Mining a new fault-tolerant pattern type as an alternative to formal concept discovery

Jérémy Besson, Céline Robardet and Jean-François Boulicaut

Institut National des Sciences Appliquées de Lyon - France

Friday July 21

Outline

Gene expression data analysis

Question

What are the sets of genes that are simultaneously over expressed in some biological situations?

Boolean gene expression data

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
S_1	1	0	1	0	1	0	0
S_2	1	1	1	1	0	1	0
S_3	1	1	1	1	1	0	0
S_4	1	1	1	1	1	0	0
S_5	0	1	1	1	1	1	1
S_6	0	0	0	1	1	1	0
S_7	1	0	0	0	0	0	0

D-Miner: an algorithm to compute formal concepts

Characteristics

Extracts all formal concepts from (G, M, I):

- in dense formal contexts
- using constraints on the intent and/or extent:
  - minimum size constraint
  - membership constraint

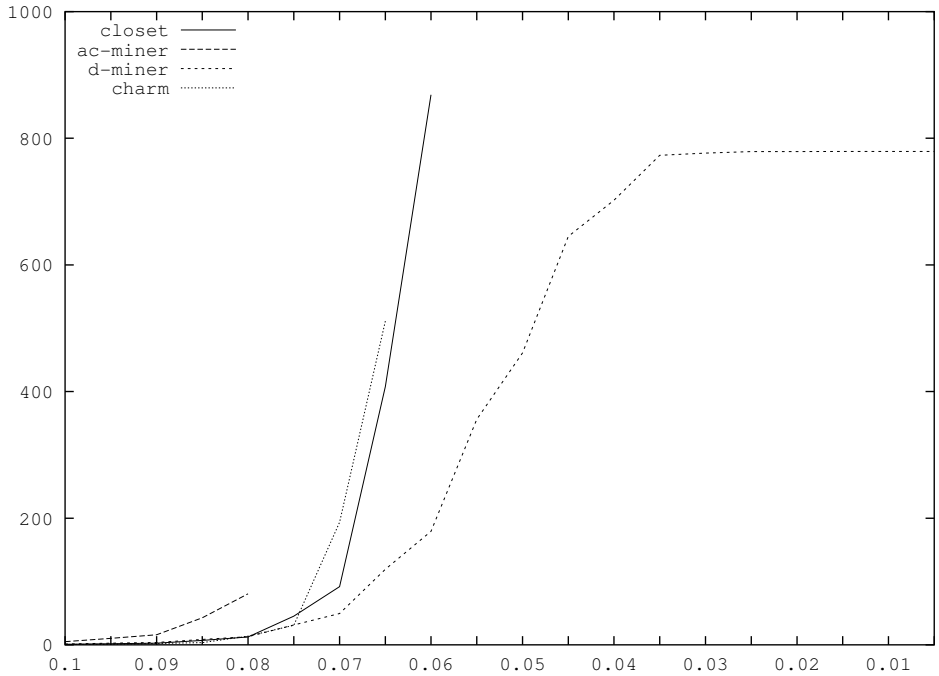
Efficiency

- Polynomial delay in worst case: $O(|G|^2.|M|)$
- Polynomial delay in average: $(|G| - \log_2(K) + 1)O(|G|.|M|)$

($K =$ is the number of formal concepts)

Experimental evaluation

Biological Data (104×304)



D-Miner succeeds in extracting all the 5 millions of formal concepts.

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	1	1	0	0
o_4	1	1	1	1
o_5	0	0	1	1
o_6	0	0	1	1

	a_1	a_2	a_3	a_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	1	1	0	1
o_4	1	1	1	1
o_5	0	0	1	0
o_6	0	0	1	1

A formal context K_1 (left), K_2 with 17% of noise (right)

FCA: Pros and Cons - 2

Goods characteristics of formal concepts that should be preserved:

- The numbers of zero values are bounded on objects and attributes.
- They are maximal bi-sets on both dimensions.
- No outside pattern object (resp. attribute) is identical to an inside pattern object (resp. on attribute).
- The intent and extent are associated by functions.

FCA: Pros and Cons - 2

Goods characteristics of formal concepts that should be preserved:

- The numbers of zero values are bounded on objects and attributes.
- They are maximal bi-sets on both dimensions.
- No outside pattern object (resp. attribute) is identical to an inside pattern object (resp. on attribute).
- The intent and extent are associated by functions.

