Contributions to Pattern Mining and Formal Concept Analysis

Habilitation de l'INSA Lyon et de l'Université Claude Bernard LYON I, Villeurbanne, 12 Feb. 2020

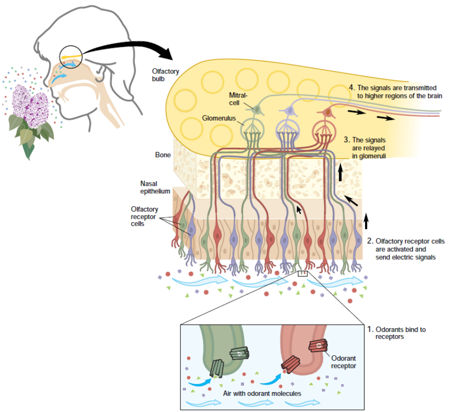
Mehdi Kaytoue

Dr. Karell Bertet	Maître de conférences (HDR), Université de la Rochelle
Dr. Florent Masseglia	Directeur de recherche, INRIA
Pr. Christel Vrain	Professeure, Université d'Orléans
Pr. Michael Berthold	Professeur, Konstanz Universität - CEO Knime AG
Pr. Angela Bonifati	Professeure, Université Claude Bernard Lyon 1
Pr. Jean-François Boulicaut	Professeur, INSA Lyon
Pr. Johannes Fürnkranz	Professeur, University of Linz
Dr. Amedeo Napoli	Directeur de recherche, CNRS

A scientific question

Understanding the olfactory system

- Olfaction is the ability to perceive odors
- Complex phenomenon from molecule to perception[1]
- Challenges
  - Established links between physico-chemical properties and olfactory qualities[2, 3]
  - Difficulties to formulate/propose rules
- Impact
  - Fundamental neuroscience research
  - Industry (food, perfume)
  - Health (anosmia, ...)



[1] “A novel multigene family may encode odorant receptors: A molecular basis for odor recognition”. *Cell (Nobel Prize in Medicine 2004)* (1991).

[2] U. J. Meierhenrich et al. “The Molecular Basis of Olfactory Chemoreception”. *Angewandte Chemie International Edition* 43.47 (2004).

[3] A. Keller et al. “Predicting human olfactory perception from chemical features of molecules”. *Science* (2017).

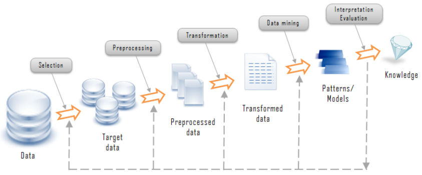
Eliciting hypotheses from data: A KDD task

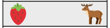
Data collection

- Descriptions: Physico-chemical properties[4], molecular structure (1D, 2D, 3D, smile), etc.
- Targets: odor(s), valence...[5]

Mining (a few and good) hypotheses (in a large search space)

- Clustering, biclusters, association rules, redescriptions, ...
- Mining subgroups discriminating a target attribute[6]
  Data query: $s = nAT \geq 24 \wedge nC \leq 11$
  Query result: $support(s) = \{24, 48, 1633\}$
  $Quality(s \rightarrow pear)$ is high: all the pears, only the pears



ID	MW	nAT	nC	Odor
1	150.19	21	11	
24	128.24	29	11	
48	136.16	24	9	
60	152.16	23	11	
82	151.28	27	12	
1633	142.22	27	10	

4 I. V. Tetko et al. "Virtual Computational Chemistry Laboratory - Design and Description". *Journal of Computer-Aided Molecular Design* 19.6 (2005).

5 S. Arctander. *Perfume and flavor materials of natural origin*. Vol. 2. 1994.

6 S. Wrobel. "An Algorithm for Multi-relational Discovery of Subgroups". *PKDD*. 1997.

Outline

- Data & Pattern Formalization
  - Numerical Pattern Mining
  - Biclustering
  - Data Dependencies
- Pattern Mining and Subgroup Discovery
  - Mining a small set of diverse patterns
  - Iteratively mine finer data representations
- Knowledge Discovery in Practice
  - Neuroscience & Olfaction
  - Social Network Analysis
  - Video Game Analytics
- Perspectives

Our investigations

- **What do we mine?**
- **How do we mine the best patterns?**
- **For what purpose?**

Outline

- Data & Pattern Formalization
  - Numerical Pattern Mining
  - Biclustering
  - Data Dependencies
- Pattern Mining and Subgroup Discovery
  - Mining a small set of diverse patterns
  - Iteratively mine finer data representations
- Knowledge Discovery in Practice
  - Neuroscience & Olfaction
  - Social Network Analysis
  - Video Game Analytics
- Perspectives

A short introduction to Formal Concept Analysis[7]

From a binary table to a concept lattice

- **Formal context** (G, M, I): A binary relation I between *objects* G and attributes M
- **Galois connection**: a pair of closure operators $(.)''$

$$A' = \{m \in M \mid \forall g \in A \subseteq G : (g, m) \in I\}$$

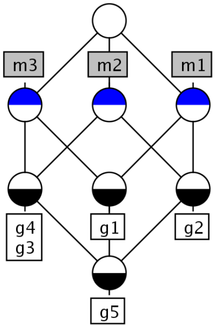
$$B' = \{g \in G \mid \forall m \in B \subseteq M : (g, m) \in I\}$$

- **Concepts** (A, B): Fixpoints, extent $A = B'$ and intent $B = A'$
- **Concept lattice**: a poset,

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \; (\Leftrightarrow B_2 \subseteq B_1)$$

- (Partial) **implications bases**

	m_1	m_2	m_3
g_1	×		×
g_2	×	×	
g_3		×	×
g_4		×	×
g_5	×	×	×



$$\{g_3\}' = \{m_2, m_3\}$$

$$\{m_2, m_3\}' = \{g_3, g_4, g_5\}$$

$$(\{g_3, g_4, g_5\}, \{m_2, m_3\})$$

$$(\{g_1, g_5\}, \{m_1, m_3\}) \leq (\{g_1, g_2, g_5\}, \{m_1\})$$

[7] B. Ganter et al. *Formal Concept Analysis*. 1999.

A short introduction to Formal Concept Analysis

UNIVERSITÉ DE LYON

Some key properties for data analysis

- A natural structure of the data
- Maximality of concepts as rectangles
- Overlapping of concepts
- Specialization/generalization hierarchy
- Synthetic representation of the data without loss of information
- Data implications
- Data navigation
- Knowledge base representation

but...

- FCA hardly deals with (large) numerical data
- FCA advances unknown in many fields where these properties are key indeed
  - The community of pattern mining rediscovered several notions from FCA and then got strongly dedicated into algorithms, but interestingly, not much interest in "pure" numerical patterns
  - Concepts are very similar to biclusters, yet new algorithms
  - Implications can be mapped to functional dependencies in the database field

First axis of research: formalize problems with FCA

Interval Patterns

A little interest in "pure" Numerical Pattern Discovery

- Pre-processing to discretize the data[8]
- Greedy cut-points selection during the exploration[9]

	m_1	m_2	m_3	m_4	m_5
g_1	1	2	2	1	6
g_2	2	1	1	5	6
g_3	2	2	1	7	6
g_4	8	9	2	6	7

$\Rightarrow$

	$m_1 \in [0; 5]$	$m_1 \in]5; 15]$	$m_2 \geq 6$	$m_3 \geq 6$	$m_4 \geq 6$	$m_5 \geq 6$
g_1	×					×
g_2	×					×
g_3	×				×	×
g_4		×	×		×	×

- Even with discretization, numerical patterns are simply n-intervals hidden in the same space traversed by decision trees using cut-points: "boxes/rectangles" with sides parallel to axes of Euclidean space

Can we formalize n-intervals or boxes in FCA?

8 Y. Yang et al. "Discretization Methods". *Data Mining and Knowledge Discovery Handbook, 2nd ed.* 2010.

9 H. Grosskreutz et al. "On subgroup discovery in numerical domains". *Data Min. Knowl. Discov.* 19.2 (2009).

Interordinal scaling[13]

Transform and mine

- A scale to encode intervals of attribute values

	$m_1 \leq 4$	$m_1 \leq 5$	$m_1 \leq 6$	$m_1 \geq 4$	$m_1 \geq 5$	$m_1 \geq 6$
4	×	×	×	×		
5		×	×	×	×	
6			×	×	×	×

- Transformed data with scaling is inefficient to store, to work on and visualize
- A lot of redundancy (actually, implications of the scale can be used during extraction[10])
- Closed concepts are meaningful, but there are some problems with minimal generators (detailed after)

	m_1	...
g_1	4	...
g_2	5	...
g_3	6	...
g_4	5	...

$\Rightarrow$

	$m_1 \leq 4$	$m_1 \leq 5$	$m_1 \leq 6$	$m_1 \geq 4$	$m_1 \geq 5$	$m_1 \geq 6$	...
g_1	×	×	×	×			...
g_2		×	×	×	×		...
g_3			×	×	×	×	...
...							

"Why not working directly on intervals[11] ... with pattern structures [12]?"

[10] A. Belfodil et al. "Mining Formal Concepts Using Implications Between Items". *ICFCA*. 2019.

[11] S. O. Kuznetsov. "Galois Connections in Data Analysis (...)". *ICFCA*. 2005.

[12] B. Ganter et al. "Pattern Structures and Their Projections". *ICCS*. 2001.

[13] B. Ganter et al. *Formal Concept Analysis*. 1999.

