Mining bi-sets in numerical data

Jérémy Besson, Céline Robardet, Luc De Raedt and Jean-François Boulicaut

Institut National des Sciences Appliquées de Lyon - France
Albert-Ludwigs-Universitat Freiburg - Germany

Outline

Mining numerical data

Example: Gene expression data analysis

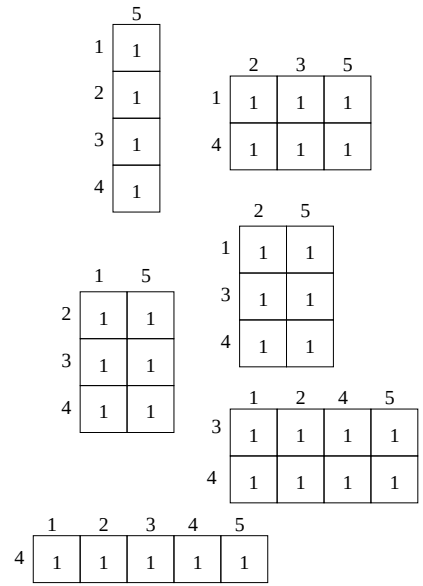
What are the sets of genes that are simultaneously over expressed in some biological situations?

1	2	2	1	6
2	1	1	0	6
2	2	1	7	6
8	9	2	6	7

$\Rightarrow$

0	1	1	0	1
1	0	0	0	1
1	1	0	1	1
1	1	1	1	1

$\Rightarrow$



Principle

$\mathcal{O}$ \ $\mathcal{P}$				
1	2	2	1	6
2	1	1	0	6
2	2	1	7	6
8	9	2	6	7

$\mathcal{M}(i,j)$ denotes the value of property $j \in \mathcal{P}$ for the object $i \in \mathcal{O}$

NBS defines a sub-matrix $\mathcal{S}$ of $\mathcal{M}$ s. t. the absolute value of the difference between the maximum value and the minimum value on $\mathcal{S}$ is less or equal to ϵ. Furthermore, none object or property can be added to the bi-set without violating this constraint.

The formal definition

Definition (Numerical bi-sets)

Given a real value ϵ, (X, Y) is a NBS iff

$$X \subseteq \mathcal{O},\ Y \subseteq \mathcal{P}$$

$$|Max_{i \in X, j \in Y} \mathcal{M}(i,j) - Min_{i \in X, j \in Y} \mathcal{M}(i,j)| \leq \epsilon$$

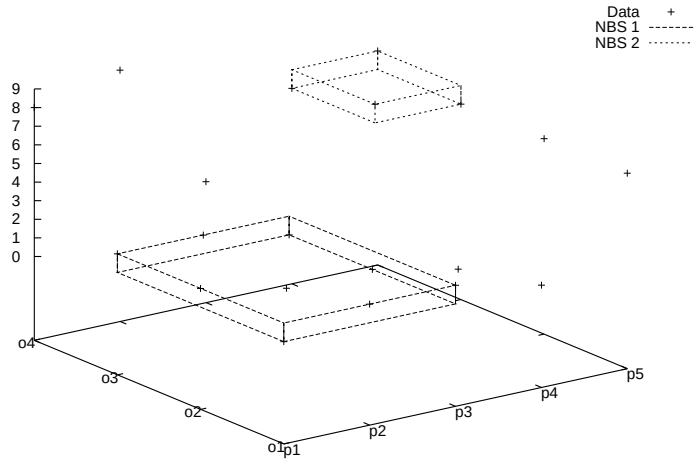
(1) $\forall y \notin Y,\ |Max_{i \in X, j \in Y \cup \{y\}} \mathcal{M}(i,j) - Min_{i \in X, j \in Y \cup \{y\}} \mathcal{M}(i,j)| > \epsilon$

(2) $\forall x \notin X,\ |Max_{i \in X \cup \{x\}, j \in Y} \mathcal{M}(i,j) - Min_{i \in X \cup \{x\}, j \in Y} \mathcal{M}(i,j)| > \epsilon$

An example

	p_1	p_2	p_3	p_4	p_5
o_1	1	2	2	1	6
o_2	2	1	1	0	6
o_3	2	2	1	7	6
o_4	8	9	2	6	7

$((o_1, o_2, o_3, o_4), (p_5))$
$((o_3, o_4), (p_4, p_5))$
$((o_4), (p_1, p_5))$
$((o_1, o_2, o_3, o_4), (p_3))$
$((o_4), (p_1, p_2))$
$((o_2), (p_2, p_3, p_4))$
$((o_1, o_2), (p_4))$
$((o_1), (p_1, p_2, p_3, p_4))$
$((o_1, o_2, o_3), (p_1, p_2, p_3))$



Properties

Definition (Specialization and monotonicity)

Our specialization relation on bi-sets denoted $\preceq$ is defined as follows: $(X_1, Y_1) \preceq (X_2, Y_2)$ iff $X_1 \subseteq X_2$ and $Y_1 \subseteq Y_2$.

