Mining over a Reliable Evidential Database: Application on amphiphilic chemical database

Ahmed Samet*, Tuan Tien Dao*
*Sorbonne University, Université de technologie de Compiègne
CNRS, UMR 7338 Biomechanics and Bioengineering
Centre de Recherche de Royallieu, Compigne, France

Abstract—In recent years, the mining of frequent itemsets from uncertain databases has attracted much attention. Several researches have been conducted using different uncertain frameworks as probabilities, fuzzy sets and, most recently, evidence theory. There is very little study paid to mining pertinent knowledge from data where reliability is questionable. In this paper, we study and extend the evidential database framework in accounting data reliability. We propose new measures of support and confidence under uncertainty that consider the reliability and extend the state-of-the-art works. The proposed framework is thoroughly experimented on a real case problem for developing classification model from a chemical database.

Keywords—Reliability, Evidential database, Frequent patterns, Associative classification.

I. INTRODUCTION

Uncertain data mining has become a hot topic in the computer and information science communities in recent years because of the important number of new application fields related to this specific kind of data [1], [2]. Such data sets are often be represented by an imperfect modelling framework such as probability [1], fuzzy set theory [3], or evidence theory [4]. Recently, this discipline has flourished with new mining algorithms such as UApriori, UFP-Growth and UH-Mine [1]. However, the uncertainty is not the only origin of imperfection. In [5], Dubois and Prade highlighted two possible origins for imperfection as imprecision and uncertainty. In [6], Lee detailed both sides of imperfection that may manifest in data and proposed a new database handling imperfection. This database uses the evidence theory as a formalism to represent information [6]. This specific data structure is named the Evidential database. Mining from evidential database has caught community attention [4], [7]. In fact, the evidential databases are assimilated to expert opinions upon several questions. Interestingly though, evidential databases generalize binary and probabilistic databases [8].

In addition, when working with experimental and computational sciences, predictive models derived from data mining methods deal with another important question: reliability of observable and measurable data. In fact, each row of the database may be considered as reliable to a certain extent. Thus, when mining from this new kind of database, the prediction outcome depends strongly on the reliability of the input data. Note that the reliability of the data may be due to the choice of selected experimental techniques, data processing methods, or data source identification. For example, suppose that we aim to develop a predictive model for lung cancer based on medical imaging techniques (e.g. X-Ray or Computed Tomography (CT) scan) and laboratory tests. A reference diagnosis database must be established a priori from a set of medical diagnoses performed by a set of heterogeneous medical doctors (e.g. inexperienced and experienced doctors) from different clinics. Thus, the database includes a set of observable and measurable data and their relative diagnoses. A naive approach would consider all the data as fully reliable and would mine patterns and rules from this database. However, the problem is that each established diagnosis result depends upon the imaging technique used and the personal expertise of the doctor involved. In particular, the use of CT scans gives more reliable data for diagnosis than the use of X-Ray techniques. Consequently, the data related to CT scans should be favoured for predictive model development. Such a case is often seen within evidential databases where rows represent experts opinions with respect to several questions (i.e., database columns). The noise is then seen as a special case where only the two extreme values for the reliability are chosen. Despite its important impact and usefulness, such reliability consideration has never been tackled within an evidential database framework.

In this paper, we study and extend the evidential database framework. From a methodological point of view, this paper includes the following key contributions: (i) modelling of the evidential database structure to integrate the reliability of each transaction (i.e. data row); (ii) extension of state-of-the-art measures of support and confidence [4], [7] that consider the reliability of each row. Two variants are proposed: the optimistic and the pessimistic approaches; (iii) implementation of new algorithms of frequent patterns and association rules mining and rule-based prediction on a real-world application. From an applicative point of view, we tackle the problem of frequent patterns and valid association rules1 mining within a real chemical database. The objective is to establish predictive relationships between the structural characteristics of amphiphile molecules and their physico-chemical properties, in consideration of the data reliability level. This database includes data extracted from literature reviews (e.g. scientific papers) and experiments. For this kind of database, the data reliability derives commonly from source reliability (e.g. selected experimental protocol, used measuring technique, and human manipulation error).

