Expert Opinion Extraction from a Biomedical Database

A. Samet[1] T. Guyet[1] B. Negrevergne[2]
T-T. Dao[3] T. N. Hoang[3] M. C. Ho-Ba-Tho[3]

[1] IRISA-UMR6074
[2] LAMSADE - Université Paris Dauphine
[3] UTC – UMR 7338 Biomechanics and Bioengineering

ECSQARU – 12/07/2017

UMR IRISA

Problematic

Example: opinions about efficiency of treatments

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

→ $Bad_1^{0.3}\ Average_1^{0.7}$: *Bad* for 30% of cases, *Average* for 70%

→ $\{Average_1 \cup Bad_1\}^1$: undistinguishable opinion (but certainly not *Good*)

Our objective: Extracting "shared opinions" from such database of opinions (uncertain data)

Challenges

→ Developing an efficient algorithm to extract "shared opinions"

→ Considering opinions as a whole

State-of-the-art methods

State-of-the-art

- Frameworks and algorithms based on probabilities, fuzzy set and evidence theory
- Approaches that consider subsets of the opinions
  - False knowledge: a fraction of the opinion is not representative
  - Too many useless patterns

  $\Rightarrow$ our proposal lies in extracting opinions that exist in the database

Example: efficiency of treatments (patterns with $\sigma = 0.7$

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

- $\{Treatment_1 = Average_1\}$ is a frequent evidential pattern
- $\{Treatment_1 = Bad_1^{0.3}\ Average_1^{0.7}\}$: is it representative of σ opinions?

State-of-the-art methods

State-of-the-art

- Frameworks and algorithms based on probabilities, fuzzy set and evidence theory
- Approaches that consider subsets of the opinions
  - False knowledge: a fraction of the opinion is not representative
  - Too many useless patterns

  $\Rightarrow$ our proposal lies in extracting opinions that exist in the database

Example: efficiency of treatments (patterns with $\sigma = 0.7$

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

- $\{Treatment_1 = Average_1\}$ is a frequent evidential pattern
- $\{Treatment_1 = Bad_1^{0.3}\ Average_1^{0.7}\}$: is it representative of σ opinions?

State-of-the-art methods

State-of-the-art

- Frameworks and algorithms based on probabilities, fuzzy set and evidence theory
- Approaches that consider subsets of the opinions
  - False knowledge: a fraction of the opinion is not representative
  - Too many useless patterns

  $\Rightarrow$ our proposal lies in extracting opinions that exist in the database

Example: efficiency of treatments (patterns with $\sigma = 0.7$

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

- $\{Treatment_1 = Average_1\}$ is a frequent evidential pattern
- $\{Treatment_1 = Bad_1^{0.3}\ Average_1^{0.7}\}$: is it representative of σ opinions?

Our proposal

Opinion mining framework

- Evidential databases for opinion modelling
- Define a measure of inclusion between opinions
- Define a measure of support based on the inclusion metric
- Developing a level-wise algorithms for opinion mining
- Apply on biomedical data reliability problem

Evidence theory

Belief function theory [Dempster 67 & Shafer 76]

- Let it be $\theta = \{H_1, H_2, ..., H_N\}$ the set all possible answers for a question Q: Frame of discernment
- A Basic Belief Assignment (BBA) is a $m : 2^\theta \rightarrow [0, 1]$ such that:

$$\sum_{A \subseteq \theta} m(A) = 1$$

- A BBA m represents the state of knowledge of a rational agent Ag at an instant t
- $m(A)$: part of belief accorded to A
- $m(\theta)$: represents the ignorance mass
- A is a focal element if and only if $m(A) > 0$

Evidential database

Definition

- An evidential database is a triplet $\mathcal{EDB} = (\mathcal{A}, \mathcal{O}, R_{\mathcal{EDB}})$.
- $\mathcal{A}$ is a set of attributes.
- $\mathcal{O}$ is a set of d transactions (i.e., rows).
- $R_{\mathcal{EDB}}$ expresses the relation between the j^{th} line (i.e., transaction T_j) and the i^{th} column (i.e., attribute A_i) by a normalized BBA.

Item & itemset

- An item is a BBA for a given attribute
- An itemset is a conjunction of BBAs (one per attribute)
  - an itemset contain an item for all attributes
  - $m_{ij} \in \mathcal{M}^{\Theta}$ denotes the opinion of j-th attribute for i-th transaction
  - where $\mathcal{M}^{\Theta}$ set of all BBAs in $\mathcal{EDB}$

Example

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: Example of evidential database

- $m_{11} = \begin{cases} m_{11}(Good_1) = 0.7 \\ m_{11}(\Theta_1) = 0.3 \end{cases}$ is an item.
- $\{m_{11}, m_{12}\}$ is an itemset.

