

Analyse de la Vraisemblance des Liens Relationnels :

Une méthodologie d'analyse classificatoire des données

Israël-César Lerman

Projet *Symbiose* IRISA / INRIA

Campus Universitaire de Beaulieu, 35042 Rennes Cedex

lerman@irisa.fr

Préambule

Une méthodologie d'Analyse des Données se conçoit *toujours* par rapport à une algorithmique d'un type donné.

L'algorithmique en question est fournie par la
Classification Ascendante Hiérarchique (CAH)

Plan

- 1 – Introduction : principes généraux et exemple**
- 2 - Le principe algorithmique de la CAH et son objet**
- 3 – La méthode de l'Analyse de la Vraisemblance des Liens Relationnels (*Une présentation à partir d'exemples*)**
- 4 – Les grands types logico-mathématiques de données**
- 5 – Les programmes *CHAVLh*, *AVARE* et *LLAhclust***
- 6 – Quelques grandes applications récentes**
- 7 - Références**

1 - Introduction

1.1 – Le paradigme de la CAH, schéma d'arbre et exemples

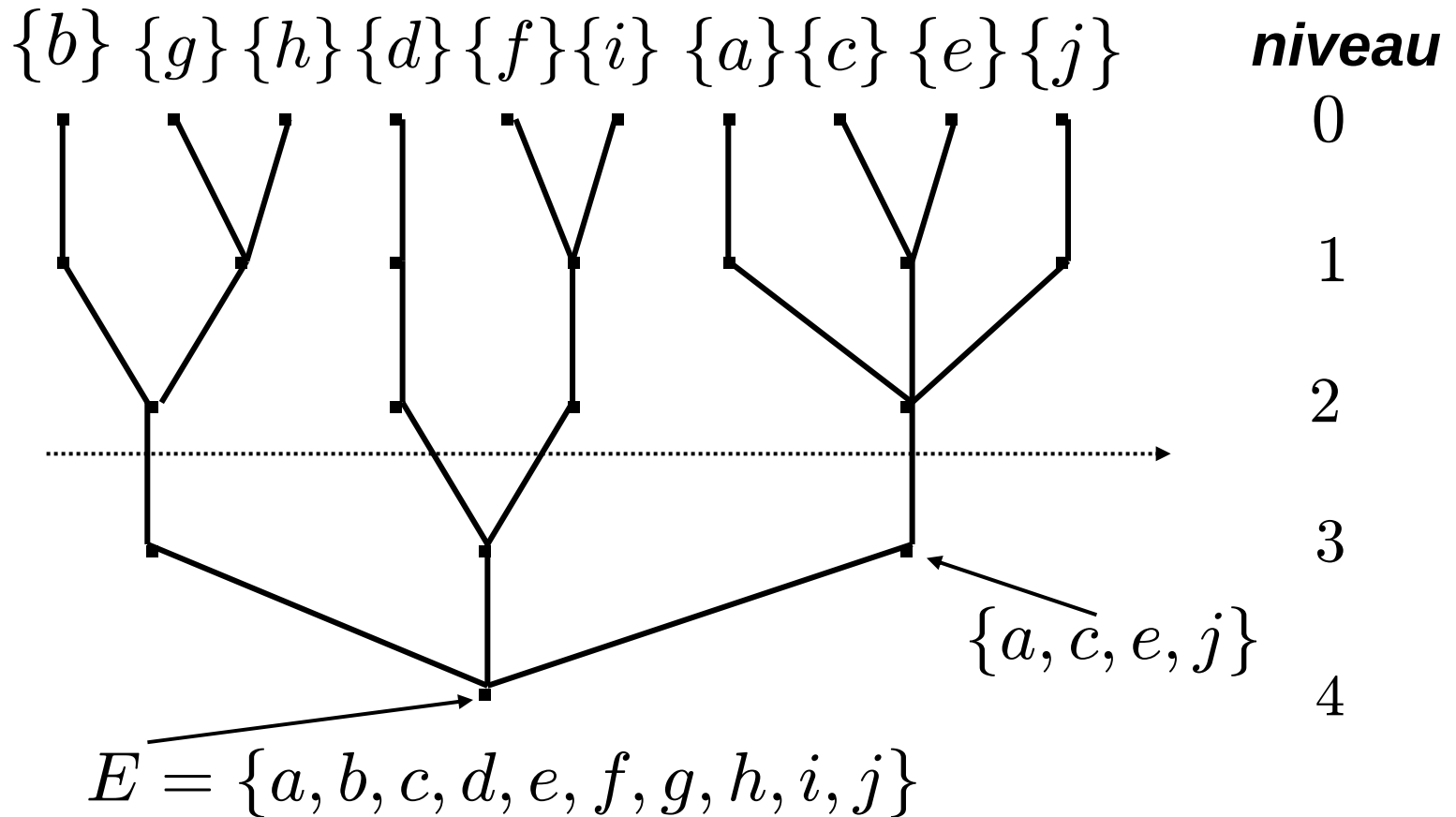
Le principe :

« Je me contenterai de rapprocher les objets suivant le plus grand nombre de degrés de leurs rapports et de leurs ressemblances ... Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble afin d'en faire un tout dont les parties soient unies et liées intimement. »

***Michel Adanson ; Histoire naturelle du Sénégal,
Bauche, Paris, 1757***

« Rapprochement des objets » $\longrightarrow$ schéma d'arbre

$$E = \{a, b, c, d, e, f, g, h, i, j\}$$



Une illustrations réelle

- **Enquête d' « Agoramétrie » (1989) (CEA-EDF)**

- **19 variables qualitatives de la forme :**

« Souhaitez-vous voir jouer un rôle important à tel homme politique ? »

Codes des résultats à la question posée :

1 : OUI

2 : NON

3 : SANS RÉPONSE

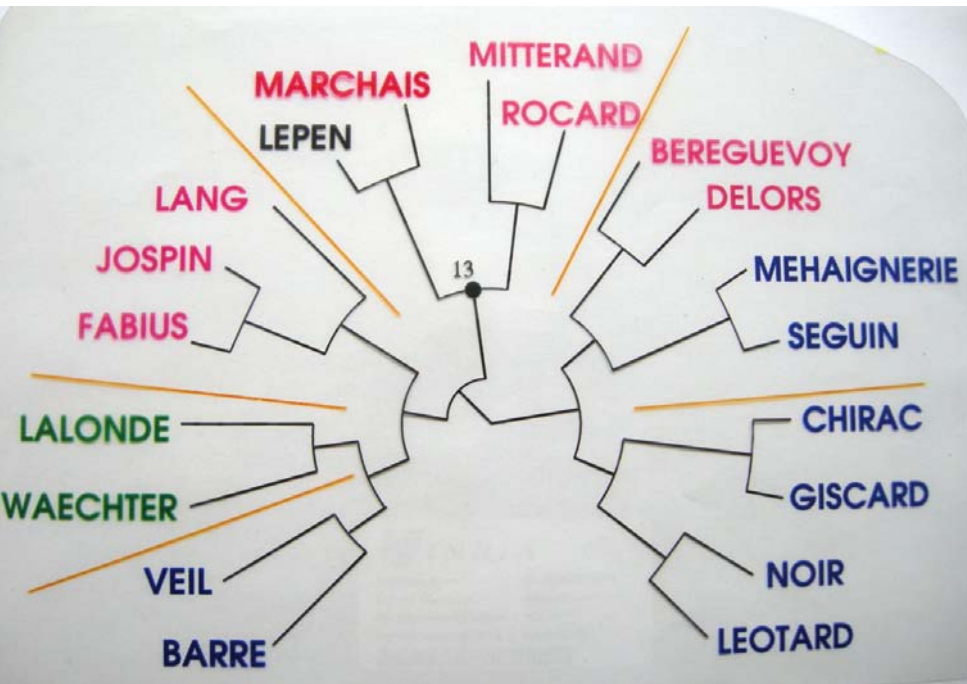
- **Deux interprétations de la structure munissant l'ensemble des valeurs de la variable qualitative :**