The constraints are respectively anti-monotonic and monotonic w.r.t. $\preceq$

Properties

Let $\mathcal{W}_\epsilon$ be the whole collection of NBS patterns for ϵ.

- Each NBS pattern (X, Y) from $\mathcal{W}_\epsilon$ is maximal w.r.t. $\preceq$.
- If there exists a bi-set (X, Y) with similar values (belonging to an interval of size ϵ), then there exists a NBS (X', Y') from $\mathcal{W}_\epsilon$ such that $(X, Y) \preceq (X', Y')$

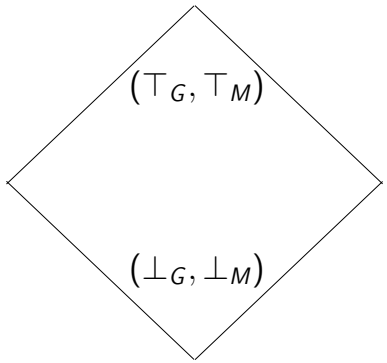
Properties

- When ϵ increases, the size of NBS pattern increases too, whereas some new NBS patterns which are not extensions of previous one can appear.
- The collection of numerical bi-sets is paving the dataset.

DR-Miner

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

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



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

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

NBS-Miner

$\mathcal{M}$ is a real valued matrix, $\mathcal{C}$ a conjunction of monotonic and anti-monotonic constraints on $2^{\mathcal{O}} \times 2^{\mathcal{P}}$ and ϵ is a positive value.