Inclusion between itemsets I

- Assuming an itemset $X = \{m_{ij} \in \mathcal{M}^{\theta}\}$, and $\mathcal{EDB}$ a database of opinions, "how much" the pattern X appears in transaction?

  $\rightarrow$ require to evaluate the inclusion of X in a transaction of $\mathcal{EDB}$
- Determining whether a pattern X is sufficiently frequent (given a threshold σ).

  $\rightarrow$ expected monotonicity property
  - def: if an itemset is not frequent, any super-itemset is frequent
  - enables efficient pruning of a priori unfrequent patterns

$\Rightarrow$ main idea: use commitment measures and its induced ordering

Inclusion between itemsets II

Belief ordering

- Let m_1 and m_2 be two BBA's on Θ. $m_1 \sqsubseteq m_2$ denotes that "m_1 is at least as committed as 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 $m_1 \sqsubseteq_s m_2$,

Inclusion between itemsets III

Plausibility based commitment measure

- Assuming two BBAs m_1 and m_2 such that $m_1 \sqsubseteq_{pl} m_2$.
- Assuming that $C(\cdot,\cdot)$ is a commitment measure between two BBAs.

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

where

$$Pl_{12}(A) = Pl_1(A) - Pl_2(A).$$

and

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

Support measure

- Assuming an itemset $X = \{m_{ij} \in \mathcal{M}^{\Theta}\}$
- The support of an item $x = m_{i'j}$ in a transaction T_i :

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

- The support of itemset X in T_i:

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

$\rightarrow$ $Sup_{T_i}(X) \in [0,1]$

- The support of itemset X in $\mathcal{EDB}$

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

$\rightarrow$ $Sup_{\mathcal{EDB}}$ is anti-monotonic (see article)

Example

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

Table: Example of evidential database

- Assuming $X = \left\{ \begin{Bmatrix} m(Good_1) = 0.8 \\ m(\Theta_1) = 0.2 \end{Bmatrix} \right\}$,
  $Sup_{\mathcal{EDB}}(X) = \frac{C(m,m_{11}) \times C(m,m_{21})}{2} = 0.56$ (frequent pattern).
- Assuming $X = \left\{ \left\{ m'(Good_2) = 1 \right\} \right\}$,
  $Sup_{\mathcal{EDB}}(X) = \frac{C(m',m_{12}) \times C(m',m_{22})}{2} = 0.22$ (infrequent pattern).

Example

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

Table: Example of evidential database

- Assuming $X = \left\{ \left\{ \begin{array}{l} m(Good_1) = 0.8 \\ m(\Theta_1) = 0.2 \end{array} \right. \right\}$, $Sup_{\mathcal{EDB}}(X) = \frac{C(m,m_{11}) \times C(m,m_{21})}{2} = 0.56$ (frequent pattern).
- Assuming $X = \left\{ \left\{ m'(Good_2) = 1 \right. \right\}$, $Sup_{\mathcal{EDB}}(X) = \frac{C(m',m_{12}) \times C(m',m_{22})}{2} = 0.22$ (infrequent pattern).

OpMiner

OpMiner

- Input: a table that contains precomputed plausibilities of all BBAs
- Two parameters:
  - *maxlen*: the maximum size of patterns
  - σ: the frequency threshold
- Level-wise mining algorithm
  - Generate candidates of size n from frequent patterns of size $n-1$
  - Evaluate the support of candidate patterns of size n
  - $n \leftarrow n+1$ until there is frequent patterns of size n

Search space

The generation of candidates is based on $\mathcal{EDB}$ content

- only existing opinions are used
- avoid to explore a too wide search space (set of BBA)

OpMiner

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**
    $\mathcal{EIFF} \leftarrow \mathcal{EIFF} \cup pat$
  end if
end for
$size \leftarrow size + 1$
$candidate \leftarrow$ 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**
      $pl \leftarrow mtopl(pat)$\\\\ *computes the plausibility out of a BBA*
      if $pl_{ij} \geq pl$ **then**
        $Sup_{Trans} \leftarrow Sup_{Trans} \times 1 - ||pl_{ij} - pl||$
      end if
    end for
    $Sup \leftarrow Sup + Sup_{Trans}$
  end for
  return $\frac{Sup_I}{d}$
end function
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* $BBA \not\subseteq_{pl} it$ **do**
        if $Items = \emptyset$ **then**
          $Add(BBA, Item)$
        else
          $Replace(BBA, it, Item)$
        end if
      end while
    end for
    return *Items*
  else
    for all $BBA \in \mathcal{EIFF}$ **do**
      for all $it \in Items$ **do**
        if $!same\_attribute(it, BBA)$ **then**
          $Cand \leftarrow Cand \cup \{BBA \cup it\}$
        end if
      end for
    end for
    return *Cand*
  end if
end function