- Variable qualitative *nominale*

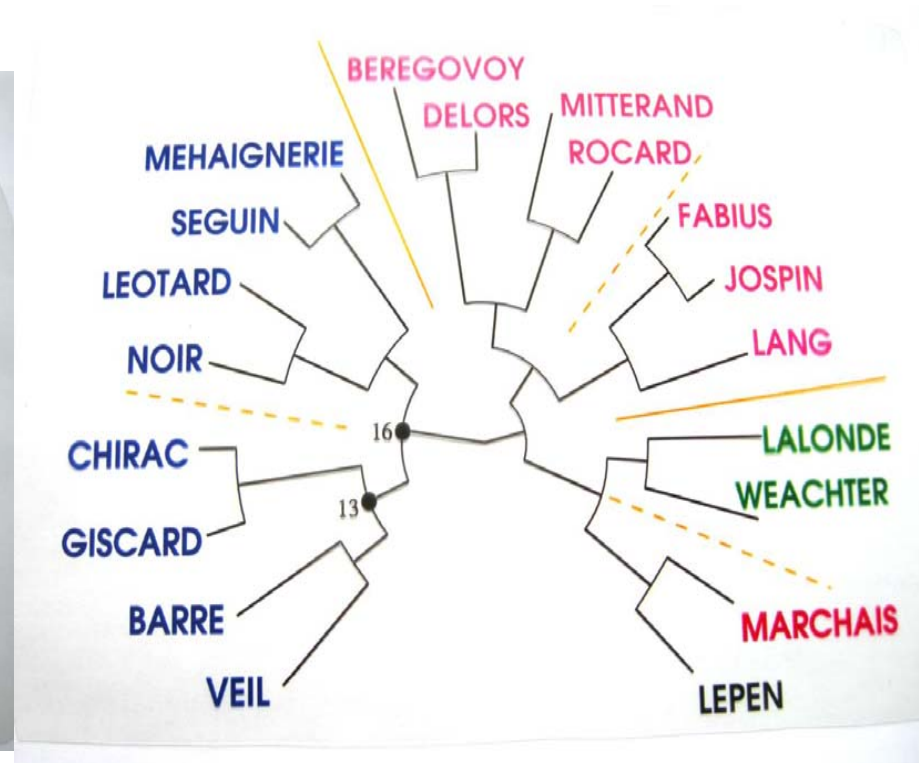
- Variable qualitative « *préordonnance* » :

$$r_{12} < r_{23} < r_{13} < r_{11} = r_{22} = r_{33}$$

Deux résultats d'une méthode de CAH sur les données d'enquête (500 individus)



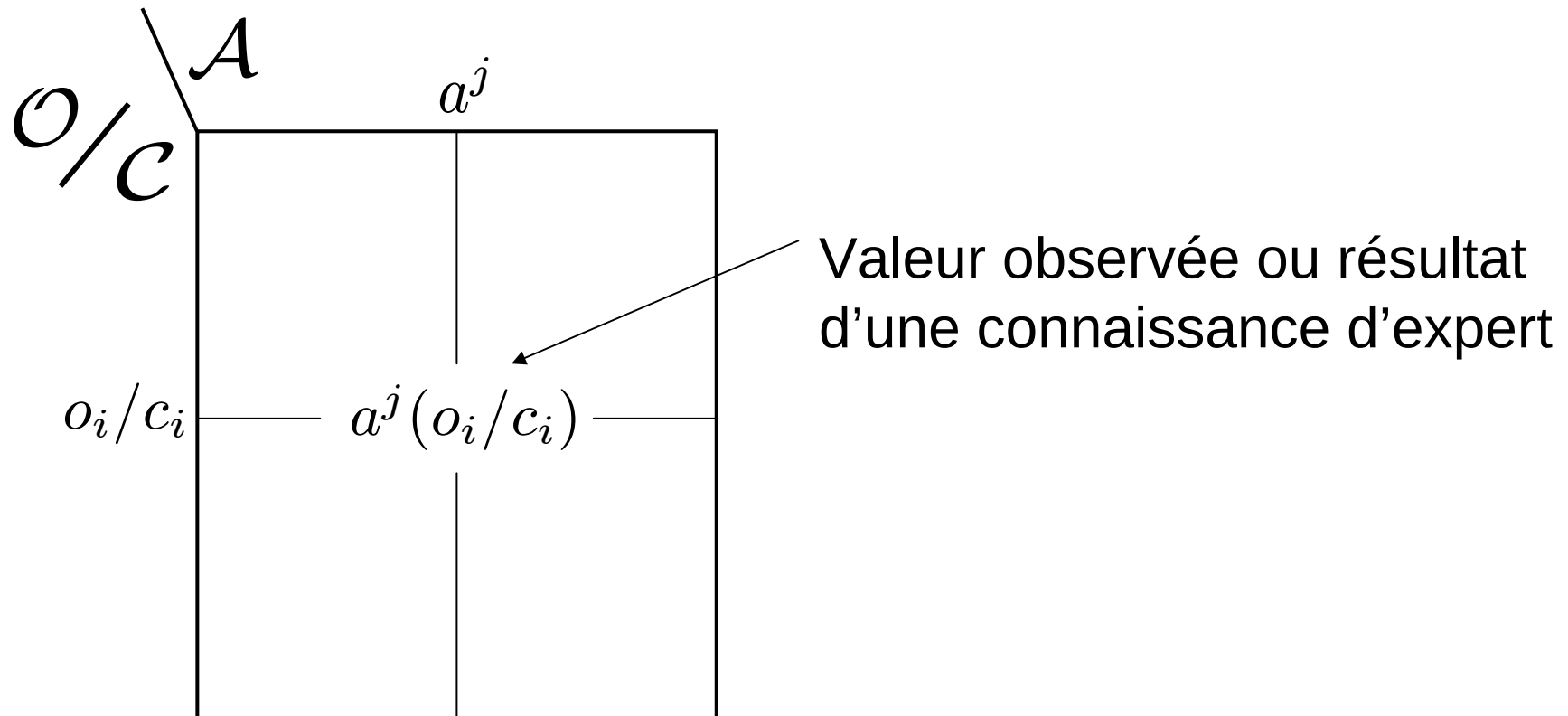
Qualitatives « nominales »



Qualitatives « préordonnances »

1.2. - Les données sont organisées au moyen d'un tableau

Le tableau des données croise l'ensemble $\mathcal{A}$ des attributs et l'ensemble $\mathcal{O}$ des objets (resp., $\mathcal{C}$ des catégories ou concepts).



2 – Le principe algorithmique de la CAH et son objet

2.1. Une expression formalisée

$$(E, \delta)$$

δ : notion de dissimilarité entre parties disjointes de E :
ensemble à organiser par la CAH.

D'un niveau au suivant de l'arbre des classifications agréger les paires de classes « les plus proches » au sens de δ .

- Initialisation : chaque classe contient un élément
- Condition d'arrêt : il y a une seule classe

δ ?

$$(E, d)$$

d : indice de dissimilarité entre éléments de E :

$$\{d(x, y) | (x, y) \in E \times E\}$$

$$E \leftarrow \mathcal{A}$$

$$E \leftarrow \mathcal{O}/\mathcal{C}$$

Comment définir d compte tenu de la nature mathématique et statistique du tableau des données ?

