Palacky University, Olomouc, Czech Republic
Blaise-Pascal University, Clermont-Ferrand, French
Tunis-El-Manar University, Tunis, Tunisia

Boolean factors as a means of clustering of interestingness measures of association rules

Authors: *Radim Belohlavek, Dhouha Grissa, Sylvie Guillaume, Engelbert Mephu Nguifo, Jan Outrata*

October 18, 2011

Presentation Outline

1. Problem
2. Properties evaluation on the measures
3. Clustering
4. Interpretation and comparison to other approaches
5. Conclusion and Perspectives

I- Problem

Objectives of associations analysis

Unsupervised learning technique, which allows you to :

- Identify patterns or associations between items or objects in a transactional, relational databases, or data warehouses.
- In other words, it consists in identifying items that appear often together at an event.

Association rules

The extraction of association rules $X \rightarrow Y$

- $X \cap Y = \emptyset$
- X, Y are conjunctions of binary variables.

$$\textbf{Valid rules} \begin{cases} Support(X \rightarrow Y) \geqslant min_{sup} & (frequency) \\ Confidence(X \rightarrow Y) \geqslant min_{conf} & (strength) \end{cases}$$

Association rules

The extraction of association rules $X \rightarrow Y$

- $X \cap Y = \emptyset$
- X, Y are conjunctions of binary variables.

$$\textbf{Valid rules} \begin{cases} Support(X \rightarrow Y) \geqslant min_{sup} & (frequency) \\ Confidence(X \rightarrow Y) \geqslant min_{conf} & (strength) \end{cases}$$

Advantage : Accelerator algorithmic virtues
Inconvenient : Irrelevant rules.

Interestingness measures

Irrelevant rules

↓

Additional step of analyzing the extracted rules

- The proposition of many objective interestingness measures
- About sixty measures.

Interestingness measures

Irrelevant rules

↓

Additional step of analyzing the extracted rules

- The proposition of many objective interestingness measures
- About sixty measures.

Which measure to choose ?

Which measure to choose ?

- Study of the "*good*" properties of measures
- 21 properties

Assist the user in choosing complementary measures

(*elimination of uninteresting rules*)

Assist the user in choosing complementary measures

Detection of groups of measures

- Interestingness measures clustering (*Tan et al. 2004, Huynh et al. 2005, Vaillant 2007, Guillaume et al. 2011*)
- Interestingness measures clustering using *Boolean Factor Analysis*.

Goal

The aim of this work is :

- To help the user to choose the best measure by exploring the possibility of obtaining overlapping clusters of measures using Boolean factor analysis
- To compare the results with those obtained by the *AHC* and *k-means* methods (*Guillaume et al. 2011*).

II- Background : *Properties evaluation on the measures*

Measures properties

- 21 properties are listed in the literature
- 2 properties found subjective

  (*based on the user knowledge in Statistics*)

  1. Measure comprehensibility
  2. Easiness to fix a threshold

19 properties retained

19 properties

- Non symmetrical
- Fixed values for different levels of implication
- Measure evolution based on parameters
- Relations between positive and negative rules
- Discrimination in the presence of large data

19 properties

- Non symmetrical
- Fixed values for different levels of implication
- Measures evolution based on parameters
- Relations between positive and negative rules
- Discrimination in the presence of large data

