

# Expert Opinion Extraction from a Biomedical Database

Ahmed Samet, Thomas Guyet, Benjamin Negrevergne, Tien-Tuan Dao, Tuan Nha Hoang, Marie-Christine Ho Ba Tho

▶ **To cite this version:**

Ahmed Samet, Thomas Guyet, Benjamin Negrevergne, Tien-Tuan Dao, Tuan Nha Hoang, et al.. Expert Opinion Extraction from a Biomedical Database. Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), Jul 2017, Lugano, Switzerland. pp.1 - 12, 10.1016/S0888-613X(02)00066-X . hal-01584984

**HAL Id: hal-01584984**
**https://inria.hal.science/hal-01584984**

Submitted on 11 Sep 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Expert Opinion Extraction from a Biomedical Database

Ahmed Samet[1], Thomas Guyet[1], Benjamin Negrevergne[3], Tien-Tuan Dao[2], Tuan Nha Hoang[2], and Marie Christine Ho Ba Tho[2]

[1] Université Rennes 1/IRISA-UMR6074
firstname.lastname@irisa.fr
[2] Sorbonne University, Université de technologie de Compiègne
CNRS, UMR 7338 Biomechanics and Bioengineering, Compiègne, France
{firstname.lastname}@utc.fr
[3] LAMSADE, Université Paris-Dauphine
firstname.lastname@dauphine.fr

**Abstract.** In this paper, we tackle the problem of extracting frequent opinions from uncertain databases. We introduce the foundation of an opinion mining approach with the definition of pattern and support measure. The support measure is derived from the commitment definition. A new algorithm called OpMiner that extracts the set of frequent opinions modelled as a mass functions is detailed. Finally, we apply our approach on a real-world biomedical database that stores opinions of experts to evaluate the reliability level of biomedical data. Performance analysis showed a better quality patterns for our proposed model in comparison with literature-based methods.

**Keywords:** Uncertain database, Data mining, Opinion, OpMiner.

## 1 Introduction

Data uncertainty has challenged nearly all types of data mining tasks, creating a need for uncertain data mining. Uncertainty is inherent in data from many different domains, including social networks and cheminformatics [1]. The problem of pattern mining, or finding frequent patterns in data, has been extensively studied in deterministic databases [2] since its introduction by Aggrawal et al. [3] as well as in the field of uncertain databases [4]. The uncertain databases have brought more flexibility in data representation [5]. For instance, mass function of evidence theory are comparable to expert's opinion since it details answer to a question over a set of response elements. It also allows to model someone's degree of belief regarding an asked question. Therefore databases storing mass functions (commonly called evidential databases), are seen as a data support for expert opinions and imperfect data.

What classical approaches have in common is that they extract answers. They extract answer elements (fragment of the expert answer) as long they are redundant in the database. Therefore, the extracted information is limited and

does not describe what experts have expressed. To illustrate this point, let us consider the example of several practitioners that have been asked to give their opinion regarding new treatments for a disease. We intend to extract knowledge from a set of experts' opinions asked about their evaluations of these treatments. Each practitioner gives his opinion regarding the efficiency of a treatment $j$ among a set of evaluation possibilities $\{Good_j, Average_j, Bad_j\}$.,

| Practitioner | Treatment 1 | Treatment 2 |
|---|---|---|
| $P_1$ | $Bad_1^{0.3}\ Average_1^{0.7}$ | $Good_2^1$ |
| $P_2$ | $\{Average_1 \cup Bad_1\}^1$ | $Good_2^{0.5}\ Average_2^{0.5}$ |

Table 1: Example of uncertain database.

The first practitioner hesitates between bad and average evaluation with a higher confidence to average. The second practitioner can not decide whether the treatment is average or bad. A classical pattern mining approach as [6] would extract answers as pattern. For instance, for a threshold of 0.7, $\{Treatment1 = Average_1\}$ is a frequent pattern[4]. Looking at Table 1, the pattern $\{Treatment1 = Average_1\}$ is a fraction of the opinion expressed by the practitioner $P_1$ and therefore the extracted information is not complete. Unfortunately, this type of output is generated by uncertain mining approaches [7,8,9]. An opinion pattern would be $Treatment_1 = Bad^{0.3}\ Average^{0.7}$ and is considered as frequent since it does not contradict with the opinion of $P_2$. In this work, we intend to shake this notion of answer pattern of uncertain databases and we aim to evaluate a pattern as a whole opinion.

