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 of objects, $M$ is a set of attributes, and $I$ is a relation between them.

The derivation operator ($\gamma$) is defined for $Y \subseteq G$ and $Z \subseteq M$ as follows:

$$\gamma(Y) = \{ g \in G \mid \forall y \in Y. \exists g \in G. \forall y \in Y. \exists g \in G \}$$

The domain of objects and the range of attributes are called the lower and upper bounds, respectively.

A (minimal) concept is a pair $(Y, Z)$, where $Y \subseteq G$, $Z \subseteq M$, and $Y \subseteq Z$. $Z'$ is the set of attributes in $Z$.

A concept lattice (Galos lattice) is a partially ordered set of concepts, the order is defined as follows: $(Y, Z) \leq (C, D)$ if $Y \subseteq C$ and $Z \subseteq D$.

Formal concepts are ordered by generality relation $(A, B) = (A, B) \in \mathcal{A} \subseteq \mathcal{B}$, a lattice called concept lattice.

Types of patterns (defined for concept $(A, B)$):

- Closed itemsets (intents): $B$.
- Minimal generators: minimal subsets $B \subseteq B'$.
- Generators: any patterns between minimal generators and closed itemsets.

Pattern Mining.

What kind of patterns we should compute?

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 $I$: object $g \in G$ has attribute $m \in M$.

Minimal Description Length (MDL) Principle.

Basic Definitions

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

Objective: $LD(C) = LD(C) + \lambda \cdot D(LD(C))$, where $LD(C)$ is the length of the pattern set $C$.

Key notions:

- Encoding length: the shortest encoding length.
- Disjoint covering: $D(C)$ is the minimal number of patterns required to cover $G$.

Total length:

$$LD(C) = \sum_{A \subseteq C} (m_a \cdot \log |A|)$$

MDL: is there a place for background knowledge?

New redundancy:

Distance to the 1st NN.

Non-redundancy:

Average path length from different levels.

Attributes relation.

Pattern mining with area $\lambda$.

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 attributes. Values close to 1 correspond to a lossless compression.

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

MDL in practice: greedy algorithm (Krimp)

Initial state

Final 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 final concept).

* datasets from LUCS-KDD repository [4]

References

8. Witten, I.H.: Mining Frequent Patterns. Springer (2011)