Comment opérer le passage (crucial) entre d et δ ?

$$\begin{array}{c}
 (E, \mu_E, d) \\
 \Updownarrow \\
 (\{d(x, y) | \{x, y\} \in P_2(E)\}) , \mu_E = \{\mu_x | x \in E\} \\
 \Downarrow \\
 (\mathcal{P}(E), \mu_{\mathcal{P}}, \delta)
 \end{array}$$

$\mathcal{P}(E)$: ensemble des parties de E , $\mu_{\mathcal{P}}$: extension additive de μ_E sur $\mathcal{P}(E)$

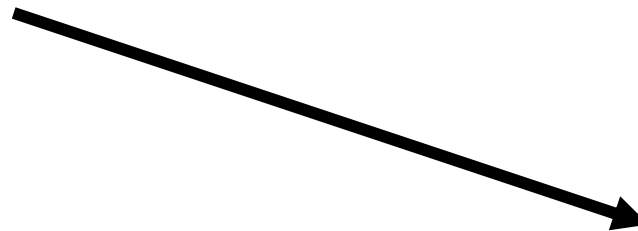
$$\begin{aligned}
 \delta &: (\mathcal{P}(E) \times \mathcal{P}(E), \mu_{\mathcal{P}}) \longrightarrow \mathcal{R}_+ \\
 \delta(X, Y) &= f[\{d(x, y) | \{x, y\} \in (X \cup Y)^2\}, \mu_{X \cup Y}] \\
 \mu_Z &= \{\mu_z | z \in Z\} , Z \in \mathcal{P}(E)
 \end{aligned}$$

2.2. – Ce qu'on souhaiterait ; mais, de façon très générale ...

$\mathcal{O}$

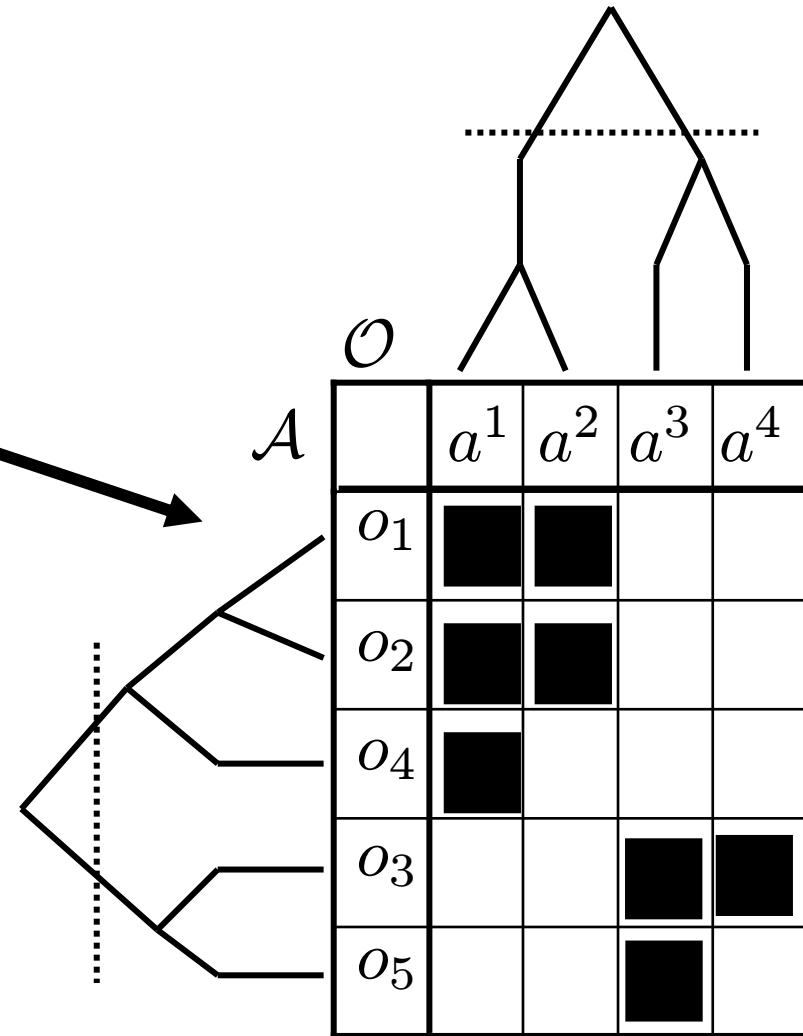
$\mathcal{A}$

	a^1	a^2	a^3	a^4
o_1	1	1	0	0
o_2	1	1	0	0
o_3	0	0	1	1
o_4	1	0	0	0
o_5	0	0	1	0



E : ensemble des unités élémentaires (attributs ou objets).

Indice de Jaccard entre éléments de E . Indice du « lien simple » entre parties disjointes de E .



... mais, de façon très générale ...

CAH(E)



Nœuds et Niveaux « significatifs »

$E \leftarrow \mathcal{A};$

$CAHA \leftarrow E.CAH;$

$E \leftarrow \mathcal{O}(\text{ resp., } E \leftarrow \mathcal{C});$

$CAHB \leftarrow E.CAH;$

$Croiser(CAHA, CAHB)$

CAH ← CHAVLh : Classification Hiérarchique par Analyse de la Vraisemblance des Liens Relationnels

3 - La méthode de l'Analyse de la Vraisemblance des Liens Relationnels (*Une présentation à partir d'un exemple traité à la main*)

3.1 – Caractéristiques principales de la méthode

- Représentation ensembliste puis relationnelle des variables descriptives $\Rightarrow$ extrême généralité de la méthode par rapport à la complexité de la description**
- Notion très générale de similarité se référant à une échelle probabiliste établie par rapport à une hypothèse d'absence de liaison entre les variables (attributs) de description, « respectant » les distributions marginales respectives de ces variables**
- Guidage statistique dans l'interprétation de l'arbre des classifications, détermination d'agrégats « intéressants » et de partitions (classifications) « intéressantes » $\Rightarrow$ Nœuds et Niveaux « significatifs » de l'arbre des classifications**

