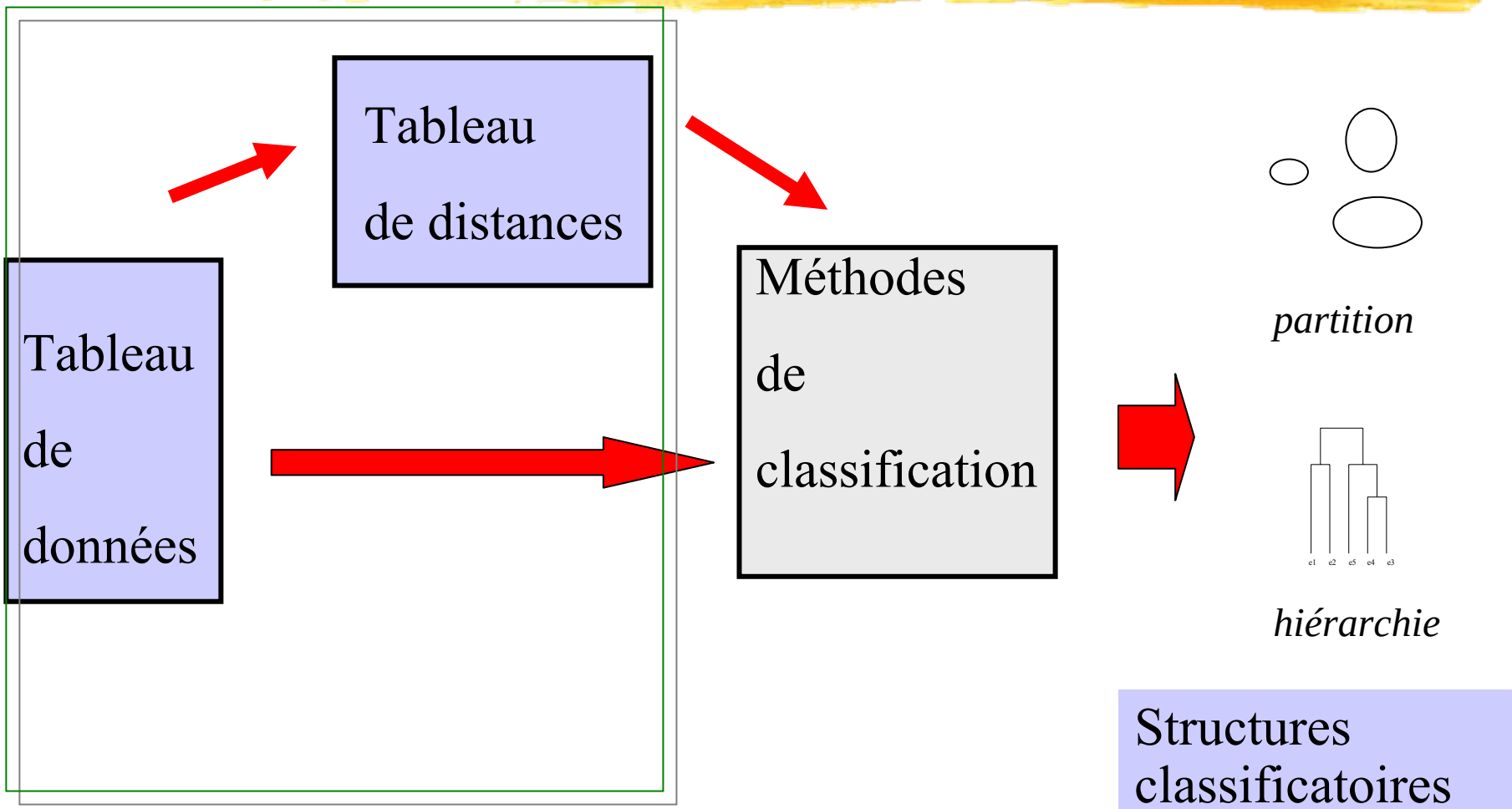
L'imprécision et la classification automatique

Yves Lechevallier
INRIA-Rocquencourt
78153 Le Chesnay Cedex
E_mail : Yves.Lechevallier@inria.fr

