How Your Supporters and Opponents Define Your Interestingness

Bruno Crémilleux, Arnaud Giacometti, Arnaud Soulet

QCM-BioChem - Paris - 30 août 2018
Interested in discovering knowledge with pattern mining?  

An example: patterns and their frequency values

\[ \text{freq}(X, \mathcal{D}) = |\{t \in \mathcal{D} : X \subseteq t\}| \]
Interested in discovering knowledge with pattern mining? (2/3)

Another example: patterns and their all-confidence values

\[
\text{all-conf}(X, D) = \frac{\text{freq}(X, D)}{\max_{i \in X} \text{freq}(i, D)}
\]
Interested in discovering knowledge with pattern mining? (3/3)

Our long-term questions:

- How can one determine whether a data mining method extracts interesting patterns?
- How can one know and evaluate if a data mining method is better than another for a given task?

How to identify the advantages and limitations of each approach?
Related work

The quest of a theory of pattern interestingness is an old challenge... (Han et al 07, Fayyad et al 03)

- **Axiomatic approaches**: define criteria to well-behaved measures
  - Association rules / supervised context (Piatetsky-Shapiro 91, Tan et al 04)
  - Unsupervised context (Hämäläinen 10)

- **Statistical approaches**: compare the statistical tests used by correlation measures (Vreeken and Tatti 14)

State of the art is restricted to measures evaluating the interestingness of correlated patterns.

How to consider other/any kind(s) of pattern mining methods including constraint-based pattern mining methods in a general framework?
Outline

- Framework of supporters and opponents
- Typology of mining methods: what do you want to mine?
- Evaluation complexity of mining methods: how to measure the quality of a data mining method?
- Experiments
- Conclusion/perspectives
Scope of our framework

- **Any method to select patterns:**
  
  selector \((s)\): a function such that \(s(X)\) increases when \(X\) is more interesting.

  - **Interestingness measures:** all-confidence and bond (Omiecinski 03), lift (Vreeken and Tatti 14), p-value (Gallo et al 07)
  - **Constraint-based patterns:** productive itemsets (Webb and Vreeken 13)
  - **Condensed representations of patterns:** free itemsets (Boulicaut et al 03), closed itemsets (Pasquier et al 99), NDI (Calders et al 02)
  - **top-\(k\) patterns** (Fu et al 00)
  - ...  

- Context of **unsupervised problems with binary data**
**Key principle**: studying the relationships between a selected pattern \( X \) and the other patterns when selecting \( X \).

**Assumption**: the higher the number of *necessary* comparisons between \( X \) and the other patterns to select \( X \), the higher the quality of the selector.

Example:

- **all-confidence**: involves the individual items of \( X \)
- **productive itemset**: involves all subsets of \( X \)

**Remark**: we will see that the semantics axis of the typology based on the type (subset, superset, incomparable) of a pattern \( Y \) in relationships with \( X \) brings meaningful information on a selector.
How to evaluate the relationships between $X$ and the other patterns?

- compare the interestingness of $X$ with respect to two very similar datasets $D$ and $D'$, where only the frequency of another pattern $Y$ varies.

In this example: $Y = A$

$X$: the assessed pattern

**Property:**
Given $X$, it is always possible to build $D$ and $D'$ such that $D <_Y D'$. 

![Diagram](image-url)
Example: all-confidence

\[ \text{all-conf}(X, D) = \frac{\text{freq}(X, D)}{\max_{i \in X} \text{freq}(i, D)} \]

Here:

- **A**: the single pattern with a different frequency value between \(D\) and \(D'\)
- **ABC**: the assessed pattern
- in this example, **A** is an opponent (the all-confidence decreases in **ABC**
An **opponent** $Y$ of an assessed pattern $X$ is a pattern that may *decrease* the interestingness of $X$ when only the support of $Y$ increases.

\[ \text{freq}(Y, D) < \text{freq}(Y, D') \]

\[ s(X, D) > s(X, D') \]
Supporters

A supporter $Y$ of an assessed pattern $X$ is a pattern that may increase the interestingness of $X$ when only the support of $Y$ increases.
Supporters and opponents
(example of the all-confidence)
Typology: polarity

What do you want to mine?

= over-represented phenomena in the data
  ➫ positive selector iff any pattern is its own supporter.

(support, all-confidence, bond, lift, productive itemsets, closed itemsets, top-k itemsets, ...)

= under-represented phenomena in the data
  ➫ negative selector iff any pattern is its own opponent

(free itemsets, negative border, FP-Outlier Factor, ...)

Typology: polarity

Recommendation (soundness): a well-behaved pattern mining method should not mix interestingness selectors with opposite polarities or make possible the existence of patterns that are supporters and opponents of a same pattern.

Could the violation of this recommendation explain the very few uses of minimal condensed representations?
Typology: semantics

What do you want to mine?

- correlation space: subsets
- condensed representation space: superset
- model-like space: incomparable sets
Typology: semantics

**Recommendation** (completeness): all patterns should be either supporters or opponents in a well-behaved pattern mining method.

Example of **MINI** approach (Gallo et al 07):

- p-value as correlation measure
Typology: semantics

**Recommendation** (completeness): all patterns should be either supporters or opponents in a well-behaved pattern mining method.

Example of **MINI** approach (Gallo et al 07):

- p-value as correlation measure
- closed patterns for removing comparable redundancies
Typology: semantics

**Recommendation (completeness):** all patterns should be either supporters or opponents in a well-behaved pattern mining method.

Example of MINI approach (Gallo et al 07):

- p-value as correlation measure
- closed patterns for removing comparable redundancies
- iterative algorithm for removing incomparable redundancies
Examples (1/2): positive correlations

A selector mines correlated itemsets iff it has subsets as opponents.