NBS-Miner

Generate$((\emptyset, \emptyset), (\mathcal{O}, \mathcal{P}))$

End NBS-Miner
Generate$(\mathcal{L})$

Let $\mathcal{L} = ((\bot_O, \bot_P), (\top_O, \top_P))$

$\mathcal{L} \leftarrow \mathit{Prop}(\mathcal{L})$

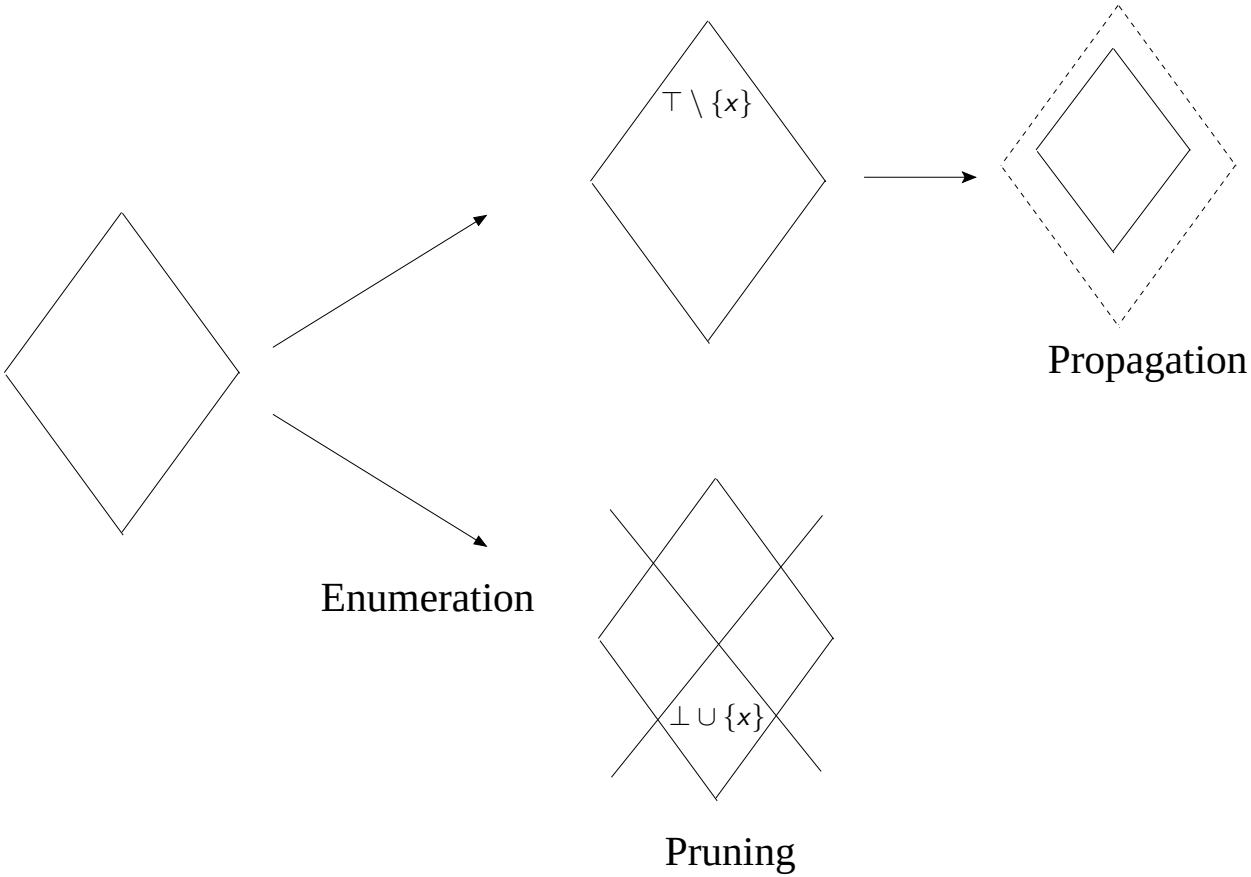
If $\mathit{Prune}(\mathcal{L})$ then

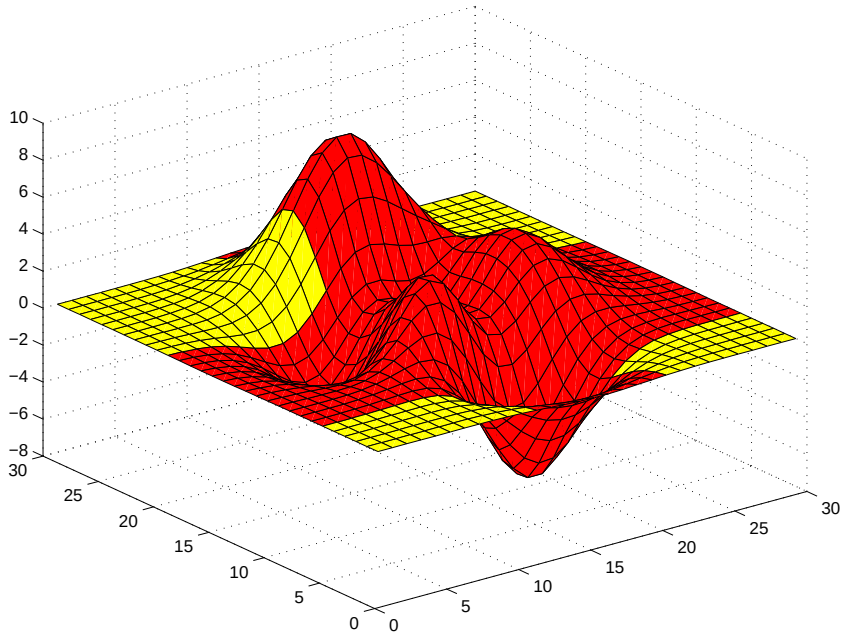
If $(\bot_O, \bot_P) \neq (\top_O, \top_P)$ then
$(\mathcal{L}_1, \mathcal{L}_2) \leftarrow \mathit{Enum}(\mathcal{L}, \mathit{Choose}(\mathcal{L}))$

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 $x \in \top \setminus \bot$
  - $(\bot \cup \{x\}, \top)$
  - $(\bot, \top \setminus \{x\})$

