00$
	$Pr^{\Theta_B}(\Theta_B) = 1.00$

fact, UApriori which is the uncertain probabilistic version of Apriori, actually performs rather well among the other tree-based algorithms and is usually faster one in dense uncertain dataset [34]. The evidential databases are naturally dense. Even though, more efficient algorithms were so far introduced, Apriori is of extensive use owe the efficiency of its pruning of candidates, that relies on the anti-monotony property [1]. Indeed, the majority of approaches in imperfect data mining uses Apriori [1], undoubtedly due to the difficulty to replicate binary simplification approaches. Finally, as a result for the lack of research in this thematic, evidential data mining works rely on a level-wise approach for frequent patterns generation [4,13]. As the UApriori, EDMA includes a trimming part [7]. The basic idea behind it is to trim away items with low existential presence from the evidential database and then to mine the trimmed structure. As a result, a structure called *Trim_Table* is constructed that stores either the precise values (i.e., Pr(.)) or the belief function (i.e., Bel(.)) of interesting items. The trimming module has two benefits. First, it allows removing items with low existential presence within the database and therefore a low probability of being either frequent or generates ones. More importantly, it allows to remove uninteresting items. For example, considering a medical predictive model, some items should be removed before any mining process. Indeed, predicting a patient disease is a critical application. Therefore, items as disjunction focal elements or total ignorance items (e.g Θ_i) should be removed.

The extraction of the frequent patterns relies on two main functions. *Support_estimation*(), in line 11, computes the precise support of an itemset taken as an input. *Frequent_itemset*() (line 24) determines whether an itemset is frequent based on the *minsup* and the precise support values. Furthermore, support computing within evidential databases has a cost. Indeed, for an evidential database of n attributes and d rows, an approximative complexity of support estimation can be expressed as follows: $\mathcal{C} = d \times (C_{BBAtreatment} \times n)$

where $C_{BBAtreatment}$ indicates the complexity of a BBA treatment (i.e., pignistic probability computing, belief function computing, etc). For the precise support, the computational complexity is higher than the belief one. For an attribute of l focal elements, the belief-based support has $O(l^2)$ as complexity since for each focal element the inclusion is studied with other elements. This complexity drops to $O(l \times log(l))$ with some heuristics. However, for the precise support, the computational complexity is $2 \times O(l^2)$.

Algorithm 1 Evidential Data Mining Apriori (EDMA) algorithm

```
Require: $\mathcal{EDB}, minsup, PT, Size\_\mathcal{EDB}$
Ensure: $\mathcal{EIFF}$
$Trim\_Table \leftarrow construct\_trim(PT, minsup)$
$\mathcal{EIFF} \leftarrow \emptyset$
$size \leftarrow 1$
$candidate \leftarrow candidate\_apriori\_gen(Trim\_Table, size)$
While $(candidate \neq \emptyset)$
$freq \leftarrow Frequent\_itemset(candidate, minsup, Trim\_Table, Size\_\mathcal{EDB})$
$size \leftarrow size + 1$
$\mathcal{EIFF} \leftarrow \mathcal{EIFF} \cup freq$
$candidate \leftarrow candidate\_apriori\_gen(\mathcal{EDB}, size, freq)$
End While
function SUPPORT_ESTIMATION($Trim\_Table$ , $I$, $d$)
    $Sup_I \leftarrow 0$
    for j=1 to d do
        $Sup_{Trans} \leftarrow 1$
        for all $i \in Trim\_Table(j).focal\_element$ do
            if $Trim\_Table(j).focal\_element \in I$ then
                $Sup_{Trans} \leftarrow Sup_{Trans} \times Trim\_Table(j).value$
            end if
        end for
        $Sup_I \leftarrow Sup_I + Sup_{Trans}$
    end for
    return $\frac{Sup_I}{d}$
end function
function FREQUENT_ITEMSET($candidate$, $minsup$, $Trim\_Table$, $Size\_\mathcal{EDB}$)
    $frequent \leftarrow \emptyset$
    for all $X$ in $candidate$ do
        if $Support\_estimation(Trim\_Table, X, Size\_\mathcal{EDB}) \geq minsup$ then
            $frequent \leftarrow frequent \cup \{X\}$
        end if
    end for
    return $frequent$
end function
```

4 Associative classification in evidential databases

In the following, we introduce a new association rules-based classifier. This classifier is based on valid association rules found by the use of the newly defined support measure.

4.1 Precise confidence measure

As highlighted in subsection 2.3, measuring the confidence of an evidential association rule with a conditional belief function has many shortcomings. In the following, we introduce a new evidential measure of confidence for association rules that relies on probabilistic fundamentals. This measure is denoted as the *precise measure* and is equal to:

$$Confidence(R) = \frac{\sum\limits_{j=1}^{d} Pr_{T_j}(R_a) \times Pr_{T_j}(R_c)}{\sum\limits_{j=1}^{d} Pr_{T_j}(R_a)} \tag{27}$$

where d is the number of transactions in the evidential database. In addition, the proposed metric sustains previous confidence measure such that introduced in [3].

Example 5 Let us consider the example of the evidential database given in Table 1. The confidence of the association rule $R_1 : A_1 \to B_1$ is computed as follows:

$$Confidence(R_1) = \frac{Pr_{T_1}(A_1) \times Pr_{T_1}(B_1) + Pr_{T_2}(A_1) \times Pr_{T_2}(B_1)}{Pr_{T_1}(A_1) + Pr_{T_2}(A_1)} = 0.75$$

The generated rules, with their respective values of confidence, could be useful in many applications. In the following, we tackle the classification problem case, in which an associative classifier is introduced. In this section, we study how to select valid association rules for classification purposes. Two types of rule are introduced: the generic and the precise association rules.

4.2 Generic and precise rules

To perform a classification with association rules, we retain only those concluding on a class label. Indeed, from a rule such that $\prod\limits_{i\in[1,I],I<n} X_i \to \prod\limits_{j\in[1,J],J<n} Y_j$, we only keep those having in the conclusion part, a class hypothesis (i.e., $Y_j \in \Theta_C$ where Θ_C is the frame of discernment).

Example 6 Considering the following set of the association rules $S = \{A_1 \to C_1; A_1, B_2 \to C_1; A_1 \to B_1\}$ and the class frame of discernment $\Theta_C = \{C_1, C_2\}$. After the classification rule reduction step, the set S is reduced to $S' = \{A_1 \to C_1; A_1, B_2 \to C_1\}$.