La classification automatique



1. Les données

- **Les données**
  - Données imprécises
  - Représentation des données
- **Structure complexe des données**
  - Tableau de données
  - Approche « symbolique »

Prise en compte de l'imprécision

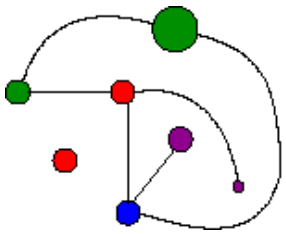
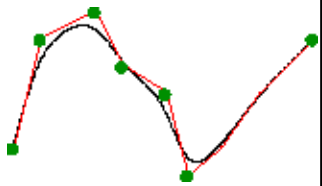
- ***Données imprécises***: Liées à la notion de précision d'un appareil de mesure. Elle est souvent employée pour désigner le manque d'exactitude ou de précision de la mesure due à l'appareil de mesure.
- ***Données incertaines***: Intrinsèque à la mesure. Souvent définie par une zone d'incertitude associée à la donnée.
- ***Données agrégées***: résultat d'une ou de plusieurs formules calculées à partir de plusieurs mesures.

Données complexes

- Données classiques : tableau de données (modèle « vectoriel »)

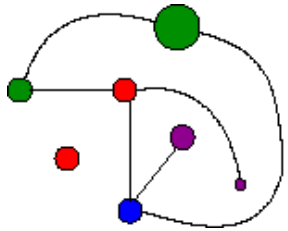
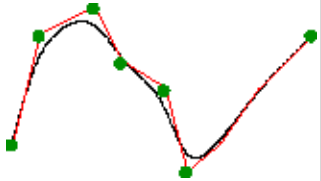
Iris d'Anderson/Fisher

#	Sépale		Pétale		Espèce
	long.	larg.	long.	larg.	
1	5.1	3.5	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
114	5.7	2.5	5.0	2.0	virginica

Graphe	Fonction	Texte	Image	Œuvre d'art
		Un texte complet		

- **Un modèle pour les données complexes**

Données complexes

Graphe	Fonction	Texte	Image
		Un texte complet	

- Graphe : Tableau de dissimilarité, réseaux sociaux
- Fonction : courbes de consommation, parcours, séquences, séries temporelles.
- Texte structuré ou semi-structuré : XML (documents, données, fichier LOG), requêtes SQL
- Image : Histogramme des couleurs, texture

Pourquoi avoir une représentation complexe ?

❖décrire ces grands volumes de données sans perte d'information. Les descriptions sont complexes afin de transmettre le maximum d'information mais en veillant à avoir une description compréhensible pour l'utilisateur ;

❖les descriptions doivent être écrites dans un formalisme proche de la représentation habituelle d'un ensemble de données mais elles doivent aussi servir d'étape intermédiaire pour d'autres analyses sans revenir aux données initiales.

Les données symboliques

E. Diday propose une extension des données uni-valuées aux données multi-valuées.

L'objectif est de donner un modèle de représentation aux données agrégées mais cela peut s'appliquer aux données incertaines ou imprécises;

Ce modèle sera différent du modèle de représentation des données.

.

- Généralisation de la notion de tableau de données
- Définition de nouveaux types de variables.

Requête de synthèse/ Résumé

Individus	Groupe	X1	X2
I1	G1	3	1
I2	G1	6	1
I3	G1	1	6
I4	G1	5	1
I5	G2	6	3
I6	G2	0	2
I7	G2	3	4

