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^n.

Types of patterns in terms of Formal Concept Analysis

FCA: Basic Notions

A formal context [Ganter and Wille, 1999; Wille, 1982] is a triple (G,M,I), where G is a set objects, M is a set attributes, and I is a relation between them.

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

Objective: LD, CT = LD ⊕ CT = LCT (D), where LD (CT) is the length of the database encoded with the code table CT and LCT (D) is the length of the code table CT computed w.r.t. D.

Key notions:
- Encoding length: new length that “compresses”, i.e. the most frequently used ones have the shortest encoding length.
- Code table: a set of selected patterns with their encoding lengths.
- Disjoint covering: principle of compression by patterns.

Total length:

Total length = total length of the code table + total length of all elements of the code table.

The Minimal Description Length (MDL) Principle.

Basic Definitions

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

Objective: LD, CT = LD ⊕ CT = LCT (D), where LD (CT) is the length of the database encoded with the code table CT and LCT (D) is the length of the code table CT computed w.r.t. D.

Key notions:
- Encoding length: new length that “compresses”, i.e. the most frequently used ones have the shortest encoding length.
- Code table: a set of selected patterns with their encoding lengths.
- Disjoint covering: principle of compression by patterns.

Total length:

Total length = total length of the code table + total length of all elements of the code table.

Non-redundancy

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

A long path is an indicator of redundancy, since in that case patterns capture 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 lift, lift len, fr can be significantly improved by the application of MDL.

MDL: is there a place for background knowledge?

MDSL: as an additional filtering stage in pattern selection.

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

MDSL: is there a place for background knowledge?

MDL in practice: greedy algorithm (Krimp)

Initial state

Candidate set, area

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

used measures for ordering candidate sets

The ordered list of candidates is used for greedy covering of data in Krimp

 References