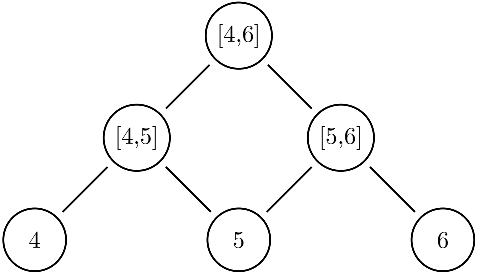
Interordinal scaling[13]

Transform and mine

- A scale to encode intervals of attribute values

	$m_1 \leq 4$	$m_1 \leq 5$	$m_1 \leq 6$	$m_1 \geq 4$	$m_1 \geq 5$	$m_1 \geq 6$
4	×	×	×	×		
5		×	×	×	×	
6			×	×	×	×

- Transformed data with scaling is inefficient to store, to work on and visualize
- A lot of redundancy (actually, implications of the scale can be used during extraction[10])
- Closed concepts are meaningful, but there are some problems with minimal generators (detailed after)



"Why not working directly on intervals[11] ... with pattern structures [12]?"

[10] A. Belfodil et al. "Mining Formal Concepts Using Implications Between Items". *ICFCA*. 2019.

[11] S. O. Kuznetsov. "Galois Connections in Data Analysis (...)". *ICFCA*. 2005.

[12] B. Ganter et al. "Pattern Structures and Their Projections". *ICCS*. 2001.

[13] B. Ganter et al. *Formal Concept Analysis*. 1999.

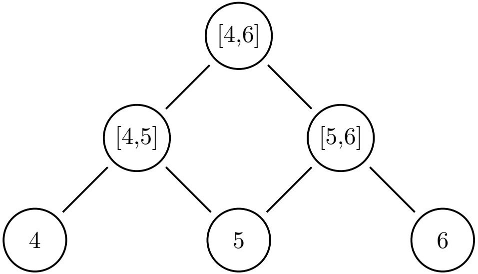
Interordinal scaling[13]

Transform and mine

- A scale to encode intervals of attribute values

	$m_1 \leq 4$	$m_1 \leq 5$	$m_1 \leq 6$	$m_1 \geq 4$	$m_1 \geq 5$	$m_1 \geq 6$
4	×	×	×	×		
5		×	×	×	×	
6			×	×	×	×

- Transformed data with scaling is inefficient to store, to work on and visualize
- A lot of redundancy (actually, implications of the scale can be used during extraction[10])
- Closed concepts are meaningful, but there are some problems with minimal generators (detailed after)



"Why not working directly on intervals[11] ... with pattern structures [12]?"

[10] A. Belfodil et al. "Mining Formal Concepts Using Implications Between Items". *ICFCA*. 2019.

[11] S. O. Kuznetsov. "Galois Connections in Data Analysis (...)". *ICFCA*. 2005.

[12] B. Ganter et al. "Pattern Structures and Their Projections". *ICCS*. 2001.

[13] B. Ganter et al. *Formal Concept Analysis*. 1999.

Interval Pattern Structures[16]

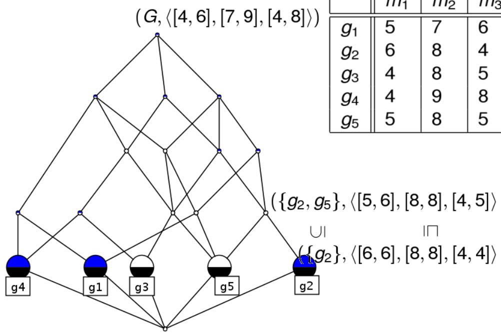
Directly mine n-intervals

- $(G, (D, \sqcap), \delta)$
  - For n-intervals, $\sqcap$ returns the convex-hull (other choices[14],[15], just need a meet-semi-lattice)
  - if D is the powerset of a set, we fall back to FCA.
- A pair of closures $(.)^{\square\square}$ forming a Galois connection

$$
\begin{aligned}
\{g_1, g_2\}^{\square} &= \bigsqcap_{g \in \{g_1, g_2\}} \delta(g) \\
&= \langle 5, 7, 6\rangle \sqcap \langle 6, 8, 4\rangle \\
&= \langle [5,6], [7,8], [4,6]\rangle \\
\langle [5,6], [7,8], [4,6]\rangle^{\square} &= \{g \in G | \langle [5,6], [7,8], [4,6]\rangle \sqsubseteq \delta(g)\} \\
&= \{g_1, g_2, g_5\}
\end{aligned}
$$

$(\{g_1, g_2, g_5\}, \langle [5,6], [7,8], [4,6]\rangle)$ is a (pattern) concept

	m_1	m_2	m_3
g_1	5	7	6
g_2	6	8	4
g_3	4	8	5
g_4	4	9	8
g_5	5	8	5



Top-down: Hyper-rectangle inclusion

[14] M. Kaytoue et al. "Embedding tolerance relations in FCA: an application in information fusion". *CIKM*. 2010.

[15] Z. Assaghir et al. "Managing Information Fusion with Formal Concept Analysis". *MDAI*. 2010.

[16] M. Kaytoue et al. "Mining gene expression data with pattern structures in FCA". *Inf. Sci.* 181.10 (2011).

Closed Interval Patterns and their Minimal Generators

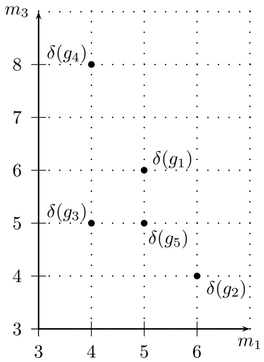
Why are closed interval patterns important in pattern mining?[17]

- $\langle [a,b],[c,d] \rangle$ with $a,b \in W_{m_1} = \{4,5,6\}$ and $c,d \in W_{m_2} = \{5,4,6,8\}$
- Total number of possible n-intervals

$$\prod_{i \in \{1,\ldots,|M|\}} \frac{|W_{m_i}| \times (|W_{m_i}|+1)}{2}$$

- An equivalence class has
  - a unique closed pattern: the smallest rectangle
  - one or several generators: the largest rectangles
  - no bijection between interval minimal generators and minimal itemsets from interordinally scaled data, it holds only for closed patterns
- Closed interval patterns offer a concise representation (10^7 to 10^9 on Bilkent's), but are not considered in the SD algorithms
- Generators may be interesting for rule-based classifiers as they "cover more"

	m_1		m_3
g_1	5		6
g_2	6		4
g_3	4		5
g_4	4		8
g_5	5		5



---

17 M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Closed Interval Patterns and their Minimal Generators

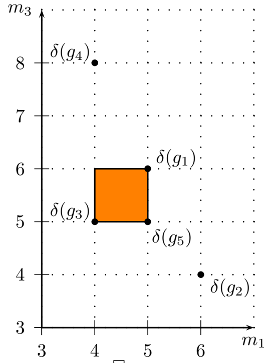
Why are closed interval patterns important in pattern mining?[17]

- $\langle [a,b],[c,d] \rangle$ with $a, b \in W_{m_1} = \{4,5,6\}$ and $c, d \in W_{m_2} = \{5,4,6,8\}$
- Total number of possible n-intervals

$$\prod_{i \in \{1,\dots,|M|\}} \frac{|W_{m_i}| \times (|W_{m_i}| + 1)}{2}$$

- An equivalence class has
  - a unique closed pattern: the smallest rectangle
  - one or several generators: the largest rectangles
  - no bijection between interval minimal generators and minimal itemsets from interordinally scaled data, it holds only for closed patterns
- Closed interval patterns offer a concise representation (10^7 to 10^9 on Bilkent's), but are not considered in the SD algorithms
- Generators may be interesting for rule-based classifiers as they "cover more"

	m_1		m_3
g_1	5		6
g_2	6		4
g_3	4		5
g_4	4		8
g_5	5		5



$\langle [4,5],[5,6] \rangle^\square = \{g_1, g_3, g_5\}$

[17] M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Closed Interval Patterns and their Minimal Generators

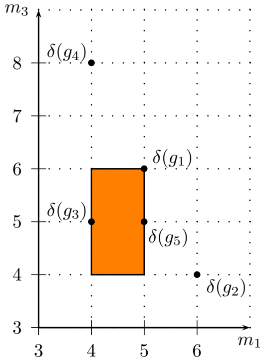
Why are closed interval patterns important in pattern mining?[17]

- $\langle [a,b],[c,d] \rangle$ with $a,b \in W_{m_1} = \{4,5,6\}$ and $c,d \in W_{m_2} = \{5,4,6,8\}$
- Total number of possible n-intervals

$$\prod_{i \in \{1,\dots,|M|\}} \frac{|W_{m_i}| \times (|W_{m_i}|+1)}{2}$$

- An equivalence class has
  - a unique closed pattern: the smallest rectangle
  - one or several generators: the largest rectangles
  - no bijection between interval minimal generators and minimal itemsets from interordinally scaled data, it holds only for closed patterns
- Closed interval patterns offer a concise representation (10^7 to 10^9 on Bilkent's), but are not considered in the SD algorithms
- Generators may be interesting for rule-based classifiers as they "cover more"

	m_1		m_3
g_1	**5**		**6**
g_2	6		4
g_3	**4**		**5**
g_4	4		8
g_5	**5**		**5**



$\langle [4,5],[4,6] \rangle^{\square} = \{g_1, g_3, g_5\}$

17 M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Closed Interval Patterns and their Minimal Generators

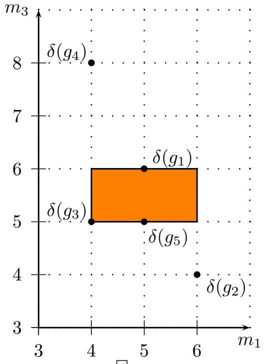
	m_1		m_3
g_1	**5**		**6**
g_2	6		4
g_3	**4**		**5**
g_4	4		8
g_5	**5**		**5**