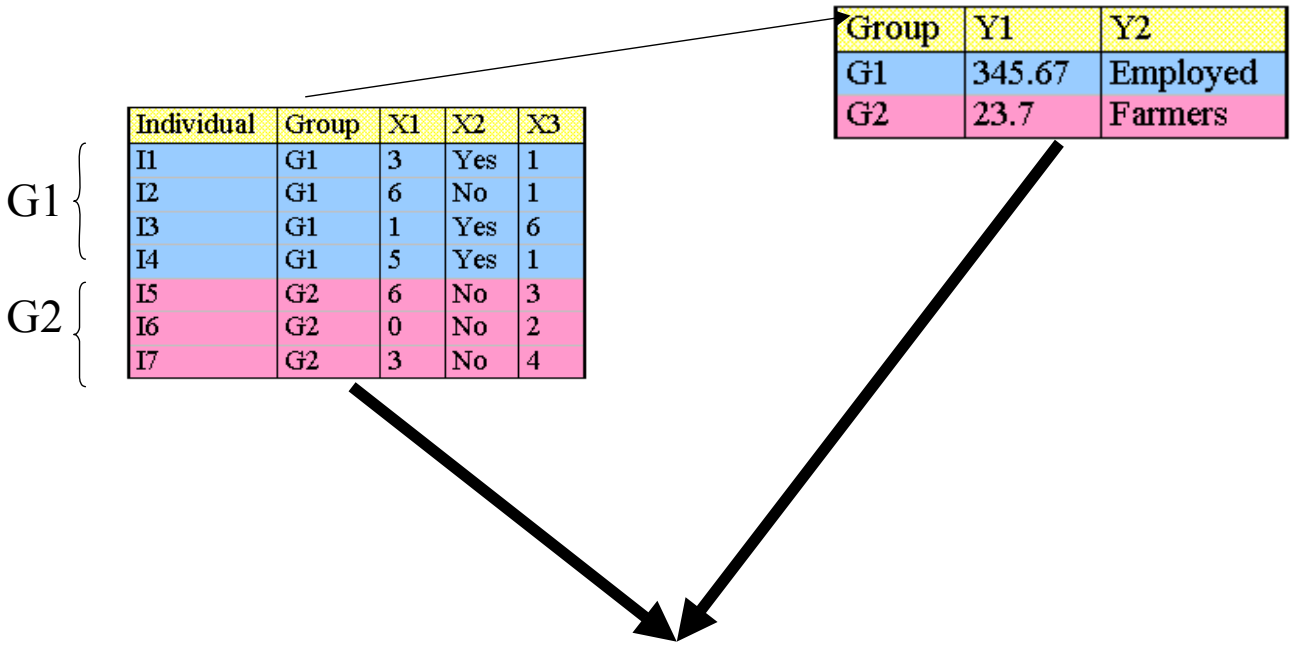
Requête de synthèse

	X1	X2
G1	3,25	2,5
G2	3	3

Opérateurs de généralisation

A partir d'un groupe d'individus

directement



Group	Y1	Y2
G1	345.67	Employed
G2	23.7	Farmers

	Individual	Group	X1	X2	X3
G1	I1	G1	3	Yes	1
	I2	G1	6	No	1
	I3	G1	1	Yes	6
	I4	G1	5	Yes	1
G2	I5	G2	6	No	3
	I6	G2	0	No	2
	I7	G2	3	No	4

	X1	X2	X3	Y1	Y2
G1	[1 :5]	{Yes,No}	[1:6]	345.67	Employed
G2	[0:3]	{No}	[2:4]	23.7	Farmers

Tableaux de données

- Ensemble des individus
- Ensemble des variables ou caractères
- Mise en correspondance
- Mesure de similarité ou de dissimilarité

Ensemble des individus et des variables

Un individu est l'unité de base sur laquelle des mesures vont être réalisées. Le terme individu peut désigner un client d'un magasin, un animal, une ville.

Sur ces individus on relève un certain nombre de mesures .

Afin de décrire de manière cohérente un ensemble d'individus l'utilisateur va devoir choisir un ensemble de variables ou de descripteurs

Le choix des variables est étroitement lié au problème posé.

- Pour une enquête les variables sont les questions.
- Pour la description d'une plante c'est un ensemble d'attributs

