What MDL can bring to Pattern Mining
Tatiana Makhalova, Sergei O. Kuznetsov, Amedeo Napoli

To cite this version:
Tatiana Makhalova, Sergei O. Kuznetsov, Amedeo Napoli. What MDL can bring to Pattern Mining. ISWS 2018 - International Semantic Web Research Summer School, Jul 2018, Bertinoro, Italy. hal-01889792

HAL Id: hal-01889792
https://hal.science/hal-01889792
Submitted on 8 Oct 2018

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.
**What MDL can bring to Pattern Mining**

Tatiana Makhalova, Sergei O. Kuznetsov, Amedeo Napoli

National Research University Higher School of Economics,
3 Kochevinsky Prospekt, Moscow, Russia
LORIA, CNRS – Inria – U. of Lorraine
BP 239, Vandœuvre-lès-Nancy, France

---

**Introduction**

Pattern Mining, Objective: find a small set of patterns that are well interpretable by experts.

Input data: binary table $G \times M$, where $G$ is a set of objects, $M$ is a set of attributes, and $I$ is a relation between them.

Interpretation of glm: object $g \in G$ has attribute $m \in M$.

---

**Pattern Mining. What kind of patterns we should compute?**

Total number of patterns is $2^{|G| \times |M|}$.

Types of patterns in terms of Formal Concept Analysis

**FCA: Basic Notions**

A formal context $\triangleright \triangleright$ is a triple $(G, M, I)$, where $G$ is a set of objects, $M$ is a set of attributes, $I \subseteq G \times M$ is a relation called incidence relation.

The derivation operator $\triangleright \triangleright$ is defined for $Y \subseteq G$ and $Z \subseteq M$ as follows:

$I \rightarrow Y = \{ m \in M \mid \forall g \in Y \cdot (g, m) \in I \}, \forall g \in Y \in \triangleright \triangleright$.

$(\triangleright \triangleright)^{-1}$ is a partial order on objects and attributes.

**Key notions:**

- **Encoding length:** new length that "compresses", i.e. the most frequently used ones have the shortest encoding length.
- **Code table:** a set of selected patterns with their encoding length.
- **Disjoint covering:** principle of compression by patterns.

**Total length:**

$L(D, CT) = L(D \mid CT) + L(CT \mid D) = |X \setminus Y| + \sum_{C \in CT} \sum_{Y \subseteq C} |Y| \cdot \log \frac{|Y|}{|\triangleright \triangleright^{-1}(Y)|}$.

**Examples:**

- **MDL in practice:**
  - **greedy algorithm (Krimp):**
  - Minimal generators $m \in M, m \in M, m \in M$.
  - Generators $m \in M, m \in M, m \in M$.

**MDL: is there a place for background knowledge?**

**Ideal:** MDL as an additional filtering stage in pattern selection.

MDL-optimal (blue) vs top-n (green) closed items

**Non-redundancy**

- Distance to the 1st FN
- Top-n concepts have a lot of "twins", while MDL-optimal ones are patronize distinctive (w.r.t. Euclidean distance).

**Non-redundancy**

- Average length of the longest paths built from possets (lattices)
- A long path is an indicator of redundancy, since in that case patterns characterize the same objects at different levels of abstraction. Short paths correspond to "flat" structures with more various patterns.

**Pattern mining with area_len, sep and area_sep**

- Lift, len, fr can be significantly improved by the application of MDL.

**Non-redundancy**

- Average number of items contained with children
- Characterizes the uniqueness of patterns in a set. It indicates just an amount of itemssets having at least one more general itemset.

**Typicality (representativeness)**

- It is measured by the usage of patterns, i.e. the frequency of the occurrence of patterns in the greedy covering, so the usage does not exceed the frequency.
- It is not obvious which values correspond to a subset of common patterns, while low values indicates that a subset contains less typical, but still interesting (w.r.t. interestingness measures) patterns.

**Data coverage**

- The rate of covered "crosses" in object-attribute relation
- A subset of selected patterns can be considered as a concise representation of a dataset. Thus, it is important to know how much information is lost by compression. It can be measured by the rate of covered attributes. Values close to 1 correspond to the lossless compression; MDL ensures better covering and allows for the biggest gain for area-based orderings.

**References**