3.2. Principe de l'indice de fusion des classes dans AVL

L'exigence première d'une CAH : établir un ordre des agrégations

Un critère *numérique* pour sélectionner la « meilleure » agrégation



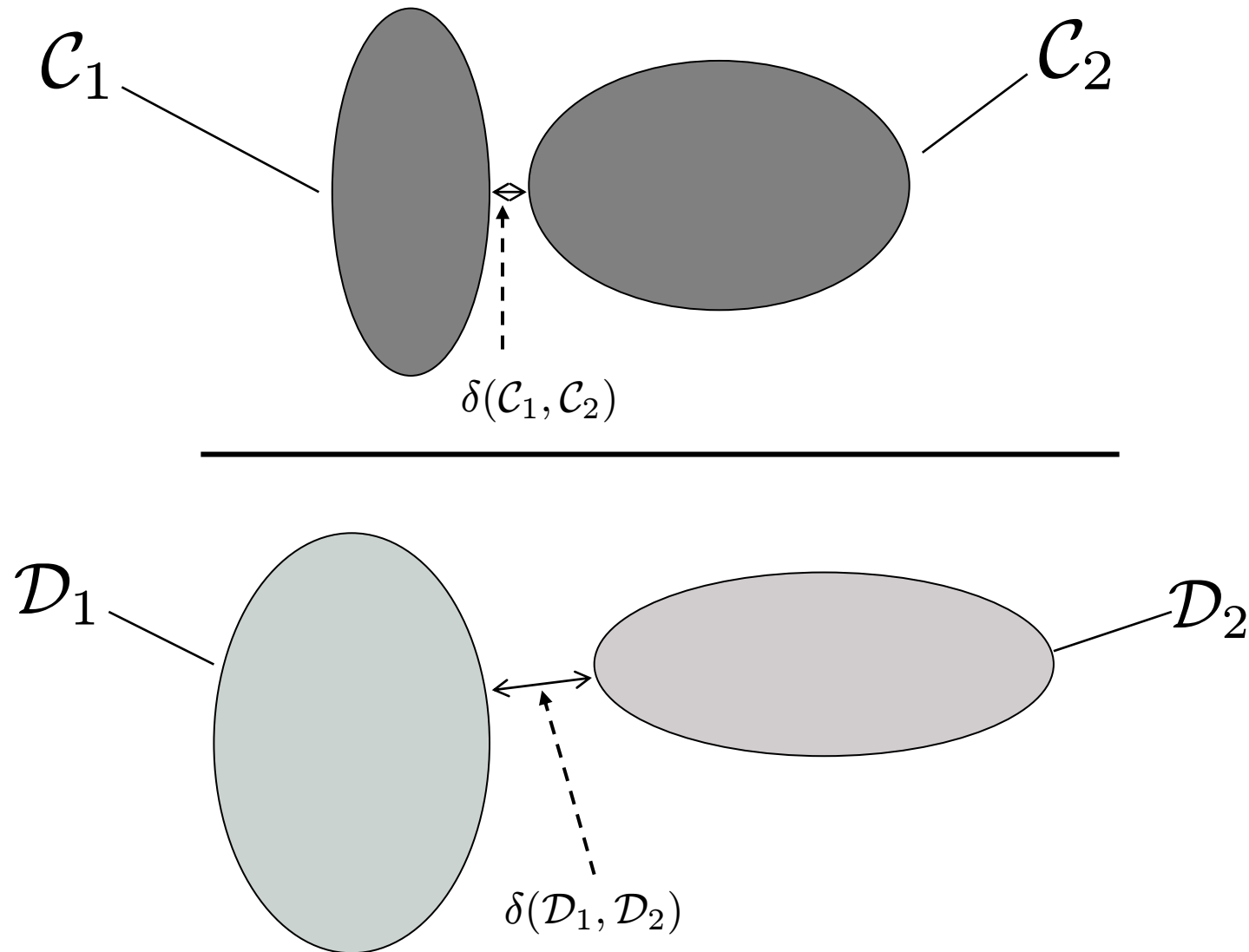
Critère de la distance minimale : $\mathcal{D} \leftarrow \mathcal{D}_1 \vee \mathcal{D}_2$

Critère de la Vraisemblance du Lien défini par cette distance :

$\mathcal{C} \leftarrow \mathcal{C}_1 \vee \mathcal{C}_2$

Exceptionnalité de la petitesse : philosophie de la théorie de l'information

Une autre illustration



**Pour AVL
réunir
d'abord**

D_1 et D_2

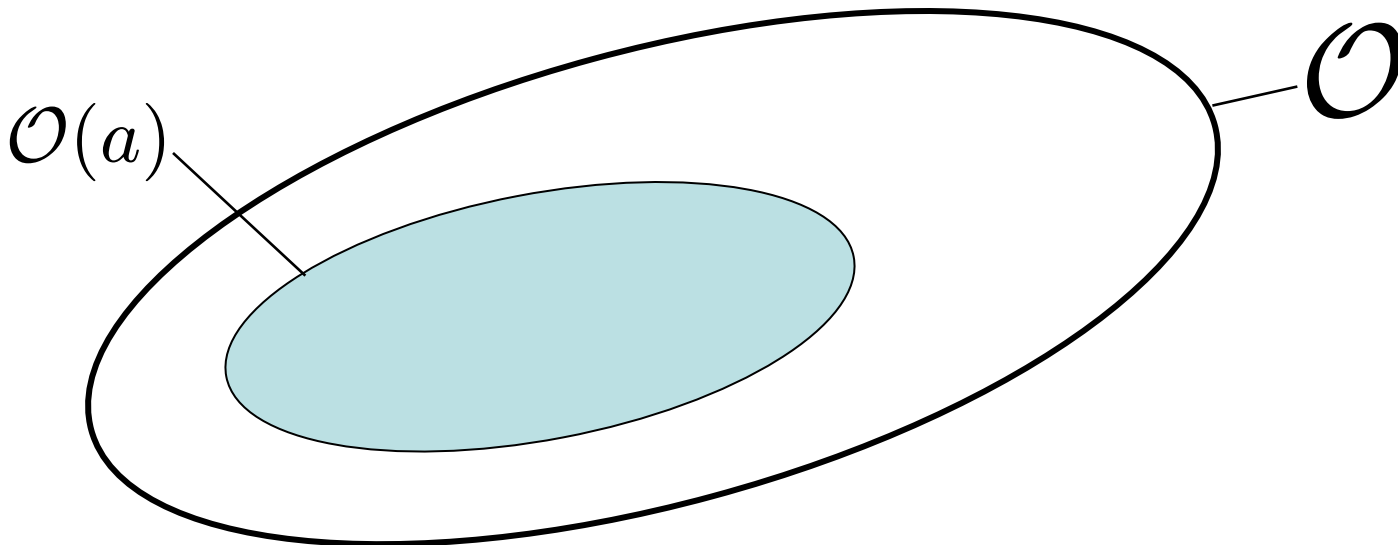
3.3. Illustration des comparaisons par paires dans la méthode AVL_R, en cas de données booléennes

- *Représentation ensembliste de l'ensemble des attributs de description*

$\mathcal{A}$: ensemble d'attributs booléens de description

$\mathcal{O}$: ensemble des objets décrits

Représentation d'un attribut booléen



	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
o_1	0	0	1	1	0	1	0	1
o_2	0	0	0	1	0	0	1	1
o_3	1	1	0	1	1	0	0	0
o_4	0	0	1	0	0	0	1	1
o_5	1	0	1	0	1	0	0	0
o_6	1	0	0	1	1	0	0	0
o_7	1	0	0	0	1	1	1	0
o_8	0	0	1	1	0	0	1	1
o_9	1	0	0	1	1	1	0	0
o_{10}	1	1	0	1	0	0	1	1

**Tableau de
données
booléennes**

$$O(a^1) = \{o_3, o_5, o_6, o_7, o_9, o_{10}\}$$

$$O(a^2) = \{o_3, o_{10}\}$$

$$O(a^3) = \{o_1, o_4, o_5, o_8\}$$

$$O(a^4) = \{o_1, o_2, o_3, o_6, o_8, o_9, o_{10}\}$$

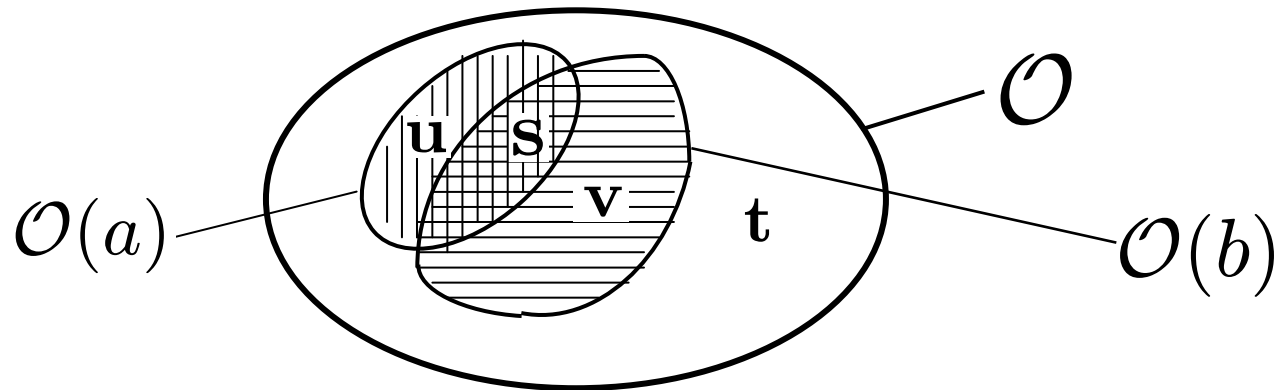
$$O(a^5) = \{o_3, o_5, o_6, o_7, o_9\}$$