This paper is organized as follows: in section II, the main principles of evidence theory and evidential data mining are recalled. In section III, we introduce a new evidential database that considers the reliability of its records. In addition, exten-

1patterns and rules having respectively support and confidence value above a fixed threshold
sions of related works on support measure and confidence are performed on reliable database cases. The mining algorithms and their performances are studied in section IV and V on the amphiphile molecule database. Finally, we conclude and sketch potential issues of future work.

II. PRELIMINARIES

In this section, we highlight the main definitions within the field of evidential data mining. The state-of-the-art support and confidence measure are presented.

An evidential database is a triplet $\mathcal{EDB} = (A, \Theta, R_{\mathcal{EDB}})$. $A$ is a set of attributes. Compared to the binary databases, even imperfect ones, the attribute (i.e., column) reflects a question that has several possible answers. The answers of a question $A_i$ ($1 \leq i \leq n$) constitute the frame of discernment $\Theta_i$. $\Theta$ is a set of $d$ transactions (i.e., rows) and could be seen as distinct information sources. $R_{\mathcal{EDB}}$ expresses the relation between the $j^{th}$ row (i.e., transaction $T_j$) and the $i^{th}$ column (i.e., attribute $A_i$) by a normalized Basic Belief Assignment (BBA) function $m_{ij} : 2^{\Theta_i} \to [0, 1]$ uyas follows:

$$m_{ij}(\emptyset) = 0, \quad \sum_{A \subseteq \Theta_i} m_{ij}(A) = 1.$$ 

In an evidential database, an item corresponds to a focal element$^3$. An itemset corresponds to a conjunction of focal elements belonging to different attributes. Two different itemsets can be related via either the inclusion or the intersection operator. Indeed, the inclusion operator for evidential itemsets [4] is defined as follows, where $X$ and $Y$ are two evidential itemsets:

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

$x_i$ and $y_j$ are respectively the $i^{th}$ and the $j^{th}$ element of $X$ and $Y$. For the same evidential itemsets $X$ and $Y$, the intersection operator is defined as follows:

$$X \cap Y = Z \iff \forall z_k \in Z, z_k \subseteq x_i \text{ and } z_k \subseteq y_j.$$  

An association rule $R$ is a causal relationship between two itemsets that can be written in the following form $R : X \to Y$ fulfilling $X \cap Y = \emptyset$.

Example 1: Let us consider a company that aims at recording the profiles of its customers. Several customers have been asked about the company products that they are using and their ages. The answers to this questionnaire have been processed and captured in the following evidential database (see Table I). The first transaction shows that costumer C1 has bought the product $P_1 \in \Theta_P = \{P_1, P_2\}$. The second attribute reflects the age of the costumers discretized in the following frame of discernment $\Theta_A = \{O \text{ (Old)}, A \text{ (Adult)}, Y \text{ (Young)}\}$. C1 is at a transition age where he is not old yet but still an adult. The product is considered as a categoric attribute whereas the age is a numeric one. In Table I, $O$ is an item and $P_1 \times O \cup A$ is an itemset such that $P_1 \subset P_1 \times O \cup A$ and $O \cap P_1 \times O \cup A = O$. Therefore, $P_1 \rightarrow O$ is an association rule.