Methodologically, we build the foundation of an opinion mining approach. We develop our mining approach on evidential databases. Evidential databases offer more knowledge representation with its simple formalism [10]. They bring more flexibility thanks to mass functions. In fact, it is possible to model all level of uncertainty from absolute certainty to total ignorance. From applicative point of view, we experiment our OpMiner algorithm on a real-world biomedical expert database. The results show the quality of retrieved patterns comparatively to classical ones. In addition, our algorithm shows interesting performances.

## 2 Preliminaries

The evidence theory or Dempster-Shafer theory [11,12] proposes a robust formalism for modeling uncertainty. 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 $2^\Theta$ onto [0, 1]:

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

[4] A pattern is called frequent if its computed support (i.e. frequency in the database) is higher than or equal to a fixed threshold set by an expert

where $\Theta$ is the *frame of discernment*. It is the set of possible answers for a addressed problem and is composed of $N$ exhaustive and exclusive hypotheses:

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

A BBA $m$ is constrained by:

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

Each subset X of $2^\Theta$ fulfilling $m(X) > 0$ is called focal element. Constraining $m(\emptyset) = 0$ is the normalized form of a BBA and this corresponds to a closed-world assumption, while allowing $m(\emptyset) > 0$ corresponds to an open world assumption[13].

Dubois and Prade [14] have made three proposals to order BBAs. Let $m_1$ and $m_2$ be two BBA's on $\Theta$. The statement that $m_1$ is at least as committed as $m_2$ is denoted $m_1 \sqsubseteq m_2$. Three types of ordering have been proposed:

- *pl-ordering* (plausibility ordering) if $Pl_1(A) \leq Pl_2(A)$ for all $A \subseteq \Theta$, we write $m_1 \sqsubseteq_{pl} m_2$,
- *q-ordering* (communality ordering) if $q_1(A) \leq q_2(A)$ for all $A \subseteq \Theta$, we write $m_1 \sqsubseteq_q m_2$,
- *s-ordering* (specialization ordering) if $m_1$ is a specialization of $m_2$, we write $m_1 \sqsubseteq_s m_2$,

In this paper, we develop our approach using the plausibility based commitment. The plausibility function $Pl(.)$ is defined as follows:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2}$$

Among all belief functions on $\Theta$, the least committed belief function is the vacuous belief function (i.e. $m(\Theta) = 1$).

Finally, it is possible to store imperfect data modelled as BBAs into a database. This kind of database is commonly called evidential database. Formally, an *evidential database* is a triplet $\mathcal{EDB} = (\mathcal{A}_{\mathcal{EDB}}, \mathcal{O}, R_{\mathcal{EDB}})$. $\mathcal{A}_{\mathcal{EDB}}$ is a set of attributes and $\mathcal{O}$ is a set of $d$ transactions (i.e., rows). Each column $A_j$ $(1 \leq j \leq n)$ has a domain $\Theta_j$ of discrete values. $R_{\mathcal{EDB}}$ expresses the relationship between the $i^{th}$ transaction (i.e., row $T_i$) and the $j^{th}$ column (i.e., attribute $A_j$) by a normalized BBA $m_{ij} : 2^{\Theta_j} \rightarrow [0, 1]$.

*Example 1.* We intend to extract knowledge from a set of experts' opinions asked about their evaluations of several treatment efficiencies for a disease. Each practitioner gives his opinion regarding a treatment $j$ from a set of evaluation possibilities $\Theta_j = \{Good_j, Average_j, Bad_j\}$.