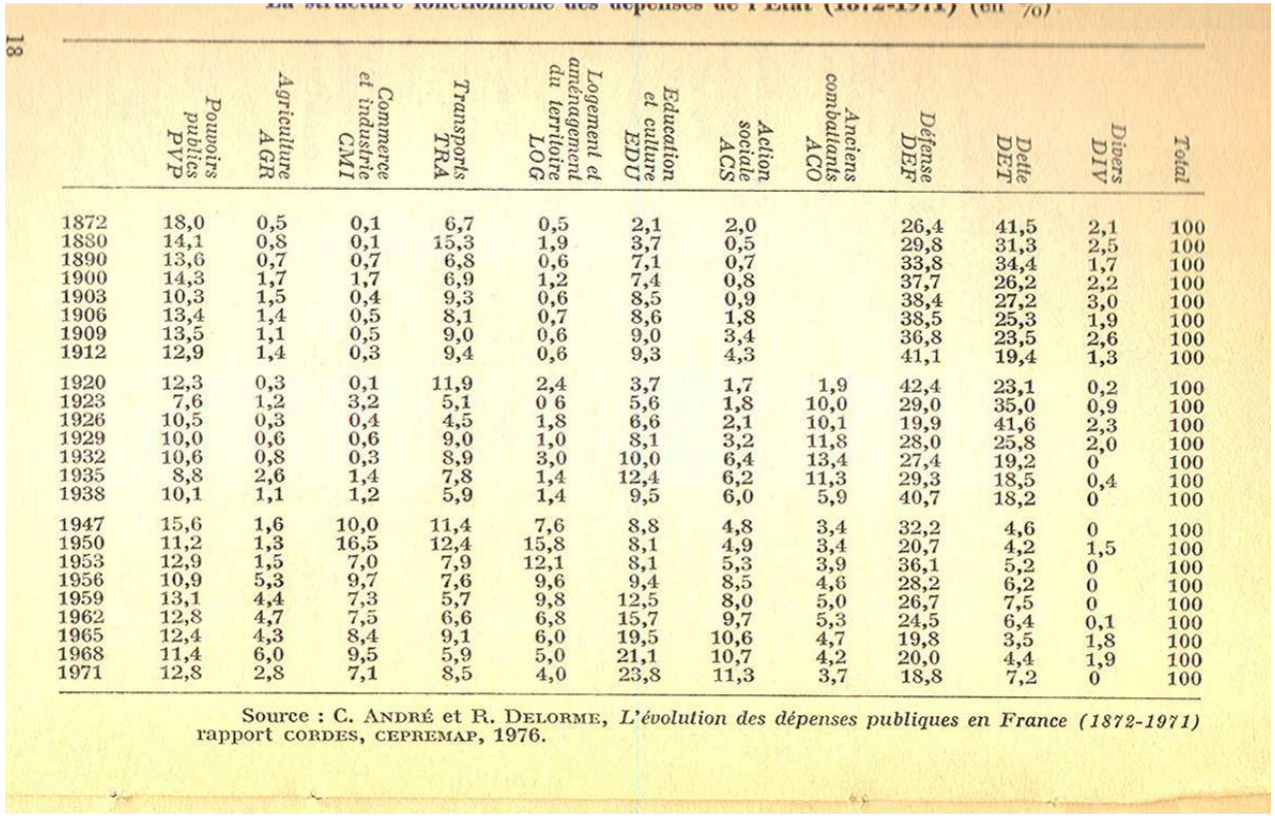
Mise en correspondance des individus avec les variables

La réalisation d'un tableau de données se fait par la mise en correspondance ou la mise en relation de l'ensemble Ω des individus avec l'ensemble Y des variables

$$T : E \times Y \rightarrow D \qquad T(x_i, Y_j) = Y_j(x_i)$$

D est inclus dans l'ensemble des réels, des entiers, des intervalles, des courbes ou des distributions…

Un exemple



La structure fonctionnelle des dépenses de l'État (1872-1971) (en %)