- **Support**: \( \emptyset \)
- **Items of \( X \)**: \( X \)
- **Subsets**

\[
supp(X) > supp(Y) \times supp(X \setminus Y)
\]
for all \( Y \subset X \)
Examples (2/2): condensed representations

A selector mines a condensed representation iff it has subsets as supporters (minimal) or supersets as opponents (maximal)

- Minimal itemsets, free itemsets
- Non-derivable itemsets
- Maximal itemsets, closed itemsets
The typology in a nutshell

<table>
<thead>
<tr>
<th>SEMANTICS</th>
<th>POLARITY</th>
<th>POLARITY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td>$X \in s^+(X)$</td>
<td><strong>Negative</strong></td>
</tr>
<tr>
<td>Subsets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \downarrow$</td>
<td>C1: $X \downarrow \cap s^-(X) \neq \emptyset$ (all-confidence, bond, productive itemsets, NDI, FPOF, lift)</td>
<td>C2: $X \downarrow \cap s^+(X) \neq \emptyset$ (free itemset, NDI, negative border)</td>
</tr>
<tr>
<td>Supersets</td>
<td>C3: $X \uparrow \cap s^-(X) \neq \emptyset$ (closed itemsets, maximal itemsets)</td>
<td>$X \uparrow \cap s^+(X) \neq \emptyset$ (FPOF)</td>
</tr>
<tr>
<td>Incomparable sets</td>
<td>$X \leftrightarrow \cap s^-(X) \neq \emptyset$ (top-k frequent itemsets)</td>
<td>$X \leftrightarrow \cap s^+(X) \neq \emptyset$ (FPOF)</td>
</tr>
</tbody>
</table>

QC 1: Soundness

$s^+(X) \cap s^-(X) = \emptyset$

QC 2: Completeness

$s^+(X) \cup s^-(X) = \mathcal{L}$
How to measure the quality of a selector?

**Definition:** the evaluation complexity of an interestingness selector $s$ is the asymptotic behavior of the cardinality of its supporters and opponents.

⇒ the higher the evaluation complexity, the better the selector
# Evaluation complexity of selectors (2/2) examples

<table>
<thead>
<tr>
<th>Selector</th>
<th>supp.</th>
<th>opp.</th>
<th>complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>support</td>
<td>{X}</td>
<td>{∅}</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>all-confidence</td>
<td>{X}</td>
<td>sing.</td>
<td>$O(k)$</td>
</tr>
<tr>
<td>bond</td>
<td>{X}</td>
<td>sing.</td>
<td>$O(k)$</td>
</tr>
<tr>
<td>lift</td>
<td>{X}</td>
<td>X↓</td>
<td>$O(2^k)$</td>
</tr>
<tr>
<td>prod. itemset</td>
<td>{X}</td>
<td>X↑</td>
<td>$O(n-k)$</td>
</tr>
<tr>
<td>max. itemset</td>
<td>{X}</td>
<td>X↑</td>
<td>$O(n-k)$</td>
</tr>
<tr>
<td>free itemset</td>
<td>X↓</td>
<td>{X}</td>
<td>$O(k)$</td>
</tr>
<tr>
<td>closed itemset</td>
<td>{X}</td>
<td>X↑</td>
<td>$O(n-k)$</td>
</tr>
<tr>
<td>NDI</td>
<td>X↓</td>
<td>X↓</td>
<td>$O(2^k)$</td>
</tr>
<tr>
<td>top-k frequent</td>
<td>{X}</td>
<td>X↔</td>
<td>$O(2^n - 2^k)$</td>
</tr>
<tr>
<td>FPOF</td>
<td>$X^\uparrow \cup X^\downarrow$</td>
<td>$X^\downarrow \cup {X}$</td>
<td>$O(2^n)$</td>
</tr>
</tbody>
</table>

- **correlated pattern selectors:** productive itemsets > all-confidence/lift > support
- **condensed representations:** NDI > closed/free itemsets

$k$: number of items in $X$ – $n$: number of items
Experiments (1/2)

**Goal:** verifying whether the quality of the correlated pattern selectors follow the evaluation complexity

- Swap randomization protocol (Gionis et al 14)
  1. Build a randomized dataset $D^*$ by shuffling $D$
  2. Mine interesting patterns according to the selector $s$ from $D$ and $D^*$
  3. Calculate the FP rate (proportion of patterns mined at the same time in $D$ and $D^*$)

- Selectors: frequency, all-confidence, lift, productivity

- Datasets coming from UCI ML repository (Dheeru and Karra Taniskidou 17)
Experiments (2/2)

- exponential complexity (productivity) **better than** linear complexity (lift, all-confidence) **better than** constant complexity (support)
Supporters and opponents materialize the required access to the data to calculate the interestingness of a pattern.

Our typology of interestingness is defined on 2 axis:

- **Semantics** = location of supporters and opponents
- **Quality** = number of supporters and opponents
Perspectives

- **Generalization of the framework:**
  - **Supervised context:** using the frequencies in $D^+$ and $D^-$ instead of the frequency
  - **Other pattern languages:** sequences, graphs,...

- **Studying the overall behavior of a combination of selectors**
  Example: Given two known selectors $s_1$ and $s_2$, what can we say about $s_1 \land s_2$ (w.r.t semantics or complexity)?

- **Considering a set of itemsets** as supporter (or opponent), and not an individual itemset: $D \prec_Y D' \implies D \prec_{Y_1, \ldots, Y_k} D'$
  Example: with our framework, the lift has each item as opponent, but in practice, the lift may vary when the frequency of several items varies at the same time
Special thanks to:

Albrecht Zimmermann (Université de Caen Normandie, France)

This work is partly supported by CNRS (Mastodons QCM-BioChem project)