## 3 Extraction opinion patterns over evidential databases

In the following subsection, we study the plausibility based commitment relation between two BBAs in the evidence theory.

| Practitioner | Treatment 1 | Treatment 2 |
|---|---|---|
| $P_1$ | $m_{11}(Good_1) = 0.7$ | $m_{12}(Good_2) = 0.4$ |
| | $m_{11}(\Theta_1) = 0.3$ | $m_{12}(Average_2) = 0.2$ |
| | | $m_{12}(\Theta_2) = 0.4$ |
| $P_2$ | $m_{21}(Good_1) = 0.6$ | $m_{22}(Good_2) = 0.3$ |
| | $m_{21}(\Theta_1) = 0.4$ | $m_{22}(\Theta_2) = 0.7$ |

Table 2: Example of evidential database

### 3.1 Plausibility based commitment measure

Let us consider two BBAs $m_1$ and $m_2$ such as $m_1 \sqsubseteq_{pl} m_2$. We intend to develop a measure to estimate the commitment level of $m_2$ wrt $m_1$.

**Definition 1.** *Given the plausibility functions $Pl_1$ and $Pl_2$ of two BBAs $m_1$ and $m_2$, the plausibility $PL_{12}(.)$ expresses the difference between two plausibility functions and is computed as follows:*

$$PL_{12}(A) = Pl_1(A) - Pl_2(A). \tag{3}$$

**Definition 2.** *Assuming two BBAs $m_1$ and $m_2$. Assuming that $C(\cdot,\cdot)$ is a commitment measure between two BBAs. It is computed as follows,*

$$\begin{aligned} &C : 2^\Theta \times 2^\Theta \mapsto [0,1] \\ &(m_2, m_1) \rightarrow \begin{cases} 1 - ||PL_{21}|| = 1 - \sqrt{\sum\limits_{A \subseteq \Theta} PL_{21}(A)^2} \; if \; m_1 \sqsubseteq_{pl} m_2 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad Otherwise \end{cases} \end{aligned} \tag{4}$$

*Property 1.* Assuming two BBAs $m_1$ and $m_2$ such as $m_2 \sqsubseteq_{pl} m_1$, Equation 4 verifies the following properties:

- $C(m_2, m_1) \geq 0$ *(separation axiom)*;
- $C(m_2, m_1) = 1 \Leftrightarrow m_1 = m_2$ *(identity of indiscernible)*;
- $C(m_2, m_1) = C(m_1, m_2)$ *(symmetry)*;
- $C(m_2, m_3) \leq C(m_2, m_1) + C(m_1, m_3)$ *(triangle inequality).*

### 3.2 Mining opinions over evidential databases

In an evidential database, an *item* corresponds to a BBA. An *itemset* (so called *pattern*) corresponds to a conjunction of several BBAs having different domains $X = \{m_{ij} \in \mathcal{M}^\Theta\}$. We recall that $i$ is the transaction id and $j$ is the attribute id. $\mathcal{M}^\Theta$ denotes the set of all BBAs in $\mathcal{EDB}$.

Let us consider an evidential database $\mathcal{EDB}$ and the itemset $X$ made of a set of BBAs. The frequency of appearance of an item $x = m_{i'j}$ in a transaction $T_i$ can be computed as follows:

$$\begin{aligned} Sup_{T_i} : \mathcal{M}_i^{\Theta_j} &\rightarrow [0,1] \\ x &\mapsto C(x, m_{ij}) \text{ where } m_{ij} \in \mathcal{M}_i^{\Theta_j}. \end{aligned} \tag{5}$$

$\mathcal{M}_i^{\Theta_j}$ is the set of BBAs in the row $T_i$ of the attribute $j$. As illustrated above, the $Sup_{T_i}$ is a measure that computes whether $x$ is in the row $T_i$. Even if the BBA is not in the studied row, we analyse if there is a BBA that generalizes it. Then, the support of an itemset $X$ over the transaction $T_i$ is computed as

$$Sup_{T_i}(X) = \prod_{x \in X} Sup_{T_i}(x). \tag{6}$$

Therefore, the support of $m_{ij}$ over the database is computed as,

$$Sup_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{i=1}^{d} Sup_{T_i}(X). \tag{7}$$

