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

I

FCA. Basic Notions

Pattern Mining. What kind of patterns we should compute?

⊆

Total length: compresses the database [Vreeken et al., 2011].

Minimal Description Length (MDL) Principle.

Idea: use measures that reflect knowledge of experts about “interestingness” of patterns

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.

Example

Formal context

Concept lattice (partially ordered full set of formal concepts)

For a formal concept (g, g, (m, m, m, m, g), - closed patterns (m, m, m, m, g);

- minimal generators (m, m, m, m, g, m, m, m); (m, m, m, g, m, m, m, m);

- area = 6

frequency ≥ 2 (g, g, m, m, m, m, m);

= 3 (separation): (g, g, (m, m, m, m, g, m, g, g, g, m, m, m)

area = 6, 13.

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

Background Knowledge: Assumptions on Interestingness

Input data

Compute patterns

Reorder patterns

Filter patterns

MDL in practice: greedy algorithm (Krimp)

An intermediate state

Final state

MDL: is there a place for background knowledge?

Minimal Description Length (MDL) Principle.

Basic Definitions

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 | is the length of the dataset 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 encoded length.

- Code table: a set of selected patterns with their encoding length.

- Disjoint covering: principle of compression by patterns.

Total length:

L(D) = \sum_{i=1}^{n} |\text{itemset}_i| + |\text{itemset}|

Code table length w.r.t. data:

Data length w.r.t. code table:

MDL: is there a place for background knowledge?

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

Non-redundancy

Distance to the 1st NN

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

Non-redundancy

Average length of the longest paths built from posssets (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 len. lift can be significantly improved by the application of MDL.

Non-redundancy

Average number of itemsets having with children

Characterizes the uniqueness of patterns in a set. It indicates just an amount of itemsets having at least one more general itemset.

Non-redundancy

Number of itemsets with children

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 top-n is dependent on ordering.

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 “crosses” in object-attribute relation. Values close to 1 correspond to the lossless compression.

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

References


