Extraction de classes homogènes et données symboliques

M^elle Aïcha EL GOLLI

Projet *AXIS* INRIA-Rocquencourt

Lise CEREMADE université Paris IX, Dauphine

aicha.elgolli@inria.fr

Mr Yves LECHEVALLIER

Projet *AXIS* INRIA-Rocquencourt

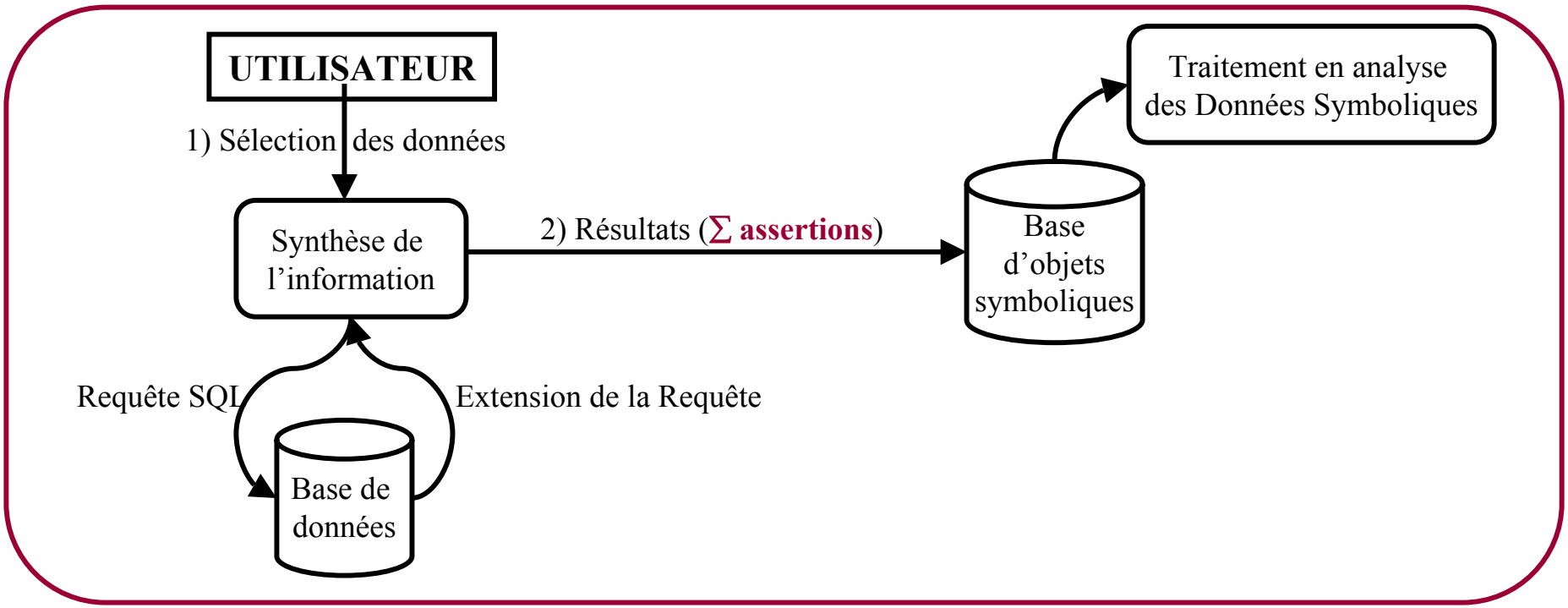
Yves.Lechevallier@inria.fr

INRIA

PLAN

- Outil de création des données symboliques par généralisation
- Problèmes de la généralisation existante
- Solution: Décomposition
- Conclusions et perspectives

DB2SO /ASSO



DB2SO (Véronique Stéphan) (Data Base To Symbolic Objects) réalise une généralisation à partir des groupes d'individus dont la description est issue de BD relationnelles.