The support within the entire database is computed by using Equation (2). The second support measure is proposed by Samet et al. [7]. The precise support measure relies on probabilistic assumptions. More precisely, for an item $x_i$, the authors have defined a new function that computes the precise value as follows:

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

Now, the transactional support of an itemset $X$ is computed from the obtained evidential database $\mathcal{EDB}$ as follows:

$$Sup_{T_j}(X) = \prod_{i \in [1..n]} Pr(x_i).$$

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Product?</th>
<th>Age?</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>$m_{11}(P_1) = 1.0$</td>
<td>$m_{21}(O \cup A) = 0.4$</td>
</tr>
<tr>
<td>C2</td>
<td>$m_{12}(P_2) = 1.0$</td>
<td>$m_{22}(Y) = 0.8$</td>
</tr>
</tbody>
</table>

whereas the age of the costumers discretized in the following frame of discernment $\Theta_A = \{O \text{ (Old)}, A \text{ (Adult)}, Y \text{ (Young)}\}$. C1 is at a transition age where he is not old yet but still an adult. The product is considered as a categoric attribute whereas the age is a numeric one. In Table I, O is an item and $P_1 \times O \cup A$ is an itemset such that $P_1 \subset P_1 \times O \cup A$ and $O \cap P_1 \times O \cup A = O$. Therefore, $P_1 \rightarrow O$ is an association rule.

The support of an itemset $X = \prod_{i \in [1..n]} x_i$, where $x_i$ is an item belonging to $\Theta_i$, can be written as follows:

$$Sup_{T_j}(X) = \prod_{i \in [1..n]} Sup_{T_j}(x_i) = \prod_{i \in [1..n]} P(x_i)$$

$$Sup_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Sup_{T_j}(X)$$

where $Sup_{T_j}$ is the support within the transaction $T_j$ (transactional support) whereas $Sup_{\mathcal{EDB}}$ is the one over the entire evidential database. $P(.)$ is a function that evaluates the pertinence of an item within a single transaction. Two major directions have been used to estimate the support of an itemset depending on the formulation of the function $P(.)$. The first approach, introduced by [4], is a pessimistic estimation of the support based on the belief function:

$$Sup_{T_j}^{B_e}(X) = \prod_{i \in [1..n]} Bel(x_i)$$

where $Bel(.)$ is belief function and is computed as follows:

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

The support within the entire database is computed by using Equation (2). The second support measure is proposed by Samet et al. [7]. The precise support measure relies on probabilistic assumptions. More precisely, for an item $x_i$, the authors have defined a new function that computes the precise value as follows:

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

Now, the transactional support of an itemset $X$ is computed from the obtained evidential database $\mathcal{EDB}$ as follows:

$$Sup_{T_j}^{Pr}(X) = \prod_{i \in [1..n]} Pr(x_i).$$

whereas the age of the costumers discretized in the following frame of discernment $\Theta_A = \{O \text{ (Old)}, A \text{ (Adult)}, Y \text{ (Young)}\}$. C1 is at a transition age where he is not old yet but still an adult. The product is considered as a categoric attribute whereas the age is a numeric one. In Table I, O is an item and $P_1 \times O \cup A$ is an itemset such that $P_1 \subset P_1 \times O \cup A$ and $O \cap P_1 \times O \cup A = O$. Therefore, $P_1 \rightarrow O$ is an association rule.

The support of an itemset $X = \prod_{i \in [1..n]} x_i$, where $x_i$ is an item belonging to $\Theta_i$, can be written as follows:

$$Sup_{T_j}(X) = \prod_{i \in [1..n]} Sup_{T_j}(x_i) = \prod_{i \in [1..n]} P(x_i)$$

$$Sup_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Sup_{T_j}(X)$$

where $Sup_{T_j}$ is the support within the transaction $T_j$ (transactional support) whereas $Sup_{\mathcal{EDB}}$ is the one over the entire evidential database. $P(.)$ is a function that evaluates the pertinence of an item within a single transaction. Two major directions have been used to estimate the support of an itemset depending on the formulation of the function $P(.)$. The first approach, introduced by [4], is a pessimistic estimation of the support based on the belief function:

$$Sup_{T_j}^{B_e}(X) = \prod_{i \in [1..n]} Bel(x_i)$$

where $Bel(.)$ is belief function and is computed as follows:

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

The support within the entire database is computed by using Equation (2). The second support measure is proposed by Samet et al. [7]. The precise support measure relies on probabilistic assumptions. More precisely, for an item $x_i$, the authors have defined a new function that computes the precise value as follows:

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

Now, the transactional support of an itemset $X$ is computed from the obtained evidential database $\mathcal{EDB}$ as follows:

$$Sup_{T_j}^{Pr}(X) = \prod_{i \in [1..n]} Pr(x_i).$$
A new metric for confidence computing based on the precise support measure is introduced in [7]. For an association rule \( R : R_a \rightarrow R_c \), the confidence is computed as follows:

\[
Conf(R) = \frac{\sum_{j=1}^{d} Sup_{T_j}(R_a) \times Sup_{T_j}(R_c)}{\sum_{j=1}^{d} Sup_{T_j}(R_a)}.
\]

The precise support measure generalizes several other support measures in imperfect and even binary databases [8]. Unfortunately, databases do not consider the reliability of the data that it contains. It is somehow assimilated to noise within databases [9]. However, such problems are more frequently seen within evidential database with unreliable experts (transactions). For instance, in Example 1, a company may be interested in looking for the profile of its relatively young costumers. To do so, the informations brought by C1 are considered unreliable. In order to consider such a constraint, a new modelling paradigm of the database is required.

### III. SUPPORT AND CONFIDENCE OVER RELIABLE DATABASE

Let us consider an evidential database representing the answers of several information sources for several questions. Each information source has a particular level of reliability. Such a database \( EDB^\alpha \) is represented by a quadruplet \((\mathcal{A}, \Theta, R, \alpha)\). Compared to \( EDB \), \( EDB^\alpha \) is also an evidential database that contains a further indication \( \alpha \) on the reliability of each transaction. \( \alpha_j \in [0, 1], j \in [1, d] \), indicates the reliability of the \( j^{th} \) transaction. Two extreme values for \( \alpha_j \) can be distinguished. An \( \alpha_j = 0 \) means that the \( j^{th} \) transaction (information source) is fully reliable and is considered for the statistical study. On the contrary, an \( \alpha_j = 1 \) means that the \( j^{th} \) transaction is unreliable.

**Example 2**: Let us consider the evidential database of a company that wishes to study the profile of its young costumers. Table II is the reliable evidential database obtained from Table I. Since C1 is not young, he is considered unreliable and a reliability value equal to 1 is attributed. C2 is young and is considered as reliable.

<table>
<thead>
<tr>
<th>Reliability ( \alpha )</th>
<th>Transaction</th>
<th>Product?</th>
<th>Age?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C1</td>
<td>m11(P1) = 1.0</td>
<td>m21(O) = 0.4</td>
</tr>
<tr>
<td>0.15</td>
<td>C2</td>
<td>m12(P2) = 1.0</td>
<td>m22(Y) = 0.8</td>
</tr>
</tbody>
</table>

The new measure of support must sustain several axioms.

Let us assume an evidential itemset \( X \), where the support measure has to satisfy the following properties:

1) Consistency: \( Sup_{T_j}^{\alpha_1}(X) = 0 \iff \alpha_j = 1 \).
2) Conservatism: \( Sup_{T_j}^{\alpha_1}(X) = Sup_{T_j}(X) \iff \alpha_j = 0 \).
3) Reliability monotony: \( Sup_{T_j}^{\alpha_1}(X) \geq Sup_{T_j}^{\alpha_2}(X) \iff \alpha_1 \leq \alpha_2 \).

where \( Sup_{T_j}^{\alpha_j} \) denotes the support of an itemset in transaction \( T_j \). The consistency axiom sustains the definition of noise within databases [9]. In fact, an unreliable transaction is not considered over the support computing process. The conservatism axiom ensures the preservation of the support for an itemset within a fully reliable transaction. Finally, the reliability monotony ensures that the more a transaction is reliable, the more its items gather support.

In this work, we dissociate two main approaches for considering the reliability of a transaction. The pessimistic and the optimistic approaches. The pessimistic approach drops the unreliable transactions from the database. On the other hand, the optimistic approach assumes that even if the transaction is unreliable, it still sustains the ignorance items (a.k.a items \( \Theta_i \)).