FCA: Pros and Cons - 2

Goods characteristics of formal concepts that should be preserved:

- The numbers of zero values are bounded on objects and attributes.
- They are maximal bi-sets on both dimensions.
- No outside pattern object (resp. attribute) is identical to an inside pattern object (resp. on attribute).
- The intent and extent are associated by functions.

FCA: Pros and Cons - 2

Goods characteristics of formal concepts that should be preserved:

- The numbers of zero values are bounded on objects and attributes.
- They are maximal bi-sets on both dimensions.
- No outside pattern object (resp. attribute) is identical to an inside pattern object (resp. on attribute).
- The intent and extent are associated by functions.

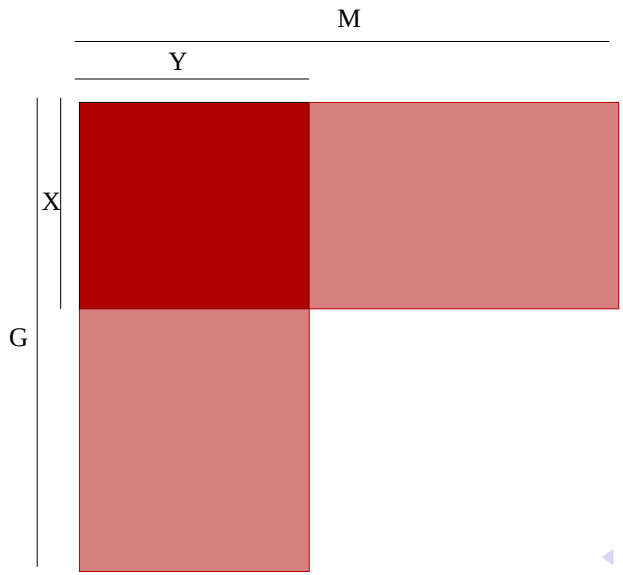
Principle

Fault-tolerant bi-sets

- the number of 0 values is upper-bounded on each element.

Relevancy

- the rows and columns outside the pattern contain more 0 values than the inside ones.



A generalization of formal concepts

(G, M, I) a formal context:

Notation

$\mathcal{Z}_o(x, Y)$ is the number of 0 values of an object x on the attributes in Y.
Similarly $\mathcal{Z}_a(y, X)$ is the number of 0 values of an attribute y on the objects in X.

Formal concepts can now be characterized by the following lemma:

Lemma

A bi-set (X, Y) is a formal concept:
(1) $\forall x \in X,\ \mathcal{Z}_o(x, Y) = 0$ *or similarly* $\forall y \in Y,\ \mathcal{Z}_a(y, X) = 0$
(2) $(\forall x \in G \setminus X,\ \mathcal{Z}_o(x, Y) \geq 1)$ *and* $(\forall y \in M \setminus Y,\ \mathcal{Z}_a(y, X) \geq 1)$.

Dense and relevant bi-sets

Definition

Given $(X, Y) \in 2^G \times 2^M$ and an integer value α, (X, Y) is said dense iff it satisfies $\mathcal{C}_d(\alpha, (X, Y)) \equiv (\forall x \in X,\ \mathcal{Z}_o(x, Y) \leq \alpha)$ and $(\forall y \in Y,\ \mathcal{Z}_a(y, X) \leq \alpha)$.

