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

Total number of patterns is 2^|M|.

Types of patterns in terms of Formal Concept Analysis

FCA: Basic Notions

A formal context is a triple (G, M, I), where G is a set of objects, M is a set of attributes, I ⊆ G × M is a relation called incidence relation. The derivation operator y is defined for y ⊆ G and z ⊆ M as follows:

\[ y = \{ m \in M \mid \forall g \in y : (g, m) \in I \} = \{ g \in G \mid \forall m \in z : (g, m) \in I \} \]

A (formal) concept is a pair (Y, Z), where Y ⊆ G and Z ⊆ M and Y \[ Y \] and Z \[ Z \] is called the (formal) extent and Z is called the (formal) intent of the concept (Y, Z).

A concept lattice (or Galois lattice) is a partially ordered set of concepts, the order ≼ is defined as follows: (Y, Z) ≼ (C, D) if Y ⊆ C and Z \[ Z \] D. A pair (Y, Z) is a subconcept of (C, D) if and only if C \[ C \] D, and (C, D) is a superconcept of (Y, Z). Formal concepts ordered by generality relation (A, B) ≼ (A_1, B_1) if A_1 \[ A_1 \] B_1 make a lattice, called concept lattice.

Types of patterns (defined for concept (A, B));

- Closed itemsets (intents): B.
- Minimal generators are minimal subsets \[ B \setminus B' = A \].

Generators are any patterns between minimal generators and closed itemsets.

MDL. is there a place for background knowledge?

MIDL: MDL as an additional filtering stage in pattern selection.

MDL-optimal (blue) vs top-\( n \) (green) closed itemsets

Non-redundancy

Distance to the 1st FN

Top-n concepts have a lot of “twins,” while MDL-optimal ones are pairwise 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 correlate to “flat” structures with more varied patterns.

Pattern mining with area \( \text{len}_{\text{sep}} \) and area\(_{\text{sep}}\) lift. \( \text{len}_{\text{3}} \) can be significantly improved by the application of MDL.

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