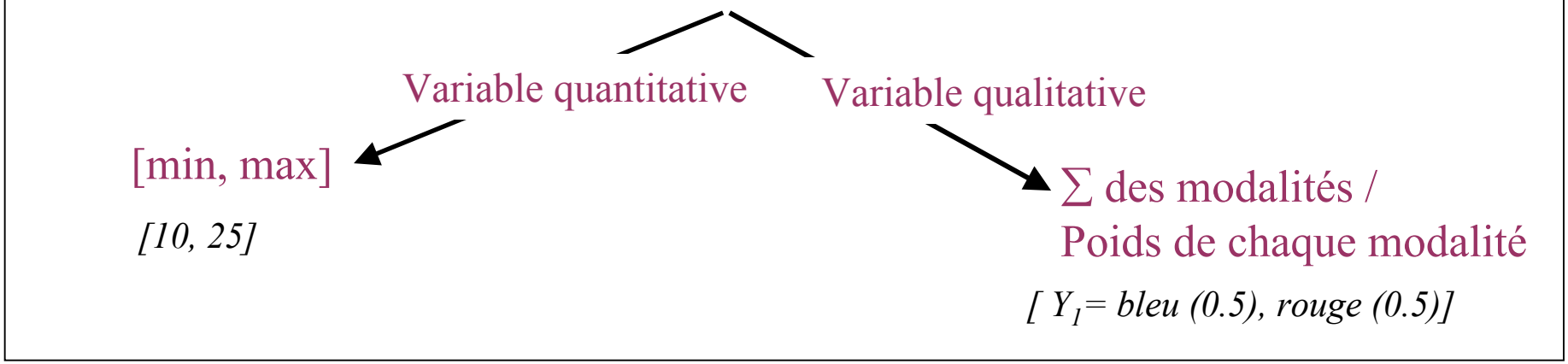
Utilise des opérateurs de généralisation et définit l'ensemble des variables symboliques et des objets symboliques et construit le tableau de données symboliques

Assertion / Opérateur de généralisation

Assertion : une conjonction d'événements

$A = \wedge_j [Y_j = d_j]$ où d_j est une valeur ou un ensemble de de valeurs

Opérateur de généralisation



L'intérêt des Objets Symboliques pour la modélisation de la base de connaissances est :

❖ De prendre en compte la notion de variabilité intervenant au sein d'un groupe de données.

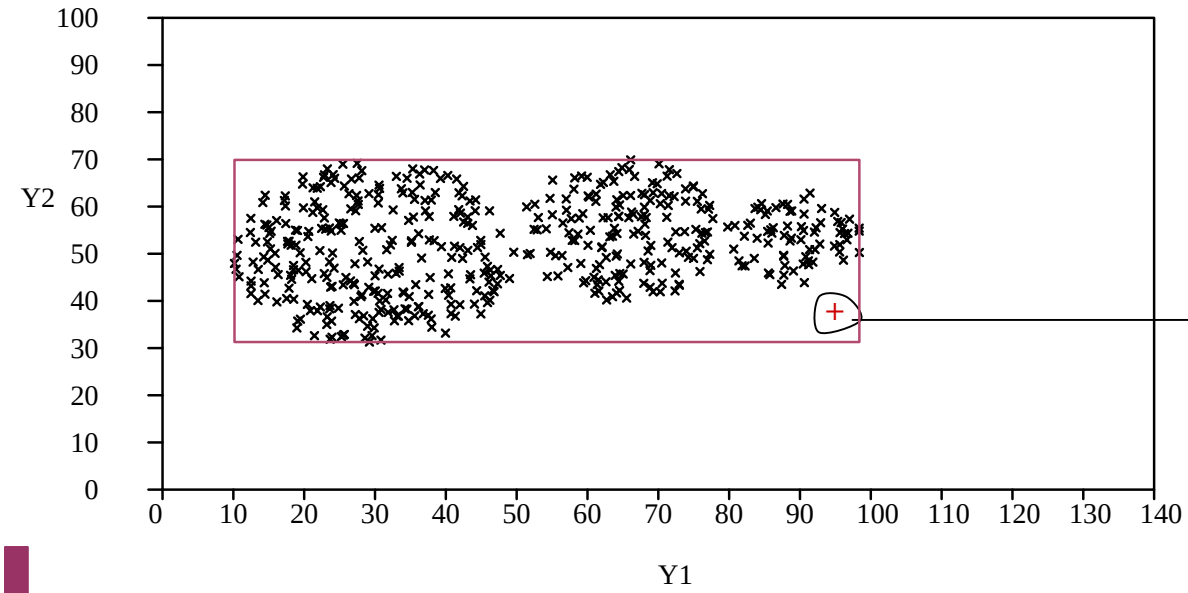
❖ De permettre des analyses ultérieures sur les résumés obtenus en permettant d'associer aux données une structure complexe.