*Property 2.* Assuming an itemset $X$, the measure of support fulfils the anti-monotony property, i.e.,

$$Sup_{\mathcal{EDB}}(X) \leq Sup_{\mathcal{EDB}}(X \cup m_{ij}). \tag{8}$$

*Proof.* Assuming an evidential database $\mathcal{EDB}$, let us consider two evidential itemsets $X$ and $X \cup m_{ij}$. We aim at proving this relation $Sup_{\mathcal{EDB}}(X) \leq Sup_{\mathcal{EDB}}(X \cup m_{ij})$:

$$\begin{aligned}
&Sup_{T_i}(X \cup m_{ij}) = \prod_{m_{i'j'} \in X \cup m_{ij}} Sup_{T_i}(m_{i'j'}) \\
&Sup_{T_i}(X \cup m_{ij}) = \prod_{m_{i'j'} \in X} Sup_{T_i}(m_{i'j'}) \times Sup_{T_i}(m_{ij}) \\
&Sup_{T_i}(X \cup m_{ij}) \leq Sup_{T_i}(X) \quad \text{since} \quad Sup_{T_i}(m_{ij}) \in [0,1] \quad \text{then} \\
&Sup_{\mathcal{EDB}}(X \cup m_{ij}) \leq Sup_{\mathcal{EDB}}(X).
\end{aligned}$$

*Example 2.* Assuming the evidential database given in Example 1. For a $minsup = 0.7$, the pattern $\{m_{11}, m_{12}\}$ have a support of $\frac{C(m_{11},m_{11}) \times C(m_{12},m_{12}) + C(m_{11},m_{21}) \times C(m_{12},m_{22})}{2} = 0.765$ and is then considered as frequent. Semantically, having a relatively good opinion on treatment 1 (i.e. $m_{11}$) and hesitant one regarding the treatment 2 (i.e. $m_{12}$) is redundant over 76.5% of asked practitioners. Moreover, comparatively to patterns of an evidential data mining algorithm, our output is more informative. In fact, a classical algorithm would provide the frequent pattern $\{Good_1, Good_2\}$ which contain less details than $\{m_{11}, m_{12}\}$.

In this section, we develop a new level-wise algorithm to mine opinions over evidential databases. OpMiner, shown in Algorithm 1 generates all BBAs of size one by favouring the most specific ones. Formally, for all $m_{ij}, m_{i'j} \in \mathcal{M}^{\Theta}$, we retain $m_{ij}$ as long as $m_{ij} \sqsubseteq_{pl} m_{i'j}$. Thus, function *candidate_gen* reduces the set of frequent patterns to the set of the more specific ones. The other less specific BBAs are used to compute the support as described in Equation 7. In addition, this selection aims at reducing time computing since candidate generation and support computing depends on the set of items (i.e. pattern with a single BBA). The patterns that have a support lower than the *minsup* are pruned in line 5. The process stops until no candidate is left.

