



HAL
open science

What MDL can bring to Pattern Mining

Tatiana Makhalova, Sergei O. Kuznetsov, Amedeo Napoli

► **To cite this version:**

Tatiana Makhalova, Sergei O. 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.science/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 Proezd, 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 glm: object $g \in G$ has attribute $m \in M$.

Background Knowledge: Assumptions on Interestingness

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 where $I(\cdot)$ is the indicator*, q is a threshold.

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

combined measures, etc.

$$area(A, B) = |A| \cdot |B|$$

$$length(A, B) = |B| I(|A| \geq q)$$

$$sep(A, B) = \frac{|A| |B|}{\sum_{g \in A} |g'| + \sum_{m \in B} |m'| - |A| \cdot |B|}$$

$$I(cond) = \begin{cases} 1 & \text{if cond is True} \\ 0 & \text{otherwise} \end{cases}$$

Pattern Mining. What kind of patterns we should compute?

Total number of patterns is 2^M

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, $I \subseteq G \times M$ is a relation called **incidence relation**.

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

$$Y' = \{m \in M \mid glm \text{ for all } g \in Y\}; \quad Z' = \{g \in G \mid glm \text{ for all } m \in Z\}$$

A **(formal) concept** is a pair (Y, Z) , where $Y \subseteq G, Z \subseteq M$ and $Y' = Z, Z' = Y$. Y is called the **(formal) extent** and Z is called the **(formal) intent** of the concept (Y, Z) .

A **concept lattice** (or Galois lattice) is a partially ordered set of concepts, the order \ll is defined as follows: $(Y, Z) \ll (C, D)$ iff $Y \subseteq C$ ($D \subseteq Z$), a pair (Y, Z) is a **subconcept** of (C, D) and (C, D) is a **superconcept** of (Y, Z) .

Formal concepts ordered by **generality relation** $(A_1, B_1) \ll (A_2, B_2)$ iff $A_1 \subseteq A_2$ make a lattice, called **concept lattice**.

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

Closed itemsets (intents): B .

Minimal generators are minimal subsets $B_i \subseteq B : B_i' = A$.