Non symmetrical

$$\mathrm{m}(X \rightarrow Y) \neq \mathrm{m}(Y \rightarrow X)$$

$$\mathrm{m}(X \rightarrow Y) \neq \mathrm{m}(X \rightarrow \overline{Y})$$

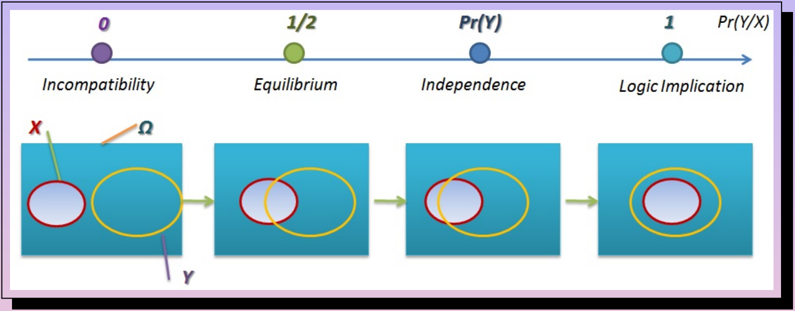
Yes : 1

No : 0

Exemple

$$\begin{aligned} Support(X \rightarrow Y) &= Support(Y \rightarrow X) \Rightarrow P(XY) = P(YX) \\ Confidence(X \rightarrow Y) &\neq Confiance(Y \rightarrow X) \Rightarrow P(Y/X) \neq P(X/Y) \end{aligned}$$

Fixed values for different levels of implication

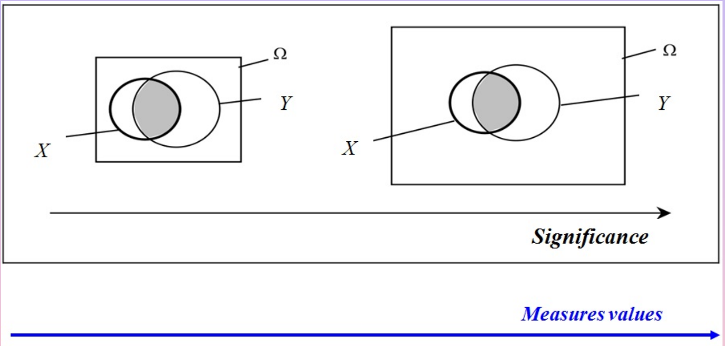


$$P_{10}(m) = 0 \;\; if \; \forall\, b \in \mathcal{R} \; \exists\, X \rightarrow Y / P(Y/X) = 1 \; and \;\; m(X \rightarrow Y) \neq b$$

$$P_{10}(m) = 1 \;\; if \; \forall\, b \in \mathcal{R} \; / \; \forall\, X \rightarrow Y \; P(Y/X) = 1 \;\Rightarrow\; m(X \rightarrow Y) = b$$

Yes : 1 / No : 0

Evolution of measures based on parameters



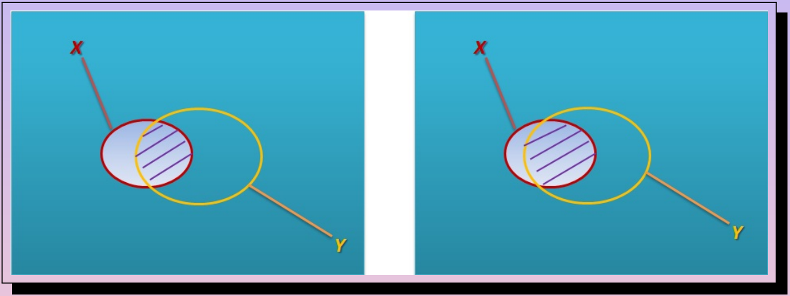
Relations between positive and negative rules

$$m(\overline{X} \rightarrow Y) = -m(X \rightarrow Y)$$
$$m(X \rightarrow \overline{Y}) = -m(X \rightarrow Y)$$
$$m(\overline{X} \rightarrow \overline{Y}) = m(X \rightarrow Y)$$

Yes : 1

No : 0

Discrimination in the presence of large data



Measures returning different values for distinct levels of implication

Properties examples

19 properties

- Non symmetrical
- Fixed values for different levels of implication
- Measure evolution based on parameters
- Relations between positive and negative rules
- Discrimination in the presence of large data

Properties evaluation on the measures

Abstract Interestingness measures

Study of 62 interestingness measures !