Sur-généralisation

Les principaux problèmes de notre généralisation sont :

❖ De généraliser variable par variable ➔ **perte d'informations (corrélation entre les variables définies sur les observations)**

❖ Hypothèse forte sur la répartition des observations dans l'espace de description de l'extension de la description symbolique ➔ **descriptions non homogènes**



Observation potentielle

Une décomposition de chaque groupe basée sur un algorithme de classification [EGC02] [SFC03]

INRIA

Extraction à partir de bases de données: Décomposition

Cette décomposition est une méthode divisive de classification (**Marie Chavent 1997, 1998**)

- construction de classes par **divisions successives** de l'ensemble des individus
- Une classe est divisée en **2 classes** en fonction d'une **question binaire**
- La quantification de cette division est basée sur le critère d'inertie donc de la distance ente les individus

INRIA

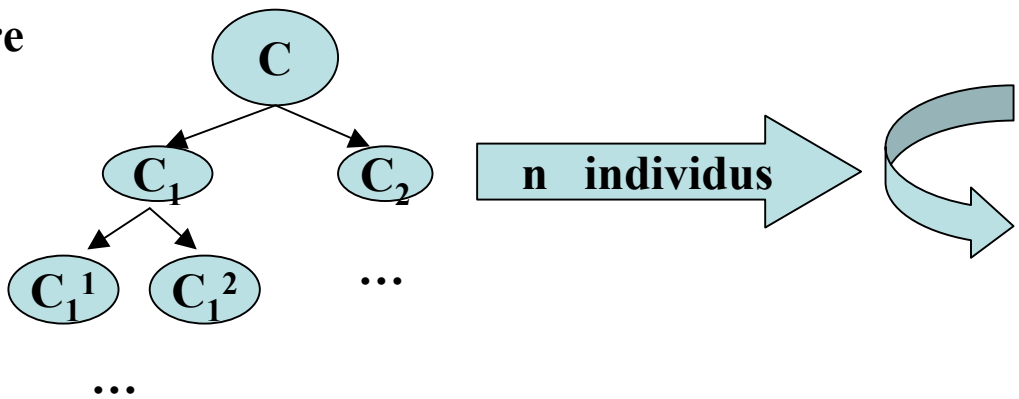
Algorithme divisif et récursif

à chaque itération il faut:

• Déterminer la question binaire associée à un événement qui permet de réduire hétérogénéité de la partition.

•Réaliser le partitionnement en 2 classes de cette classe

Arbre binaire



$(2^{n-1} - 1)$ possibilités

(n -1) possibilités

Détermination de la classe à diviser

- Donner un critère d'homogénéité H qui permet d'évaluer les bipartitions construites, dans notre cas c'est le critère d'inertie

$$H(c) = \sum_{x_i \in C} p_i d_M^2(x_i, g_C)$$

- Le critère d'évaluation d'une partition est donc l'inertie intraclasse W *(critère additif)*

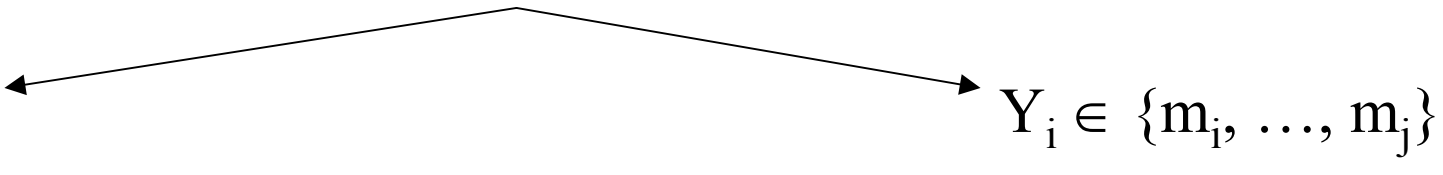
$$W(P_k) = \sum_{l=1}^{k} H(c_l)$$

- Le choix de la classe à diviser est la classe qui maximise la différence:

$$W(P_k) - W(P_{k+1}) = \underline{H_C - H_{C1} - H_{C2}} = \Delta$$

Réalisation du partitionnement en 2 classes

- Une bipartition admissible est une bipartition induite par une question binaire dépendant d'un événement .