Generators are any patterns between minimal generators and closed itemsets

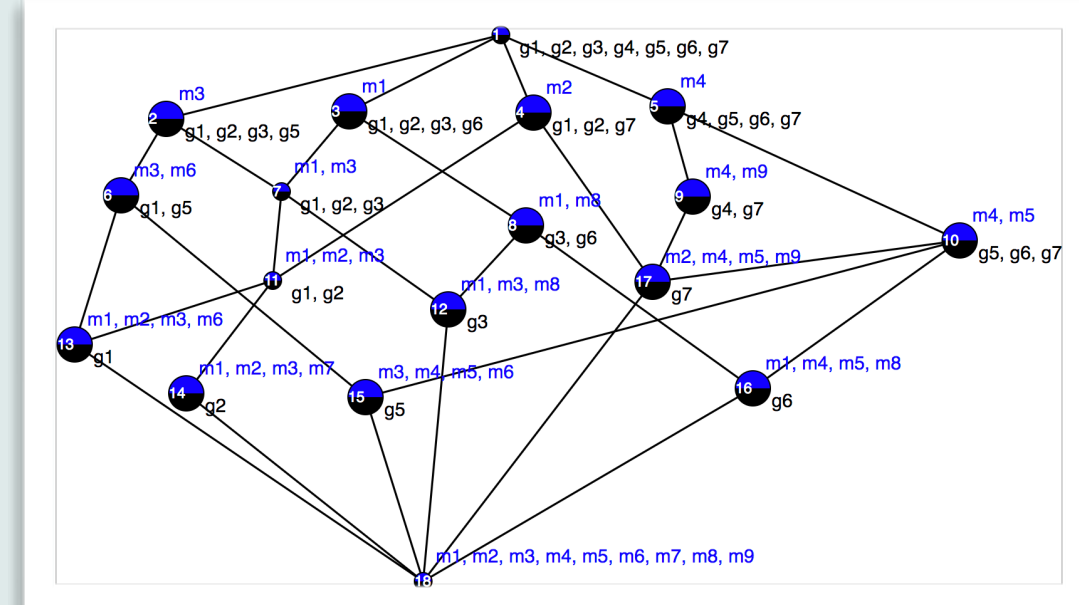
Example

Formal context

Objects	m1	m2	m3	m4	m5	m6	m7	m8	m9
g1: dog	X	X	X	X					
g2: cat	X	X	X		X				
g3: frog	X	X			X				
g4: car		X				X			
g5: ball		X	X	X					
g6: chair	X	X	X		X				
g7: fur coat	X	X	X		X				

m1: 4 legs
m2: wool
m3: change size
m4: cold-resistant
m5: do release CO2
m6: black-white
m7: yellow-brown
m8: green
m9: gray

Concept lattice (partially ordered full set of formal concepts)



For a formal concept $(\{g_1, g_2\}, \{m_1, m_2, m_3\})$

- closed patterns $\{m_1, m_2, m_3\}$;
- minimal generators $\{m_1, m_2\}, \{m_2, m_3\}$;
- generators $\{m_1, m_2\}, \{m_2, m_3\}, \{m_1, m_2, m_3\}$.

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

(area) $(\{g_1, g_2\}, \{m_1, m_2, m_3\}), (\{g_1, g_2, g_3\}, \{m_1, m_3\}), (\{g_5, g_6, g_7\}, \{m_4, m_5\})$; area = 6
(length, frequency ≥ 2): $(\{g_1, g_2\}, \{m_1, m_2, m_3\})$; length = 3
(separation): $(\{g_1, g_2\}, \{m_1, m_2, m_3\}), (\{g_1, g_2, g_3\}, \{m_1, m_3\})$; separation = 6/13.



Background Knowledge

Input data

Compute patterns

Reorder patterns

Filter patterns

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: $L(D, CT) = L(D | CT) + L(CT | D)$, where $L(D | CT)$ is the length of the dataset encoded with the code table CT and $L(CT | D)$ 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: $L(D, CT) = L(D | CT) + L(CT | D)$
Code table length w.r.t. data: $L(CT | D) = \sum_{i \in CT} code(i) + len(i)$
Data length w.r.t. code table: $L(D | CT) = \sum_{d \in D} \sum_{i \in cover(d)} len(i)$

CT computed w.r.t. D

D computed w.r.t. CT

Itemsets	Encoding length	Data with covering	Encoding length
m_3	1	$(m_1)(m_2)(m_3)(m_6)$	4
m_1	1	$(m_1)(m_2)(m_3)(m_7)$	4
m_2	1	$(m_1)(m_2)(m_3)(m_8)$	4
m_4	1	$(m_1)(m_2)(m_3)(m_9)$	4
m_6	1	$(m_1)(m_3)(m_6)$	3
m_7	1	$(m_1)(m_3)(m_7)$	3
m_8	1	$(m_1)(m_3)(m_8)$	3
m_9	1	$(m_1)(m_3)(m_9)$	3
m_5	1	$(m_3)(m_4)(m_9)$	3

MDL in practice: greedy algorithm (Krimp)

Initial state

CT		Data with covering	Candidate set, area
Itemsets	Usage		
m_3	4	$(m_1)(m_2)(m_3)(m_6)$	$m_1 m_2 m_3, 6$
m_1	3	$(m_1)(m_2)(m_3)(m_7)$	$m_1 m_3, 6$
m_2	2	$(m_1)(m_3)(m_8)$	$m_1 m_2 m_3 m_6, 4$
m_4	1	$(m_3)(m_4)(m_9)$	$m_1 m_2 m_3 m_7, 4$
m_6-m_7	1		$m_1 m_3 m_8, 3$
m_5	0		$m_3 m_4 m_9, 3$

