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

To cite this version:
Tatiana Makhalova, Sergei 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.archives-ouvertes.fr/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.
Pattern Mining. What kind of patterns we should compute?

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 \( g \): object \( g \in G \) has attribute \( m \in M \).

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

Types of patterns in terms of Formal Concept Analysis

**FCA. Basic Notions**

A formal context \((G, M, i)\) is a triple \((G, M, i)\), where \( G \) is a set of objects, \( M \) is a set of attributes, and \( i : G \times M \) is a relation called incidence relation.

The derivation operator \( y' = \text{deriv}(y) \) is defined for \( y = (x \in G \mid \forall g : g \in G \Rightarrow x \in i(g) \) as follows:

\[ y' = (m \in M \mid \forall g : g \in G \Rightarrow m \in i(g)) \]

A formal concept \((Y, Z)\) is a pair \((Y, Z)\), where \( Y \subseteq G \) is a set of objects and \( Z \subseteq M \) is a set of attributes. \( Y \) is called the \((formal)\) extent and \( Z \) is called the \((formal)\) intent of the concept \((Y, Z)\).

A concept lattice \((Galois lattice)\) is a partially ordered set of concepts, the order \( \preceq \) is defined as follows: \( (Y, Z) \preceq (Y', Z') \) if \((Y' \supseteq Y \cap Z', Z' \supseteq Z \cap Y')\).

**Key notions:**
- Encoding length: new length that "compresses", i.e., the most frequently used ones have the shortest (encoded) length.
- Closed concepts: minimal concepts having non-empty intents.
- Disjoint covering: principle of compression by concepts.
- Total length: length of the derivation operator.
- Code table: length of closed concepts.

**MDL**

**Minimal Description Length (MDL) Principle. Basic Definitions**

The main principle: the best set of patterns is the set that best compresses the database [Vlecken et al., 2011].

**Objective:** \( \text{Minimize} \sum_{m=1}^{M} \sum_{g=1}^{G} - \log_{2} P(g|m) \)

**Key notions:**
- Encoding length: new length that "compresses", i.e., the most frequently used ones have the shortest (encoded) length.
- Code table: length of closed concepts.

**MDL in practice:** greedy algorithm (Krimp)

**Initial state**

Final state

**Reduction in the number of patterns**

Significant reduction in the number of patterns (up to 5% of the final concept).

**Examples**

**Formal context**

**Concept lattice** (partially ordered full set of formal concepts)

For a formal concept \((g, g_i, (m_m, m_m, m_m))\)
- closed patterns \((m_m, m_m)\);
- minimal generators \((m_m, m_m, m_m)\);
- generators \((m_m, m_m, m_m, m_m, m_m, m_m, m_m)\).

**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 are better. The high values of usage 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.

**References**