18

	Pouvoirs publics PVP	Agriculture AGR	Commerce et industrie CMI	Transports TRA	Logement et aménagement du territoire LOG	Education et culture EDU	Action sociale ACS	Anciens combattants ACO	Défense DEF	Dette DET	Divers DIV	Total
1872	18,0	0,5	0,1	6,7	0,5	2,1	2,0		26,4	41,5	2,1	100
1880	14,1	0,8	0,1	15,3	1,9	3,7	0,5		29,8	31,3	2,5	100
1890	13,6	0,7	0,7	6,8	0,6	7,1	0,7		33,8	34,4	1,7	100
1900	14,3	1,7	1,7	6,9	1,2	7,4	0,8		37,7	26,2	2,2	100
1903	10,3	1,5	0,4	9,3	0,6	8,5	0,9		38,4	27,2	3,0	100
1906	13,4	1,4	0,5	8,1	0,7	8,6	1,8		38,5	25,3	1,9	100
1909	13,5	1,1	0,5	9,0	0,6	9,0	3,4		36,8	23,5	2,6	100
1912	12,9	1,4	0,3	9,4	0,6	9,3	4,3		41,1	19,4	1,3	100
1920	12,3	0,3	0,1	11,9	2,4	3,7	1,7	1,9	42,4	23,1	0,2	100
1923	7,6	1,2	3,2	5,1	0 6	5,6	1,8	10,0	29,0	35,0	0,9	100
1926	10,5	0,3	0,4	4,5	1,8	6,6	2,1	10,1	19,9	41,6	2,3	100
1929	10,0	0,6	0,6	9,0	1,0	8,1	3,2	11,8	28,0	25,8	2,0	100
1932	10,6	0,8	0,3	8,9	3,0	10,0	6,4	13,4	27,4	19,2	0	100
1935	8,8	2,6	1,4	7,8	1,4	12,4	6,2	11,3	29,3	18,5	0,4	100
1938	10,1	1,1	1,2	5,9	1,4	9,5	6,0	5,9	40,7	18,2	0	100
1947	15,6	1,6	10,0	11,4	7,6	8,8	4,8	3,4	32,2	4,6	0	100
1950	11,2	1,3	16,5	12,4	15,8	8,1	4,9	3,4	20,7	4,2	1,5	100
1953	12,9	1,5	7,0	7,9	12,1	8,1	5,3	3,9	36,1	5,2	0	100
1956	10,9	5,3	9,7	7,6	9,6	9,4	8,5	4,6	28,2	6,2	0	100
1959	13,1	4,4	7,3	5,7	9,8	12,5	8,0	5,0	26,7	7,5	0	100
1962	12,8	4,7	7,5	6,6	6,8	15,7	9,7	5,3	24,5	6,4	0,1	100
1965	12,4	4,3	8,4	9,1	6,0	19,5	10,6	4,7	19,8	3,5	1,8	100
1968	11,4	6,0	9,5	5,9	5,0	21,1	10,7	4,2	20,0	4,4	1,9	100
1971	12,8	2,8	7,1	8,5	4,0	23,8	11,3	3,7	18,8	7,2	0	100

Source : C. André et R. Delorme, *L'évolution des dépenses publiques en France (1872-1971)* rapport CORDES, CEPREMAP, 1976.