**Algorithm 1** OpMiner algorithm

**Require:** $\mathcal{EDB}, minsup, \mathcal{EDB}_{pl}, maxlen$
**Ensure:** $\mathcal{EIFF}$
$\mathcal{EIFF}, Items \leftarrow \emptyset, size \leftarrow 1$
$Items \leftarrow$ CANDIDATE_GEN$(\mathcal{EDB}, \mathcal{EIFF}, Items)$
**While** $(candidate \neq \emptyset \ and \ size \leq maxlen)$
**for all** $pat \in candidate$ **do**
  **if** SUPPORT$(pat, minsup, \mathcal{EDB}_{pl}, Size_\mathcal{EDB}) \geq minsup$ **then**
    $\mathcal{EIFF} \leftarrow \mathcal{EIFF} \cup pat$
$size \leftarrow size + 1$
$candidate \leftarrow$ CANDIDATE_GEN$(\mathcal{EDB}, \mathcal{EIFF}, Items)$
**End While**
**function** SUPPORT($pat$, $minsup$, $\mathcal{EDB}_{pl}$,$d$)
  $Sup \leftarrow 0$
  **for** i=1 to d **do**
    **for all** $pl_{ij} \in \mathcal{M}_i$ **do**
      $pl \leftarrow mtopl(pat)\backslash\backslash$ *computes the plausibility out of a BBA*
      **if** $pl_{ij} \geq pl$ **then**
        $Sup_{Trans} \leftarrow Sup_{Trans} \times 1 - ||pl_{ij} - pl||$
    $Sup \leftarrow Sup + Sup_{Trans}$
  **return** $\frac{Sup_I}{d}$
**function** CANDIDATE_GEN($\mathcal{EDB}$, $\mathcal{EIFF}$, $Items$)
  **if** $size(Items) = 0$ **then**
    **for all** $BBA \in \mathcal{EDB}$ **do**
      **while** $Items \neq \emptyset \ and \ BBA \not\sqsubseteq_{pl} it$ **do**
        **if** $Items = \emptyset$ **then**
          $Add(BBA, Item)$
        **else**
          $Replace(BBA, it, Item)$
    **return** $Items$
  **else**
    **for all** $BBA \in \mathcal{EIFF}$ **do**
      **for all** $it \in Items$ **do**
        **if** $!same_attribute(it, BBA)$ **then**
          $Cand \leftarrow Cand \cup \{BBA \cup it\}$
    **return** $Cand$

## 4 Experiments: Data reliability assessment using biomedical expert opinion

The investigation of the effects of muscles morphology and mechanics on motion, and the risks of injury, has been at the core of many studies, sometimes with conflicting results. Often different measurement methods have been used, making comparison of the results and drawing sound conclusions impossible [15]. In this section, we aim at studying the opinion of several experts on collected measurement data. To do so, we collected data by a systematic review pro-

cess of 20 data sources (papers) from reliable search engines (PubMed and ScienceDirect). Data is described over 7 parameters regarding muscle morphology, mechanics and motion analysis. Four main questions were asked to experts about measuring technique ($Q_1$), experimental protocol ($Q_2$), number of samples ($Q_3$) and range of values ($Q_4$). An expert opinion database was built from an international panel of 20 contacted experts with different expertise (medical imaging, motion analysis). Five evaluation degrees were possible $\{Very\ high, High, Moderate, Low, Very\ low\}$. Each given degree was associated to a confidence value. In this study, as a first goal, we aim at finding frequent opinions in the database. Frequent opinions show correlation between opinions. The second goal is to evaluate the reliability of sources. To do so, the algorithm selects from the frequent set of patterns those that express a positive opinion regarding the same source.

Table 3 shows a small set of recorded answers from experts. For instance, row 1 details the opinions of expert 1 regarding the source S1 (data measures retrieved from a source). The expert expresses his opinion over 4 questions. The column $Conf_i$ shows the confidence of the expert regarding his given opinion for the question $Q_i$.

| Expert | S1 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $Q_1$ | $Conf_1$ | $Q_2$ | $Conf_2$ | $Q_3$ | $Conf_3$ | $Q_4$ | $Conf_4$ |
| 1 | Hig | Hig | Hig | Hig | Mo | Hig | Hig | Mo |
| 2 | Hig | Ver | Mo | Ver | Hig | Ver | Mo | Ver |
| 3 | Hig | Hig | Hig | Hig | Hig | Hig | Hig | Hig |
| 4 | Hig | Hig | Mo | Hig | Hig | Hig | Mo | Hig |
| 5 | Lo | Ver | Lo | Ver | Mo | Ver | Mo | Ver |
| 6 | Mo | Mo | Mo | Mo | Lo | Hig | Lo | Hig |
| 7 | Mo | Ver | Mo | Ver | Hig | Ver | Mo | Ver |
| 8 | Mo | Ver | Lo | Hig | Hig | Ver | Lo | Ver |
| 9 | Mo | Ver | Mo | Hig | Hig | Ver | Mo | Hig |
| 10 | Mo | Hig | Mo | Hig | Mo | Hig | Mo | Hig |
| 11 | Ver | Ver | Ver | Ver | Ver | Ver | Ver | Ver |

| | |
|---|---|
| Very high | Very high confidence |
| High | High confidence |
| Moderate | Moderate confidence |
| Low | Low confidence |
| Very low | Very low confidence |

Table 3: Sample of the expert opinion data.