Why are closed interval patterns important in pattern mining?[17]

- $\langle [a,b],[c,d] \rangle$ with $a,b \in W_{m_1} = \{4,5,6\}$ and $c,d \in W_{m_2} = \{5,4,6,8\}$
- Total number of possible n-intervals

$$\prod_{i \in \{1,\dots,|M|\}} \frac{|W_{m_i}| \times (|W_{m_i}|+1)}{2}$$

- An equivalence class has
  - a unique closed pattern: the smallest rectangle
  - one or several generators: the largest rectangles
  - no bijection between interval minimal generators and minimal itemsets from interordinally scaled data, it holds only for closed patterns
- Closed interval patterns offer a concise representation (10^7 to 10^9 on Bilkent's), but are not considered in the SD algorithms
- Generators may be interesting for rule-based classifiers as they "cover more"



$\langle [4,6],[5,6] \rangle^{\square} = \{g_1, g_3, g_5\}$

17 M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Closed Interval Patterns and their Minimal Generators

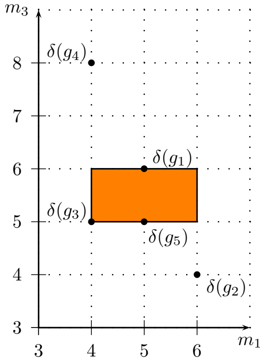
Why are closed interval patterns important in pattern mining?[17]

- $\langle [a,b],[c,d] \rangle$ with $a,b \in W_{m_1} = \{4,5,6\}$ and $c,d \in W_{m_2} = \{5,4,6,8\}$
- Total number of possible n-intervals

$$\prod_{i \in \{1,\ldots,|M|\}} \frac{|W_{m_i}| \times (|W_{m_i}|+1)}{2}$$

- An equivalence class has
    - a unique closed pattern: the smallest rectangle
    - one or several generators: the largest rectangles
    - no bijection between interval minimal generators and minimal itemsets from interordinally scaled data, it holds only for closed patterns
- Closed interval patterns offer a concise representation (10^7 to 10^9 on Bilkent's), but are not considered in the SD algorithms
- Generators may be interesting for rule-based classifiers as they "cover more"

	m_1		m_3
g_1	**5**		**6**
g_2	6		4
g_3	**4**		**5**
g_4	4		8
g_5	**5**		**5**



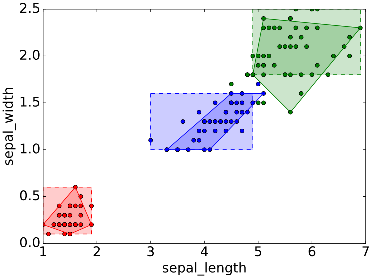
$\langle [4,6],[5,6] \rangle^\square = \{g_1, g_3, g_5\}$

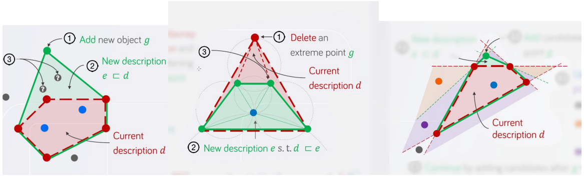
[17] M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Convex Polygon Patterns

Towards more expressive numerical patterns[18]

- Interval patterns consider each attribute independently
- Convex polytope patterns combine numerical attributes with conjunctions of linear inequalities $12.m_1(g) + 3.m_2(g) \leq 12 \wedge \ldots$
- Several algorithms to mine the structure $(G, (D, \sqcap), \delta)$ (as $\sqcap$ comm., reflx. assoc.)
  - Basic bottom up CbO: computes closures & test canonicity: we can avoid this
  - Top-down enumeration on a Delaunay triangulation: incrementally computes convex hulls
  - Bottom-up "vision-based algorithm": only points seeing one side can be added





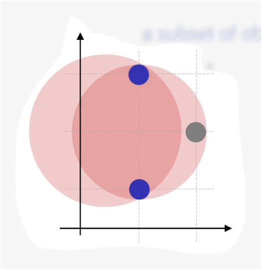
Only 2D! Integrate spatial attributes in a pattern structure

18 A. Belfodil et al. "Mining Convex Polygon Patterns with Formal Concept Analysis". *IJCAI*. 2017.

Beyond Pattern Structures

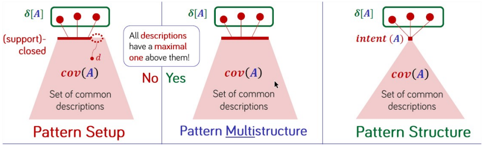
Pattern structures model only meet-semi-lattices

- Several minimal enclosing disks of a set of points
- Intersection of graph/sequence patterns is not unique but pattern structures can be used[19]!)
- What are the necessary conditions to apply this trick, what is the trick exactly?



First step towards understanding[20, 21]

- Pattern setups: D is just a poset
- Pattern-multi-structures: D is a multi-semi-lattice, can be turned to a pattern structure (anti-chain completion)



Can we design generic algorithms?

[19] S. O. Kuznetsov. "Learning of Simple Conceptual Graphs from Positive and Negative Examples". *PKDD*. 1999.

[20] Aimene Belfodil. "An Order Theoretic Point-of-view on Subgroup Discovery.". PhD thesis. 2019.

[21] A. Belfodil et al. "On Pattern Setups and Pattern Multistructures". *Int. J. General Systems (revision)* (2019).

Biclusters

Biclusters just look like concepts!

- Recommender systems, gene expression data, NN compression and mining
- a **local** phenomena in the data: "rectangles of values", differ with clustering
- **connectedness**: equality, similarity...
- **overlapping** of rectangles
- a partial **order** of biclusters
- **maximality** of rectangles

Many definitions, *ad hoc*/heuristic search[22]

- Iterative Row/Column Clustering Combination
- Divide and Conquer
- Greedy Iterative Search vs. Exhaustive Enumeration

	m_1	m_2	m_3	m_4	m_5
g_1	1	2	2	1	6
g_2	2	1	**1**	**0**	6
g_3	2	2	**1**	**7**	6
g_4	8	9	**2**	**6**	7

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

1.0	1.0	1.0	0.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	40	40

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

22 S.C. Madeira et al. "Biclustering algorithms for biological data analysis: a survey". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1.1 (2004).

Some bridges between FCA and biclustering (1/4)

Scaling may be enough![23]

- A bicluster (A, B) of similar values is s.t. $m_i(g_j) \simeq_\theta m_k(g_l), \forall g_j, g_l \in A, \forall m_i, m_k \in B$ and maximal if no object/attribute can be added
- $\simeq_\theta$ is a tolerance relation: reflexive, symmetric, but not transitive, from which classes of tolerance are defined as maximal (convex) sets of pairwise similar values, thus, also closed sets

$\simeq_1$	0	1	2	6	7	8	9
0	×	×					
1	×	×	×				
2		×	×				
6				×	×		
7				×	×	×	
8					×	×	×
9						×	×

Classes of tolerance
$\{0, 1\}$
$\{1, 2\}$
$\{6, 7\}$
$\{7, 8\}$
$\{8, 9\}$

Class of tolerance	Formal context	Concepts	Bicluster corresponding to first concept on left list
$[0, 1]$	$\begin{array}{c\|ccc} & m_2 & m_3 & m_4 \\ \hline g_1 & & & \times \\ g_2 & \times & \times & \times \end{array}$	$(\{g_1, g_2\}, \{m_4\})$ $(\{g_2\}, \{m_2, m_3, m_4\})$	$\begin{array}{c\|ccccc} & m_1 & m_2 & m_3 & m_4 & m_5 \\ \hline g_1 & 1 & 2 & 2 & \mathbf{1} & 6 \\ g_2 & 2 & 1 & 1 & \mathbf{0} & 6 \\ g_3 & 2 & 2 & 1 & 7 & 6 \\ g_4 & 8 & 9 & 2 & 6 & 7 \end{array}$
$[1, 2]$	$\begin{array}{c\|cccc} & m_1 & m_2 & m_3 & m_4 \\ \hline g_1 & \times & \times & \times & \times \\ g_2 & \times & \times & \times & \\ g_3 & \times & \times & \times & \\ g_4 & & & \times & \end{array}$	$(\{g_1, g_2, g_3\}, \{m_1, m_2, m_3\})$ $(\{g_1\}, \{m_1, m_2, m_3, m_4\})$ $(\{g_1, g_2, g_3, g_4\}, \{m_3\})$	$\begin{array}{c\|ccccc} & m_1 & m_2 & m_3 & m_4 & m_5 \\ \hline g_1 & \mathbf{1} & \mathbf{2} & \mathbf{2} & 1 & 6 \\ g_2 & \mathbf{2} & \mathbf{1} & \mathbf{1} & 0 & 6 \\ g_3 & \mathbf{2} & \mathbf{2} & \mathbf{1} & 7 & 6 \\ g_4 & 8 & 9 & 2 & 6 & 7 \end{array}$
$[6, 7]$	$\begin{array}{c\|cc} & m_4 & m_5 \\ \hline g_1 & & \times \\ g_2 & & \times \\ g_3 & \times & \times \\ g_4 & \times & \times \end{array}$	$(\{g_3, g_4\}, \{m_4, m_5\})$ $(\{g_1, g_2, g_3, g_4\}, \{m_5\})$	$\begin{array}{c\|ccccc} & m_1 & m_2 & m_3 & m_4 & m_5 \\ \hline g_1 & 1 & 2 & 2 & 1 & 6 \\ g_2 & 2 & 1 & 1 & 0 & 6 \\ g_3 & 2 & 2 & 1 & \mathbf{7} & \mathbf{6} \\ g_4 & 8 & 9 & 2 & \mathbf{6} & \mathbf{7} \end{array}$

Existing algorithms, natural distributed computing, ...