Les individus sont les années, les variables représentent les dépenses

Un autre exemple

Nom	Pays	Type	PG	CA	MG	NA	K	SUL	NO3	HCO3	CL
Evian	F	M	P	78	24	5	1	10	3.8	357	4.5
Montagne des Pyrénées	F	S	P	48	11	34	1	16	4	183	50
Cristaline-St-Cyr	F	S	P	71	5.5	11.2	3.2	5	1	250	20
Fiée des Lois	F	S	P	89	31	17	2	47	0	360	28
Volcania	F	S	P	4.1	1.7	2.7	0.9	1.1	0.8	25.8	0.9
Saint Diéry	F	M	G	85	80	385	65	25	1.9	1350	285
Luchon	F	M	P	26.5	1	0.8	0.2	8.2	1.8	78.1	2.3
Volvic	F	M	P	9.9	6.1	9.4	5.7	6.9	6.3	65.3	8.4
Alpes/Moulettes	F	S	P	63	10.2	1.4	0.4	51.3	2	173.2	1
Orée du bois	F	M	P	234	70	43	9	635	1	292	62
Arvie	F	M	G	170	92	650	130	31	0	2195	387
Alpes/Roche des Ecrins	F	S	P	63	10.2	1.4	0.4	51.3	2	173.2	10
Ondine	F	S	P	46.1	4.3	6.3	3.5	9	0	163.5	3.5
Thonon	F	M	P	108	14	3	1	13	12	350	9
Aix les Bains	F	M	P	84	23	2	1	27	0.2	341	3
Contrex	F	M	P	486	84	9.1	3.2	1187	2.7	403	8.6
La Bondoire Saint Hippolite	F	S	P	86	3	17	1	7	19	256	21
Dax	F	M	P	125	30.1	126	19.4	365	0	164.7	156
Quézac	F	M	G	241	95	255	49.7	143	1	1685.4	38
Salvetat	F	M	G	253	11	7	3	25	1	820	4
Stamna	GRC	M	P	48.1	9.2	12.6	0.4	9.6	0	173.3	21.3
Iolh	GR	M	P	54.1	31.5	8.2	0.8	15	6.2	267.5	13.5
Avra	GR	M	P	110.8	9.9	8.4	0.7	39.7	35.6	308.8	8
Rouvas	GRC	M	P	25.7	10.7	8	0.4	9.6	3.1	117.2	12.4
Alisea	IT	M	P	12.3	2.6	2.5	0.6	10.1	2.5	41.6	0.9
San Benedetto	IT	M	P	46	28	6.8	1	5.8	6.6	287	2.4
San Pellegrino	IT	M	G	208	55.9	43.6	2.7	549.2	0.45	219.6	74.3
Levissima	IT	M	P	19.8	1.8	1.7	1.8	14.2	1.5	56.5	0.3
Vera	IT	M	P	36	13	2	0.6	18	3.6	154	2.1
San Antonio	IT	M	P	32.5	6.1	4.9	0.7	1.6	4.3	135.5	1
La Française	F	M	P	354	83	653	22	1055	0	225	982
Saint Benoit	F	S	G	46.1	4.3	6.3	3.5	9	0	163.5	3.5
Plancoët	F	M	P	36	19	36	6	43	0	195	38
Saint Alix	F	S	P	8	10	33	4	20	0.5	84	37
Puits Saint Georges/Casino	F	M	P	46	33	430	18.5	10	8	1373	39
St-Georges/Corse	F	S	P	5.2	2.3	14.05	1.15	6	0	30.5	25
Hildon bleue	GB	M	P	97	1.7	7.7	1	4	26.4	236	16
Hildon blanche	GB	M	G	97	1.7	7.7	1	4	26.4	236	16

M: minérale S: source

P: plate G: gazeuse

ions CA en mg/litre

Les individus sont les marques d'eau en bouteilles

Les variables décrivent l'étiquette de cette eau

Les données sont «hétérogènes »

Chapitre 1 : G. Saporta et N. Niang

Analyse des données

G. Govaert Hermes 2003

Exemple de tableau de données symboliques

Table

	specific gravit	freezing point	iodine value	saponification	Major Fatty aci
linseed oil	[0.93 : 0.94]	[-27.00 : -18.00]	[170.00 : 204.00]	[118.00 : 196.00]	Linol, , Oleic, Palmi, Myris
perilla oil	[0.93 : 0.94]	[-5.00 : -4.00]	[192.00 : 208.00]	[188.00 : 197.00]	Linol, , Oleic, Palmi, Seari
cotton seed oil	[0.92 : 0.92]	[-6.00 : -1.00]	[99.00 : 113.00]	[189.00 : 198.00]	Linol, Oleic, Palmi, Myris, Seari
sesame oil	[0.92 : 0.93]	[-6.00 : -4.00]	[104.00 : 116.00]	[187.00 : 193.00]	Linol, Oleic, Palmi, Seari, Arach
camellia oil	[0.92 : 0.92]	[-21.00 : -15.00]	[80.00 : 82.00]	[189.00 : 193.00]	Linol, Oleic
olive oil	[0.91 : 0.92]	[0.00 : 6.00]	[79.00 : 90.00]	[187.00 : 196.00]	Linol, Oleic, Palmi, Seari
beef tallow	[0.86 : 0.87]	[30.00 : 38.00]	[40.00 : 48.00]	[190.00 : 199.00]	Oleic, Palmi, Myris, Seari, Capri
hog fat	[0.86 : 0.86]	[22.00 : 32.00]	[53.00 : 77.00]	[190.00 : 202.00]	Linol, Oleic, Palmi, Myris, Seari, Lauri

Données d'Ichino sur les huiles

Variables ou descripteurs symboliques

- Définir les variables intervalle ou multi-valuées (distributions, courbes)
- Nous définissons, à partir de $Y_j : \Omega \to \Im_j$, par agrégation, des variables symboliques $\tilde{Y}_j : E \to D_j$ qui sont des fonctions multi-valuées

H.-H. Bock et E. Diday « Analysis of Symbolic Data » Springer, 2000.