Experiments

- Comparison with two alternative approaches:
  - U-Apriori: probabilistic itemset miner
  - EDMA: Evidential itemset miner
- Evaluation criteria
  - number of patterns
  - computing time
  - qualitative evaluation
- Evaluation dataset: use of a real use-case

Application details

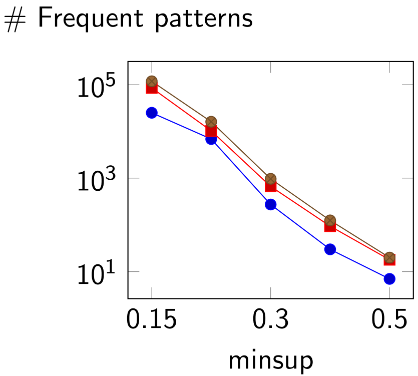
Application and dataset description [3]

- Objective: biomedical data reliability (reliable clinical decision support).
- Data collection: systematic review process
  - 7 parameters: muscle morphology and mechanics and motion analysis
  - 20 data sources (papers) from reliable search engines (PubMed and ScienceDirect) : multiple sources (2-7) for one parameter
- Questionnaire: Google Form (remote assessment)
  - Four main questions: measuring technique, experimental protocol, number of samples, range of values
  - Four complementary questions: confidence levels
- Expert opinion database: international panel: 20 contacted and 11 (opinions received) from experts with different expertise (medical imaging, motion analysis)

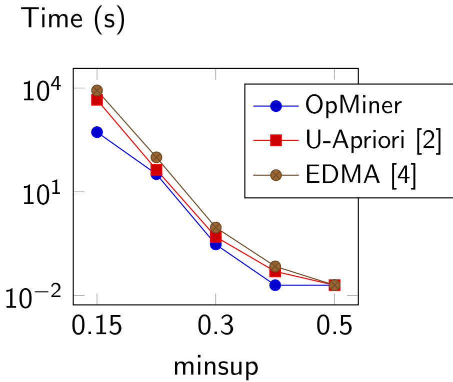
Opinion dataset

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

Mining performance



Frequent patterns



Computational time

Pattern comparison

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

Classical pattern Vs. OpMiner pattern

Pattern comparison

EDMA S1 best pattern	OpMiner S1 best pattern
$\{m_1(Hig_1 \cup Mod_1) = 1,\ m_2(Hig_2 \cup Mod_2) = 1,$ $m_3(Hig_3 \cup Mod_3) = 1, m_4(Hig_4 \cup Mod_4) = 1\}$	$\{m_1(Mo_1) = 1, \begin{cases} m_2(Mo_2) = 0.8 \\ m_2(\Theta_2) = 0.2 \end{cases}$ $m_3(Hig_3) = 1, \begin{cases} m_4(Mo_4) = 0.8 \\ m_4(\Theta_4) = 0.2 \end{cases} \}$

Classical pattern Vs. OpMiner pattern

Conclusion and Perspectives

Conclusion

- We tackle the extraction of shared opinion patterns from uncertain database (evidential databases)
- We proposed to use a measure based on commitment to encode itemset inclusion (use of plausibility)
- We derived a support measure for BBAs
- Application on expert opinion biomedical database

Perspectives

- Refining the inclusion and support measure using the specialization matrix of Smets [5]
- Improving the scalability of OpMiner by decremental pruning [1]

Thank You
for your attention.

Charu C Aggarwal.
Managing and Mining Uncertain Data, volume 3.
Springer, 2010.

C-K Chui, B. Kao, and E. Hung.
Mining frequent itemsets from uncertain data.
in Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Nanjing, China, pages 47–58, 2007.

Tuan Nha Hoang, Tien-Tuan Dao, and Marie-Christine Ho Ba Tho.
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, pages 359–370, 2016.

Ahmed Samet, Eric Lefèvre, and Sadok Ben Yahia.
Evidential data mining: precise support and confidence.
Journal of Intelligent Information Systems, pages 1–29, 2016.

Philippe Smets.
The application of the matrix calculus to belief functions.
International Journal of Approximate Reasoning, 31(1–2):1–30, 2002.