$$O(a^6) = \{o_1, o_7, o_9\}$$

$$O(a^7) = \{o_2, o_4, o_7, o_8, o_{10}\}$$

$$O(a^8) = \{o_1, o_2, o_4, o_8, o_{10}\}$$

- **Comparaison entre attributs booléens**



$$s = n(a \wedge b) = \text{card}(\mathcal{O}(a \wedge b)) = \text{card}(\mathcal{O}(a) \cap \mathcal{O}(b))$$

$$u = n(a \wedge \neg b) = \text{card}(\mathcal{O}(a \wedge \neg b)) = \text{card}(\mathcal{O}(a) \cap \mathcal{O}(\neg b))$$

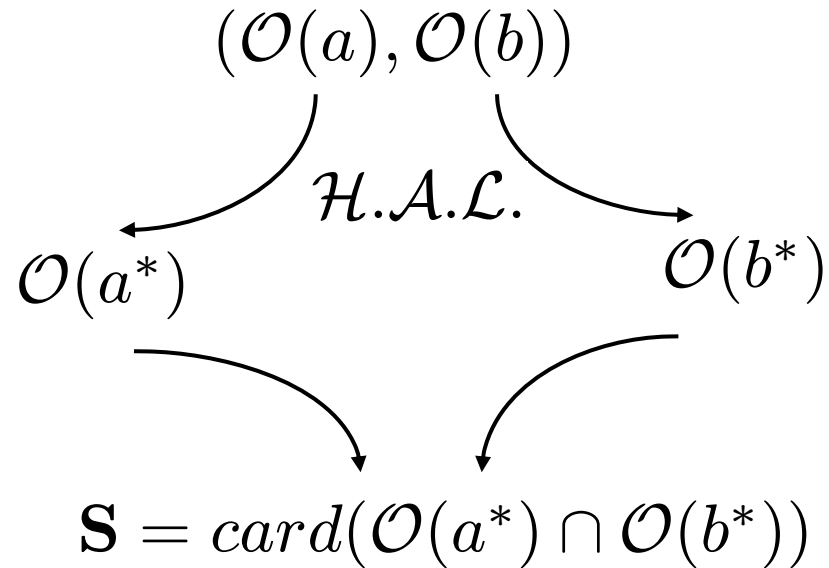
$$v = n(\neg a \wedge b) = \text{card}(\mathcal{O}(\neg a \wedge b)) = \text{card}(\mathcal{O}(\neg a) \cap \mathcal{O}(b))$$

$$t = n(\neg a \wedge \neg b) = \text{card}(\mathcal{O}(\neg a \wedge \neg b)) = \text{card}(\mathcal{O}(\neg a) \cap \mathcal{O}(\neg b))$$

Degré d'invraisemblance de la grandeur de s relativement au contexte

Contexte local de la comparaison entre a et b puis contexte global relativement à l'ensemble de toutes les paires d'attributs

Contexte local



$$P_l(a, b) = \text{Pr}\{\mathbf{S} \leq \mathbf{s}\}$$

Un indice statistiquement normalisé

$$\mathbf{Q}(a, b) = (\mathbf{s} - \mathcal{E}(\mathbf{S})) / \sqrt{\text{var}(\mathbf{S})}$$

$$P_l(a, b) = \Phi(\mathbf{Q}(a, b))$$

Cas du tableau précédent des données booléennes

$\mathcal{A}$	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
a^1	6	0.89	0.27	0.58	0.95	0.74	0.38	0.15
a^2	2	2	0.45	0.84	0.74	0.55	0.74	0.74
a^3	1	0	4	0.44	0.38	0.66	0.68	0.88
a^4	4	2	2	7	0.51	0.65	0.51	0.75
a^5	5	1	1	3	5	0.82	0.24	0.06
a^6	2	0	1	2	2	3	0.54	0.54
a^7	2	1	2	3	1	1	5	0.92
a^8	1	1	3	4	0	1	4	5

$s \backslash P_l$

Inversions entre S et P_l

$$s(a^4, a^7) = 3 \text{ et } P_l(a^4, a^7) = 0.51$$

$$s(a^5, a^6) = 2 \text{ et } P_l(a^5, a^6) = 0.82$$

Inversions entre l'indice de Jaccard et P_l

$$J(a, b) = \frac{n(a \wedge b)}{n(a \vee b)}$$

$$J(a^1, a^6) = \frac{2}{7} \text{ et } P_l(a^1, a^6) = 0.74$$

$$J(a^4, a^7) = \frac{3}{9} \text{ et } P_l(a^4, a^7) = 0.51$$

Contexte global

$$(a^1, a^2, \dots, a^j, \dots, a^p) \longrightarrow (a^{1*}, a^{2*}, \dots, a^{j*}, \dots, a^{p*})$$

$\mathcal{H.A.L.}$

$$\mathcal{P}_2(\mathcal{A}) = \{\{a^j, a^k\} | 1 \leq j \leq p\}$$

$$Q(a^j, a^k) \longleftarrow Q_g(a^j, a^k) = (Q(a^j, a^k) - moy_e(Q)) / \sqrt{var_e(Q)}$$

$$m_e(Q) \text{ et } var_e(Q) \text{ sur } \mathcal{P}_2(\mathcal{A})$$

$$P_g(a^j, a^k) = Pr\{Q_g(a^{j*}, a^{k*}) \leq Q_g(a^j, a^k)\} = \Phi(Q_g(a^j, a^k))$$

- **Comparaison entre objets décrits par des attributs (booléens)**

Les différentes étapes

(i) Contribution brute de l'attribut a^j à la comparaison de deux objets

$$s^j(o_i, o_{i'})$$

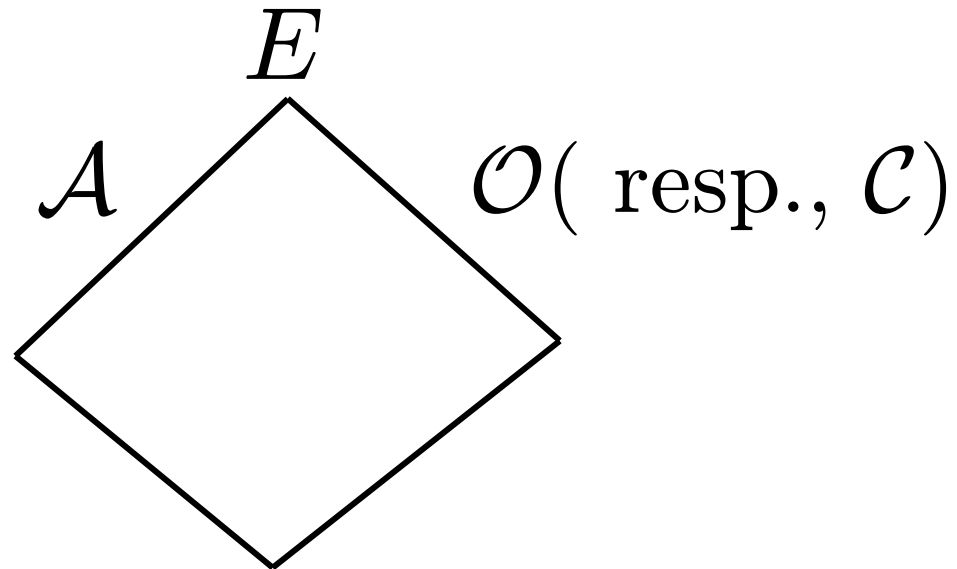
(ii) Normalisation statistique de $s^j(o_i, o_{i'})$ relativement à la distribution empirique :

$$\{s^j(o_l, o_{l'}) | 1 \leq l, l' \leq n\}$$
$$S^j(o_i, o_{i'}) = (s^j(o_i, o_{i'}) - m_e(s^j)) / \sqrt{\text{var}_e(s^j)}$$