Even with the classification rule reduction, their number is still overwhelming requiring further filtrating strategies. Two main ideas can be distinguished. A first one consists in retaining only the *generic rules* (i.e., the most generic in terms of premise). The second one is the opposite. It relies on pruning all rules and keeping those having the largest premise (those types of rule are denoted *precise rules*). In the following, we explicitly define those two heuristics for rule filtration.

The generic rule reduction approach consists in retaining only classification rules with a minimal premise. These rules can be viewed as a generalization for other redundant ones. Indeed, a rule R_1 is considered as a redundant one if and only if it does not bring any additional information having at hand a rule R_2. The retained rules from the reduction process constitute the set of generic rules extracted from the set of frequent itemsets $\mathcal{FI}$.

Another heuristic consists in retaining only the rule of the largest premise. Those kind of rules are considered as the most precise and the brought information is considered as reliable.

Example 7 Considering the previous set of the association rules $S = \{A_1 \rightarrow C_1; A_1, B_2 \rightarrow C_1; A_1 \rightarrow B_1\}$. After redundant rule reduction, the set S becomes equal to $S' = \{A_1 \rightarrow C_1; A_1 \rightarrow B_1\}$. The rule $A_1, B_2 \rightarrow C_1$ is not retained since it brings no further information having already the rule $A_1 \rightarrow C_1$. The set S of precise rules is, then, equal to $S'' = \{A_1, B_2 \rightarrow C_1\}$.

In the next subsection, we describe our approach of classification based on association rules within evidential databases.

4.3 Classification with association rules

Let us suppose the existence of an instance X, to classify, represented by a set of BBAs belonging to the evidential database $\mathcal{EDB}$ such that:

$$X = \{m_i | m_i \in X, x_i^j \in \Theta_i\} \tag{28}$$

where x_i^j is a focal element of the BBA m_i. Each retained association rule, from the rule set $\mathcal{R}$, is considered as a potential piece of knowledge that could be helpful for the class retrieval of X. In order to select rules that may lead to the adequate classification, we look for rules having a non null intersection with X such that:

$$\mathcal{RI} = \{R \in \mathcal{R}, \exists x_i^j \in \Theta_i,\ x_i^j \in R_a\}. \tag{29}$$

Each rule found in the set $\mathcal{RI}$ constitutes a piece of information concerning the membership of the instance X. Since several rules can be found and fulfilling the intersection condition, it is of importance to benefit from them all. In our work, we assume that all information is valuable and should be handled within the information fusion problem. From the set $\mathcal{RI}$, we extract the set of generic or precise classification rules. Indeed, each rule from the computed set $R_l \subset \mathcal{RI}$, $l \in [1 \dots L]$ and $L < |\mathcal{RI}|$, that brings a new information (different R_a) is transformed into a BBA with respect to the frame of discernment Θ_C (i.e.,frame of discernment of R_c):

$$\begin{cases} m_{R_l}^{\Theta_C}(\{R_c\}) = Confidence(R_l) \\ m_{R_l}^{\Theta_C}(\Theta_C) = 1 - Confidence(R_l) \end{cases} \tag{30}$$

Table 3: The evidential transaction X under classification

	Attribute A	Attribute B
X	$m_1(A_1) = 0.6$ $m_1(A_2) = 0.2$ $m_1(\Theta_A) = 0.2$	$m_2(B_1) = 0.5$ $m_2(B_2) = 0.1$ $m_2(\Theta_B) = 0.4$

where R_c is the conclusion part of the rule R_l. The L constructed BBAs are then fused following the Dempster's rule of combination [8] as follows:

$$m_{\oplus} = \oplus_{l=1}^{L} m_{R_l}^{\Theta_C}. \tag{31}$$

Equation (31) combines all association rules with the same consideration. Indeed, each secant rule is considered and summed with the other ones having even more intersection with X. From this point, it is interesting to make distinction between rules during the combination. To do so, it is important to distinguish between the reliability of a rule and the confidence found through the use of the function $Confidence(.)$. The confidence expresses the pertinence of a rule in the studied database. However, the reliability of a rule describes the weight accorded to a rule during the fusion process.

In order to compute the reliability of a rule, enumerating all possible criteria for rule distinction held a great importance. In the following, we highlight the importance of weighting up the classification association rules in use (i.e., $\mathcal{RI}$). The rule's selection is an important step in our classification process. However, it does not guarantee the presence of only quality rules. Indeed, since the set $\mathcal{RI}$ of classification rules relies on secant rules (i.e., rules having an intersection with the instance under classification), several non pertinent rules could be retained for fusion. Besides that fact, the performance of the orthogonal sum (Equation (31)) could be deteriorated by the number of retained rules (i.e., rule's BBA in Equation (30)) as well as by its combination property. The property 2 shows both properties that must be used for weighting the association rules.

Property 2 Let us assume a set of association rules $\mathcal{RI}$ having a non-null intersection with X, two properties must be fulfilled:

- P$_1$: A significant weight must be assigned to the highest precise rule with regard to the instance X under classification, i.e., $R_{1a} \subset R_{2a}$.
- P$_2$: The mass assigned to each focal element within the instance X should be a criteria for rule weighting

The following example sheds light on the encountered problem in rules' aggregation.

Example 8 Let us consider the evidential transaction X under classification shown in Table 3.

Let $\mathcal{RI} = \{R_1 : A_1, B_1 \rightarrow C_1; R_2 : \Theta_A, B_1 \rightarrow C_1; R_3 : A_1 \rightarrow C_1; R_4 : B_2 \rightarrow C_2\}$ be a set of classification association rules. R_1 would get a higher weight than R_2 with respect to the property P_1. R_4 is not P_2-pertinent since $m_2(B_2)$ is too low.