Measure	Formula
Cohen	$2\,\frac{p(XY)-p(X)p(Y)}{p(X)p(\bar{Y})+p(\bar{X})p(Y)}$
Causal confidence	$1-\frac{1}{2}\left(\frac{1}{p(X)}+\frac{1}{p(\bar{Y})}\right)p(X\bar{Y})$
Bayes factor	$\frac{p(XY)p(\bar{Y})}{p(X\bar{Y})p(Y)}$
Implication intensity	$p[Poisson(np(X)p(\bar{Y}))\geq p(X\bar{Y})]$
Loevinger	$1-\frac{p(X\bar{Y})}{p(X)p(\bar{Y})}$
Ochiai	$\frac{p(XY)}{\sqrt{p(X)p(Y)}}$
Pearl	$p(X)\lvert\frac{p(XY)}{p(X)}-p(Y)\rvert$
Y Yule	$\frac{\sqrt{p(XY)p(\bar{X}\bar{Y})}-\sqrt{p(X\bar{Y})p(\bar{X}Y)}}{\sqrt{p(XY)p(\bar{X}\bar{Y})}+\sqrt{p(X\bar{Y})p(\bar{X}Y)}}$

Study of 62 interestingness measures $\times$ 19 properties

$\Downarrow$

Matrix construction !

Measure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Evaluation measures example

Measure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Evaluation measures example

Measure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Non symmetrical measures.

Evaluation measures example

Measure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Measures decreasing according to the consequent size.

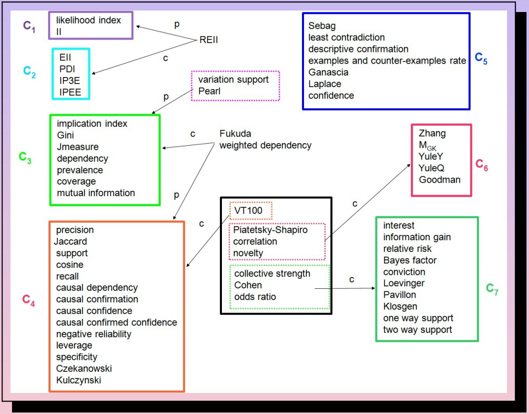
III- Clustering

Clustering of interestingness measures.

1. Interestingness measures clustering using *AHC* and *k-means* methods
2. Interestingness measures clustering using *Boolean Factor Analysis*.

Clustering of IMs using AHC and k-means methods.

- *consensus for 7 clusters*
- *Divergence for 12 measures*



Clustering of IMs using Boolean factor analysis.

Boolean Factor Analysis (BFA) = decomposition of binary object-attribute data matrix I to Boolean product of object-factor matrix A and factor-attribute matrix B :

$$I_{ij} = (A \circ B)_{ij} = \max_{l=1}^{k} \min(A_{il}, B_{lj})$$

$A_{il} = 1$ ... factor l applies to object i
$B_{lj} = 1$ ... attribute j is one of the manifestations of factor l

$(A \circ B)_{ij}$ ... "object i has attribute j if and only if there is a factor l such that l applies to i and j is one of the manifestations of l"

PROBLEM : find the number k of factors as small as possible !

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} = \overbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}^{k} \circ \left.\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}\right\} k$$

Boolean factor analysis – Solution using FCA

Belohlavek R., Vychodil V. : Discovery of optimal factors in binary data via a novel method of matrix decomposition. *J. Comput. System Sci* **76**(1)(2010), 3–20.