(iii) Somme des contributions normalisées :

$$S(o_i, o_{i'}) = \sum \{S^j(o_i, o_{i'}) | 1 \leq j \leq p\}$$

(iv) Normalisation statistique globale de $S(o_i, o_{i'})$ par rapport à $P_2(\mathcal{O})$ conduit à l'indice probabiliste de la vraisemblance du lien



$$\{P(x, y) | \{x, y\} \in P_2(E)\}$$

Matrice des dissimilarités « informationnelles »

$$\{\Delta(x, y) = -\text{Log}_2(P(x, y)) | \{x, y\} \in P_2(E)\}$$

3.4. Famille de critères de la vraisemblance du lien maximal et construction de l'arbre

C et D deux parties disjointes (classes) de E

- **Indice brut :**

$$p(C, D) = \max\{P(c, d) \mid (c, d) \in C \times D\}$$

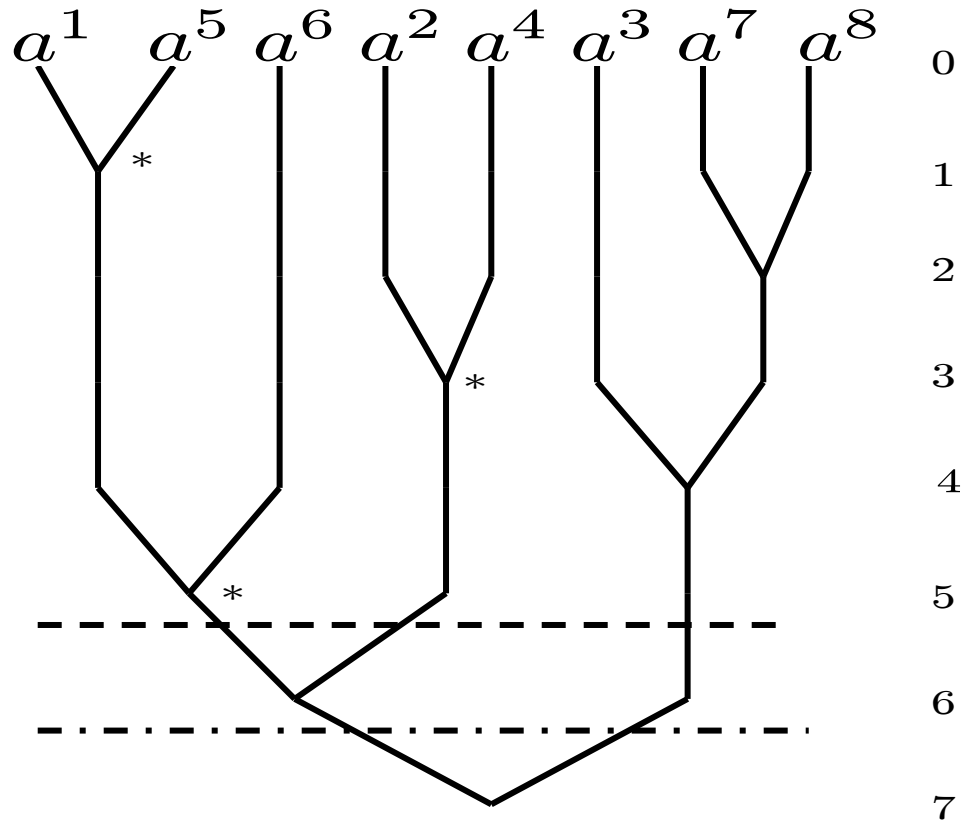
$$P(C, D) = \Pr\{p(C^*, D^*) \leq p(C, D)\} = (p(C, D))^{(|C| \times |D|)}$$

- **Une famille plus large :**

$$P_\epsilon(C, D) = (p(C, D))^{(|C| \times |D|)^\epsilon}, \quad 0 \leq \epsilon \leq 1$$

$$S_\epsilon = -\log[-\log(P_\epsilon(C, D))]$$

- **Les deux arbres de classification issus de l'exemple**



ARBA : Arbre de classification sur $\mathcal{A}$

----- : niveau « significatif », * : nœud « significatif »

3.5. Niveaux et nœuds les plus « significatifs » d'un arbre de classification

- *Représentation d'une partition*

(ici produite à un niveau donné de l'arbre de classification)

$$F = P_2(E)$$

$$\{\pi_1, \pi_2, \dots, \pi_k, \dots, \pi_l\}$$

$$\pi_k = \{E_{k1}, E_{k2}, \dots, E_{ki}, \dots, E_{kc_k}\}$$

$$R(\pi_k) = \sum \{P_2(E_{ki}) | 1 \leq i \leq c_k\}$$

Exemple dans le cas du premier arbre où $E = \mathcal{A}$

$$\pi_5 = \{\{1, 5, 6\}, \{2, 4\}, \{3, 7, 8\}\}$$

$$R(\pi_5) = \{\{1, 5\}, \{1, 6\}, \{5, 6\}, \{2, 4\}, \{3, 7\}, \{3, 8\}, \{7, 8\}\}$$

- Critère d'association entre une Similarité (numérique ou ordinale) et une partition

- *Indice brut :*

$$s(Q, \pi) = \sum \{Q(x, y) | \{x, y\} \in R(\pi)\}$$

$\mathcal{H.A.L.}$

$\pi^* \in \mathcal{P}(n; t(\pi))$: partition aléatoire

- *Indice brut aléatoire :*

$$s(Q, \pi^*) \longrightarrow \mathcal{N}(M, V)$$

- *Statistique globale évaluant l'adéquation:*

$$C(Q, \pi) = \frac{s(Q, \pi) - M}{\sqrt{V}}$$

• Niveaux et nœuds les plus significatifs

- *Distribution de la Statistique « globale » des niveaux :*

$$\{C(Q, \pi_i) | 1 \leq i \leq l\}$$

- *Taux d'accroissement entre deux niveaux consécutifs :*

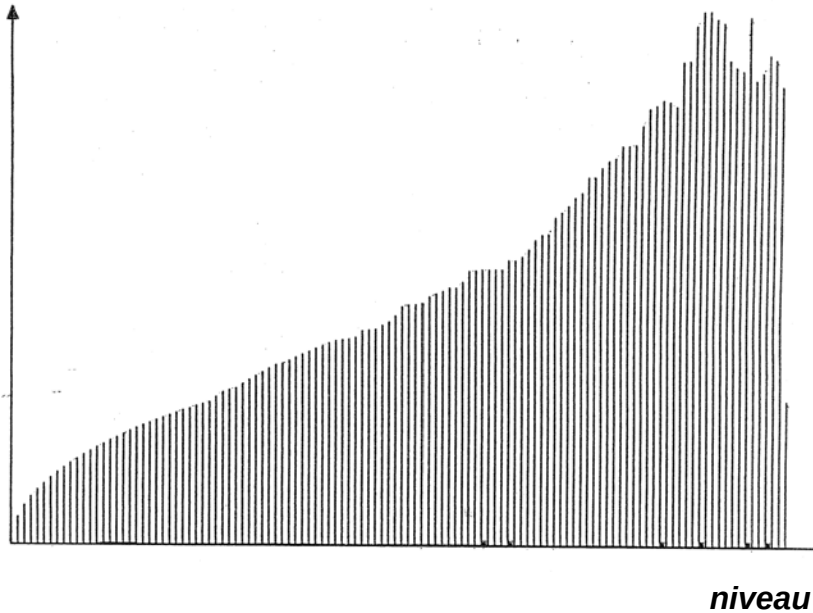
$$\tau(Q, \pi_i) = C(Q, \pi_i) - C(Q, \pi_{i-1}) , 1 \leq i \leq l$$