In the following, we introduce a new approach for weighting the classification association rules. This method has to fulfil both criteria risen in Example 8. Let us assume an association rule $R : R_a \rightarrow R_c$ that we aim to assess the relevance of the rule's premise part with the instance X under classification. For each item part of the considered premise $\{x_i^j \in R_a | x_i^j \in \Theta_i, i \in [1, I], j \in [1, J]\}$, we compute its distance with the appropriate part of the instance under classification. From each $x_i^j \in R_a, i \in [1, I], j \in [1, J]$, we build a categorical BBA $m_i^c(\{x^j\})$ (Equation (2)). The resulting BBA is compared to m_i to assess their separating distance in terms of the P_1-criterion. The distance is computed as follows:

$$d_i(m_i^c, m_i) = \sqrt{\frac{1}{2}(m_i^c - m_i)^t.D.(m_i^c - m_i)} \tag{32}$$

where:

$$D(A, B) = \begin{cases} 1 \quad \text{if} \quad A = B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} \quad \text{if} \quad A, B \subseteq 2^{\Theta}. \end{cases} \tag{33}$$

d_i is the Jousselme's distance [17]. The matrix $D(A, B)$ establishes the inclusion relationship between superset elements. The rule's weight is found by considering all computed distances $\{d_i | i \in I\}$ as follows:

$$weight(R) = 1 - \frac{\sum\limits_{i \in [1,I]} d_i}{I}. \tag{34}$$

Thus, Equation (30) becomes:

$$\begin{cases} {}^{\alpha}m_{R_l}^{\Theta_C}(\{R_c\}) = weight(R_l) \times Confidence(R_l) \\ {}^{\alpha}m_{R_l}^{\Theta_C}(\Theta_C) = 1 - weight(R_l) \times Confidence(R_l) \end{cases} \tag{35}$$

Example 9 Let us assume the set of rules shown in Example 8. Table 4 is a numerical example of association rules' fusion. Considering the case of four association rules, the column *Reliability* shows association rules' weight based on the computed distance d_l (Equation (34)). The reliability factors flag out the expected results, e.g., R_1 is more P_1-reliable than R_2 and that sustains the P_1 property. On the other hand, R_3 is more P_2-reliable than R_4. The rules are then modelled and weighted following Equation (35). The decision with pignistic probability gives the C_1 class which is naturally the case.

4.4 Evidential Associative Classifier: EvAC

In the following, we introduce the Evidential Associative Classifier (EvAC) that extracts all valid association rules from frequent patterns and classifies evidential instances. The EvAC algorithm, whose pseudo-code is sketched by

Table 4: Numerical example of rule's weighting and fusion

Rule	Reliability (*weight*)	$m_{R_l}^{\Theta_C}$	${}^{\alpha}m_{R_l}^{\Theta_C}$	$m_{\oplus}$	BetP
$R_1 : A_1, B_1 \rightarrow C_1$	0.66	$m_{R_1}^{\Theta_C}(C_1) = 0.59$	${}^{\alpha}m_{R_1}^{\Theta_C}(C_1) = 0.39$		
		$m_{R_1}^{\Theta_C}(\Theta_C) = 0.41$	${}^{\alpha}m_{R_1}^{\Theta_C}(\Theta_C) = 0.61$	$m_{\oplus}(C_1) = 0.59$	
$R_2 : \Theta_A, B_1 \rightarrow C_1$	0.60	$m_{R_2}^{\Theta_C}(C_1) = 0.32$	${}^{\alpha}m_{R_2}^{\Theta_C}(C_1) = 0.19$		
		$m_{R_2}^{\Theta_C}(\Theta_C) = 0.68$	${}^{\alpha}m_{R_2}^{\Theta_C}(\Theta_C) = 0.81$	$m_{\oplus}(C_2) = 0.06$	$BetP(C_1) = 0.77$
$R_3 : A_1 \rightarrow C_1$	0.69	$m_{R_3}^{\Theta_C}(C_1) = 0.32$	${}^{\alpha}m_{R_3}^{\Theta_C}(C_1) = 0.22$		
		$m_{R_3}^{\Theta_C}(\Theta_C) = 0.68$	${}^{\alpha}m_{R_3}^{\Theta_C}(\Theta_C) = 0.78$	$m_{\oplus}(\Theta_C) = 0.35$	$BetP(C_2) = 0.23$
$R_4 : B_2 \rightarrow C_2$	0.27	$m_{R_4}^{\Theta_C}(C_2) = 0.66$	${}^{\alpha}m_{R_4}^{\Theta_C}(C_2) = 0.18$		
		$m_{R_4}^{\Theta_C}(\Theta_C) = 0.34$	${}^{\alpha}m_{R_4}^{\Theta_C}(\Theta_C) = 0.82$		

Algorithm 2, extracts all interesting classification association rules by computing their confidence. The rule extraction is caried out at the beginning of the algorithm. In fact, the user can select to work with either generic or precise rules. The confidence is computed thanks to $Find\_Confidence()$ function (line 4) that implements the precise confidence, given in section 4, and only retains valid rules. In addition, rules are filtrated by redundancy. Interested reader may refer to [28] for further details about rule's filtration. From each secant rule R to the instance under classification X, EvAC models a BBA through invoking the *construct_BBA*() function (line 19). The resulting BBA is studied and weighted via a reliability factor. Those reliability factors are retrieved in the *compute_reliability*() function (line 25), which integrates the Jousselme's distance. Finally, the decision is made upon the use of the pignistic probability and the class of the instance under classification is returned. EvAC algorithm computational complexity depends highly on the association rules used for classification. The computational complexity is polynomial $C_{comb}O(l^2)$. C_{comb} is the complexity of combining l BBAs with the Dempster's rule of combination.

5 Experiments

In this section, we assess the performance of EDMA and EvAC algorithms on evidential databases. The evidential databases are obtained through dataset transformation. In the following, evidential databases' construction is highlighted. Different practical uses of evidential database can exist. We can imagine a medical database in which we store patients medical records. Those records are doctors diagnostics expressed with BBAs relatively to their certainty about patient's sickness. Even though some works are worth of cite, e.g [4, 13], none of them worked on a real evidential database. In [4], tests were carried out on a synthetic database. On the other hand, Hewawasam et al. [13] worked on a simplified naval anti-surface warfare scenario and such kind of databases is hardly accessible. In the following, we propose a method that makes it possible to straighforwardly construct an evidential database from a