[23] M. Kaytoue et al. "Biclustering Numerical Data in Formal Concept Analysis". *ICFCA*. vol. 6628. LNCS. 2011.

Some bridges between FCA and biclustering (2/4)

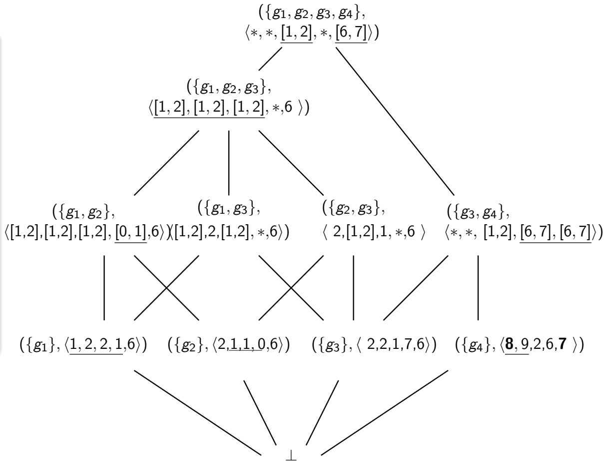
Interval Pattern Structures[24]

- Consider the tolerance relation when (i) computing intersections

$$[a_1, b_1] \sqcap [a_2, b_2] = \begin{cases} [min(a_1, a_2), max(b_1, b_2)] \\ \text{if} \leq \theta \\ *, \text{otherwise} \end{cases}$$

- (ii) checking subsumption $* \sqsubseteq [a, b]$
- Check maximality during the exploration

	m_1	m_2	m_3	m_4	m_5
g_1	**1**	**2**	**2**	**1**	6
g_2	2	1	1	0	6
g_3	2	2	1	7	6
g_4	8	9	2	6	7



24 M. Kaytoue et al. "Biclustering Numerical Data in Formal Concept Analysis". *ICFCA*. vol. 6628. LNCS. 2011.

Some bridges between FCA and biclustering (2/4)

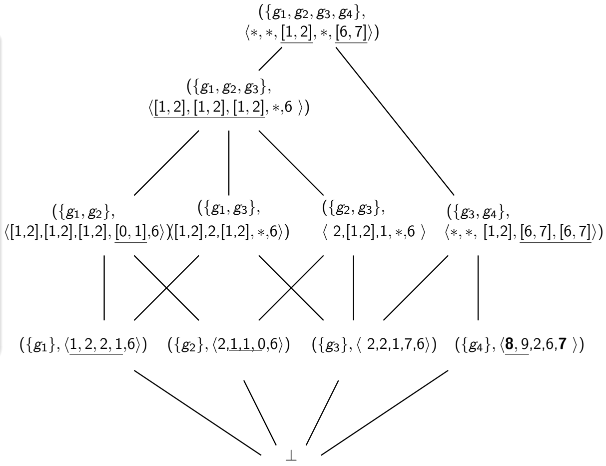
Interval Pattern Structures[24]

- Consider the tolerance relation when (i) computing intersections

$$[a_1, b_1] \sqcap [a_2, b_2] = \begin{cases} [min(a_1, a_2), max(b_1, b_2)] \\ \text{if} \leq \theta \\ *, \text{otherwise} \end{cases}$$

- (ii) checking subsumption $* \sqsubseteq [a, b]$
- Check maximality during the exploration

	m_1	m_2	m_3	m_4	m_5
g_1	**1**	**2**	**2**	1	6
g_2	2	1	1	0	6
g_3	**2**	**2**	**1**	7	6
g_4	8	9	2	6	7



24 M. Kaytoue et al. "Biclustering Numerical Data in Formal Concept Analysis". *ICFCA*. vol. 6628. LNCS. 2011.

Some bridges between FCA and biclustering (2/4)

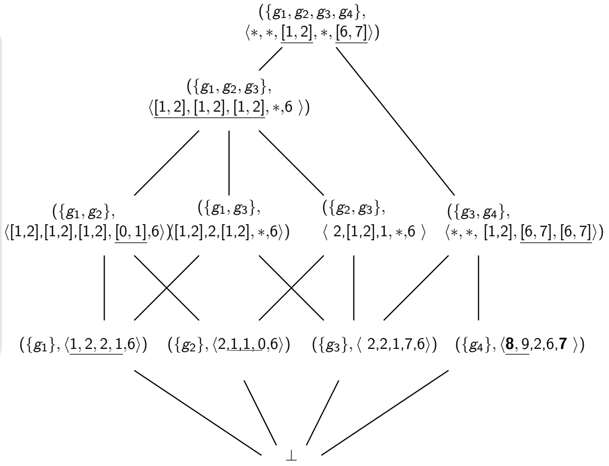
Interval Pattern Structures[24]

- Consider the tolerance relation when (i) computing intersections

$$[a_1, b_1] \sqcap [a_2, b_2] = \begin{cases} [min(a_1, a_2), max(b_1, b_2)] \\ \text{if} \leq \theta \\ *, \text{otherwise} \end{cases}$$

- (ii) checking subsumption $* \sqsubseteq [a, b]$
- Check maximality during the exploration

	m_1	m_2	m_3	m_4	m_5
g_1	**1**	**2**	**2**	1	6
g_2	**2**	**1**	**1**	0	6
g_3	**2**	**2**	**1**	7	6
g_4	8	9	2	6	7



24 M. Kaytoue et al. "Biclustering Numerical Data in Formal Concept Analysis". *ICFCA*. vol. 6628. LNCS. 2011.

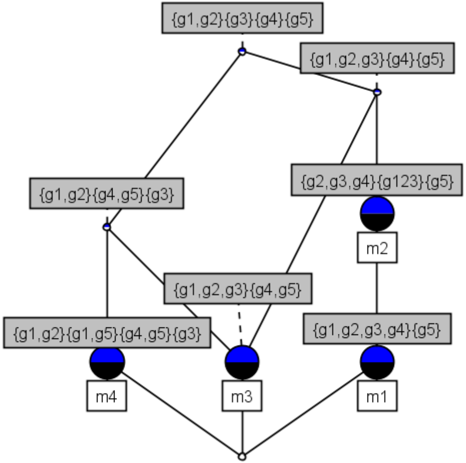
Some bridges between FCA and biclustering (3/4)

Partition Pattern Structures[25]

- $(M, (P(G,\theta), \sqcap), \delta)$: Each attribute is partitioned with hard/soft partitions (depending if θ is nonzero)
- $\sqcap$ and $\sqsubseteq$ are classic partition intersection and ordering

	m_1	m_2	m_3	m_4
g_1	1	2	2	8
g_2	2	1	2	9
g_3	2	1	1	2
g_4	1	0	7	6
g_5	6	6	6	7

$$\delta(m_1) = \{\{g_1, g_2, g_3, g_4\}\{g_5\}\}$$
$$\delta(m_2) = \{\{g_2, g_3, g_4\}\{g_1, g_2, g_3\}\{g_5\}\}$$
$$\delta(m_3) = \{\{g_1, g_2, g_3\}\{g_4, g_5\}\}$$
$$\delta(m_4) = \{\{g_4, g_5\}\{g_1, g_5\}\{g_1, g_2\}\{g_3\}\}$$



We feel that there is a hidden dimension somewhere...

25 M. Kaytoue et al. "FCA Methods for Mining Biclusters of Similar Values on Columns". *CLA*. 2014.

Some bridges between FCA and biclustering (4/4)

An additional dimension?[26]

- Triadic/Polyadic Concept Analysis[27]
- Efficient implementations to mine polyadic concepts[28]
- A bijection between the collection of biclusters (A,B) and the collection of triadic concepts (A, B, C) for some θ
- Generalizes to n-dimensional datasets, i.e. "n-clusters"

	$t_1 = [0,0]$					$t_2 = [0,1]$					$t_3 = [0,2]$					$t_4 = [0,6]$					$t_5 = [0,7]$				
	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5
g_1						×			×		×	×	×	×		×	×	×	×	×	×	×	×	×	×
g_2				×			×	×	×		×	×	×	×		×	×	×	×	×	×	×	×	×	×
g_3								×			×	×	×			×	×	×		×	×	×	×	×	×
g_4													×					×	×				×	×	×

	$t_6 = [0,8]$					$t_7 = [0,9]$					$t_8 = [1,9]$					$t_9 = [2,9]$					$t_{10} = [6,9]$				
	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5
g_1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×		×					×
g_2	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×				×					×
g_3	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×				×	×
g_4	×		×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×

	$t_{11} = [7,9]$					$t_{12} = [8,9]$					$t_{13} = [9,9]$				
	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5
g_1															
g_2															
g_3				×											
g_4	×	×			×	×	×					×			

We are still puzzled here on how to avoid the data scaling and work directly on what could be called a multi-dimensional pattern structure

26. M. Kaytoue et al. "Biclustering meets triadic concept analysis". *Ann. Math. Artif. Intell.* 70.1-2 (2014).
27. F. Lehmann et al. "A Triadic Approach to Formal Concept Analysis". *ICCS.* 1995.
28. L. Cerf et al. "Closed patterns meet *n*-ary relations". *TKDD* 3.1 (2009).

Data Dependencies

Functional dependencies (FD)...