An intermediate state

CT		Data with covering	Candidate set, area
Itemsets	Usage		
$m_1 m_2 m_3$	2	$(m_1 m_2 m_3)(m_6)$	$m_1 m_3 m_8, 3$
m_3	2	$(m_1 m_2 m_3)(m_7)$	$m_3 m_4 m_9, 3$
m_1, m_4	1	$(m_1)(m_3)(m_8)$	$m_1 m_3, 2$
m_6-m_9	1	$(m_3)(m_4)(m_9)$	
m_2, m_5	0		

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

Final state

CT		Data with covering
Itemsets	Usage	
$m_1 m_2 m_3$	2	$(m_1 m_2 m_3)(m_6)$
$m_1 m_3 m_8$	2	$(m_1 m_2 m_3)(m_7)$
$m_3 m_4 m_9$	1	$(m_1 m_3 m_8)$
m_6, m_7	1	$(m_3 m_4 m_9)$
m_1-m_5	0	
m_8-m_9	0	

Reduction in the number of patterns*

dataset	nmb. of obj.	nmb. of attr.	nmb. of concepts total	MDL	dataset	nmb. of obj.	nmb. of attr.	nmb. of concepts total	MDL
auto	205	135	67 557	19.26	horse colic	368	83	173 808	101
breast	699	16	642	9.04	iris	150	19	107	13
car	1 728	25	12 617	94	led7	3 200	24	1 937	152
chess	28 056	58	152 753	1 675	mushroom	8 124	90	181 945	211
dermatology	366	49	16 324	47	nursery	12 960	30	176 536	392
ecoli	336	29	694	25	page blocks	5 473	44	715	45
flare	1 389	38	16 303	106	pima indians	768	38	1 609	50
glass	214	46	4 704	50	ticTacToe	958	29	42 685	160
heart	303	50	36 708	54	wine	178	68	13 170	52
hepatitis	155	52	199 954	44	zoo	101	42	4 563	17

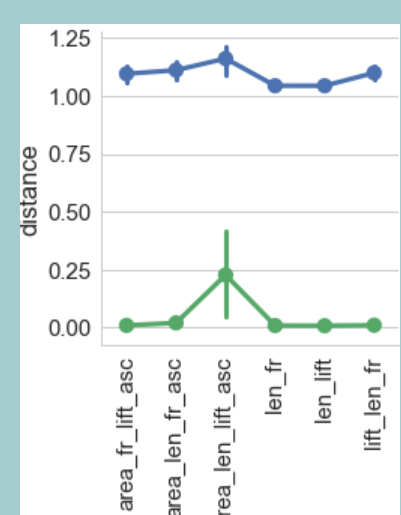
Significant reduction in the number of patterns (up to 5% of the formal concepts).

* datasets from LUCS-KDD repository [4]

MDL: is there a place for background knowledge?

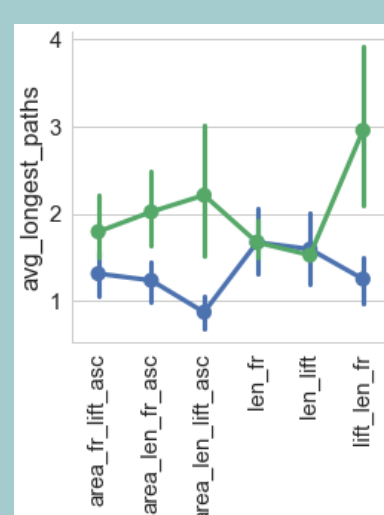
Idea: MDL as an additional filtering stage in pattern selection.