Algorithm 2 Evidential Associative Classifier (EvAC)

```
Require: $Pr\_Table, minconf, \Theta_C, X, \mathcal{EIFF}$
Ensure: $Class$
for all $x \in \mathcal{EIFF}$ do
    $R \leftarrow Construct\_Rule(x, \Theta_C)$
    if $R \neq \emptyset$ then
        $confidence \leftarrow Find\_Confidence(R, Pr\_Table, minconf)$
        $\mathcal{R} \leftarrow Redundancy(\mathcal{R}, R, confidence)$
    end if
end for
for all $R \in \mathcal{R}$ do
    if $X \cap R \neq \emptyset$ then
        $RI \leftarrow RI \cup R$
    end if
end for
for all $R \in RI$ do
    $weight_I \leftarrow compute\_reliability(R, X)$
    ${}^{\alpha}BBA \leftarrow construct\_BBA(R, weight_I)$
    $m \leftarrow m \oplus^{\alpha} BBA$
end for
$Class \leftarrow argmax_{H_k \in \Theta_C} BetP(H_k)$
function CONSTRUCT_BBA($R, weight_I$)
    $m_R(R.conclusion) \leftarrow weight_I \times R.confidence$
    $m_R.(\Theta) \leftarrow 1 - weight_I \times R.confidence$
    $BBA \leftarrow m_R$
    return $BBA$
end function
function COMPUTE_RELIABILITY($R$,$X$)
    $d \leftarrow 0$
    for all $r \in R_a$ do
        $d \leftarrow d + Jousselme\_distance(m_r, X^i)$
    end for
    $weight_I \leftarrow \frac{d}{sizeof(R_a)}$
    return $\alpha_I$
end function
```

numerical dataset. We based our evidential database construction on the ECM clustering approach [23]. It is an FCM-like-based algorithm on the concept of credal partition, extending those of fuzzy and possibilistic ones. To derive such a structure, we minimized the proposed objective function:

$$J_{ECM}(M, V) \triangleq \sum_{i=1}^{d} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} c_j^{\alpha} m_{ij}^{\beta} dist_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \tag{36}$$

subject to:

$$\sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1, \ldots, d \tag{37}$$

where $m_{i\emptyset}$ and m_{ij} respectively denote $m_i(\emptyset)$ and $m_i(A_j)$. M is the credal partition $M = (m_1, \ldots, m_d)$ and V is a cluster centers matrix. c_j^{α} is a weighting coefficient and $dist_{ij}$ is the Euclidean distance. In our case, we use the default values prescribed by the authors in [23], i.e. $\alpha = 1$, $\beta = 2$ and $\delta = 10$.

The ECM algorithm was successfully applied on several UCI benchmarks [11] in order to construct our evidential databases. The characteristics of the constructed evidential databases are summarized in Table 5 in terms of number of instances and attributes. For each dataset, the number of focal elements, after the ECM application, was addressed. The number of focal element is related to the objective function J_{ECM} that was minimized (the reader is referred to the appendix A for further details). The fourth column of Table 5 illustrates the sum of all generated focal elements that indicates the actual width size of each evidential database. Indeed, the actual column's size of the database is computed as follows: $\sum_{i=1}^{n} f_i$. The variable f_i is the number of focal element of the i^{th} attribute found by the minimization of the subjective equation (Equation (37)). We used two types of benchmarks. The largest databases such as Skin Segmentation_EDB, KEGG_EDB and MAGIC_EDB were used to assess the scalability of the mining algorithm. The smallest databases such as Iris_EDB, Wine_EDB, Vertebral column_EDB and Diabetes_EDB were tested to assess the accuracy of the classifier.

Table 5: Database characteristics

Database	**#Instances**	**#Attributes**	**#Focal elements**
Iris_EDB	150	5	40
Vertebral Column_EDB	310	7	116
Diabetes_EDB	767	9	132
Wine_EDB	178	14	196
Magic_EDB	19020	11	30
KEGG_EDB	53414	24	96
Skin Segmentation_EDB	245057	4	32

5.1 Pattern extraction performance

In the following, we compare the precise support measure introduced in section 3 that of the belief-based support [4]. The performance and the quality of both support measures are scrutinized. The quality is highlighted in terms of the number of frequent patterns. In addition, the performance is shown by the computational time. Table 6 shows the performance of the precise support (denoted *precise*) and the belief-based support (denoted *Bel*). For our experiments, we integrated the ramification support (proposed in subsection 3.1) into the mining itemsets algorithm.

In terms of quality, as expected, the precise support discovers more frequent patterns than do the belief-based one. This result corroborates the theoretical bases found in subsection 3.2. Indeed, precise support evaluates properly the support. On the contrary, the belief-based support under-evaluates the support because of the intrinsic $Bel(.)$ function nature. The number of frequent patterns increases linearly as far as the considered *minsup* threshold

Table 6: Comparative results in terms of the number of frequent extracted patterns

Support	Iris_EDB		Diabete_EDB		Vertebral_Column_EDB		Wine_EDB		Magic_EDB	
	Precise	Bel	Precise	Bel	Precise	Bel	Precise	Bel	Precise	Bel
0.9	23	23	1013	757	56	56	4850	4082	1011	1011
0.8	42	23	3831	1013	88	56	7155	4082	1011	1011
0.7	107	23	12408	1397	411	56	24565	8179	1011	1011
0.6	244	91	38887	3958	795	88	71258	12275	3574	3574

decreases. In addition to the study of generated frequent patterns, Table 7 shows the difference in terms of performance between precise and two variants of the belief-based support. A first version applies the Cartesian product (denoted Cart-Bel) in order to find the support which computes all possible BBAs needed for support measure. The second one is the proposed ramification of the belief support that uses the Table_Bel (denoted Bel) which is the belief version of the precise table.

Table 7: Comparative results in terms of execution time (seconds)

Support	Iris_EDB			Diabete_EDB			Vertebral Column_EDB			Wine_EDB			Magic_EDB		
	Precise	Bel	Cart-Bel ≈	Precise	Bel	Cart-Bel ≈	Precise	Bel	Cart-Bel ≈	Precise	Bel	Cart-Bel ≈	Precise	Bel	Cart-Bel ≈
0.9	0,18	0.17	6.38E+12	14.60	9.35	3.29E+72	0.77	0.21	4.60E+24	58.88	37.32	1.83E+188	163.86	94.92	2.75E+50
0.8	0.21	0.17	6.38E+12	120.36	9.36	3.29E+72	0.82	0.21	4.60E+24	109.39	40.70	1.83E+188	159.40	92.92	2.75E+50
0.7	0.58	0.17	6.38E+12	851.70	17.77	3.29E+72	4.28	0.22	4.60E+24	1536.93	89.42	1.83E+188	161.59	90.87	2.75E+50
0.6	2.29	0.50	6.38E+12	11586	71.01	3.29E+72	9.05	0.30	4.60E+24	16172.59	179.88	1.83E+188	867.72	771.78	2.75E+50

The extraction performances, running time-wise, of the belief-based support is better than those of the precise one. This observation can be explained by the number of extracted patterns. The more frequent candidates generated are, higher the consumed time is. In addition, the precise support studies more subsets than does the belief-based one that badly influences the obtained performances.