- *Distribution de la Statistique « locale » des niveaux :*

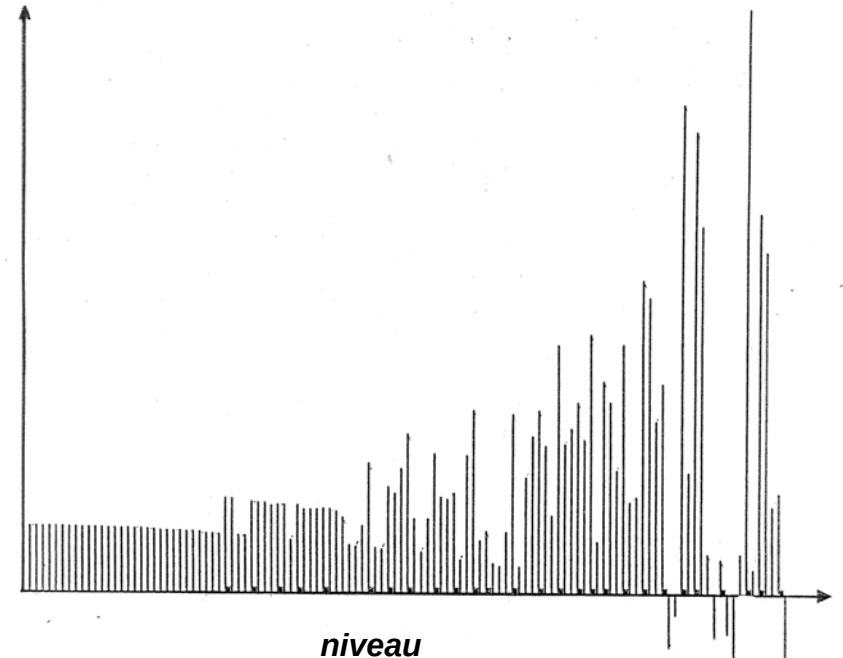
$$\{\tau(Q, \pi_i) | 1 \leq i \leq l\}$$

- Niveau « significatif » : *maximum local de* $C(Q, \pi_i)$
- Nœud « significatif » : *maximum local de* $\tau(Q, \pi_i)$

- Diagrammes des deux distributions



Distribution de la Statistique « globale »



Distribution de la Statistique « locale »

3.6. Croisement entre deux classifications duales

		$\mathcal{O}$			A_1		A_2		A_3	
$\mathcal{A}$			a^1	a^5	a^6	a^2	a^4	a^3	a^7	a^8
C_1	o_2									
	o_8									
	o_4									
	o_{10}									
	o_1									
C_2	o_6									
	o_9									
	o_3									
	o_7									
	o_5									

*Et s'il s'agit d'un
« grand » tableau
de données ?*

- **Croisement entre deux classes d'attributs « logiquement » indépendants**

$$(A, B)$$

$$x \longrightarrow \Phi_A(x) = \frac{1}{c(A)} \sum \{\Phi_a(x) | x \in A\}$$

$$x \longrightarrow \Phi_{\bar{A}}(x) = \frac{1}{c(A)} \sum \{\Phi_{\bar{a}}(x) | x \in A\}$$

$$\Phi_A(x) + \Phi_{\bar{A}}(x) = 1$$

- **Indice brut :**

$$(A, B) \longrightarrow \nu(A, B) = \nu(A \wedge B) = \sum \{\Phi_A(x) \times \Phi_B(x) | x \in \mathcal{O}\}$$

$$(A, B) \longrightarrow (A^*, B^*)$$

- **Indice statistiquement normalisé :**

$$\chi(A, B) = \frac{1}{\sqrt{n-1}} (\nu(A \wedge B) - \mathcal{E}[\nu(A^* \wedge B^*)]) / \sqrt{\text{var}[\nu(A^* \wedge B^*)]}$$

• **Tableau des coefficients :** $\chi(A_j, c^i)$, $1 \leq j \leq 3$, $i = 1, 2$.

	A_1	A_2	A_3
c^1	-0.833	$+0.143$	$+0.833$
c^2	$+0.833$	-0.143	-0.833

Comment faire pour les autres types de données relationnelles et d'abord, comment caractériser synthétiquement les différents types de données ?

5 - Les grands types mathématico-logiques de données

• DEUX SYSTÈMES : $\mathcal{T}$ et $\mathcal{S}$

$\mathcal{T}$: Système de Tarsky

$$\mathcal{T} = (\mathcal{O}; R_1, R_2, \dots, R_j, \dots, R_p)$$

p relations sur $\mathcal{O}$ d'aritées respectives quelconques

Système $\mathcal{S}$

$$\mathcal{S} = (\mathcal{C}; R_1, R_2, \dots, R_j, \dots, R_p)$$

Distributions statistiques des R_j , $1 \leq j \leq p$ sur chaque c de $\mathcal{C}$

Tableau de données numériques ? Tableau de contingence ?

5 – Les programmes CHAVLh, AVARE et LLAhclust

- CHAVLh (Classification Hiéarchique par Aanalyse de la Vraisemblance des Liens en cas de données hétérogène) (P. Peter, H. Leredde et I-C. Lerman)

Structures des données prises en compte

- 1 – Numériques-Quantitatives
- 2 - Booléennes
- 3 - Qualitatives nominales
- 4 - Qualitatives ordinales
- 5 - Qualitatives préordonnances ou graphes valués binaires
- 6 - Juxtaposition « horizontale » de tableaux de contingence
- 7 – Tableau de dissimilarités fourni par l'expert

Classification de l'ensemble  des objets décrits par des variables d'un même type ou par un mélange de types de variables

- Classification de l'ensemble $\mathcal{V}$ des variables d'un même type :
 - *numérique, booléen, qualitatif nominal, qualitatif ordinal*

AVARE (Association entre Variables qualitatives pREordonnance)
(M. Ouali-Allah)



- *Classification AVL d'un ensemble de variables qualitatives des différents types codées en termes de préordonnances*

Variable taxonomique : préordonnance très spécifique

CHAVLh, dépôt APP décembre 2005

Environnement *R*

LLAhclust (Likelihood Linkage Analysis hierarchical clustering)
(I. Kojadinovic, I.C. Lerman et P. Peter)

6 – Quelques applications récentes en grandeur réelle

- Simulation du comportement de l'exécution de « très gros » programmes à partir d'un échantillonnage « représentatif »

T. Lafage ; « *Étude, réalisation et application d'une plate-forme de collecte de traces d'exécution de programmes* », Thèse de doctorat, Université de Rennes 1, Décembre 2000.

- Détermination de classes sémantiques dans le *Traitement Automatique de Langues Naturelles*

M. Rossignol ; « *Acquisition automatique de lexiques sémantiques pour la recherche d'information* », Thèse de doctorat, Université de Rennes 1, Octobre 2005.

