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.
What MDL can bring to Pattern Mining

Tatiana Makhalova, Sergei O. Kuznetsov, Amedeo Napoli

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

Introduction
Patterns are subsets of attributes that describe an object.

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 \( g \): object \( g \) has attribute \( m \).

Types of patterns in terms of Formal Concept Analysis

Pattern Mining. What kind of patterns we should compute?

Example

Discussion of interessingness: some proofs have been published recently.

Background Knowledge: Assumptions on Interestingness

MDL ensures better covering and allows for the biggest gain for area-based orderings.

MDL in practice: greedy algorithm (Krimp)

Minimal Description Length (MDL) Principle.

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 itemsets

Non-redundancy

Distance to the 1st NN

Non-redundancy

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

Non-redundancy

Average number of itemssets with children

Non-redundancy

Average number 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 are better. The high values of usage correspond to a subsets 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 top-\( n \) is dependent on ordering.

References