Figure 1 shows the runtime performances of several algorithms on the largest datasets. In fact, we compared EDMA algorithm to B-Apriori which is the Apriori-based algorithm that use the belief-based support [13]. It is important to notice that B-Apriori outperforms EDMA in terms of computation times as far as the size of the database increases. This can be explained by two reasons. First, by the cost of using the precise support rather than the belief-based one. In fact, the precise-based support computes the intersections between sets. Finally, since EDMA generates more candidates in a level-wise manner than B-Apriori, the support computation is a costly task.

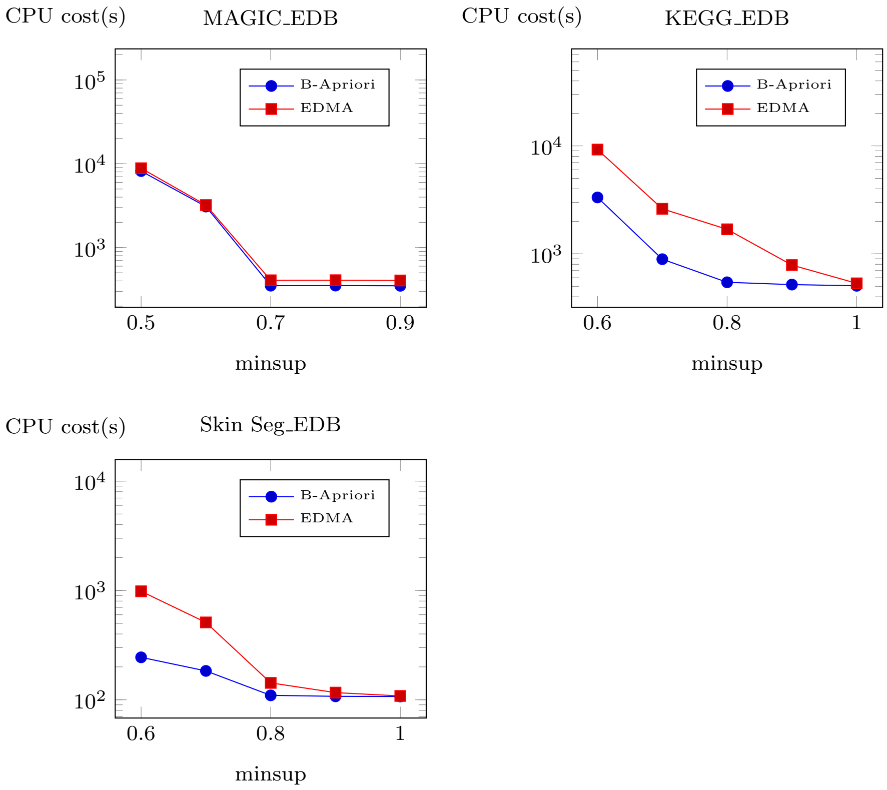
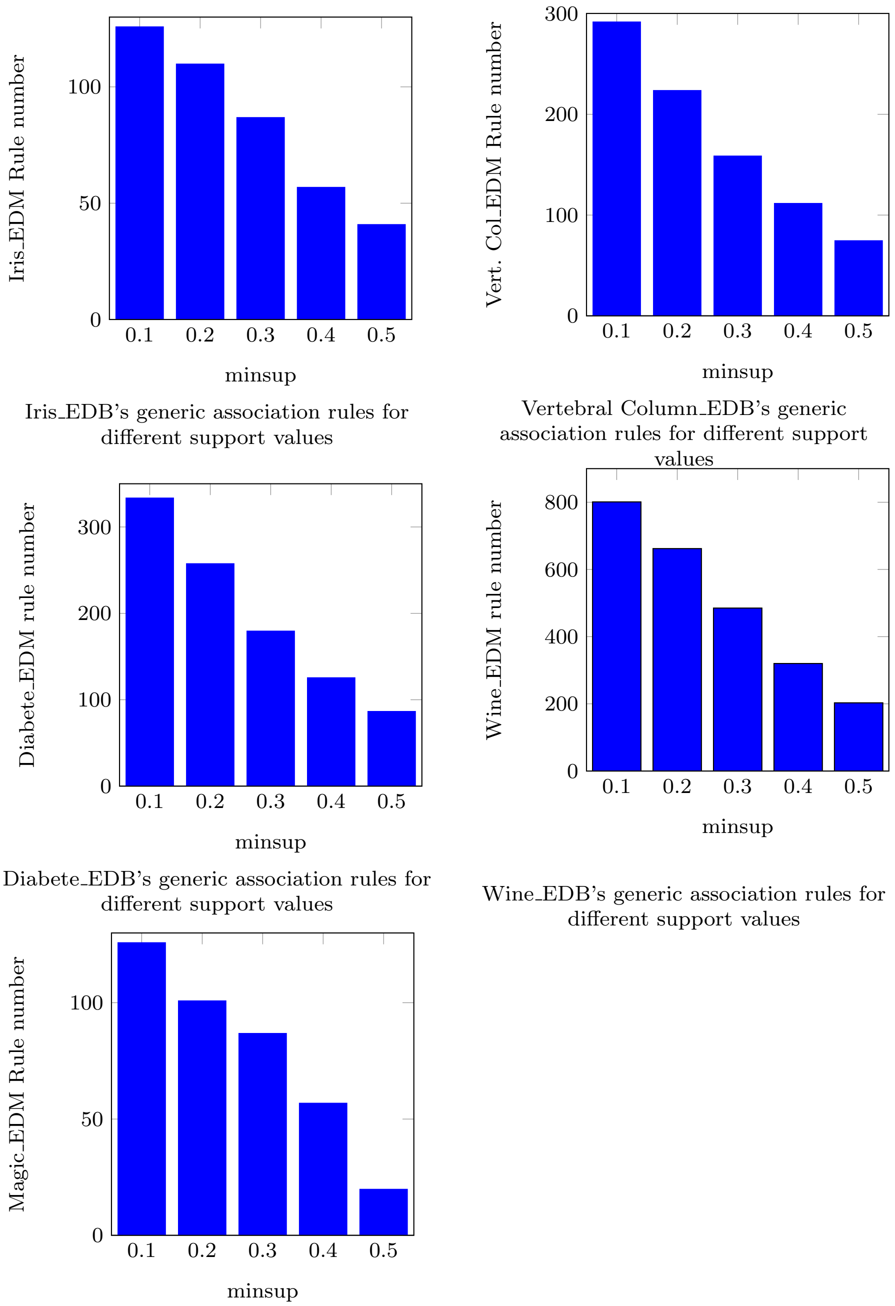