- Correspondance entre profils génotypiques et profils phénotypiques dans l'hémochromatose

I.-C. Lerman ; « Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques », *Revue de Statistique Appliquée*, 2006, vol. 2, p. 33-63.

- **Structuration d'une famille spécifique de séquences d'ADN**

S. Tempel ; « *Dynamique des hélitrons dans le genome d'Arabidopsis thaliana : développement de nouvelles stratégies d'analyse des éléments transposables* », Thèse de doctorat, Université de Rennes 1, 2007.

- **Organisation d'espèces de « phlébotomes » pour une description biologique très complexe**

I.-C. Lerman, P. Peter ; « Representation of Concept Description by Multivalued Taxonomic Preordonance Variables », in: *Selected Contributions in Data Analysis and Classification Studies in Classification, Data Analysis, and Knowledge organization*, P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (editors), *Studies in Classification, Data Analysis, and Knowledge organization*, Springer, 2007, p. 271-284

- **Segmentation d'images numérisés**

I.-C. Lerman et K. Bachar ; « Comparaison de deux critères en Classification Ascendante Hiérarchique sous contrainte de contiguïté. Application en imagerie numérique », *Revue de Statistique Appliquée*, 2008 (sous presse)

6 - Références importantes pour le fondement de la méthodologie

6.1. – Références générales

I.C. Lerman ; *Classification et Analyse Ordinale des Données*, Dunod, Paris, 1981.

I.C. Lerman ; Foundations of the Likelihood Linkage Analysis (LLA) classification method, *Applied Stochastic Models and Data Analysis*, vol. 7, # 1, march 1991, pp. 63-76.

I.C. Lerman ; Likelihood linkage analysis (LLA) classification method (Around an example treated by hand) *Biochimie* 75, Elsevier editions, 1993, pp. 379-397.

6.2. – Coefficients d'association (de similarité) entre variables relationnelles

I.C. Lerman, Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles, I : *Mathématique (, Informatique) et Sciences Humaines*, 30ème année, n° 118, 1992, pp. 35-52, II : *Mathématique (, Informatique) et Sciences Humaines*, 30ème année, n° 119, 1992, pp. 75-100, Paris.

M. Ouali-Allah ; *Analyse en préordonnance des données qualitatives. Application aux données numériques et symboliques*, thèse de doctorat de l'Université de Rennes 1, 5 décembre 1991.

F. Daudé ; *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*, thèse de doctorat de l'Université de Rennes 1, 24 juin 1992.

I.C. Lerman ; Indices d'association partielle entre variables qualitatives nominales, *RAIRO* série verte, vol. 17, n° 3, août 1983, pp. 213-259.

I.C. Lerman ; Indices d'association partielle entre variables qualitatives ordinales, *Publications Institut Statistique de Paris*, XXVIII, fasc. 1, 2, 1983, pp. 7-46.

I.C. Lerman ; Justification et validité statistique d'une échelle $[0,1]$ de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées, *Publications Institut Statistique Universités de Paris*, XXIX, fasc. 3-4, 1984, pp. 27-57.

I.C. Lerman ; Comparing partitions. Mathematical and Statistical aspects, 1ère conférence internationale des fédérations des sociétés de classification. Aix-la-Chapelle, juin 1987, in *Classification and related methods of data analysis*, Edited by H. H. Bock, North Holland, 1988, pp. 121-132.

I.C. Lerman ; Comparing classification tree structures: a special case of comparing q-ary relations, *RAIRO-Operations Research*, vol. 33, 1999, sept., pp. 339-365.

I.C. Lerman, F. Rouxel ; Comparing classification tree structures: a special case of comparing q-ary relations II, *RAIRO-Operations Research*, vol. 34, 2000, pp. 251-281.

6.3. – Indices de similarité entre objets décrits par des variables relationnelles de types quelconques

I.C. Lerman et Ph. Peter ; Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en parallèle, *Data Analysis and Informatics IV*, North Holland, 1986, pp. 121-136.

Ph. Peter ; *Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations assistée par ordinateur*, thèse de doctorat de l'Université de Rennes 1, mars 1987.

I.C. Lerman ; Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème de consensus en Classification, *Revue de Statistique Appliquée*, XXXV (2), 1987, pp. 39-60.

I.C. Lerman, Ph. Peter ; Indice probabiliste de vraisemblance du lien entre objets quelconques : analyse comparative entre deux approches, *Revue de Statistique Appliquée*, volume LI(1), 2003, pp. 5-35.

I.C. Lerman, Ph. Peter et J.L. Risler ; Matrices AVL pour la classification et l'alignement de séquences protéiques », *Publication Interne IRISA n° 866*, septembre 1994, *Rapport de Recherche INRIA n° 2466*.

I.C. Lerman and Ph. Peter ; Representation of Concept Description by Multivalued Taxonomic Preordonance Variables, in *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, P. Brito, P. Bertrand, G. Cucumel and F. de Carvalho, editors, *Springer 2007*, pp. 271-284

6.4. – Tableaux de contingence

I.C. Lerman et B. Tallur ; Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence, *Revue de Statistique Appliquée*, n° 28, 3, pp. 5-28.

I.C. Lerman, « Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence », *Mathématique & Sciences Humaines*, 21ème année, n° 83, 1983, pp. 5-30.

I.C. Lerman, « Analyse classificatoire d'une correspondance multiple, typologie et régression » in « *Data Analysis and Informatics, III* », E. Diday et al. (editors), North Holland, 1984, pp. 193-221.

B. Tallur, « *Contribution à l'analyse exploratoire de tableaux de contingence par la classification* », thèse de doctorat ès sciences, Université de Rennes 1, septembre 1988.

6.5. – Critères de fusion entre classes

I.C. Lerman, Formules de réactualisation en cas d'agrégations multiples, *RAIRO, série R.O.* vol 23, n° 2, 1989, pp. 151-163.

F. Costa Nicolau and H. Bacelar-Nicolau ; Some trends in the Classification of Variables, in *Data Science, Classification, and Related Methods*, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, Y. Baba Editors, Springer, 1998, pp. 89-98.

6.6. – Croisement de classifications

I.C. Lerman ; Croisement de classifications floues, *Publications de l'Institut de statistique des Universités de Paris*, 1979, XXIV, fasc. 1-2, pp. 13-46.

I.C. Lerman, M. Hardouin et T. Chantrel ; Analyse de la situation relative entre deux classifications « floues », *Secondes Journées Internationales Analyse des Données et Informatique, Data Analysis and Informatics*, North Holland, 1980, pp. 523-533.

I.C. Lerman ; Association entre variables qualitatives ordinales nettes ou floues, *Statistique et Analyse des Données*, n° 7, 1983, pp. 41-73.

I.C. Lerman ; Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques, *Revue de Statistique Appliquée*, 2006, LIV (2), pp. 33-63.

6.7. – Logiciels

P. Peter, H. Leredde et I.C. Lerman ; Notice du programme CHAVLh (Classification Hiéarchique par Aanalyse de la Vraisemblance des Liens en cas de variables hétérogènes), Dépôt APP (*Agence pour la Protection des Programmes*)
IDDN.FR.001.240016.000.S.P.2006.000.20700, Université de Rennes 1, Décembre 2005.

M. Ouali Allah ; Programme pour le calcul de coefficients d'association entre variables relationnelles, *La revue de modulad* n° 25, juin 2000, pp. 63-73.

I. Kojadinovic, I.C. Lerman and P. Peter ; lien LLAhclust dans *R* : [http ://cran.rproject.org/src/contrib/Descriptions/LLAhclust.html](http://cran.rproject.org/src/contrib/Descriptions/LLAhclust.html)).