Clustering Constrained Symbolic Data

Marc Csernel – Francisco de A.T. de Carvalho

INRIA – **CIn/UFPE**

BEIJING

2011 October 27th

Outline

SDA considers two kinds of rules

Hierarchicals rules

$$y_1 \in \mathcal{P}^*(D_1) \Longrightarrow y_2 = NA$$

where $D_1 \subset \mathcal{D}_1$ and NA means not applicable .
We speak sometimes of mother-daughter variables.

They induce FALSE Missing Data

$$Hand \in \{absent\} \Longrightarrow Hand\_Color = N.A.$$
$$Hand \in \{absent\} \Longrightarrow Finger = N.A.$$

Logical dependences

$$y_1 \in \mathcal{P}^*(D_1) \Longrightarrow y_2 \in \mathcal{P}^*(D_2).$$

$$Color \in \{Blue, Red\} \Longrightarrow Size \in \{Small, Very\_Small\}$$

But very few method are using them

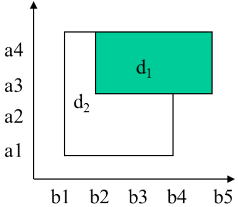
Rules induces HOLES in the description space.

If we have the two symbolic description :

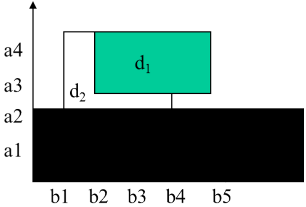
$$d_1 = [a \in \{a1, a2, a3, a4\}] \wedge [b \in \{b1, b2, b3, b4\}]$$
$$d_2 = [a \in \{a3, a4\}] \wedge [b \in \{b2, b3, b4, b5\}]$$

and the rule :

$$\text{if} = [a \in \{a3, a4\} \text{ then b = N.A.}$$



without rule



in presence of the rule

d_1 and d_2 seems more similar in the presence of the rule.

Influence of the rules on discrimination

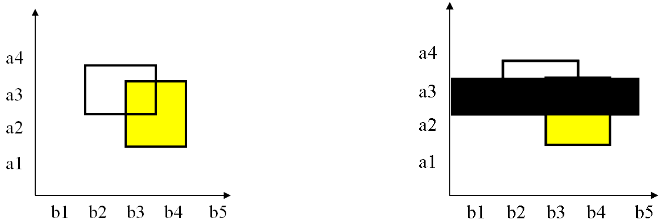
if we have the two symbolic descriptions :

$d_1 = [a \in \{a3, a4\}] \wedge [b \in \{b2, b3\}]$

$d_2 = [a \in \{a2, a3\}] \wedge [b \in \{b1, b2, b4\}]$

and the rule :

$\text{if} = [a \in \{a3\}$ then $b \in \{b1, b2, b4\}$



without rule — in presence of rules

d_1 and d_2 can be discriminated in presence of a rule.

N.A. propagation

$$\left.\begin{array}{r} if[Hand \in \{Absent\}] \Rightarrow [Finger = N.A.] \\ if[Finger \in \{Absent\}] \Rightarrow [Finger\_size = N.A.] \end{array}\right\} \Rightarrow$$

if Hand [$\in \{Absent\}$] $\Longrightarrow$ [Finger_Size = N.A.].

The dependency graph induced by the rules

With the three following rules

if [Hand ∈ {Absent}] ⇒ [Hand_size = N.A.]
if [Hand ∈ {Absent}] ⇒ [Finger = N.A.]
if [Finger ∈ {Absent}] ⇒ [Finger_Size = NA.]

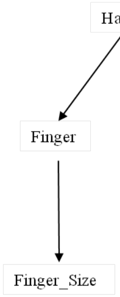


Figure: Dependency tree between variables

Notion of coherence

- An individual is *coherent* if it description respects the rules ;
- The *coherent part* of a symbolic description S_1 is the part of the virtual extension $V_{ext}(S_1)$ where the rules are respected.
- A symbolic description is *coherent* if it has a *nonempty coherent part*.
- A symbolic description S_1 is *fully coherent* if all $V_{ext}(S_1)$ is coherent.
- A symbolic description S_1 is *incoherent* if no part $V_{ext}(S_1)$ is coherent.