Fig. 1: Runtime performance on several database benchmarks

5.2 EvAC classification performance

In the following, we study the classification result performance of the generic and precise rules. Table 8 compares the classification differences in the use of precise rules. The precise rules are studied, in this table, by adding the weighting approach. The weighting approach has shown its usefulness since we improved the classification for Iris_EDB and Vertibral_EDB. We maintained the same perfect result for the Wine_EDB whereas it has dropped in Diabete_EDB. Overall, the rule weighting approach has proven its efficiency and has optimized the results of non weighted precise rules.

The results of the weighting approach has been carried out using the generic rules. The performance of EvAC associative classifier has been scrutinized by investigating the generic rules. The results are shown in Table 9. The rule's weighting approach has also proven its efficiency for classification with generic rules. Indeed, the results have been drastically improved comparatively to the classification with non weighted generic rules. Wine_EDB and Iris_EDB classification rates have been improved and we maintained the same results



Iris_EDB's generic association rules for different support values

Vertebral Column_EDB's generic association rules for different support values

Diabete_EDB's generic association rules for different support values

Wine_EDB's generic association rules for different support values

Magic_EDB's generic association rules for different support values

Fig. 2: Generic association rule's number for different support values

for the other databases. It is also worth of to mention that the proposed weighting approach for rule use in classification has shown promising results without being tested on an adequate experimental field. Indeed, the premise of generic rules are constituted with at most two items. In that case, not all criteria evoked in subsection 4.3 are considered since only the genericity and pertinence are considered. The same also holds for precise rules.

The comparison between the generic and the precise association rules is possible. The precise rules (with and without weighting) better highlight results than do the generic ones. Indeed, the larger the rule's premise is, the more pertinent the rule becomes. On the other hand, EvAC with generic rules merges much more rules than do with precise ones. In addition, all generic rules are considered with the same weight in the fusion process despite their pertinence difference. These characteristics along with the Dempster's rule of combination behaviour mislead the fusion process to errors. Indeed, as shown in Figure 2, the high number of fused rules depends highly from the *minsup* value. Unlike the generic approach, the number of precise rule is defined by number of larger premise's rule which is dependent from the treated evidential transaction.

Table 8: Comparative result with precise classification rules

Database	Iris_EDB	Vertebral Column_EDB	Diabetes_EDB	Wine_EDB	Magic_EDB
Precise rules	80.67%	88.38%	83.20%	100%	93.88%
Precise rules with weighting	82.00%	89.03%	82.81%	100%	94.39 %

Table 9: Comparative result with Generic classification rules

Database	Iris_EDB	Vertebral Column_EDB	Diabetes_EDB	Wine_EDB	Magic_EDB
Generic rules	78.67%	67.74%	65.10%	51.68%	64.83%
Generic rules with weighting	80.00%	67.74%	65.10%	76.40%	67.58%

Table 10: Classification accuracies for several evidential databases

Dataset	EvAC	EDMA [28]	CMAR [21]	SVM	Neural Networks
Iris_EDB	82.00%	80.67%	94.00%	96.00%	97.33%
Diabete_EDB	82.81%	83.20%	75.10%	77.47%	80.60%
Wine_EDB	100%	100%	95.00%	99.43%	**100%**
Vertebral column_EDB	89.03%	88.38%	81.61%	80%	87.74%

In Table 10, we confront the EvAC associative classifier to several well-known other classifiers. We compared the accuracy of classification of the introduced EvAC to other associative classifiers such as EDMA [28] and CMAR [21].

EDMA [28] is an associative classifier on evidential databases, as EvAC, deprived from association rules weighting whereas CMAR [21] is a classical association rule classifier. The results through the table show the EvAC outperforms EDMA for all databases except the Diabete_EDB. This comfort us on the importance of the association rule weighting process. We also outperform CMAR for all datasets except Iris one. A possible reason would be the non effectiveness of imprecision modelling on Iris dataset with belief functions. Finally, by comparing EvAC accuracy to those of SVM and Neural Networks, we notice that our algorithm is competitive and provides better classification accuracy on several datasets. This could be explained by the contribution of imprecision modelling since EDMA outperforms them too.

6 Conclusion

In this paper, we tackled data mining problem in evidential databases. We detailed state-of-the-art of evidential support metric and confidence. In the first part of the remainder, we proposed a simplification of existing support measure. We also introduced a new support formula that brings precision by analysing deeply the BBA's frame of discernment. The proposed precise measure extracts more hidden frequent patterns than the usual method. In addition to frequents generation, we tackled association rule's extraction from evidential databases. We proposed a new confidence measure for association rules in evidential databases. The proposed measure is based on precise support (i.e., probability measure). The rules are then filtrated to retain only classification and non redundant rules. We also introduced an algorithm, denoted EvAC, that makes it possible to classify with evidential association rules. All generated rules are scrutinized following two criteria and a new measure for rule's relevance is introduced. The classification is based on rule's fusion with regards to their relevance. As illustrated in the experimentation section, the proposed method provides an interesting performance rates. In this work, all transaction's database are considered with the same weight. In real life applications, a transaction may represents expert's opinion. In future work, we may overcome this constraint by revising the support formula if not all expert are reliable.

Acknowledgements

The authors would like to express their sincere gratitude to the anonymous reviewers for their constructive and helpful comments and suggestions which have been in help to improve the quality of this paper.

A Evidential database creation through Evidential C-Means

From a set of numerical data such as those in Table 11, it is possible to construct an evidential database with ECM. For example, the database, presented in Table 11, is composed of 30 instances and 2 features. This dataset is composed of 2 classes $\{C_1, C_2\}$. Figure 3 illustrates the representation of these data in the feature space. From this database, the case of instance #28 will be studied (in bold in Table 11). In Figure 3, this point is represented by a pentagram.