Definition

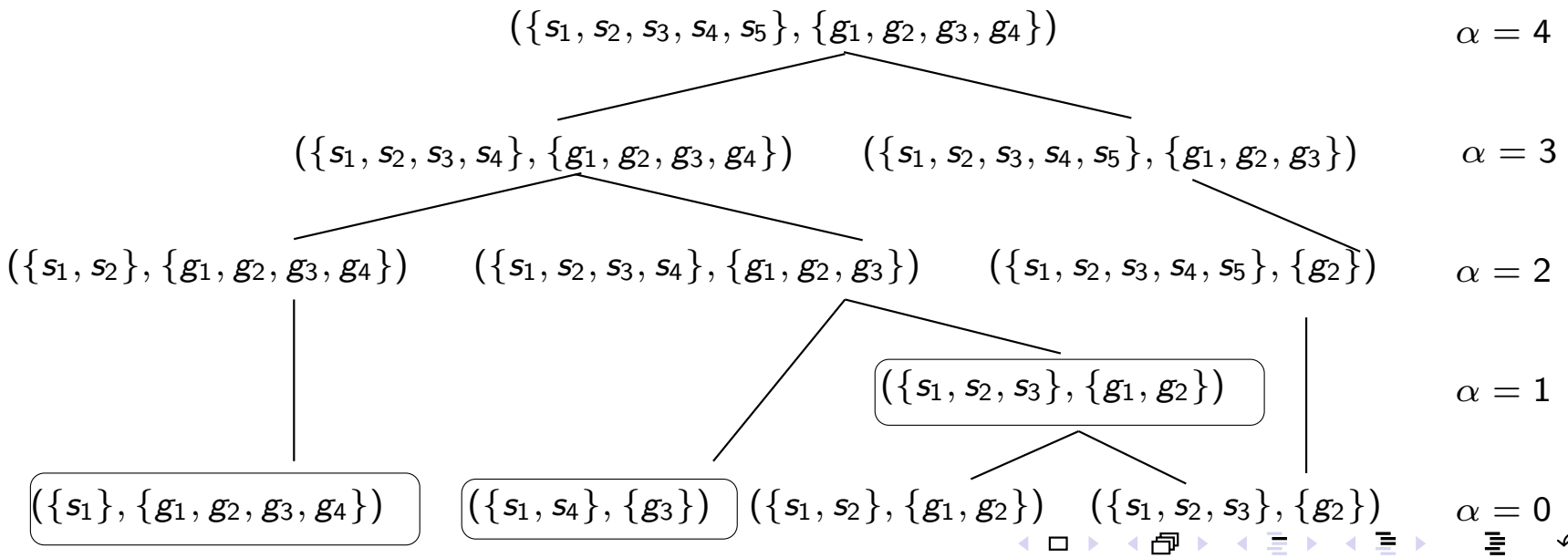
Given $(X, Y) \in 2^G \times 2^M$, and a positive integer value δ, (X, Y) is said relevant iff it satisfies:
$\mathcal{C}_r(\delta, (X, Y)) \equiv (\forall g \in G \setminus X,\ \forall x \in X,\ \mathcal{Z}_o(g, Y) \geq \mathcal{Z}_o(x, Y) + \delta)$ and $(\forall m \in M \setminus Y,\ \forall y \in Y,\ \mathcal{Z}_a(m, X) \geq \mathcal{Z}_a(y, X) + \delta)$

δ is the difference between the number of 0 values inside and outside the pattern.

The collection of bi-sets satisfying $\mathcal{C}_d$ and $\mathcal{C}_r$

	g_1	g_2	g_3	g_4
s_1	1	1	1	1
s_2	1	1	0	0
s_3	0	1	0	0
s_4	0	0	1	0
s_5	0	0	0	0

$\alpha = 5$ and $\alpha' = 4$
$\delta = 1$



An example of DR-bi-sets

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
S_1	1	0	1	0	1	0	0
S_2	1	1	1	1	0	1	0
S_3	1	1	1	1	1	0	0
S_4	1	1	1	1	1	0	0
S_5	0	1	1	1	1	1	1
S_6	0	0	0	1	1	1	0
S_7	1	0	0	0	0	1	0

$\alpha = 1$ and $\delta = 1$

DR-bi-sets properties

When $\alpha = 0$ and $\delta = 1$, DR-bi-sets are the formal concepts.

DR-bi-set size increases with parameter α.

Property

Given $0 \leq \alpha_1 \leq \alpha$, $\forall (X_1, Y_1) \in DR_{\alpha_1 \delta}$, $\exists (X, Y) \in DR_{\alpha \delta}$ *such that* $X_1 \subseteq X$ *and* $Y_1 \subseteq Y$.

DR-bi-sets are embedded by two functions.

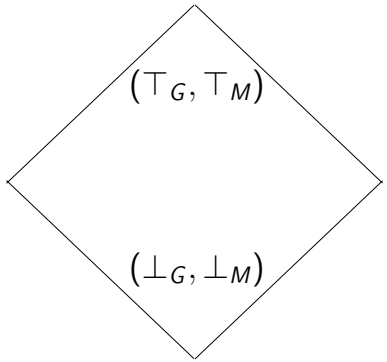
Property

For $\delta > 0$, *there exists two functions called* ψ_{DR} *and* ϕ_{DR} *such that* $\psi_{DR}\, 2^G \rightarrow 2^M$ *and* $\phi_{DR} : 2^M \rightarrow 2^G$ *such that* (X, Y) *is a DR-bi-set iff* $X = \phi_{DR}(Y)$ *and* $Y = \psi_{DR}(X)$.