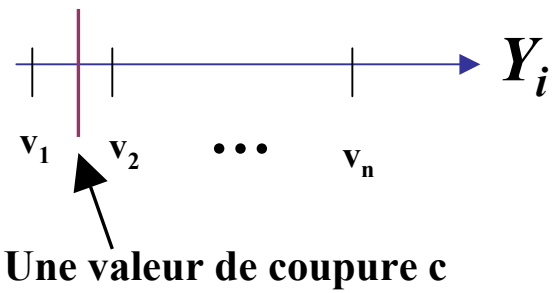
- Choisir parmi les bipartitions admissibles celle qui minimise le critère associé à la bipartition
- Complexité de l'algorithme est liée au nombre des bipartitions

Réalisation du partitionnement en 2 classes

Y_i quantitative

$\downarrow$

(n-1)

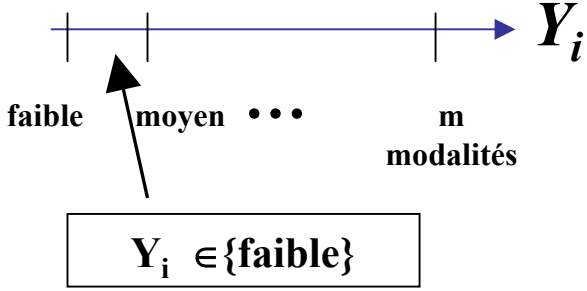


$C_1=\{w \in \Omega / Y_i(w) <= c\}$
$C_2=\{w \in \Omega / Y_i(w) > c\}$

Y_i qualitative Ordinale

$\downarrow$

(m-1)



Y_i qualitative nominale

$\downarrow$

$(2^{m-1} -1)$

Réalisation du partitionnement en 2 classes

Minimiser W de la bipartition

Th.Huygens

Maximiser inertie interclasse de la bipartition B

$$\sum_{x_i \in C_1} p_i d_M^2(x_i, g_{c_1}) + \sum_{x_i \in C_2} p_i d_M^2(x_i, g_{c_2})$$

$$\mu(C_1)\, d^2(g_C, g_{C_1}) + \mu(C_2)\, d^2(g_C, g_{C_2}) =$$

Poids de C_1

$$\frac{\mu(C_1) * \mu(C_2)}{\mu(C_1) + \mu(C_2)} d^2(g_{C_1}, g_{C_2}) \text{ [WARD]}$$

- **Traitement des valeurs manquantes**

Algorithme

Initialisation: $P_1 = \Omega$; k$\leftarrow$1;

Tant Que (k < nb class -1) alors:

- pour chaque classe $C \in P_k$, choisir parmi les bipartitions (C_1, C_2) de C induites par les questions binaires, la partition qui minimise:

$$H(C_1) + H(C_2) = \sum_{x_i \in C_1} p_i d_M^2(x_i, g_{c_1}) + \sum_{x_i \in C_2} p_i d_M^2(x_i, g_{c_2})$$

- choisir la classe $C \in P_k$ qui maximise:

$W(P_k) - W(P_{k+1}) = H(C) - H(C_1) - H(C_2)$

- $P_{k+1} = P_k \cup \{C_1, C_2\} - \{C\}$

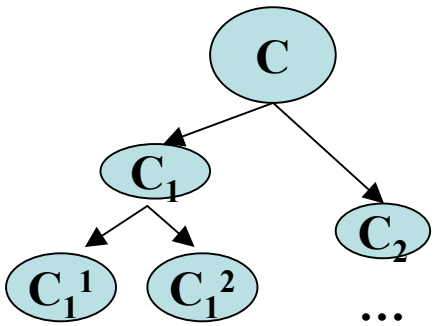
k $\leftarrow$ k+1 ;