ECM starts by creating the user requested number of cluster for each feature. In this example, we choice respectively 3 and 2 clusters for Feature n°1 and Feature n°2.

According to one feature, ECM estimates the distance between each instance and each cluster' center. A BBA is created depending on the computed distance. Afterwards, ECM tries to minimize the objective function defined in Equation (36). ECM computes recursively the cluster's center until the objective function is no more minimization is possible. From evidential data mining point of view, ECM allows us to construct for each instance, according to each feature, a BBA that represents its membership to each cluster. The clusters are different categories that we may extract for a dataset feature (column). In the proposed example, results of clustered are illustrated in Figure 4. In this figure, the studied instance is also represented by a pentagram. Thus for this instance, a BBA m_1 is obtained, with ECM, on frame of discernment $\Theta_A = \{A_1, A_2, A_3\}$ according to Feature n°1. A second BBA, m_2, is computed on frame of discernment $\Theta_B = \{B_1, B_2\}$ according to Feature n°2. These BBAs correspond to mass functions of the evidential database for each attribute (column). Table 12 shows BBAs obtained for instance #28 according to these 2 features.

Instance	Feature n°1	Feature n°2
#1	1.572823	2.178659
#2	3.374233	1.243512
#3	2.216990	1.885116
#4	1.586457	1.584009
#5	2.248950	3.157813
#6	1.603087	2.270480
#7	1.720469	2.988283
#8	2.272330	0.965524
#9	1.154969	0.998604
#10	1.276800	0.953602
#11	0.719646	1.061612
#12	1.390259	0.700845
#13	0.939414	1.251207
#14	0.519969	1.156361
#15	1.188240	1.053371
#16	2.614883	2.996433
#17	3.046593	3.467626
#18	3.331759	2.824884
#19	3.809939	2.974583
#20	2.593651	2.780790
#21	3.429305	3.097608
#22	3.444431	3.034611
#23	4.243411	1.583589
#24	1.896018	1.978944
#25	2.022313	2.364037
#26	2.054863	1.971281
#27	2.017151	2.187874
#28	**1.827643**	**2.112703**
#29	1.608744	1.980301
#30	1.764552	1.836785

Table 11: Numerical Dataset.

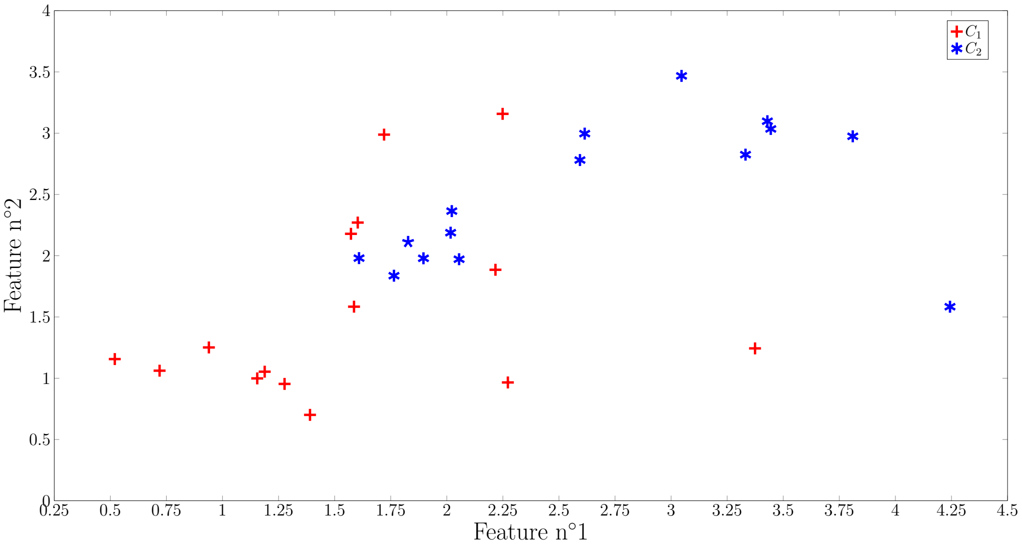


Fig. 3: Representation of data proposed in Table 11.

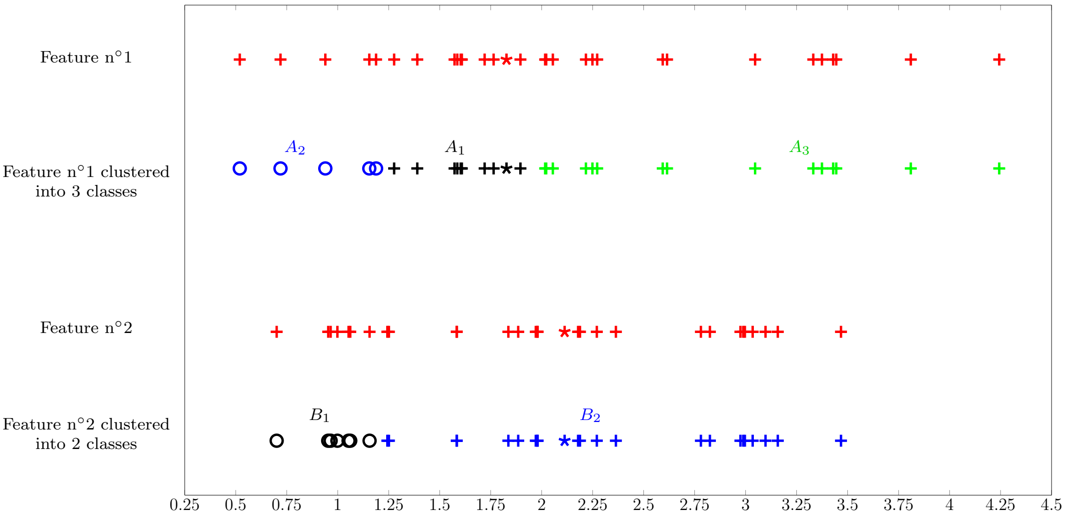


Fig. 4: ECM clustering from the given dataset of Table 11

Transaction	Attribute n°1 (from Feature n°1)	Attribute n°2 (from Feature n°2)
#28	$m_1(\{A_1\}) = 0.6855$ $m_1(\{A_2\}) = 0.0175$ $m_1(\{A_1, A_2\}) = 0.0260$ $m_1(\{A_3\}) = 0.0060$ $m_1(\{A_1, A_3\}) = 0.0147$ $m_1(\{A_2, A_3\}) = 0.0712$ $m_1(\Theta_A) = 0.1791$	$m_2(\{B_1\}) = 0.0425$ $m_2(\{B_2\}) = 0.8146$ $m_2(\Theta_B) = 0.1429$

Table 12: BBAs obtained with ECM for instance #28.