DR-Miner

Lattice of the whole collection of bi-set: $((\emptyset, \emptyset), (G, M))$

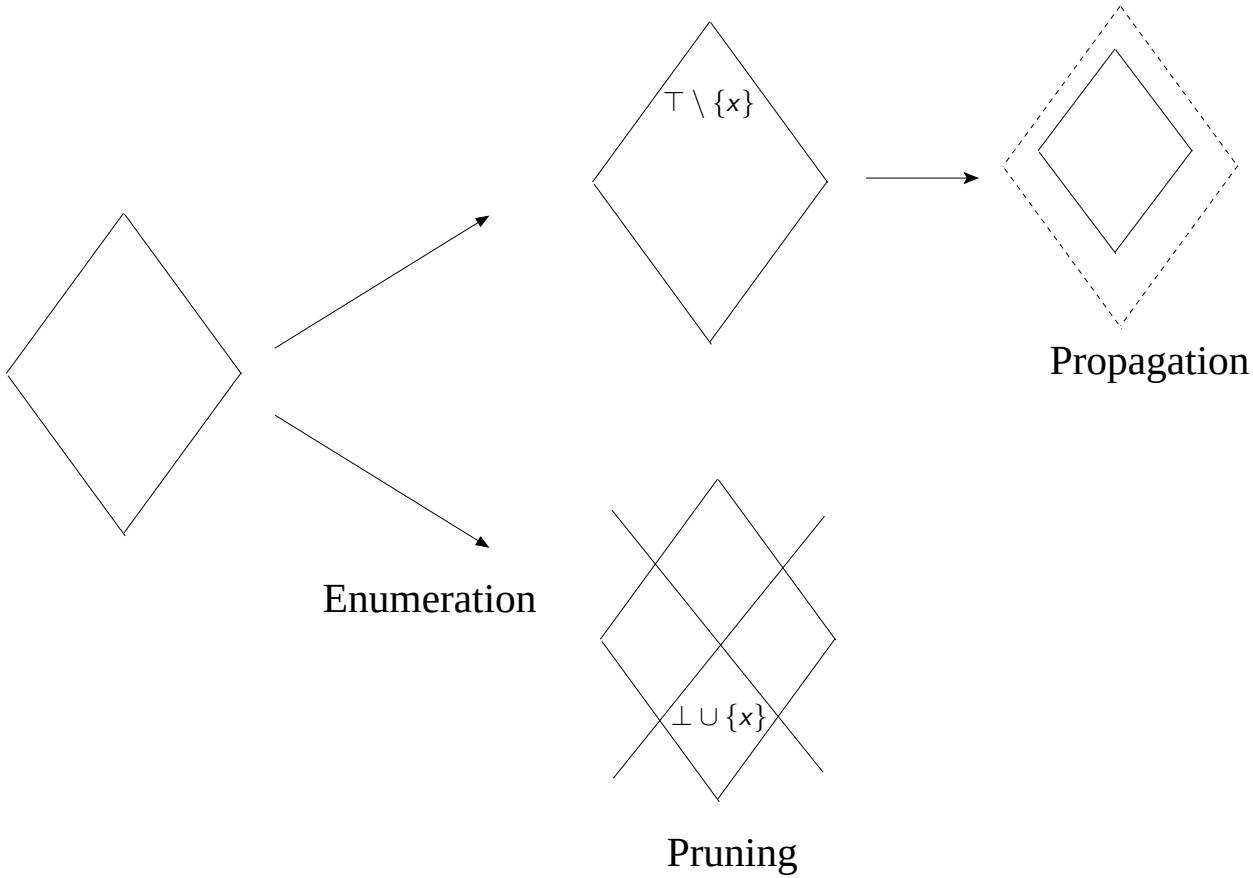
A sublattice $((\bot_G, \bot_M), (\top_G, \top_M))$



$$UB_{\mathcal{C}_r}((\bot_G, \bot_M), (\top_G, \top_M)) \equiv$$
$$\forall s \in G \setminus \top_G, \forall t \in \bot_G, \mathcal{Z}_o(s, \top_M) \geq \mathcal{Z}_o(t, \bot_M) + \delta \text{ and}$$
$$\forall s \in M \setminus \top_M, \forall t \in \bot_M, \mathcal{Z}_a(s, \top_G) \geq \mathcal{Z}_a(t, \bot_G) + \delta$$

$$UB_{\mathcal{C}_d}((\bot_G, \bot_M)(\top_G, \top_M)) \equiv$$
$$(\forall x \in \bot_G,\ \mathcal{Z}_o(x, \bot_M) \leq \alpha) \text{ and } (\forall y \in \bot_M\ \mathcal{Z}_a(y, \bot_G) \leq \alpha)$$