Fin Tant Que

Inertie intraclasse (W) de la bipartition (C_1, C_2) de C remplacer par B [WARD]

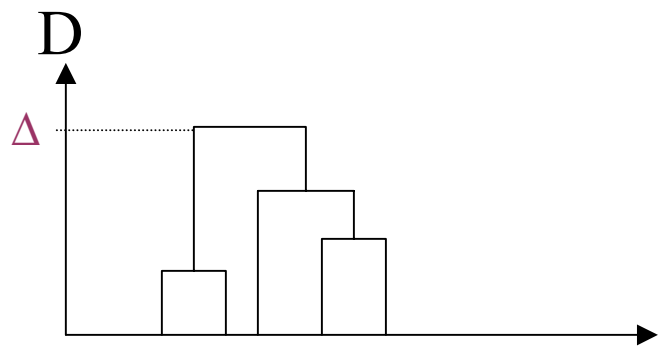
Décomposition

Pas d'ordre de découpage



Arbre de décision

Ordre de construction



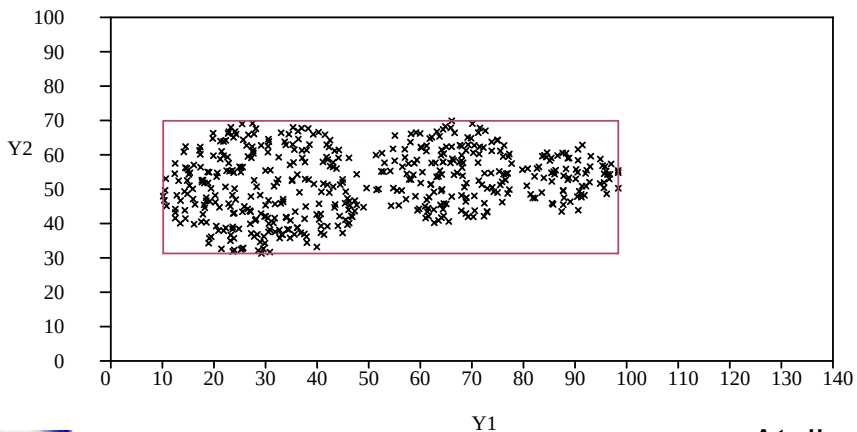
Hiérarchie

Décomposition

La qualité d'une assertion est liée à sa densité: le nombre d'individus recouverts par l'assertion par unité de volume de la description.

