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
Types of patterns in terms of Formal Concept Analysis

Pattern Mining. What kind of patterns we should compute?

Interpretation of gIm: object well interpretable by experts.

Pattern Mining. Objective:

Total length:

Key notions:

MDL Principle.

Examples of interestingness measures for concept (A,B)

[area] "interesting patterns are those that take the biggest area in dataset"

[length] "interesting patterns are the most detailed ones that are quite frequent in dataset"

[separation] "interesting patterns are separated the best from the context"

...combined measures, etc.

Non-redundancy

Distance to the 1st NN

Non-redundancy

Average length of the longest paths built from posets (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 correspond to "flat" structures with more varied patterns.

Pattern mining with area len_sep and area_sep lift. lift_len, 3 can be significantly improved by the application of MDL.

MDL: is there a place for background knowledge?

Idea: 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 posets (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 correspond to "flat" structures with more varied patterns.

Pattern mining with area len_sep and area_sep lift, lift_len, 3 can be significantly improved by the application of MDL.

MDL in practice: greedy algorithm (Krimp)

Initial state

Candidate set, area

Final state

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

Reduction in the number of patterns*

* datasets from LUCS-KDD repository [4]

Used measures for ordering candidate sets.

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

area_fn lift = frequency(B)↑lift(fn)

area_fn lift = frequency(B)↑lift(fn) sort by lift(fn)↑then by frequency(B)

References


8. Krummen, M., O'Malley, A., School of Computing and Information Systems, University of Melbourne, Australia.