### A. Pessimistic approach

In this subsection, we introduce a new support measure within the reliable database based on a pessimistic approach. The support measure consists of considering the transaction relatively to its reliability. Basically, the more the transaction reliability approaches one, the less it is considered in the support computing process. The pessimistic approach drops the noise within databases [9]. In fact, an unreliable transaction is not considered over the support computing process. The conservatism axiom ensures the preservation of the support for an itemset within a fully reliable transaction. Finally, the reliability monotony ensures that the more a transaction is reliable, the more its items gather support.

The confidence can also be derived into two versions depending on the used support measure. Indeed, we distinguish between the precise based confidence \( Conf_{Pess-Pr} \) and the belief based one \( Conf_{Pess-Bel} \).

**Example 3**: Let us consider the evidential database \( EDB^\alpha \) of Table II. The support of the itemset \( P_1 \times O \) is equal to 0 with both \( Sup_{EDB^\alpha}^{Pess-Pr} \) and \( Sup_{EDB^\alpha}^{Pess-Bel} \) support measures. It is coherent with the aims of the company study (profile of young costumers). On the other hand, the support of the itemset \( P_2 \times Y \) is equal to 0.765 and 0.68 with respectively the pessimistic precise and the pessimistic belief measures. The pessimistic confidence of \( P_2 \rightarrow Y \) is computed as \( Conf_{Pess-Pr}(P_1 \rightarrow Y) = \frac{0.15\times(1-0.15)\times(1+0.9)}{0.85\times1} = 0.9 \).
B. Optimistic approach

In this subsection, we discuss another strategy for considering the reliability of a transaction. The optimistic approach assumes that even if an information source is considered as unreliable, it can still strengthen the ignorance items. Since each transaction is considered as an information source. It is possible to discount the information, modeled as a BBA, with Equation (6). As a consequence, in case of an unreliable transaction (i.e., $\alpha = 1$), the mass accorded to focal elements is transferred to the ignorance items. The result is a vacuous BBA\(^4\). Then, the support of an itemset $X$, called $Sup_{EDB}^{Opt-Bel}(X)$, within a reliable database is computed as Equation (3) and (2) with respect to:

$$Bel(\emptyset) = \sum_{\emptyset \neq B \subseteq X} m^\alpha(B)$$

where $m^\alpha$ is the discounted BBA of $m$ with respect to:

$$\begin{cases} m^\alpha(B) = (1-\alpha) \times m(B) & \forall B \subseteq \Theta \\ m^\alpha(\emptyset) = (1-\alpha) \times m(\emptyset) + \alpha. \end{cases}$$  \hspace{1cm} (6)$$

Consequently, the confidence of an association rule $R$ can be written as follows:

$$Conf^{Opt-Bel}(R) = \frac{\sum_{j=1}^{d} Sup_{T_j}^{Bel}(R_a) \times Sup_{T_j}^{Bel}(R_c)}{\sum_{j=1}^{d} Sup_{T_j}^{Bel}(R_a)}.$$  

**Example 4:** Let us consider the evidential database of Table III obtained through $EDB^\infty$ of Table II discounting. The support of the itemset $P_1 \times O$ is equal to 0. However, the support of $\Theta P \times \Theta A$ is equal to 1. Such results sound logical since the support of singleton items within unreliable transaction is null. In addition, enhancing the ignorance items indicate the absence of information regarding the treated attribute. That is why the optimistic support partially sustains the consistency axiom. The optimistic approach remains conservative and reliability monotonic. Indeed, compared to the initial belief based support within $EDB$, where the support of $P_1 \times O$ is 0.4, the optimistic approach for the support has dropped to 0. The optimistic confidence of $P_2 \rightarrow Y$ is computed as $Conf^{Opt-bel}(P_2 \rightarrow Y) = \frac{0.85 \times 0.68}{0.85 + 0.68} = 0.68$. 