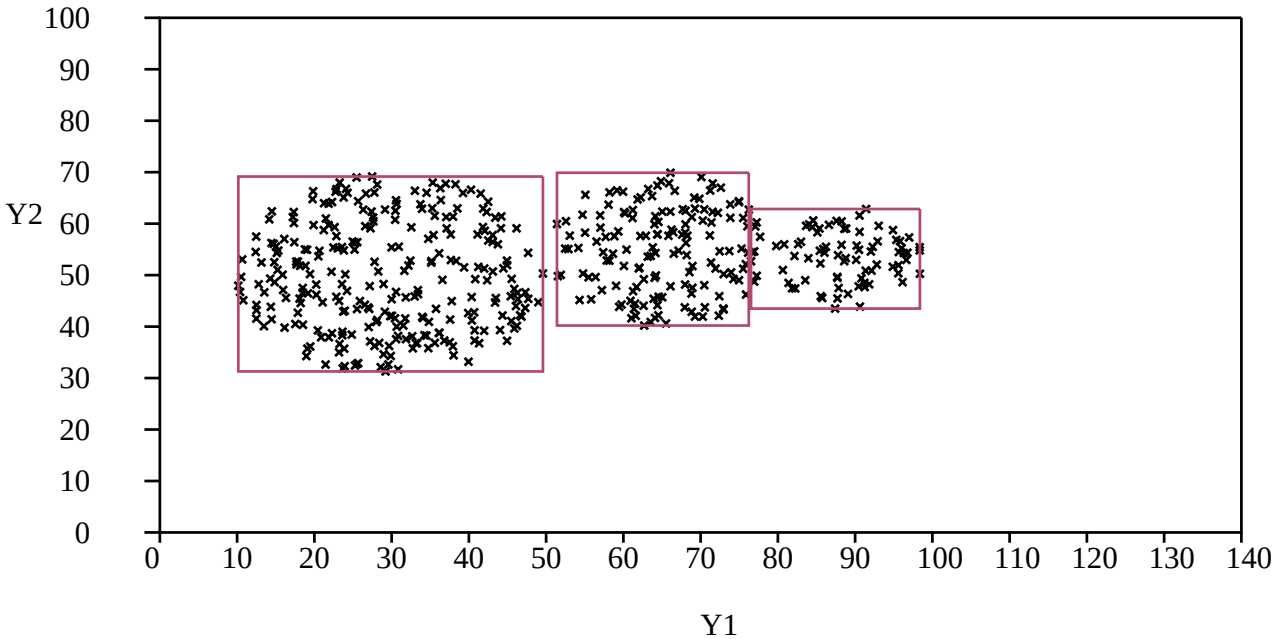
$Dens(a) = card(ext_G(a)) / vol(d)$

↳ + la densité au sein d'un hypercube est élevée et uniformément répartie par unité de volume plus la qualité de la généralisation augmente



[Y1 ∈[10.15, 98.38] ∧
Y2 ∈[31.28, 69.89]]