Notion of coherence

description	Wings	Wing_color
d_1	{Absent}	{Blue, Red, Yellow}
d_2	{Absent, Present}	{Blue, Red, Yellow}
d_3	{Present}	{Blue, Red, Yellow}
d_4	{Absent}	{N.A.}

if [Wings $\in \{Absent\}$] $\Longrightarrow$ [Wing_color = N.A.] (r_1)

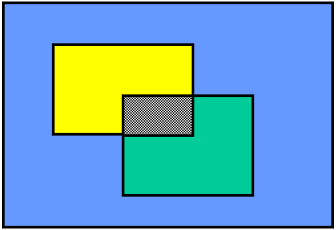
d_1 is **not coherent**,
d_3 is **fully coherent**.

d_2 is **coherent**
d_4 is **fully coherent**

Description potential

The coherent part of the hyper-volume described by a symbolic description.
denoted $\pi(d_1)$

- Frequently use to compute distance between symbolic description.
- Combinatorial computation in presence of rule.
- In the following we will try to avoid this overhead.



exemple of computation of description Potential

d_1 = a1,a2b1,b2c1,c2d1,d2
$\pi(d_1)$ = 2x2x2x2 = 16 (Without rule)

With the rules

if $a \in \{a1\}$ then $b \in \{b1\}$ (r1)
if $c \in \{c1\}$ then $d \in \{d1\}$ (r2)

a1	b1	c1	d1	Y
a1	b1	c1	d2	N(r2)
a1	b1	c2	d1	Y
a1	b1	c2	d2	Y
a1	b2	c1	d1	N(r1)
a1	b2	c1	d2	N(r1,r2)
a1	b2	c2	d2	N(r1)
a1	b2	c2	d2	N(r1)

a2	b1	c1	d1	Y
a2	b1	c1	d2	N(r2)
a2	b1	c2	d1	Y
a2	b1	c2	d2	Y
a2	b2	c1	d1	Y
a2	b2	c1	d2	N(r2)
a2	b2	c2	d1	Y
a2	b2	c2	d2	Y

Formulae of Description Potential

Without variable dependencies :

$$\pi(d) = \prod_{i=1}^{p} \mu(D_i)$$

with

$$\mu(D_i) \begin{cases} Card(D_i) & if \quad D_i \text{ is discrete} \\ Span(D_i) & if \quad D_i \text{ is continuous} \end{cases}$$

Formulae of Description Potential

With variable dependencies :

$$\pi(d/r_1 \wedge \cdots \wedge r_t) = \prod_{i=1}^{p} \mu(D_i) - \sum_{j=1}^{t} \pi(a \wedge \neg r_j)$$
$$+ \sum_{j<k} \pi((d \wedge \neg r_j) \wedge \neg r_k) + \cdots +$$
$$(-1)^{t+1} \pi((d \wedge \neg r_1) \wedge \neg r_2) \wedge \cdots \wedge \neg r_t)$$

We become combinatorial

$$\varphi(a,b) = \frac{\pi(a \oplus b) - \pi(a)}{2} + \frac{\pi(a \oplus b) - \pi(b)}{2}$$

where $a \oplus b$ is the join operator (Ichino and Yaguchi (1994)).

$a \oplus b = \bigwedge_{i=1}^{p} [y_i \in A_i \oplus B_i]$, where :

- $A_i \oplus B_i = A_i \cup B_i$ for set-valued variables.
- $A_i \oplus B_i = [\min(low(A_i), low(B_i)), \max(up(A_i), up(B_i))]$ for interval variables.

Normal Symbolic Form (The Idea)

We want to represent only the fully coherent part of a symbolic description.