In the following section, we introduce two algorithms. The Evidential Reliable Apriori (ER-Apriori) that allows to mine frequent itemset and valid association rule from reliable evidential database. Evidential Associative Classification (EvAC) algorithm allows a classification based on the valid association rules fusion.

\(^4\)The vacuous BBA, defined by $m(\emptyset) = 1$, represents complete ignorance.

IV. EVIDENTIAL RELIABLE APRIORI

To mine frequent patterns and valid association rules from reliable evidential database, a specific ER-Apriori algorithm is developed. The proposed algorithm is an Apriori-based one which has several benefits. In [4], the generation of frequent patterns within an evidential database is made with an Apriori based approach. Even in commonly used probabilistic database, UApriori which is the uncertain probabilistic version of Apriori, actually performs rather well in comparison to tree-based algorithms and is usually faster in dense uncertain datasets [2]. The evidential databases are naturally dense.

**Algorithm 1 Evidential Reliable Apriori (ER-Apriori)**

Require: $EDB$, minsup, minconf, $Size_{EDB}, \Theta_C$
Ensure: $R$, $\mathcal{EIFF}$, $m$

1: Trim Table $\leftarrow$ construct_trim($EDB$, minsup, $Size_{EDB}$)
2: $\mathcal{EIFF} \leftarrow \emptyset$
3: size $\leftarrow$ 1
4: candidate $\leftarrow$ candidate_apriori_gen($EDB$, size)
5: While (candidate $\neq \emptyset$)
6: freq $\leftarrow$ Frequent_itemset (candidate, minsup, $Trim\_Table$, $Size_{EDB}, \alpha$)
7: size $\leftarrow$ size + 1
8: $\mathcal{EIFF} \leftarrow \mathcal{EIFF} \cup freq$
9: candidate $\leftarrow$ candidate_apriori_gen($EDB$, size, freq)
10: End While
11: for all $x \in \mathcal{EIFF}$ do
12: $R \leftarrow Construct\_Rule(x, \Theta_C)$
13: if $R \neq \emptyset$ then
14: confidence $\leftarrow$ Find\_Confidence($R$, $Trim\_Table, \alpha$)
15: if confidence $\geq$ minconf then
16: $\mathcal{R} \leftarrow \mathcal{R} \cup \{R\}$
17: end if
18: end if
19: end if
20: end for

Algorithm 1 details ER-Apriori. It is a level-wise algorithm similar to UApriori [10] and the original binary Apriori [11]. The candidates generated are pruned with respect to their computed support. As UApriori, ER-Apriori includes a trimming part [10]. 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(\cdot)) or the belief function (i.e., Bel(\cdot)) of interesting items. Line 12 to 19 ensure retrieving all valid association rules. The $CONSTRUCT\_RULE(\cdot)$ function allows the construction of association rules from a frequent itemset x. Once found, the function $FIND\_CONFIDENCE(\cdot)$ computes their confidence.

Algorithm 2 details the classification process based on the largest premise rules. The evAC algorithm classifies the data with fusion techniques. Two belief functions are combined with Dempster rule of combination [12] as follows:

$$m_{\oplus}(A) = \frac{1}{1-K} \sum_{B|A=C=A} m_1(B) \times m_2(C) \forall A \subseteq \Theta, A \neq \emptyset.$$  

where $K$ is the conflictual mass [12]. The $FILTRATE\_LARGE\_PREMISE(\cdot)$ function (line 1) allows to filtrate the rules and retain only those with the largest premise, having intersection with the instance to classify or to predict X. In fact, the set of the largest premise...
Algorithm 2 Evidential Associative Classification (EvAC) algorithm