The evidential database is constructed by using the evaluation of the experts and their confidences. First, the evaluation of the expert is used to model a certain BBA[5]. Then, the confidence is used to integrate uncertainty into the BBA. To do so, the confidence is used as reliability measure and part of the mass initially given to the evaluation is then transferred to the ignorance mass.

[5] A BBA is called a certain BBA when it has one focal element, which is a singleton. It is representative of perfect knowledge and the absolute certainty.

Formally, the discounting of a mass function $m$ can be written as follows

$$\begin{cases} m^{\alpha}(B) = (1-\alpha) \times m(B) & \forall B \subseteq \Theta \\ m^{\alpha}(\Theta) = (1-\alpha) \times m(\Theta) + \alpha. \end{cases} \tag{9}$$

$\alpha$ is the reliability factor and is in the set $\{0.8, 0.6, 0.4, 0.2, 0\}$. The higher $\alpha$ is the more mass is transferred to $m(\Theta)$.

In the following, we compare a classical evidential pattern mining approaches such as EDMA [16] and U-Apriori [6] with the output of OpMiner. To do so, we compare these three algorithms in terms of number of extracted patterns and computational time. Figure 1 illustrates the number of extracted patterns with regards to the threshold *minsup*. It is evident that the pattern mining approach EDMA finds the highest number of patterns for all fixed *minsup* comparatively to probabilistic approach approach and OpMiner. In fact, EDMA computes frequent patterns from a set of $28 \times 2^5$ items (i.e. sum of the size of all superset of attributes). Therefore, EDMA extracts more patterns than the probabilistic U-Apriori that mines from a set of $28 \times 5$ items (i.e. sum of the size of all frames of discernment). OpMiner is has a different approach since an item is a BBA and therefore the number of items is the number of BBAs in the database (i.e., $28 \times 11$). In addition, this number is reduced by selecting, at first, only the more committed BBAs. As a result OpMiner is more efficient than the two other approaches since it generates less candidates. OpMiner not only generates less frequent patterns but more informative ones since it regroups several information in a single item. Even if in our application, all treated BBAs are *simple*[6], OpMiner works perfectly on *normal* BBAs[7].

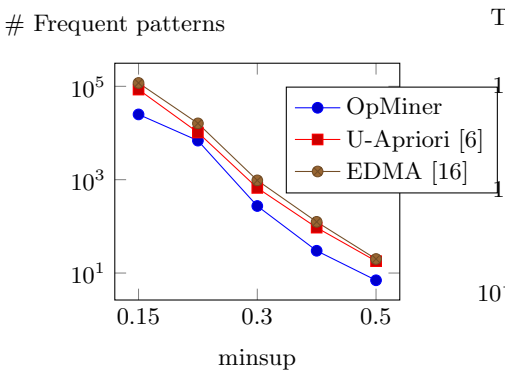



Fig. 1: Number of retrieved frequent patterns from the database.

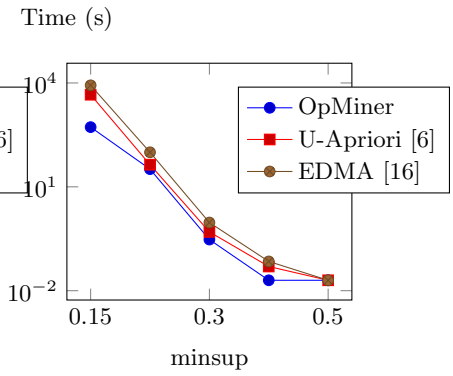



Fig. 2: Number of retrieved valid association rules from the database.

[6] A BBA is said to be simple if it has at most two focal sets and, if it has two, $\Theta$ is one of those.

[7] A BBA is said to be normal if $\emptyset$ is not a focal set.