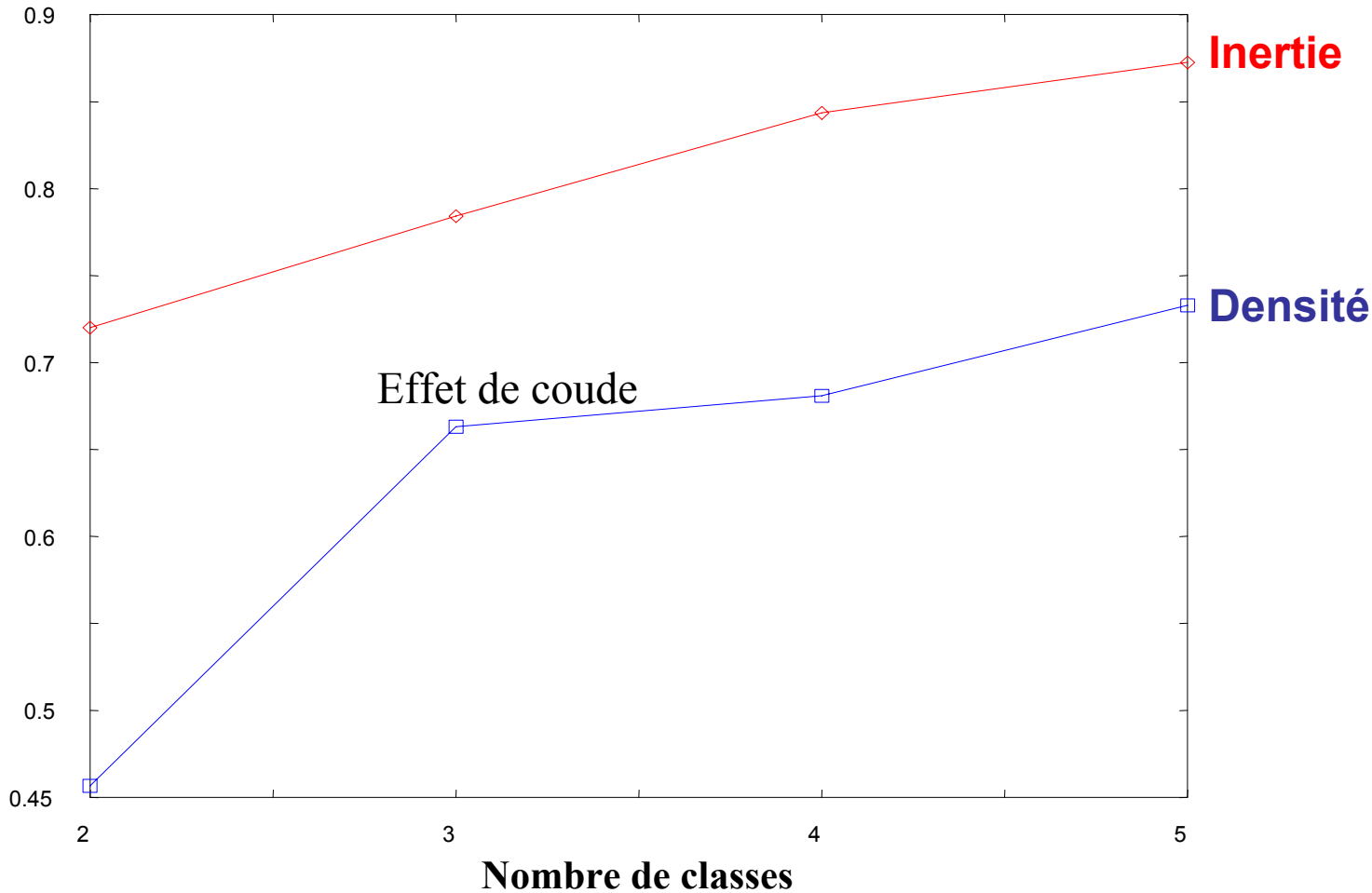
Densité ≈ 0.146



Densité≈ 0.188

$$
\begin{aligned}
[\ & Y1 \in [10.15, 49.58] \wedge Y2 \in [31.28, 69.15] \ \textbf{OU} \\
& Y1 \in [51.40, 76.23] \wedge Y2 \in [40.20, 69.89] \ \textbf{OU} \\
& Y1 \in [76.52, 98.38] \wedge Y2 \in [43.49, 62.84] \ ]
\end{aligned}
$$

Amélioration de la qualité des objets obtenus à partir de cette décomposition qui est une disjonction d'assertions ou « horde ».


0.9
0.8
0.7
0.6
0.5
0.45
2
3
4
5
Nombre de classes
Effet de coude
Inertie
Densité

Conclusions

- Méthode de classification qui réduit considérablement la complexité connue des méthodes divisives.
- Méthode permettant d'homogénéiser les descriptions obtenues par généralisation symbolique.
- Méthode permettant d'extraire des connaissances sur les groupes, utiles pour les analyses ultérieures.

Construction des Objets Symboliques

variables de classe: Date & Zone

75 Objets Symboliques

1 Jan_Matin
1 Jan_Midi
1 Jan_Apres_midi
1 Jan_Soir
...
15 Jan_Matin
15 Jan_Midi
15 Jan_Apres_midi
15 Jan_Soir

Descripteurs

IDNavigation	identificateur d'une navigation
NBRequest_OK	nombre de requêtes correctes
PRequest_SEL	pourcentage de requêtes correctes
NBrequest	nombre de requêtes essayées
DureeTotale	d'une navigation
Repetition	nombre de requêtes répétées
User_Agent	identificateur d'un *navigateur*
User_System	identificateur du *système d'exploitation*
MDurée_OK	*moyenne* de la durée d'une navigation
MSize_OK	*moyenne* de la taille des pages lues
Date	date de la navigation
Zone	période de la navigation dans la journée
Pays	identificateur du Pays

Classe : 1 Cardinal : 10 -- SEMAINE DU 1ER JANVIER --

```
================================
(  3) 05 Jan_Nuit          [1.0]  (  6) 02 Jan_Nuit          [0.3]  ( 11) 04 Jan_Apres_midi    [1.2]
( 23) 01 Jan_Soir          [1.2]  ( 27) 02 Jan_Matin         [1.2]  ( 36) 15 Jan_Soir          [1.0]
( 38) 06 Jan_Nuit          [1.1]  ( 42) 03 Jan_Nuit          [1.0]  ( 44) 01 Jan_Matin         [1.4]
( 63) 02 Jan_Apres_midi    [0.6]
```

Classe : 2 Cardinal : 11 -- WEEK END --

```
================================
(  7) 04 Jan_Midi          [0.9]  (  9) 05 Jan_Soir          [0.7]  ( 13) 11 Jan_Matin         [2.5]
( 18) 05 Jan_Matin         [0.6]  ( 29) 11 Jan_Soir          [0.4]  ( 30) 12 Jan_Midi          [0.4]
( 37) 04 Jan_Nuit          [0.5]  ( 41) 04 Jan_Soir          [0.8]  ( 43) 05 Jan_Midi          [1.1]
( 65) 04 Jan_Matin         [2.2]  ( 70) 11 Jan_Midi          [1.0]
```