- Let T be a set of tuples, and $X, Y \subseteq \mathcal{U}$, a FD $X \rightarrow Y$ holds if: $\forall t, t' \in T : t(X) = t'(X) \implies t(Y) = t'(Y)$
- A minimal generating setcan restore all FD's of T with Amstrong rules (reflexivity, augmentation, transitivity)

id	a	b	c	d
t_1	1	3	4	1
t_2	4	3	4	3
t_3	1	8	4	1
t_4	4	3	7	3

$a \rightarrow d,\ d \rightarrow a$

... look like implications in FCA

- Let (G, M, I) be a formal context, and $X, Y \subseteq M$, implication $X \rightarrow Y$ holds if $X' \subseteq Y'$: a objects from G having the attributes in X also have the attributes in Y
- Implications obey the Amstrong rules

	m_1	m_2	m_3
g_1	×		
g_2	×	×	
g_3		×	×
g_4		×	×
g_5	×	×	×

$m_3 \rightarrow m_2$

Used in DB for query optimization, normalization, data cleaning, error detection, but again, a several algorithms[29] and more & more types/relaxations of the FD definitions[30]

29 T. Papenbrock et al. "FD Discovery: An Experimental Evaluation of Seven Algorithms". *PVLDB* 8.10 (2015).

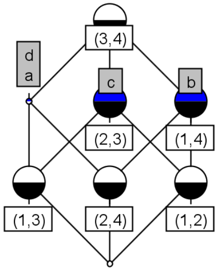
30 L. Caruccio et al. "Relaxed FD's - A Survey of Approaches". *IEEE Trans. Knowl. Data Eng.* 28.1 (2016).

Bridges between FCA and data dependencies

A first connection was proposed quite some time ago[31]

id	a	b	c	d
t_1	1	3	4	1
t_2	4	3	4	3
t_3	1	8	4	1
t_4	4	3	7	3

$\mathbb{K}$	a	b	c	d
(t_1, t_2)		×	×	
(t_1, t_3)	×		×	×
(t_1, t_4)		×		
(t_2, t_3)			×	
(t_2, t_4)	×	×		×
(t_3, t_4)				



"Quadratic transformation"! But...

Objects of the formal context encodes agree sets, i.e., the equivalence relation of a partition for each attribute... that we can intersect (on which rely algorithms such as TANE)

31 B. Ganter et al. *Formal Concept Analysis*. 1999.

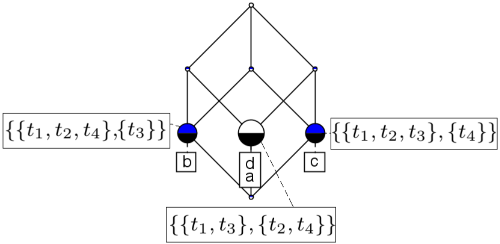
Bridges between FCA and data dependencies

Partition pattern structures

- Describe attributes with partitions and intersect: Pattern implications are in 1-1-correspondence with FD's[32]

id	a	b	c	d
t_1	1	3	4	1
t_2	4	3	4	3
t_3	1	8	4	1
t_4	4	3	7	3

m	$\delta(m) \in (D, \sqcap)$
a	$\{\{t_1, t_3\}, \{t_2, t_4\}\}$
b	$\{\{t_1, t_2, t_4\}, \{t_3\}\}$
c	$\{\{t_1, t_2, t_3\}, \{t_4\}\}$
d	$\{\{t_1, t_3\}, \{t_2, t_4\}\}$



- Relaxations are directly handled with "soft" partitions (tolerance) (same intersection/inclusion operations) ![33]
- Order dependencies are a bit trickier, Triadic Concept Analysis helped ("3rd dimension not symmetric")[34]

[32] J. Baixeries et al. "Characterizing FD's in FCA with pattern structures". *Ann. Math. Artif. Intell.* 72.1-2 (2014).

[33] J. Baixeries et al. "Characterizing approximate-matching dependencies in FCA". *Discr. Appl. Math.* 249 (2018).

[34] V. Codocedo et al. "Characterization of Order-like Dependencies with FCA". *CLA*. 2016.

Outline

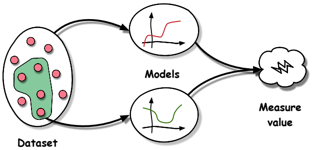
A short Introduction to Subgroup Discovery

Find subgroups of objects that behave differently[35]

- Most famous case: Weighted Relative Accuracy
  - $p = \langle MW \geq 142.22, nC \geq 11 \rangle$
  - $supp(p) = \{1, 60, 82\}$
  - $WRAcc(p, Musk) = \frac{3}{6} \times \left(\frac{2}{3} - \frac{2}{6}\right) = 0.17$
- "Generalized" with Exceptional Model Mining[36]

ID	MW	nAT	nC	Odor
1	150.19	21	11	
24	128.24	29	11	
48	136.16	24	9	
60	152.16	23	11	
82	151.28	27	12	
1633	142.22	27	10	

Discover a small set of diverse and high quality patterns?
In presence of hundreds of possibly correlated labels?[37]
Optimize directly the quality of the pattern set?[38]



35 S. Wrobel. "An Algorithm for Multi-relational Discovery of Subgroups". *PKDD*. 1997.

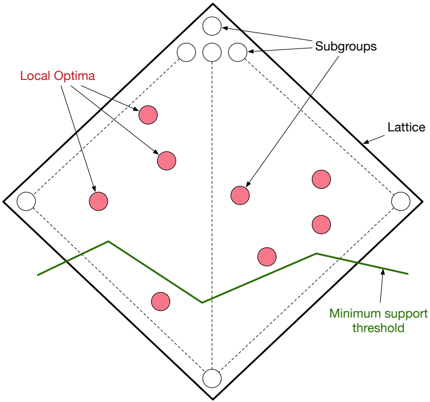
36 W. Duivesteijn et al. "Exceptional Model Mining (...)". *Data Min. Knowl. Discov.* (2016).

37 G. Bosc et al. "Local SD for Eliciting and Understanding New Structure-Odor Relationships". *DS*. 2016.

38 A. Belfodil et al. "FSSD: A Fast and Efficient Algorithm for Subgroup Set Discovery". *IEEE DSAA*. 2019.

Algorithm for (numerical) subgroup discovery

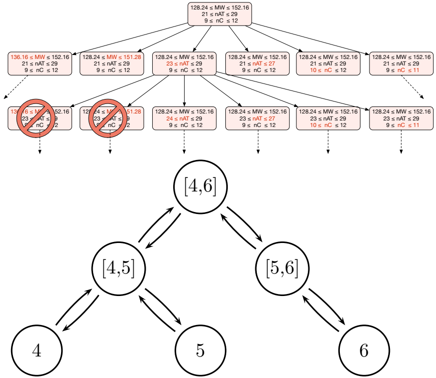
Finding (a few) (interesting) interval patterns



Algorithm for (numerical) subgroup discovery

Finding (a few) (interesting) interval patterns

- **Diversity** & non-monotonicity imply to consider a large search space
- **"Really exhaustive" search**
  - Bottom-up: objects sets from empty set (CbO)[39]
  - Top-down: "MinIntChange" as left/right shrinks[40]
  - Impossible for (not so) large data



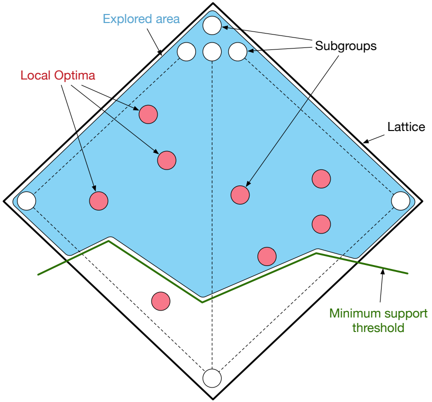
[39] M. Kaytoue et al. "Mining gene expression data with pattern structures in FCA". *Inf. Sci.* 181.10 (2011).

[40] M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Algorithm for (numerical) subgroup discovery

Finding (a few) (interesting) interval patterns

- **Diversity** & non-monotonicity imply to consider a large search space
- **“Really exhaustive” search**
  - Bottom-up: objects sets from empty set (CbO)
  - Top-down: “MinIntChange” as left/right shrinks
  - Impossible for (not so) large data
- **“Not really exhaustive” search**: discretization is applied before/during the search, impossible on large data, no information loss estimation, no idea if better discr. exist



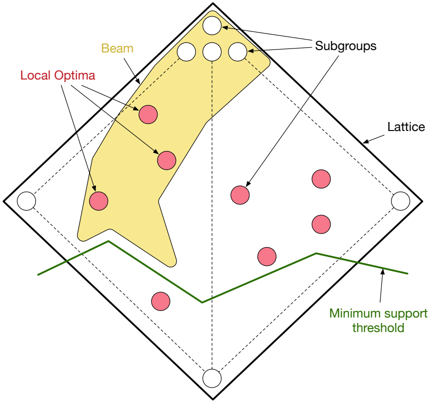
39

[39] M. Atzmüller et al. “SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery”. *PKDD*. 2006.