In order to test the quality of the patterns, we oppose the best pattern of EDMA relatively to the first four attributes shown in Table 4 to the best one provided by OpMiner. In fact, it is possible to select from the set of frequent patterns those having items of the four attributes. These patterns show the answer (opinion for OpMiner) that the majority of the experts have expressed. These patterns are representative of the quality of source $S1$ measures. As it is show in

| | EDMA S1 best pattern | OpMiner S1 best pattern |
|---|---|---|
| Pattern | {Q1=Hig or Mod, Q2=Hig or Mod, Q3=Hig or Mod, Q4=Hig or Mod} | $\{m_1(Mo_1)=1, \begin{cases} m_2(Mo_2)=0.8 \\ m_2(\Theta_2)=0.2 \end{cases} m_3(Hig_3)=1, \begin{cases} m_4(Mo_4)=0.8 \\ m_4(\Theta_4)=0.2 \end{cases} \}$ |

Table 4: EDMA's pattern Vs. OpMiner's pattern

Table 4, the construction of both patterns is not the same. EDMA's pattern is constructed from focal elements in contrary of OpMiner that contains BBAs. In addition, the interpretation of both patterns is different. EDMA's pattern shows a hesitation between *high* and *moderate* as an answer trend. Therefore, from this point, making an evaluation of source S1 is not straightforward. OpMiner pattern has a different meaning. It gives for each asked question the most shared opinion (i.e. BBA). It means that, for question 1, 2 and 4 the answer is *moderate* with *high* or *very high* confidence. For question 4, the trend is a high evaluation with a *very high* confidence. As a result, with an overall moderate evaluation of its measure, it is possible to conclude that source S1 is moderately reliable.

## Conclusion

In this paper, we introduced a new approach for mining opinion patterns from uncertain database. The uncertainty and the imprecision of the data are modelled with the evidence theory. The extraction is based on new anti-monotonic measures of support derived from the commitment relation. A mining algorithm OpMiner is then applied to retrieve frequent opinions patterns a from the database. The results on a real-world database shows more informative extracted patterns than literature-based approaches. In future work, we will be interested in refining the inclusion and support measure using the specialization matrix of Smets [17]. Furthermore, the performance of OpMiner algorithm could be improved by adding specific heuristics such as the decremental pruning[7].

**Acknowledgements** This work is a part of the PEPS project funded by the French national agency for medicines and health products safety (ANSM), and of the SePaDec project funded by Region Bretagne.

## References


1. Samet, A., Dao, T.T.: Mining over a reliable evidential database: Application on amphiphilic chemical database. In Proceeding of 14th International Conference on Machine Learning and Applications, Miami, Florida (2015) 1257–1262
2. Aggarwal, C.C., Han, J.: Frequent pattern mining. Springer (2014)
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 (1994) 487–499
4. 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 (2009) 29–38
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) (1996) 89–117
6. 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 (2007) 47–58
7. Aggarwal, C.C.: Managing and Mining Uncertain Data. Volume 3. Springer (2010)
8. Hewawasam, K.R., Premaratne, K., Shyu, M.L.: Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections. Trans. Sys. Man Cyber. Part B **37**(6) (2007) 1446–1459
9. Chen, Y., Weng, C.: Mining association rules from imprecise ordinal data. Fuzzy Set Syst **159(4)** (2008) 460–474
10. 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 (2009) 457–468
11. Dempster, A.: Upper and lower probabilities induced by multivalued mapping. AMS-38 (1967)
12. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
13. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence **66**(2) (1994) 191–234
14. Dubois, D., Prade, H.: The principle of minimum specificity as a basis for evidential reasoning. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Paris, France (1986) 75–84
15. Hoang, T.N., Dao, T.T., Ho Ba Tho, M.C.: Clustering of children with cerebral palsy with prior biomechanical knowledge fused from multiple data sources. In Proceedings of 5th International Symposium Integrated Uncertainty in Knowledge Modelling and Decision Making, Da Nang, Vietnam (2016) 359–370
16. Samet, A., Lefèvre, E., Ben Yahia, S.: Evidential data mining: precise support and confidence. Journal of Intelligent Information Systems (2016) 1–29
17. Smets, P.: The application of the matrix calculus to belief functions. International Journal of Approximate Reasoning **31**(12) (2002) 1–30