DR-Miner

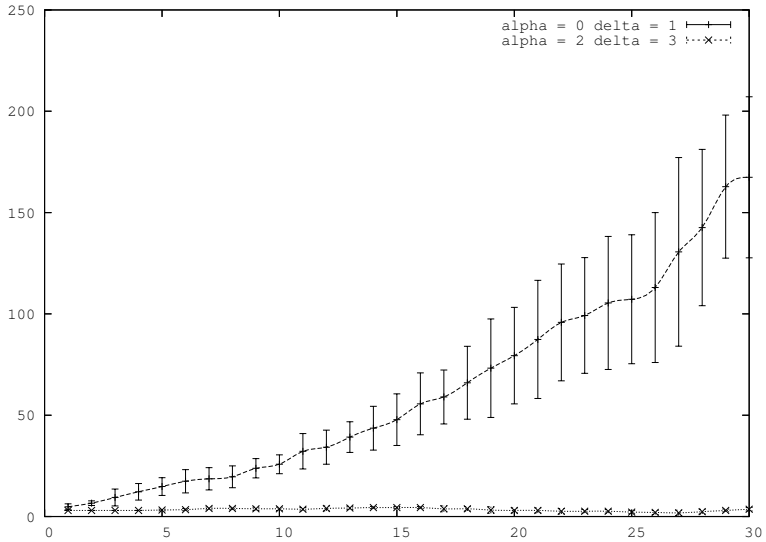


DR-Miner

- Pruning: if $UB_{\mathcal{C}_r}(\bot, \top)$ or $UB_{\mathcal{C}_d}(\bot, \top)$ are not satisfied, the sublattice $(\bot, \top)$ is pruned
- Propagation ($x \in \top \setminus \bot$):
  - if $UB_{\mathcal{C}_r}(\bot, \top \setminus \{x\})$ is not satisfied then the sublattice is modified in $(\bot \cup \{x\}, \top)$
  - if $UB_{\mathcal{C}_d}(\bot \cup \{x\}, \top)$ is not satisfied then the sublattice is modified in $(\bot, \top \setminus \{x\})$
- Enumeration: we choose $x \in \top \setminus \bot$ with the most 0 values on $\top$ to generate two sublattices
  - $(\bot \cup \{x\}, \top)$
  - $(\bot, \top \setminus \{x\})$