Algorithm for (numerical) subgroup discovery

Finding (a few) (interesting) interval patterns

- **Diversity** & non-monotonicity imply to consider a large search space
- **"Really exhaustive" search**
  - Bottom-up: objects sets from empty set (CbO)
  - Top-down: "MinIntChange" as left/right shrinks
  - Impossible for (not so) large data
- **"Not really exhaustive" search**: discretization is applied before/during the search, impossible on large data, no information loss estimation, no idea if better discr. exist
- **Beam-search**: a set of parallel directed hill climbings greedy algorithm, may get stuck in a few local optima, increasing the beam is difficult



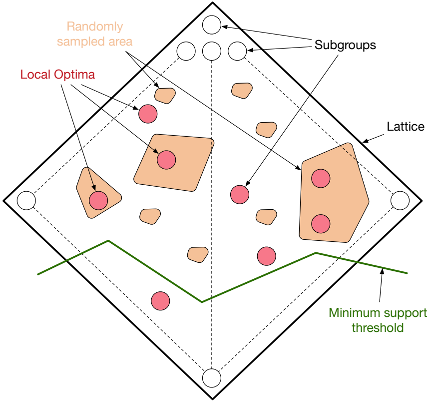
39

[39] M. van Leeuwen et al. "Diverse subgroup set discovery". *Data Min. Knowl. Discov.* 25.2 (2012).

Algorithm for (numerical) subgroup discovery

Finding (a few) (interesting) interval patterns

- **Diversity** & non-monotonicity imply to consider a large search space
- **"Really exhaustive" search**
  - Bottom-up: objects sets from empty set (CbO)
  - Top-down: "MinIntChange" as left/right shrinks
  - Impossible for (not so) large data
- **"Not really exhaustive" search**: discretization is applied before/during the search, impossible on large data, no information loss estimation, no idea if better discr. exist
- **Beam-search**: a set of parallel directed hill climbings greedy algorithm, may get stuck in a few local optima, increasing the beam is difficult
- **Sampling**: may be concerned with the long tail problem: a few patterns are interesting, many are not



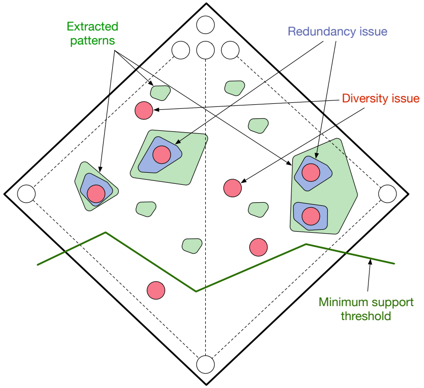
39

[39] M. Boley et al. "Direct local pattern sampling by efficient two-step random procedures". *KDD*. 2011.

Algorithm for (numerical) subgroup discovery

Finding (a few) (interesting) interval patterns

- **Diversity** & non-monotonicity imply to consider a large search space
- **"Really exhaustive" search**
  - Bottom-up: objects sets from empty set (CbO)
  - Top-down: "MinIntChange" as left/right shrinks
  - Impossible for (not so) large data
- **"Not really exhaustive" search**: discretization is applied before/during the search, impossible on large data, no information loss estimation, no idea if better discr. exist
- **Beam-search**: a set of parallel directed hill climbings greedy algorithm, may get stuck in a few local optima, increasing the beam is difficult
- **Sampling**: may be concerned with the long tail problem: a few patterns are interesting, many are not



A trade-off needs to be found between exploration and exploitation
Produce a small diverse set of patterns and avoid redundancy

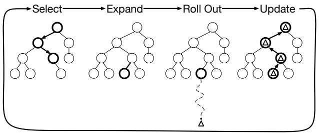
Monte Carlo Tree Search

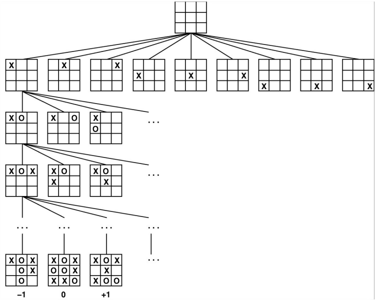
Sampling large trees/lattices of game states

MCTS[39] is an exploration method that builds iteratively the search tree according to random simulations.

- It aims at finding the best arm of a multi-armed bandit by sampling the search below each arm
- It explores the search space with random simulations to get rewards
- The more iterations, the best approximation of expected reward of each arm
- The trade-off between exploration and exploitation: Always go to the same restaurant vs. Try a new one!

$$UCT(s, s') = \frac{Q(s')}{N(s')} + 2\sqrt{\frac{\ln(N(s))}{N(s')}}$$



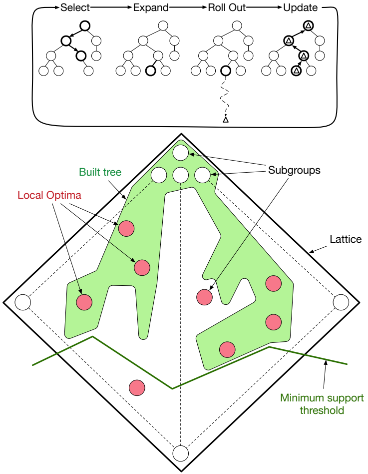


39 C. Browne et al. "A Survey of MCTS Methods". *IEEE Trans. Comput. Intellig. and AI in Games* (2012).

Diverse pattern set discovery with MCTS

MTCS4DM[40]

- Use the specialization operations to get direct lower neighbors in the lattice
- Building iteratively the search tree thanks to a fixed number of random simulations based on the exploration/exploitation trade-off
- Leads to exhaustive search if enough memory
- Ensure diversity *per se*: Extracting the top-k diverse and non redundant subgroups
- No knowledge on the measure is required
- A result is always available and improves over time
- An expert can express his preferences, used to drive the search (bias the simulations)

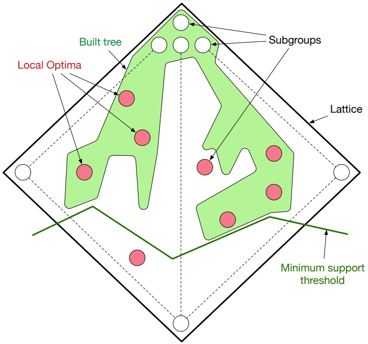


[40] G. Bosc et al. "Anytime discovery of a diverse set of patterns with MCTS". *Data Min. Knowl. Discov.* (2018).

Diverse pattern set discovery with MCTS

MTCS4DM[40]

- Use the specialization operations to get direct lower neighbors in the lattice
- Building iteratively the search tree thanks to a fixed number of random simulations based on the exploration/exploitation trade-off
- Leads to exhaustive search if enough memory
- Ensure diversity *per se*: Extracting the top-k diverse and non redundant subgroups
- No knowledge on the measure is required
- A result is always available and improves over time
- An expert can express his preferences, used to drive the search (bias the simulations)



but... Select/Expand/RollOut/Update are tricky to define, and are –to some extent– pattern language dependent[41]

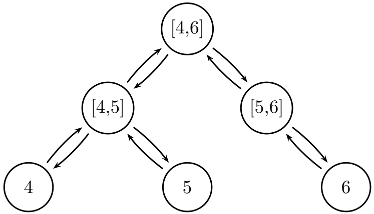
[40] G. Bosc et al. "Anytime discovery of a diverse set of patterns with MCTS". *Data Min. Knowl. Discov.* (2018).

[41] R. Mathonat et al. "A Bandit Model to Discover Interesting Subgroups in Labeled Sequences". *IEEE DSAA*. 2019.

Refine & mine

Minimal interval changes are too... minimal

- Direct specializations: minimal left/right shrinks
- Implies to search "interesting patterns very deeply"
- Can we cut anywhere in the attribute domain?



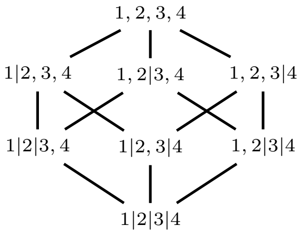
M. Kaytoue et al. "Revisiting Numerical Pattern Mining with Formal Concept Analysis". *IJCAI*. 2011.

Refine & mine

Minimal interval changes are too... minimal

- Direct specializations: minimal left/right shrinks
- Implies to search "interesting patterns very deeply"
- Can we cut anywhere in the attribute domain?
- Consider all possible discretizations, a finite lattice!
  - Top: roughest discretization holds (very rough) approximations of patterns of exhaustive search
  - Specializations: adding new cut points
  - Bottom: finest discretization holds pattern of exhaustive search!

With $domain(m_1) = \{1, 2, 3, 4\}$



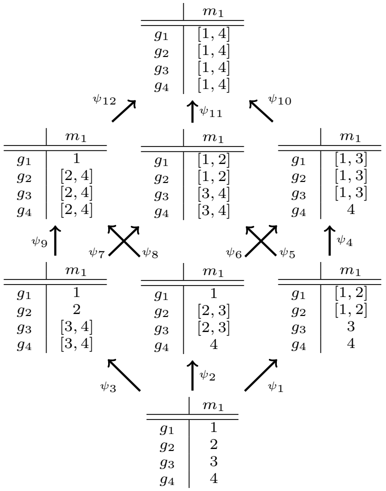
The lattice of discretizations
A meet-sub-semi-lattice of the partition lattice
A product for all attributes of the dataset

Refine & mine

Minimal interval changes are too... minimal

- Direct specializations: minimal left/right shrinks
- Implies to search "interesting patterns very deeply"
- Can we cut anywhere in the attribute domain?
- Consider all possible discretizations, a finite lattice!
    - Top: roughest discretization holds (very rough) approximations of patterns of exhaustive search
    - Specializations: adding new cut points
    - Bottom: finest discretization holds pattern of exhaustive search!

Most pattern mining algorithms consider one or several nodes of this lattice, order-isomorphic to the powerset lattice.