DR-Miner





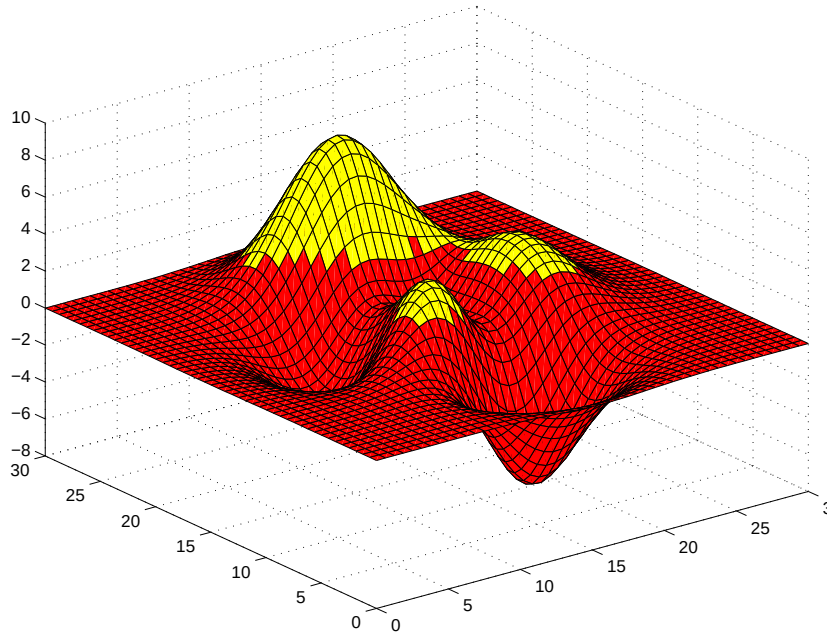
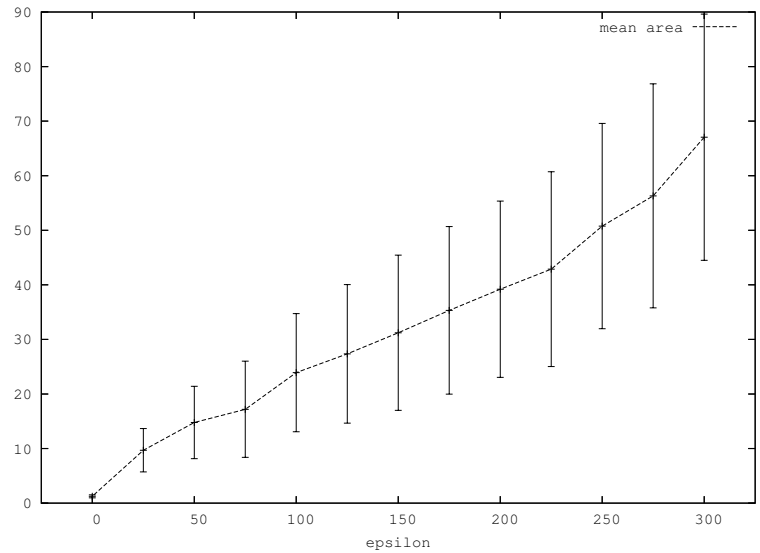


Figure: Examples of extracted NBS



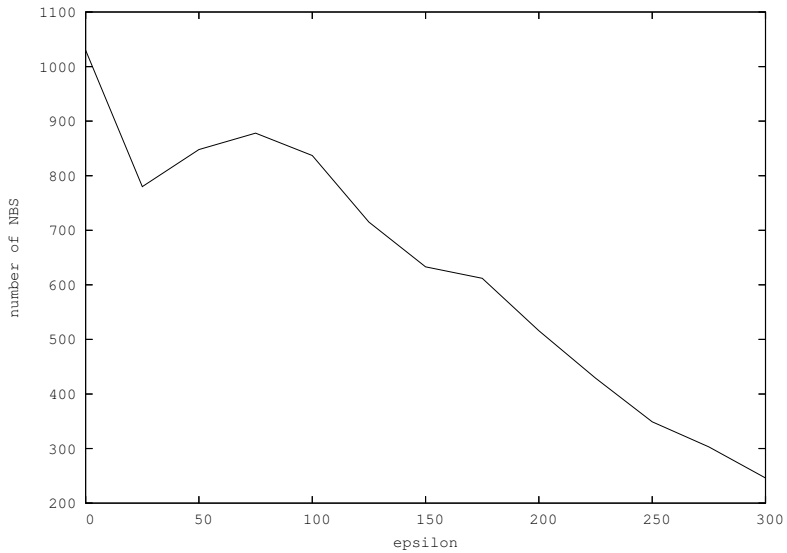


Figure: Mean area of the NBS w.r.t. ϵ

Conclusion

qsdqs