- According to this goal we will cut the description space into several subspaces.
- Each of these subspaces will correspond to a premise variable and all related conclusion variables (according to the dependency tree) .
- each subspaces will be cut into slices. For each slice all the values of the premise variable will lead to the same conclusion.

Normal Symbolic Form

if [Hand ∈ {Absent}] ⇒ [Hand_size = N.A.]
if [Hand ∈ {Absent}] ⇒ [Finger = N.A.]

if [Finger ∈ {Absent}] ⇒ [Finger_Size = NA.]

	Hand	Finger	Finger_Size	Thorax_color
d_1	{absent,present}	{absent,present}	{small,big}	{red,blue}
d_2	{absent,present}	{absent,present}	{medium}	{red,green}

Original table.

	Hand_T	Thorax_color
d_1	{1,2}	{red,blue}
d_2	{1,3}	{red,green}

Main Table

Hand_T	Hand	Finger_T
1	{absent}	N.A.
2	{present}	{1,2}
3	{present}	{1,3}
4	{present}	{1,4}

Hand_T table

Finger_T	Finger	Finger_Size
1	{absent}	N.A.
2	{present}	{ big, small }
3	{present}	{medium}
4	{present}	{small,medium,big}

Finger_T table

Tables decomposed according to the N.S.F

N.F.S. consequences

N.F.S has two consequences :

- We need to cut the data in different tables following the dependence tree. It is only possible if the dependencies between the variables form a tree or a forest
- We need to cut eache symbolic description into two parts :
  - The part where the premise is true
  - The part where the premise is false

Computing Description Potential 0

	Hand_T	Thorax_color	pot
d_1	{1,2}	{red,blue}	
d_2	{1,3}	{red,green}	
$d_1 \oplus d_2$	{1,4}	{red,green,blue}	

Main Table

Hand_T	Hand	Finger_T	pot
1	{absent}	N.A.	
2	{present}	{1,2}	
3	{present}	{1,3}	
4	{present}	{1,4}	

Hand_T table

Finger_T	Finger	Finger_Size	pot
1	{absent}	N.A.	
2	{present}	{ big, small }	
3	{present}	{medium}	
4	{present}	{small,medium,big}	

Finger_T table

INRIA ROCQUENCOURT

UFPE

Computing Description Potential 1

	Hand_T	Thorax_color	pot
d_1	{1,2}	{red,blue}	
d_2	{1,3}	{red,green}	
$d_1 \oplus d_2$	{1,4}	{red,green,blue}	

Main Table

Hand_T	Hand	Finger_T	pot
1	{absent}	N.A.	
2	{present}	{1,2}	
3	{present}	{1,3}	
4	{present}	{1,4}	

Hand_T table

Finger_T	Finger	Finger_Size	pot
1	{absent}	N.A.	
2	{present}	{ big, small }	
3	{present}	{medium}	
4	{present}	{small,medium,big}	

Finger_T table

Computing Description Potential 2

	Hand_T	Thorax_color	pot
d_1	{1,2}	{red,blue}	
d_2	{1,3}	{red,green}	
$d_1 \oplus d_2$	{1,4}	{red,green,blue}	

Main Table

Hand_T	Hand	Finger_T	pot
1	{absent}	N.A.	
2	{present}	{1,2}	
3	{present}	{1,3}	
4	{present}	{1,4}	

Hand_T table

Finger_T	Finger	Finger_Size	pot
1	{absent}	N.A.	
2	{present}	{ big, small }	
3	{present}	{medium}	
4	{present}	{small,medium,big}	

Finger_T table

	Hand_T	Thorax_color	pot
d_1	{1,2}	{red,blue}	
d_2	{1,3}	{red,green}	
$d_1 \oplus d_2$	{1,4}	{red,green,blue}	

Main Table

Hand_T	Hand	Finger_T	pot
1	{absent}	N.A.	
2	{present}	{1,2}	
3	{present}	{1,3}	
4	{present}	{1,4}	