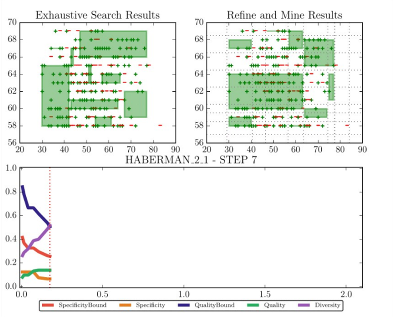


The lattice of all interval pattern structure projections
$((G, (D, \sqcap_{interval}), \psi_{i \in [1;16]} \circ \delta), \sqcap_{partition})$

Refine & mine

Algorithm: A first proposition

- No need to start from the top! Simply build an arbitrary discretization (equi-width -depth)
- (Partially) explore the chain until the bottom (if given enough budget!)
- At each step, performs a interval pattern mining, provides distance to the exploration end, guarantees if the best possible subgroup has been encountered already
- Could use a closed itemset mining algorithm (called the "nominal" property in Cortana/SD)



A. Belfodil et al. "Anytime Subgroup Discovery in Numerical Domains with Guarantees". *ECML/PKDD (best student paper in data mining award)*. 2018.

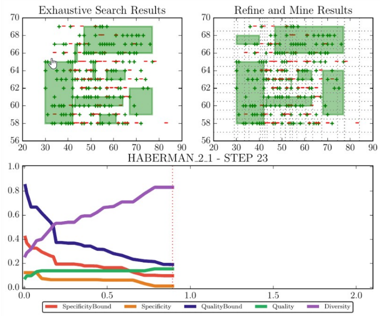
Most pattern mining algorithms consider one or several nodes of this lattice, order-isomorphic to the powerset lattice.

Refine & mine

Algorithm: A first proposition

- No need to start from the top! Simply build an arbitrary discretization (equi-width -depth)
- (Partially) explore the chain until the bottom (if given enough budget!)
- At each step, performs a interval pattern mining, provides distance to the exploration end, guarantees if the best possible subgroup has been encountered already
- Could use a closed itemset mining algorithm (called the "nominal" property in Cortana/SD)

Most pattern mining algorithms consider one or several nodes of this lattice, order-isomorphic to the powerset lattice.

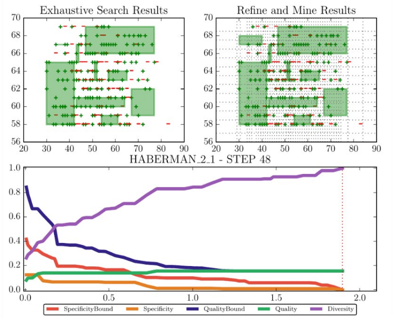


A. Belfodil et al. "Anytime Subgroup Discovery in Numerical Domains with Guarantees". *ECML/PKDD (best student paper in data mining award)*. 2018.

Refine & mine

Algorithm: A first proposition

- No need to start from the top! Simply build an arbitrary discretization (equi-width -depth)
- (Partially) explore the chain until the bottom (if given enough budget!)
- At each step, performs a interval pattern mining, provides distance to the exploration end, guarantees if the best possible subgroup has been encountered already
- Could use a closed itemset mining algorithm (called the "nominal" property in Cortana/SD)



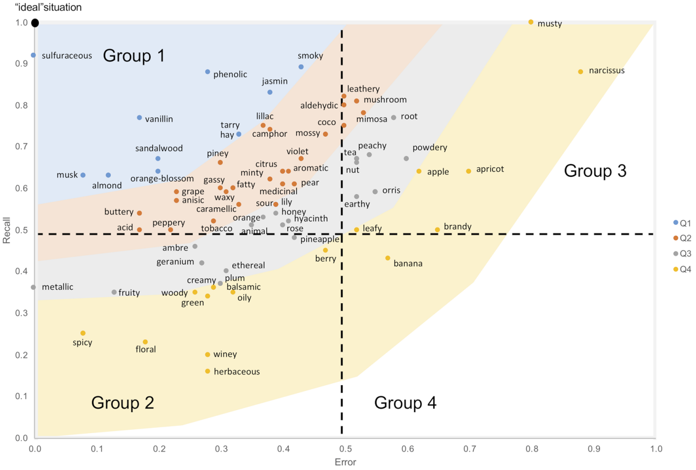
Most pattern mining algorithms consider one or several nodes of this lattice, order-isomorphic to the powerset lattice.

A. Belfodil et al. "Anytime Subgroup Discovery in Numerical Domains with Guarantees". *ECML/PKDD (best student paper in data mining award)*. 2018.

Outline

- Data & Pattern Formalization
  - Numerical Pattern Mining
  - Biclustering
  - Data Dependencies
- Pattern Mining and Subgroup Discovery
  - Mining a small set of diverse patterns
  - Iteratively mine finer data representations
- Knowledge Discovery in Practice
  - Neuroscience & Olfaction
  - Social Network Analysis
  - Video Game Analytics
- Perspectives

Olfaction in Neuroscience

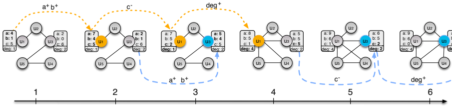


C. C. Licon et al. "Chemical features mining provides new descriptive structure-odor relationships". *PLOS Computational Biology* 15.4 (Apr. 2019).

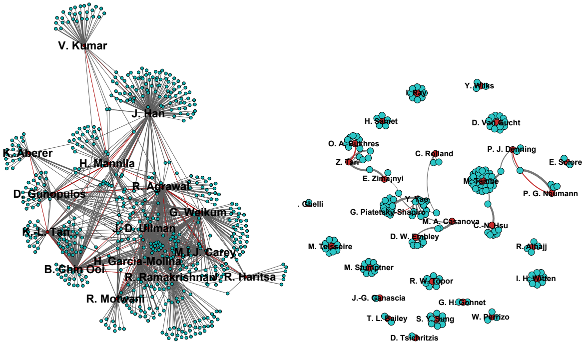
Social Network Analysis

European project with TCD & Tapastreet

- 3 years project, 8 months in the company
- Theoretical contributions in DM2L
- Applied & Engineering contributions
- Impact on teaching (internships & class)



[42]



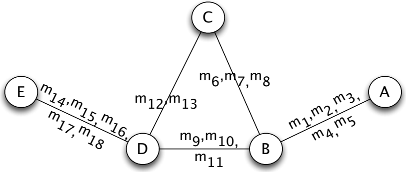
$$\{closeness_1^-\}, \{IEEETkde^+\}, \{numCliques_1^+\} \rightarrow \{numCliques_1^-\}$$
$$\{eigenvector_1^{++}\}, \{Journal^{++}, betweenness_3^{++}\} \rightarrow \{betweenness_4^{++}\}$$

[42] M. Kaytoue et al. “What effects topological changes in dynamic graphs? - Elucidating relationships between vertex attributes and the graph structure”. *Social Netw. Analys. Mining* 5.1 (2015).

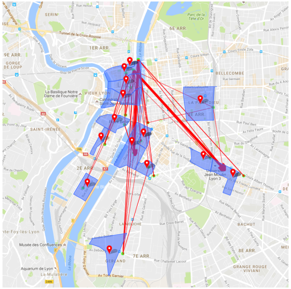
Social Network Analysis

European project with TCD & Tapastreet

- 3 years project, 8 months in the company
- Theoretical contributions in DM2L
- Applied & Engineering contributions
- Impact on teaching (internships & class)



Nodes and edges are provided with a context[42]

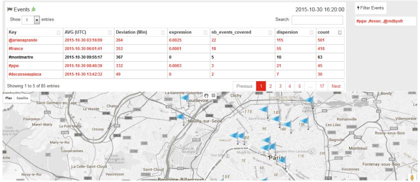


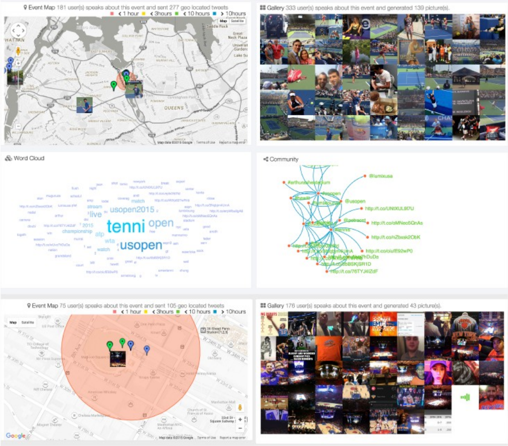
[42] M. Kaytoue et al. "Exceptional contextual subgraph mining". *Machine Learning* 106.8 (2017).

Social Network Analysis

European project with TCD & Tapastreet

- 3 years project, 8 months in the company
- Theoretical contributions in DM2L
- Applied & Engineering contributions
- Impact on teaching (internships & class)





42

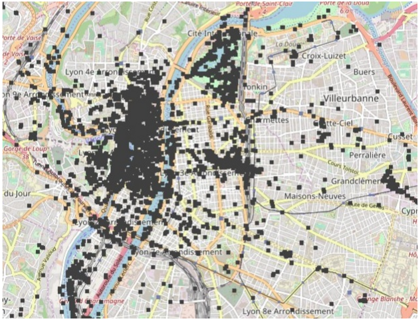
---

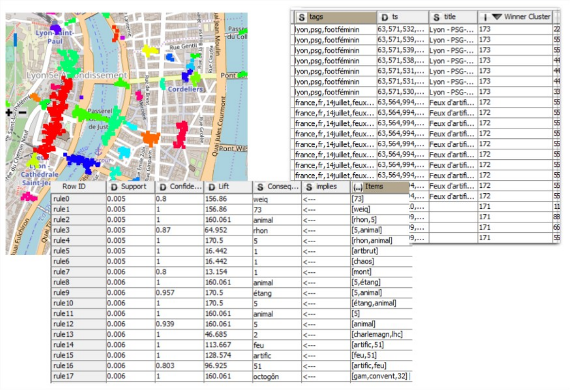
42 P. Houdyer et al. "Gazouille: Detecting and Illustrating Events from Social Media". *Demo@ECML/PKDD*. 2015.