Matrices A and B can be constructed from a set $\mathcal{F}$ of formal concepts of input data I, so-called **factor concepts** :

$$\mathcal{F} = \{\langle A_1, B_1\rangle, \ldots, \langle A_k, B_k\rangle\} \subseteq \mathcal{B}(X, Y, I)$$

- l-th column of $A_{\mathcal{F}}$ = characteristic vector of A_l
- l-th row of $B_{\mathcal{F}}$ = characteristic vector of B_l

Decomposition using formal concepts to determine factors is optimal :

Theorem

For every $n \times m$ binary matrix I, there exists $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ such that $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$ and $|\mathcal{F}| = \rho(I)$, where $A_{\mathcal{F}}$ and $B_{\mathcal{F}}$ are $n \times k$ and $k \times m$ binary matrices, $\circ$ is the Boolean product of matrices and ρ is the smallest possible number k of factors (so-called Schein rank of I).

Method

- We extended the original 62×21 measure-property matrix by adding for every property its negation, and obtained a 62×42 measure-property matrix.
- We computed the decomposition of the matrix using a greedy approximation algorithm (from the mentioned paper) and obtained 38 factors, denoted $F_1, \ldots, F_{38}$.
- We took the discovered factors for clusters and looked for the interpretation of the clusters.

IMs clustering using boolean factor analysis

I : 62 *measures x* 42 *properties* input binary matrix (with negated properties) =

	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14.1	P15	P16	P17
correlation	0	1	1	1	1	1	1	0	0	1	1	0	0	1	1
Cohen	0	1	1	1	1	1	1	0	0	1	1	0	0	0	0
confidence	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0
causal confidence	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
Pavillon	1	1	0	1	1	1	1	0	0	1	1	0	0	0	1
Ganascia	1	1	1	1	0	0	0	1	1	0	0	0	0	0	1
causal confirmation	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
descriptive confirmation	1	1	0	1	0	0	0	0	1	0	0	0	0	0	1
conviction	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0
cosine	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
coverage	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A_F : 62 *measures x* 38 *factors* binary matrix

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24
correlation	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
Cohen	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0
confidence	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
causal confidence	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
Pavillon	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
Ganascia	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
causal confirmation	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
descriptive confirmation	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
conviction	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
cosine	0	1	0	0	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0
coverage	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1

B_F : 38 *factors x* 42 *properties* binary matrix

	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14.1	P15	P16	P17	P18	P19	P20	P21	P14.2	P14.3	$\overline{P3}$	$\overline{P4}$	$\overline{P5}$	$\overline{P6}$
F1	0	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
F3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
F4	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F5	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
F6	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
F7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
F8	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
F9	0	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	1	0	0	0
F10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
F11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

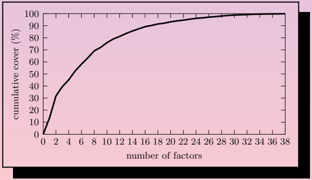
IV- Interpretation and comparison to other approaches

Interpretation of results

We computed the decomposition of the matrix I and obtained 38 factors :

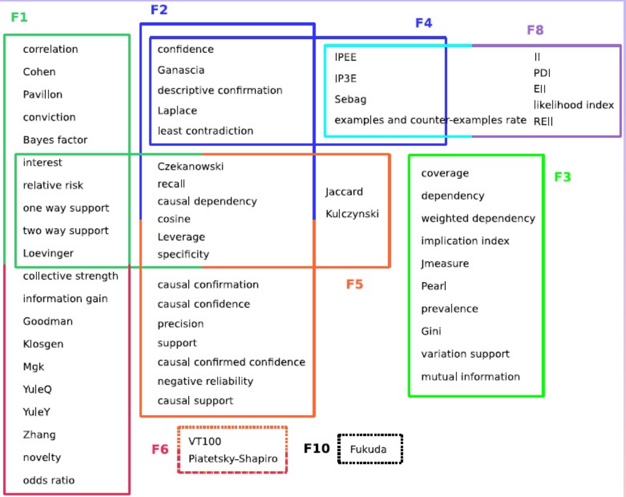
- The first 21 factors cover 94% of the input measure-property matrix.
- The first nine cover 72%.
- The first five cover 52.4%.
- The first ten cover all measures.