Require: $\mathcal{R}, X, \Theta_C$
Ensure: $\text{Class}$
1: $\mathcal{R}_{\text{large}} \leftarrow \text{FILTRATE_LARGE_PREMISE}(\mathcal{R}, X, \Theta_C)$
2: for all $r \in \mathcal{R}_{\text{large}}$ do
3: $m \leftarrow \left\{ \begin{array}{ll} m(\{r, \text{conclusion}\}) = \text{conf}(r) \\ m(\Theta_C) = 1 - \text{conf}(r) \end{array} \right.$
4: $m_\Theta \leftarrow m_\Theta \oplus m$
5: end for
6: $\text{Class} \leftarrow \text{argmax}_{H_k \in \Theta_C} \text{BetP}(H_k)$
7: function $\text{FILTRATE_LARGE_PREMISE}(\mathcal{R}, X, \Theta_C)$
8: $\max \leftarrow 0$
9: for all $r \in \mathcal{R}$ do
10: if $r, \text{conclusion} \in \Theta_C$ and $X \cap r, \text{premise} \neq \emptyset$ then
11: if $\text{size}(r, \text{premise}) > \max$ then
12: $\mathcal{R}_{\text{large}} \leftarrow \{r\}$
13: $\max \leftarrow \text{size}(r, \text{premise})$
14: else
15: if $\text{size}(r, \text{premise}) = \max$ then
16: $\mathcal{R}_{\text{large}} \leftarrow \mathcal{R}_{\text{large}} \cup \{r\}$
17: end if
18: end if
19: end if
20: end for
21: return $\mathcal{R}_{\text{large}}$
22: end function

rules $\mathcal{R}_{\text{large}}$ are more precise than those with the shortest premise. Once found, they are considered as independent sources and are combined (line 2 to 5). The function $\text{argmax}$ in line 6 allows the retention of the hypothesis that maximizes the pignistic probability $\text{BetP}(\cdot)$ that can be written as follows:

$$\text{BetP}(H_n) = \sum_{A \subseteq \Theta} \frac{|H_n \cap A|}{|A|} \times \frac{m(A)}{1 - m(\emptyset)} \quad \forall H_n \in \Theta.$$ 

V. REAL-WORLD APPLICATION: AMPHIPHILE MOLECULE DATABASE

Knowledge extraction from a chemical database is of great interest to the identification of useful molecules for a specific purpose. In this real case study, we aimed at predicting the relationships between structural characteristics and the physico-chemical properties of the amphiphile molecules. In particular, we focused on the prediction of the Critical Micelle Concentration (CMC) of each molecule by using its structural properties. The database is established from the domain literature using a systematic review process. Each retrieved paper is scanned and reviewed by two domain experts. Relevant information of structural characteristics and related physico-chemical properties are extracted and stored into a raw database for further processing. A transformation process is performed to establish a reliable evidential database from raw data. Two types of data are considered in this study: the numeric and the categoric. The categoric data are transformed into a certain BBA in case of numeric data, the transformation into a BBA is made thanks to Evidential C-Means (ECM) [13]. Interested reader in evidential database construction may refer to [7] for further details. The database after transformation and processing contains 199 amphiphile molecules (i.e., rows) detailed in 24 attributes (structural characteristics and related physico-chemical properties) (i.e., columns). The amphiphile molecule evidential database contains over $10^5$ items (i.e., focal elements) after transformation. For our amphiphile database, domain experts have participated in the trimming process where several items are pruned due to their insignificance. Therefore, the number of items has dropped to 267086.

Figure 1 shows the number of extracted frequent patterns for five measures: the precise, the belief, the precise pessimistic, the belief pessimistic and the belief optimistic support measures. The results shows that precise-based support provides the highest number of frequent patterns with a peak of 87423 comparatively to belief-based that has a peak of 50415. It is important to note that pessimistic support measures always provide a lower number of frequent patterns than those not using a reliability labelling. The same goes for the optimistic approach that retrieves the same number of frequent patterns as those of the belief pessimistic approach. This can be explained as a result of the pruning part conducted by experts where the ignorance hypotheses are not retained. In addition, all belief based support measures provide a lower number of frequent patterns compared to their corresponding precise-based one. This confirms that the belief is a pessimistic measure of support.

