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?**

Total number of patterns is 2^D.

**Types of patterns in terms of Formal Concept Analysis**

**Pattern Mining. Objective:**

- **FCM:** Basic Notions
  - A formal context \((G,M, I)\) is a triple of \((G,M, I)\), where \(G\) is a set of objects, \(M\) is a set of attributes, and \(I\) is a relation between them.
- **Disjoint covering:** Code table: \(\{\text{g}, \text{g}\}\) is a subconcept of \(\{\text{g}, \text{g}\}\) is a superconcept of \(\{\text{g}, \text{g}\}\).
- **Height** of \(\{\text{g}, \text{g}\}\) is a concept lattice \(M\) is a set of attributes, \(T\) is a set of conditions.
- **Name** of the concept \(\{\text{g}, \text{g}\}\) is a subconcept of \(\{\text{g}, \text{g}\}\)
- **Width** of \(\{\text{g}, \text{g}\}\) is the dataset encoded with the code table
- **Principle of compression by patterns.**

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

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

**Objectives:**
- **Disjoint covering:** principle of compression by patterns.
- **Relating formal concepts:** Code table.lattice \(L(D) = \sum_{(C,D)} \frac{1}{D(C)} + \sum_{(C,D)} \frac{1}{D(D)} \cdot \frac{1}{D(C)} \cdot \frac{1}{D(T)} \cdot \frac{1}{D(T)}

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

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

**Non-redundancy**
- **Distance to the 1st NN**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Non-redundancy**
- **Average length of the longest paths built from posets (lattices)**

**Data coverage**
- **The rate of covered “crosses” in object-attribute relation**

**Non-redundancy**
- **Average number of items having children**

**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.

**References**

1. Aggarwal, C.C., Han, J.: Frequent pattern mining. (2014)