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.
Formal concepts ordered by Total number of patterns is 2
Pattern Mining: What kind of patterns we should compute?

well interpretable by experts.

Key notions: the dataset encoded with the code table compresses the database [Vreeken et al., 2011].

Idea:

formal context \( \times \subseteq \) are subsets of attributes that describe an object.

\( CT \) is a pair (Y, Z) = (A, B) \# \subseteq \), which indicates just an uniqueness of patterns in a Characterizes the

Top-n concepts have a lot of “twins,” while MDL-optimal ones are partitioned distinctive (w.r.t. Euclidean distance).

Top-n concepts have a lot of “twins,” while MDL-optimal ones are partitioned distinctive (w.r.t. Euclidean distance).

Reduced by a factor of 1/6

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 varied patterns.

Pattern mining with area_len_sep and area_sep varied patterns.

Average length of the longest paths built from posets (lattices)

A partial order (Y, Z) is a, partially ordered set of concepts, the order \( \preceq \) is defined as follows: (Y, Z) \( \preceq \) (C, D) if Y \( \subseteq \) C and Z \( \subseteq \) D, a pair (Y, Z) is a subconcept of (C, D) and (D, C) is a superconcept of (Y, Z).

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.

The usage of MDL-optimal patterns is almost the same for different orders while the usage of order is depending on

The most interesting concepts w.r.t. given assumptions:

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

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

MDL: is there a place for background knowledge?

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

MDL in practice: greedy algorithm (Krimp)

Initial state

An intermediate state

Add ordered candidates one by one if they allow for reducing the total length

Reduction in the number of patterns*

Significant reduction in the number of patterns (up to 5% of the formal concepts).

* datasets from LUCS-KDD repository [4]

References