Classe : 3 Cardinal : 23 -- SEMAINE (MATIN)--

```
================================
(  0) 15 Jan_Midi          [1.1]  (  2) 08 Jan_Matin         [0.8]  (  4) 15 Jan_Matin         [0.5]
(  8) 06 Jan_Midi          [0.9]  ( 10) 13 Jan_Apres_midi    [0.6]  ( 12) 08 Jan_Apres_midi    [1.5]
( 14) 08 Jan_Midi          [0.7]  ( 21) 06 Jan_Matin         [0.6]  ( 25) 09 Jan_Midi          [0.6]
( 26) 10 Jan_Midi          [0.7]  ( 32) 13 Jan_Soir          [1.7]  ( 34) 10 Jan_Apres_midi    [1.1]
( 35) 09 Jan_Apres_midi    [3.2]  ( 40) 13 Jan_Matin         [0.9]  ( 45) 07 Jan_Matin         [0.5]
( 46) 07 Jan_Midi          [0.7]  ( 48) 14 Jan_Matin         [1.0]  ( 53) 09 Jan_Matin         [0.9]
( 54) 10 Jan_Matin         [1.0]  ( 58) 07 Jan_Apres_midi    [1.5]  ( 59) 14 Jan_Apres_midi    [1.1]
( 64) 15 Jan_Apres_midi    [0.6]  ( 66) 06 Jan_Apres_midi    [1.1]
```

Classe : 4 Cardinal : 20 -- SOIR, NUIT and WEEK END --

```
================================
(  5) 07 Jan_Nuit          [0.5]  ( 19) 12 Jan_Matin         [0.9]  ( 20) 01 Jan_Nuit          [1.1]
( 22) 11 Jan_Nuit          [1.3]  ( 24) 13 Jan_Nuit          [1.1]  ( 28) 15 Jan_Nuit          [0.4]
( 33) 01 Jan_Apres_midi    [0.7]  ( 39) 08 Jan_Nuit          [0.8]  ( 47) 06 Jan_Soir          [0.9]
( 51) 03 Jan_Soir          [1.0]  ( 55) 05 Jan_Apres_midi    [0.7]  ( 56) 12 Jan_Apres_midi    [0.9]
( 57) 03 Jan_Apres_midi    [1.6]  ( 60) 09 Jan_Nuit          [1.6]  ( 61) 10 Jan_Nuit          [1.0]
( 62) 01 Jan_Midi          [2.3]  ( 67) 11 Jan_Apres_midi    [0.8]  ( 68) 12 Jan_Nuit          [0.7]
( 69) 14 Jan_Nuit          [1.0]  ( 73) 12 Jan_Soir          [0.7]
```

Classe : 5 Cardinal : 11 -- SOIR SEMAINE --

```
================================
(  1) 14 Jan_Soir          [0.8]  ( 15) 07 Jan_Soir          [1.4]  ( 16) 03 Jan_Midi          [1.1]
( 17) 02 Jan_Soir          [1.3]  ( 31) 14 Jan_Midi          [0.8]  ( 49) 02 Jan_Midi          [1.2]
( 50) 08 Jan_Soir          [1.0]  ( 52) 03 Jan_Matin         [1.0]  ( 71) 09 Jan_Soir          [0.9]
( 72) 10 Jan_Soir          [0.5]  ( 74) 13 Jan_Midi          [1.0]
```

Description des variables les plus discriminantes

Position	Name	Bj/Tj	Wj/W	Tj/T	Quality
(5)	**PRequest_SEL**	42.47	17.35	16.67	4.12
(11)	Repetition	19.73	8.06	16.67	-51.63
(13)	**User_Agent**	68.53	28.00	16.67	68.02
(14)	**User_System**	63.73	26.04	16.67	56.24
(15)	MDurée_OK	33.63	13.74	16.67	-17.56
(17)	MSize_OK	16.64	6.80	16.67	-59.19