References

1. Aggarwal, C.C.: Managing and Mining Uncertain Data, vol. 35. Springer (2009)
2. Aggarwal, C.C., Li, Y., Wang, J., Wang, J.: Frequent pattern mining with uncertain data. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France pp. 29–38 (2009)
3. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In Proceedings of international conference on Very Large DataBases, VLDB, Santiago de Chile, Chile pp. 487–499 (1994)
4. Bach Tobji, M.A., Ben Yaghlane, B., Mellouli, K.: Incremental maintenance of frequent itemsets in evidential databases. In Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Verona, Italy pp. 457–468 (2009)
5. Bell, D.A., Guan, J., Lee, S.K.: Generalized union and project operations for pooling uncertain and imprecise information. Data & Knowledge Engineering **18**(2), 89–117 (1996)
6. Ben Yahia, S., Hamrouni, T., Mephu Nguifo, E.: Frequent closed itemset based algorithms: a thorough structural and analytical survey. SIGKDD Explorations **8**(1), 93–104 (2006)
7. Chui, C.K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. In Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Nanjing, China pp. 47–58 (2007)
8. Dempster, A.: Upper and lower probabilities induced by multivalued mapping. AMS-38 (1967)
9. Dubois, D., Prade, H.: Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
10. Fagin, R., Halpern, J.Y.: A new approach to updating beliefs. In: In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI'90, pp. 347–374. Elsevier Science Publishers (1990)
11. Frank, A., Asuncion, A.: UCI machine learning repository (2010). URL: http://archive.ics.uci.edu/ml
12. Gärdenfors, P.: Probabilistic reasoning and evidentiary value. In: Evidentiary Value: Philosophical, Judicial, and Psychological Aspects of a Theory: Essays Dedicated to Sören Halldén on His Sixtieth Birthday. C.W.K. Gleerups (1983)
13. Hewawasam, K.K.R., Premaratne, K., Shyu, M.L.: Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections. IEEE Transactions on Systems, Man, and Cybernetics, Part B **37**(6), 1446–1459 (2007)
14. Hewawasam, K.K.R., Premaratne, K., Shyu, M.L., Subasingha, S.P.: Rule mining and classification in the presence of feature level and class label ambiguities. In: SPIE 5803, Intelligent Computing: Theory and Applications III, 98 (2005)
15. Hong, T.P., Kuo, C.S., Chi, S.C.: Mining association rules from quantitative data. Intelligent Data Analysis **3(5)**, 363–376 (1999)

16. Hong, T.P., Kuo, C.S., Wang, S.L.: A fuzzy AprioriTid mining algorithm with reduced computational time. Applied Soft Computing **5**(1), 1–10 (2004)
17. Jousselme, A.L., Maupin, P.: Distance in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning **53**(2), 118–145 (2012)
18. Lee, S.K.: An extended relational database model for uncertain and imprecise information. In Proceedings of the 18th International Conference on Very Large Data Bases, VLDB92, Vancouver, British Columbia, Canada pp. 211–220 (1992)
19. Lee, S.K.: Imprecise and uncertain information in databases: an evidential approach. In Proceedings of Eighth International Conference on Data Engineering, Tempe, AZ pp. 614–621 (1992)
20. Leung, C.K.S., Mateo, M.A.F., Brajczuk, D.A.: A tree-based approach for frequent pattern mining from uncertain data. in Proceedings of 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Osaka, Japan **5012**, 653–661 (2008)
21. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. in Proceedings of IEEE International Conference on Data Mining (ICDM01), San Jose, CA, IEEE Computer Society pp. 369–376 (2001)
22. Manjusha, R., Ramachandran, R.: Web mining framework for security in e-commerce. In Proceedings of International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, India pp. 1043 –1048 (2011)
23. Masson, M.H., Denœux, T.: ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition **41**(4), 1384–1397 (2008)
24. Ordonez, C., Ezquerra, N., Santana, C.A.: Constraining and summarizing association rules in medical data. Knowledge and Information Systems **9**(3), 259–283 (2006)
25. Ordonez, C., Omiecinski, E.: Discovering association rules based on image content. In Proceedings of the IEEE Advances in Digital Libraries Conference (ADL'99), Baltimore, MD pp. 38–49 (1999)
26. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. Journal of Information Systems **24**, 25–46 (1999)
27. Samet, A., Lefevre, E., Ben Yahia, S.: Mining frequent itemsets in evidential database. In Proceedings of the fifth International Conference on Knowledge and Systems Engeneering, Hanoi, Vietnam pp. 377–388 (2013)
28. Samet, A., Lefèvre, E., Ben Yahia, S.: Classification with evidential associative rules. In Proceedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Montpellier, France pp. 25–35 (2014)
29. Samet, A., Lefevre, E., Ben Yahia, S.: Evidential database: a new generalization of databases? In Proceedings of 3rd International Conference on Belief Functions, Belief 2014, Oxford, UK pp. 105–114 (2014)
30. Smets, P.: Belief functions. in Non Standard Logics for Automated Reasoning , P. Smets, A. Mamdani, D. Dubois, and H. Prade, Eds. London,U.K: Academic pp. 253–286 (1988)
31. Smets, P.: The Transferable Belief Model and other interpretations of Dempster-Shafer's Model. In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI'90, MIT, Cambridge, MA pp. 375–383 (1990)
32. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence **66**(2), 191–234 (1994)
33. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with titanic. Data Knowl. Eng. **42**, 189–222 (2002)
34. Tong, Y., Chen, L., Cheng, Y., Yu, P.S.: Mining frequent itemsets over uncertain databases. In Proceedings of the 38th International Conference on Very Large Databases, VLDB12, Istanbul, Turkey **5**(11), 1650–1661 (2012)
35. Wu, X., Zhang, C., Zhang, S.: Database classification for multi-database mining. Information Systems **30**, 71–88 (2005)
36. Yin, J., Zhou, X., Yang, M.: Data mining in incomplete database. Computer Engineering **12**, 013 (2006)