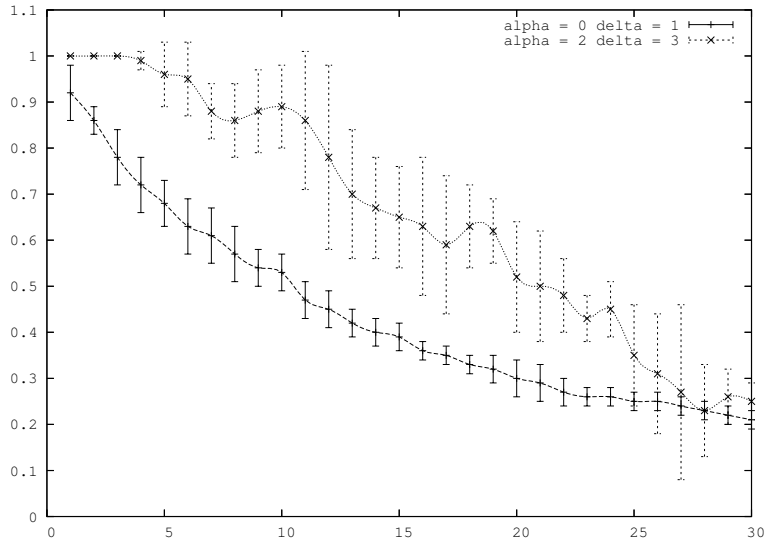
Robustness on synthetic data

Collection size



% noise

σ

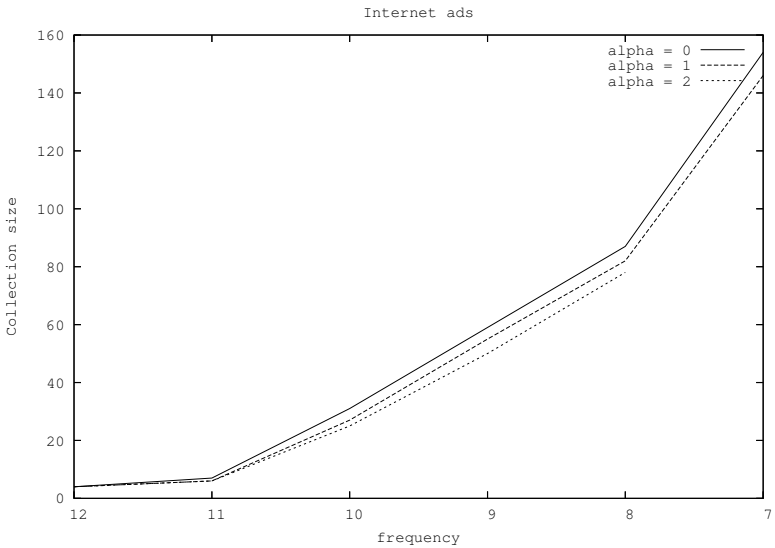


% noise

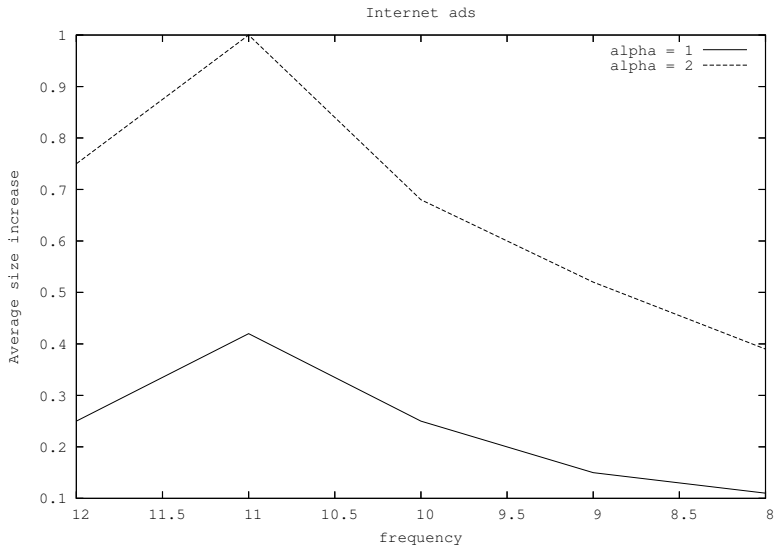
Mean and standard deviation of the number of bi-sets (5 trials) (left) and of σ (right) w.r.t. the percentage of noise.

Impact of parameter alpha

Collection size



Average size increase w.r.t. formal concepts



frequency

Impact of parameter delta

Mushroom ($\mathcal{C}_{ms}(\mathbf{r}, 500, 10)$, $\alpha = 0$)							
$\delta = \delta'$	Concepts	1	2	3	4	5	6
size	1 102	1 102	11	6	2	1	0
time	1.6s	10s	4s	4s	3s	2s	2s
Meningitis ($\mathcal{C}_{ms}(\mathbf{r}, 10, 5)$, $\alpha = 1$, $\delta = 1$)							
δ'	Concepts	1	2	3	4	5	6
size	354 366	-	75 376	22 882	8 810	4 164	2 021
time	5s	-	693s	327s	181s	109s	70s

DR-bi-set collection sizes and extraction time when δ' is varying from 1 to 6 on Mushroom and Meningitis.

Conclusion

- Defining fault-tolerant pattern types is important to support many Knowledge Discovery processes.
- Declarative specification + complete solvers $\Rightarrow$ improvement to support interpretation

Conclusion

FCA extension to fault-tolerance

+ A declarative specification with good properties on both sets
- None explicit definition of the functions between 2^G and 2^M

Algorithm

+ A generic algorithm to extract bi-sets under monotonic constraints
- Not so efficient when constraints are not monotonic

Knowledge Discovery

+ Extraction of useful patterns in real-life data