Social Network Analysis

European project with TCD & Tapastreet

- 3 years project, 8 months in the company
- Theoretical contributions in DM2L
- Applied & Engineering contributions
- Impact on teaching (internships & class)





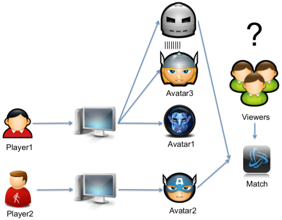
tags	ts	title	Winner Cluster	
lyon,psg,footféminin	63,571,532,...	Lyon - PSG-...	173	22
lyon,psg,footféminin	63,571,539,...	Lyon - PSG-...	173	55
lyon,psg,footféminin	63,571,539,...	Lyon - PSG-...	173	55
lyon,psg,footféminin	63,571,538,...	Lyon - PSG-...	173	44
lyon,psg,footféminin	63,571,531,...	Lyon - PSG-...	173	44
lyon,psg,footféminin	63,571,531,...	Lyon - PSG-...	173	44
lyon,psg,footféminin	63,571,530,...	Lyon - PSG-...	173	33
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
france,fr,14juillet,feux...	63,564,994,...	Feux d'artifi...	172	55
	4,...	Feux d'artifi...	172	55
	4,...	Feux d'artifi...	172	55
	0,...		171	11
	9,...		171	88
	9,...		171	66
	9,...		171	55

Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.005	0.8	156.86	weiq	<---	[73]
rule1	0.005	1	156.86	73	<---	[weiq]
rule2	0.005	1	160.061	animal	<---	[rhon,5]
rule3	0.005	0.87	64.952	rhon	<---	[5,animal]
rule4	0.005	1	170.5	5	<---	[rhon,animal]
rule5	0.005	1	16.442	1	<---	[artbrut]
rule6	0.005	1	16.442	1	<---	[chaos]
rule7	0.006	0.8	13.154	1	<---	[mont]
rule8	0.006	1	160.061	animal	<---	[5,étang]
rule9	0.006	0.957	170.5	étang	<---	[5,animal]
rule10	0.006	1	170.5	5	<---	[étang,animal]
rule11	0.006	1	160.061	animal	<---	[5]
rule12	0.006	0.939	160.061	5	<---	[animal]
rule13	0.006	1	46.685	2	<---	[charlemagn,lhc]
rule14	0.006	1	113.667	feu	<---	[artific,51]
rule15	0.006	1	128.574	artific	<---	[feu,51]
rule16	0.006	0.803	96.925	51	<---	[artific,feu]
rule17	0.006	1	160.061	octogôn	<---	[gam,convent,32]

Video Game Analytics

A growing field, with freely available data!

- Lack of data in industrial projects
- Rise of Video Game Live Streaming[42] (1-month stay at MIT Media Lab)
- Applications with minor contributions to pattern mining
  - Strategic Sequential Patterns[43]
  - Player keystroke dynamics[44]
  - Learning to play[45]
- Realistic data generation
- Impact on teaching and industry



---

42 M. Kaytoue et al. “Watch me playing, i am a professional (...)”. *MSND@WWW*. 2012 – 210 citations!.

43 G. Bosc et al. “Pattern Mining (...) to Study Strategy Balance in RTS Games”. *IEEE Trans. on Games* (2017).

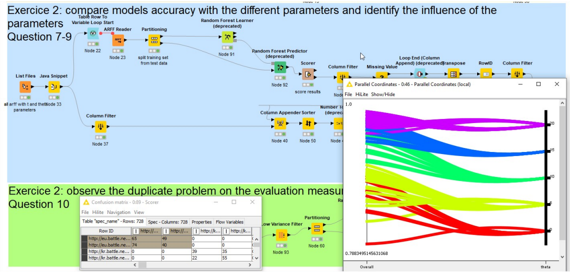
44 O. Cavadenti et al. “(...)Clustering confusion matrices to identify cyberathletes aliases”. *DSAA*. 2015.

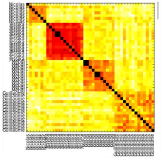
45 O. Cavadenti et al. “What is Wrong in My MOBA? Patterns Discriminating Deviant Behaviours”. *DSAA*. 2016.

Video Game Analytics

A growing field, with freely available data!

- Lack of data in industrial projects
- Rise of Video Game Live Streaming[42] (1-month stay at MIT Media Lab)
- Applications with minor contributions to pattern mining
  - Strategic Sequential Patterns[43]
  - Player keystroke dynamics[44]
  - Learning to play[45]
- Realistic data generation
- Impact on teaching and industry





[42] M. Kaytoue et al. "Watch me playing, i am a professional (...)". *MSND@WWW*. 2012 – 210 citations!.

[43] G. Bosc et al. "Pattern Mining (...) to Study Strategy Balance in RTS Games". *IEEE Trans. on Games* (2017).

[44] O. Cavadenti et al. "(...)Clustering confusion matrices to identify cyberathletes aliases". *DSAA*. 2015.

[45] O. Cavadenti et al. "What is Wrong in My MOBA? Patterns Discriminating Deviant Behaviours". *DSAA*. 2016.

Outline

Overview

Format Concept Analysis: a mean for cross-domain fertilization

- Numerical patterns: we can consider elegantly all closed n-intervals, better understanding of some patterns
- Biclusters: several types of biclusters are concepts (of a pattern structures or a (triadic) context)
- Data dependencies: Implications of a pattern structures & between formal contexts can model many types

Pattern Mining Algorithms

- Handling pattern set diversity with UCB & Monte Carlo Tree Search ("with a guarantee")
- Refine & mine: Perform exhaustive search on finer and finer data representation (with guarantees)

Research community

- FCA: Editorial board member of the int. conf. on FCA (since 2014[46]), PC member of its sister CLA
- AI: PC for many AI conferences (IJCAI/ECAI), reviewer for AI, Annals of Math. and AI, Discr. Appl. Math.
- DM: PC for ECML/PKDD, KDD, ICDM, reviewer for Machine Learning, Data Ming. Knowl. Discov. journals

[46] C.-V. Glodeanu, M. Kaytoue and C. Sacarea. "Formal Concept Analysis - 12th International Conference, ICFCA 2014, Cluj-Napoca, Romania, June 10-13, 2014. Proceedings". Vol. 8478. LNCS. 2014.

What's coming next

Data & Pattern Formalization

- Patterns: intervals for classification, polygons, circles, ... Links between TCA and pattern structures
- FDs: A systematic approach given the relation properties; pseudo-closed-sets & Algorithms

Subgroup Discovery and Algorithms

- Monte Carlo Tree Search, Refine&mine, sequence mining...: From rough to finer data representations
- Take in to account data complexity in the exploration exploitation trade-off. Formally, one simply "project" a pattern structure (multi? setup?)
- Constrained pattern mining, pattern quality measure cannot tell everything: Take into account user-feedback during the search[47], reuse his choices, learn the quality measure

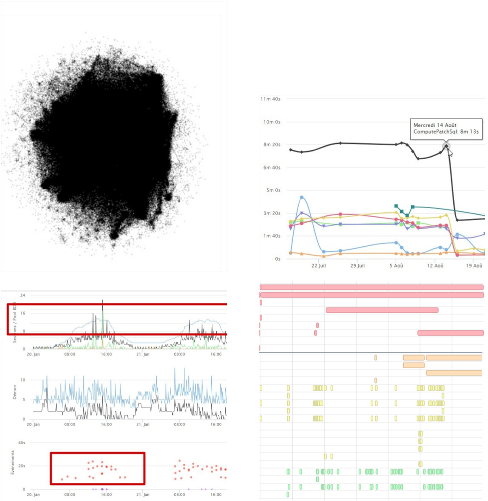
Towards a systematic actionnability, through knowledge discovery in practice...
... which face inevitably research challenges (it depends until where one wishes to go)

[47] G. Bosc et al. "h(odor): Interactive Discovery of Hypotheses on the Structure-Odor Relationship in Neuroscience". *Demo@ECML/PKDD*. 2016.

Knowledge Discovery in Practice at Infologic

A fantastic growth urges to process digitization

- For long purely client business oriented
- Now in desperate need of formalizing, understand, optimize its activities (development, integration, marketing & sellers, direction, teaching, ...)
- Collect/Consolidate data from many sources (source code, client database, usage data, sells, catalogs)
    - Static Code and Software Analytics
    - Predictive Maintenance
    - Knowledge Spaces, Graphs
    - Behavioral Data Analytics
    - Natural Language processing
    - Business rules reasoning



- Olivier Cavadenti. "Contribution de la découverte de motifs à l'analyse de collections de traces unitaires.". PhD thesis. 2016.
- Guillaume Bosc. "Anytime Discovery of a Diverse Set of Patterns with Monte Carlo Tree Search". PhD thesis. 2017.
- Aimene Belfodil. "An Order Theoretic Point-of-view on Subgroup Discovery.". PhD thesis. 2019.
- Victor Codocedo, Post-doctoral researcher (2015–2016)
- Pierre Houdyer, Research Engineer (2015–2016)
- Romain Mathonat. "Sampling patterns in sequential data. Application to Rocket League®", 2020.

Thank you for your attention!