Figure 2 highlights the number of extracted frequent patterns for five measures: the precise, the belief, the precise pessimistic, the belief pessimistic and the belief optimistic support measures. The results shows that precise-based support provides the highest number of frequent patterns with a peak of 87423 comparatively to belief-based that has a peak of 50415. It is important to note that pessimistic support measures always provide a lower number of frequent patterns than those not using a reliability labelling. The same goes for the optimistic approach that retrieves the same number of frequent patterns as those of the belief pessimistic approach. This can be explained as a result of the pruning part conducted by experts where the ignorance hypotheses are not retained. In addition, all belief based support measures provide a lower number of frequent patterns compared to their corresponding precise-based one. This confirms that the belief is a pessimistic measure of support.

Table IV: Classification Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>EvAC$_{\text{Pess}}$</th>
<th>EvAC$_{\text{Pr}}$</th>
<th>EvAC$_{\text{Pess-Pr}}$</th>
<th>EvAC$_{\text{Bel}}$</th>
<th>N. Net</th>
<th>KNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>65.83</td>
<td>63.83</td>
<td>49.20</td>
<td>49.20</td>
<td>38.66</td>
<td>43.21</td>
<td>34.84</td>
</tr>
</tbody>
</table>

Table V: Recall, Precision and F-measure for EvAC$_{\text{Pess-Pr}}$ on the CMC classification

<table>
<thead>
<tr>
<th>CMC</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.075</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>0.28</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>0.79</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>1.41</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>1.69</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>1.24</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>

$A$ BBA is called a certain BBA when it has one focal element and it is a singleton. It is representative of perfect knowledge and the absolute certainty.
of the classification. In addition, the classification accuracy depends on the quality of the discretization. We use ECM, which provides a better handling to the uncertainty with BBA. In contrast, the result of the $k$-NN, the Neural Networks and SVM using the Weka software, in Table IV, are obtained after going through a PKIDiscretization. The comparison shows that our proposed framework performs more efficiently.

In Table V, we scrutinize the performance of the EvAC$^{Pess-Ptr}$ with the Recall, Precision and F-measure relatively to each CMC class. We report the $F_1$ score which is the harmonic mean between precision and recall. Specifically, the $F_1$ score is:

$$F_1 = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}, \quad \text{Prec} = \frac{tp}{tp + fp}, \quad \text{Rec} = \frac{tp}{tp + fn}$$

with $tp$, $fp$, $fn$ denoting true positives, false positives, and false negatives. Several recall values are low such as the CMC=$\{14.16\}$ comparatively to the other classes. This results can be explained by the proximity of centroid clusters found by ECM. In fact, CMC=$\{14.16\}$ and CMC=$\{16.09\}$ could be merged into one representative class for a better detection.

VI. Conclusion

In this paper, we studied and extended the toolbox for dealing with evidential databases in order to improve the reliability of each transaction. An extension of state-of-the-art support and confidence measures is introduced. Experiments are conducted on real-world chemical databases. The gathered results highlight the importance of quantifying the reliability of each row within the evidential database in regards to improving classification performance. Despite the quality of the results, the runtime of the proposed algorithms is problematic. Even if evidential databases provide a generalization for some other databases, improvements have to be made to be more competitive runtime-wise. In future work, we will investigate the scalability of the provided mining algorithm using specific strategies such as the decremental counter [1].

ACKNOWLEDGEMENT

This work was performed, in partnership with the SAS PIVERT, within the frame of the French Institute for the Energy Transition (Institut pour la Transition Energétique (ITE) P.I.V.E.R.T. (www.institut-pivert.com) selected as an Investment for the Future ("Investissements d’Avenir"). This work was supported, as part of the Investments for the Future, by the French Government under the reference ANR-001-01.

The authors would like to thank Thophile Gaudin, Hailing Lu, Anne Wadoouachi, Gwladys Pourceau, Elisabeth Van Hecke, Isabelle Pezron for their contribution in the construction of the chemical database.

REFERENCES