Cumulative cover of input matrix



Results Interpretation

Venn Diagram of Boolean Factors



The interpretation of the first 4 factors, which cover nearly half of the matrix (45.1%), shows :

- A high *similarity* with other clusters of measures reported in the literatture.
- A clearly interpretable meaningful overlapping clusters of measures.

Interpretation : Factor 1

The interpretation of the first factor F_1, reveals :

- F_1 applies to 20 measures whose evolutionary curve increases w.r.t. the number of examples and have a fixed point in the case of independence.
- These measures share 9 properties.
- F_1 applies only to descriptive and discriminant measures that are not based on a probabilistic model.

Comparison to other approaches : Factor 1

The comparison of the first factor F_1 with the classification results shows :

- F_1 applies to two classes, C_6 and C_7, which are closely related within the dendogram obtained with the *agglomerative hierarchical clustering* method (Guillaume et al. 2011).
- $C_6 \cup C_7$ contains 15 measures.
- The 5 missing measure (in the Venn diagram of Boolean factors) form a class obtained with *K-means* method with *Euclidian* distance.

AHC : The dendogram

Interpretation : Factor 2

The interpretation of the second factor F_2, reveals :

- F_2 applies to 18 measures, whose evolutionary curve increases w.r.t. the number of examples and have a variable point in the case of independence.
- These measures share 11 properties.
- F_2 applies only to measures that are not discriminant, are indifferent to the first counter-examples, and are not based on a probabilistic model.

Comparison to other approaches : Factor 2

The comparison of the second factor F_2 with the classification results shows :

- F_2 applies to two classes, C_4 and C_5, which are also closely related within the dendogram obtained with the *agglomerative hierarchical clustering* method.
- $C_4 \cup C_5$ contains 22 measures.
- The 4 missing measure (in the Venn diagram of Boolean factors) which not covered by F_2 since they are not indifferent to the first counter-examples.

V- Conclusion and Perspectives

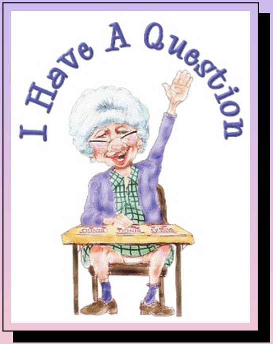
Conclusion and Perspectives

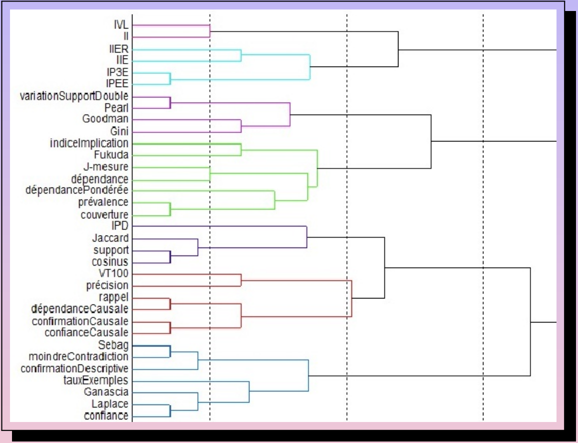
- The preliminary results on clustering the measures using Boolean factors seem promising.
- A user can benefit of the clustering of measures in using a type of measure and measures that belong to different classes of measures.

Perspectives :

- The method need not start from scratch – an interesting feature that can be explored in the future.

Thank you for your attention !




IVL
II
IIER
IIE
IP3E
IPEE
variationSupportDouble
Pearl
Goodman
Gini
indiceImplication
Fukuda
J-mesure
dépendance
dépendancePondérée
prévalence
couverture
IPD
Jaccard
support
cosinus
VT100
précision
rappel
dépendanceCausale
confirmationCausale
confianceCausale
Sebag
moindreContradiction
confirmationDescriptive
tauxExemples
Ganascia
Laplace
confiance

Return