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



Non-redundancy
Distance to the 1st NN

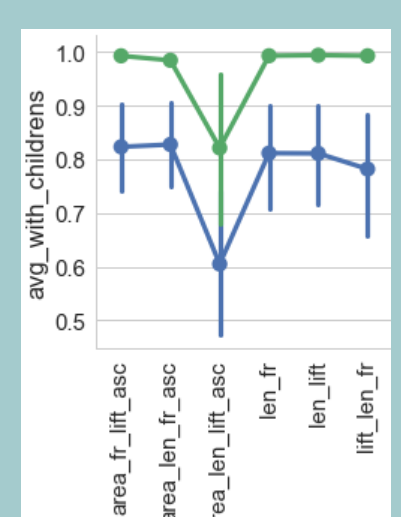
Top-n concepts have a lot of "twins", while MDL-optimal ones are pairwise distinctive (w.r.t. Euclidean distance).



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

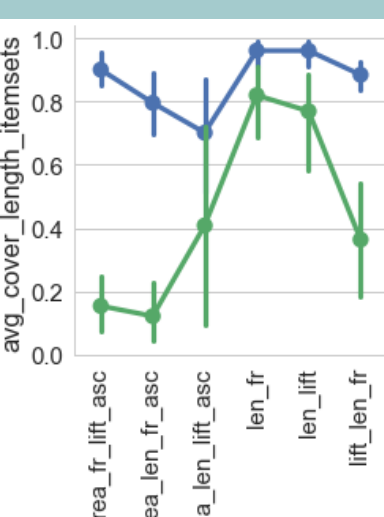
A long path is an indicator of redundancy, since in that case patterns characterize the same objects at different levels of abstraction. Short paths correspond to "flat" structures with more varied patterns.

Pattern mining with $area_len_sep$ and $area_sep$ lift, $lift_len_fr$ can be significantly improved by the application of MDL.



Non-redundancy
Average number of itemsets with children

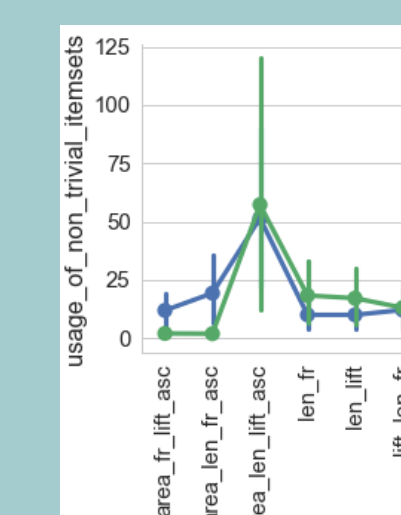
Characterizes the uniqueness of patterns in a set. It indicates just an amount of itemsets having at least one more general itemset.



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 the lossless compression

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



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.

References

- Aggarwal, C.C., Han, J.: Frequent pattern mining. Springer (2014)
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16(5), 412-424 (2000)
- Buzmakov, A., Kuznetsov, S.O., Napoli, A.: Fast generation of best interval patterns for nonmonotonic constraints. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 157-172. Springer (2015)
- Coenen, F.: TheLucs-kddiscretised/normaliseddammandcarndatallibrary(2003). [http://www.csc.liv.ac.uk/frans/KDD/Software/LUCS KDD DN](http://www.csc.liv.ac.uk/frans/KDD/Software/LUCS%20KDD%20DN)
- Ganter, B., Wille, R.: Formal concept analysis: Logical foundations (1999)
- Ganter, B., Kuznetsov, S.O.: Formalizing hypotheses with concepts. In: International Conference on Conceptual Structures. pp. 342-356. Springer (2000)
- Grünwald, P.D.: The minimum description length principle. MIT press (2007)
- Kuznetsov, S.O.: Machine learning and formal concept analysis. In: Klund, P. (ed.) Concept Lattices. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
- Kuznetsov, S.O., Makhalova, T.: On interestingness measures of formal concepts. *Information Sciences* 442-443, 202-219 (2018)
- Vreeken, J., Van Leeuwen, M., Siebes, A.: Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery* 23(1), 169-214 (2011)