Hand_T table

Finger_T	Finger	Finger_Size	pot
1	{absent}	N.A.	1
2	{present}	{ big, small }	2
3	{present}	{medium}	1
4	{present}	{small,medium,big}	3

Finger_T table

Computing Description Potential 4

	Hand_T	Thorax_color	pot
d_1	{1,2}	{red,blue}	
d_2	{1,3}	{red,green}	
$d_1 \oplus d_2$	{1,4}	{red,green,blue}	

Main Table

Hand_T	Hand	Finger_T	pot
1	{absent}	N.A.	1
2	{present}	{1,2}	3
3	{present}	{1,3}	2
4	{present}	{1,4}	4

Hand_T table

Finger_T	Finger	Finger_Size	pot
1	{absent}	N.A.	1
2	{present}	{ big, small }	2
3	{present}	{medium}	1
4	{present}	{small,medium,big}	3

Finger_T table

Computing Description Potential 5

	Hand_T	Thorax_color	pot
d_1	{1,2}	{red,blue}	8
d_2	{1,3}	{red,green}	6
$d_1 \oplus d_2$	{1,4}	{red,green,blue}	15

Main Table

Hand_T	Hand	Finger_T	pot
1	{absent}	N.A.	1
2	{present}	{1,2}	3
3	{present}	{1,3}	2
4	{present}	{1,4}	4

Hand_T table

Finger_T	Finger	Finger_Size	pot
1	{absent}	N.A.	1
2	{present}	{ big, small }	2
3	{present}	{medium}	1
4	{present}	{small,medium,big}	3

Finger_T table

The dynamic clustering algorithm

We can use the Dynamic Clustering Algorithm applied to Dissimilarity tables (Lechevallier 1974).

- The prototype of each cluster is a member of the set of examples.
- The quality of each cluster is the sum of the dissimilarities of its items and its prototype
- The quality of the partition is the sum of the quality of each cluster
- The quality is the clustering criterion.
- The classification problem is to find a partition and a set of k prototypes that minimise the clustering criterion.

The limit of memory growing

We consider here set valued variables and hierarchical rules

- First locally between mother and daughter
- globally between the root of a dependence tree an one of the leaves.

- The local growing is always less than 2
- Lines containing an N.A. can not refer to other lines They do not induces further growing.
- The global growing is bounded by local growing

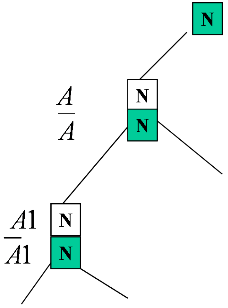


Figure: without rule

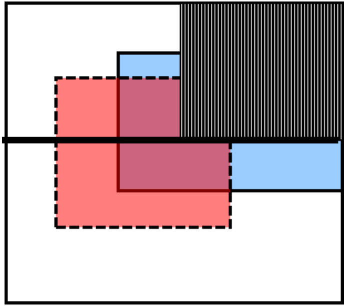


Figure: Two different parts

Operation with N.S.F. creating a new volume 2

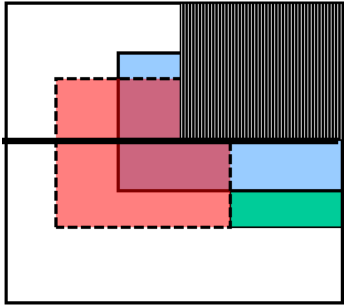


Figure: When the rule is false

Operation with N.S.F. creating a new volume 3

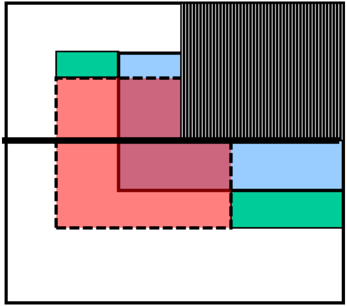


Figure: When the rule is false
Everything all right

Operation with N.S.F. creating a new volume 4

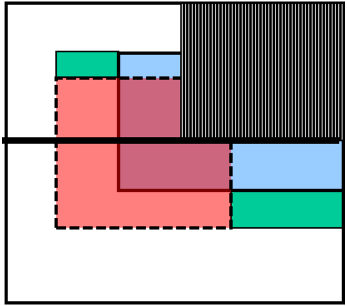


Figure: Everything all right

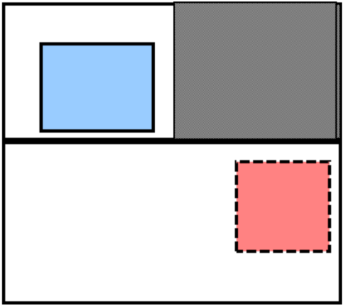


Figure: Each object in a separate Part

Operation with N.S.F. creating a new volume 5

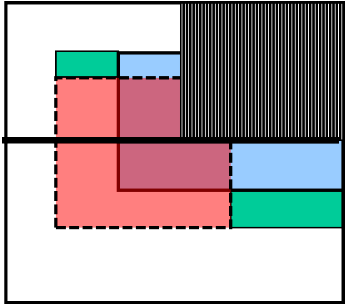


Figure: Everything all right

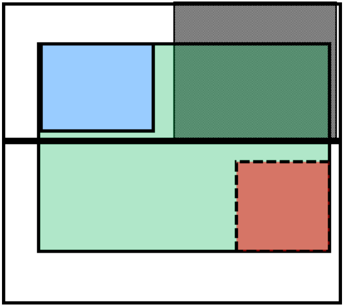


Figure: Problem

Operation with N.S.F. creating a new volume 6

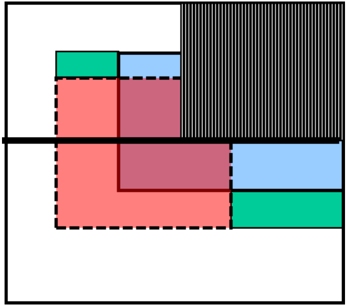


Figure: Everything all right

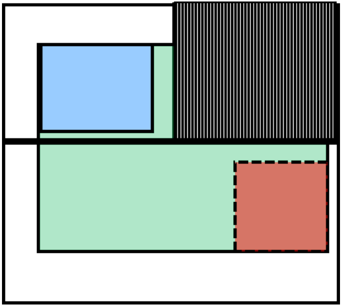


Figure: Problem
We must apply the rule again

Conclusion

We have presented a methods which allows clustering of symbolic description presence of rules in a polynomial time instead of a combinatorial one.
This methods allows to deal with "false missing values" .
We can apply the method to other classification problems.

Some Bibliography

De Carvalho F.A.T and Csernel M. and Lechevallier Y. 2009
Clustering constrained symbolic data, *Pattern Recognition Letters*, 30.11, 1037-1045

Csernel, Marc and De Carvalho, F.A.T. 1999.
Usual Operations with Symbolic Data under Normal Symbolic Form. *Applied Stochastic Models in Business and Industry*, 15, 241-257.

Brito, P., De Carvalho, F.A.T. 2002
"Symbolic Clustering of Constrained Probabilistic Data". In : *Exploratory Data Analysis in Empirical Research* , Opitz O. & Schwaiger M., Springer 12-21.

Manabu Ichino, Hiroyuki Yaguchi 1994 : *Generalized Minkowski metrics for mixed feature-type data analysis*. IEEE Transactions on Systems, Man, and Cybernetics 24(4) : 698-708

Csernel, M. and De Carvalho, F.A.T. 2002.
On memory requirement with Normal Symbolic Form, in : *Exploratory Data Analysis in Empirical Research* , Opitz O. & Schwaiger M., Springer 22-30.

De Carvalho, F. A. T. 1998.
Extension based proximities between constrained Boolean symbolic objects, in :*Data Science, Classification and Related Methods*, Hayashi, C. et al (Eds.) Springer, 370 - 378.

Thank you for your attention...