Acquisition d'objets symboliques à partir d'une base de données

Un ensemble $\Omega=\{1,..,N\}$ de N individus, appelés ***individus de premier ordre***, décrit par une variable aléatoire vectorielle $Y=(Y_1,\ldots,Y_p)$

Une fonction $G:\Omega \to E=\{1,\cdots,K\}$

E est l'ensemble des ***concepts***.

Un ensemble $\{C_1,\ldots,C_K\}$ de classes $C_i \subset \Omega$, les éléments de cet ensemble sont considérés comme des individus dits ***individus de second ordre***, vérifiant $G(\omega)=i$ si $\omega \in C_i$

Un ***objet symbolique*** est une modélisation d'un ***concept***.

Tableau de données initial

Ce tableau de données est le résultat d'une requête sur l'ensemble Ω .

Ω	G	Y_1	...	Y_j	...	Y_p
1	$G(1)$			$Y_j(1) = x_1^j$		
...				...		
i	$G(i)$	$Y_1(i) = x_i^1$	...	$Y_j(i) = x_i^j$	...	$Y_p(i) = x_i^p$
...				...		
N	$G(N)$			$Y_j(N) = x_N^j$		

A chaque w de Ω correspond un vecteur de description $(Y_1(w),\ldots, Y_p(w))$ et une classe d'appartenance $G(w)$.

Tableau de données symboliques

Des variables symboliques $\tilde{Y}_j : E \to D_j$ sont définies à partir des variables initiales $Y_j : \Omega \to \Im_j$

E	$\tilde{Y}_1$	...	$\tilde{Y}_j$	...	$\tilde{Y}_p$
1			$\tilde{Y}_j(1) = \xi_1^j$		
...			...		
k	$\tilde{Y}_1(k) = \xi_k^1$	...	$\tilde{Y}_j(k) = \xi_k^j$	...	$\tilde{Y}_p(k) = \xi_k^p$
...			...		
K			$\tilde{Y}_j(K) = \xi_K^j$		

Dans ce cas les différentes cases ne contiennent plus une valeur unique.

Les variables symboliques

- Les variables intervalles : $\xi_k^j = \left[a_k^j, b_k^j\right]$

- Les variables multi-valuées: $\xi_k^j = \left\{m_{1,k}^j, \ldots, m_{l,k}^j\right\}$ est un sous-ensemble des modalités de la variable initiale Y_j

- Les variables modales: $\xi_k^j = \left\{(m_{1,k}^j, p_{1,k}^j), \ldots, (m_{l,k}^j, p_{l,k}^j)\right\}$ est un sous-ensemble pondéré des modalités de la variable initiale. Cette pondération peut être une distribution.

Objet Symbolique

E Diday défini un objet symbolique par le triplet (d,R,a) suivant :

- d description de l'objet, élément de l'espace de description D
- R opérateur de comparaison entre descriptions
- a fonction qui permet de définir l'extension

$$a : D \rightarrow \{0,1\} \text{ ou } [0,1]$$

2 : les méthodes de classification

- **Les méthodes de classification**
  - Introduction
  - Les méthodes de partitionnement
  - Approches « fuzzy » dans les méthodes de partitionnement

La classification automatique

La *classification non supervisée* propose la recherche de *classes homogènes* à partir d'un ensemble d'observations.

Objectif : les observations les plus semblables doivent appartenir à la même classe.

C'est un objectif très intuitif mais ce n'est pas une définition précise de la notion de *classe*.

Une classe est définie par une fonction caractéristique

La classification automatique

Les principales approches

- Il existe des classes sous-jacentes et que le défi est de les découvrir,
- il faut construire les classes dans un sens structurel, à travers les structures classificatoires,
- il faut trouver les classes utiles à l'utilisateur.
- Associer un concept à chaque classe ou un ensemble flou

Difficultés

Cette classification d'objets est réalisée à partir d'un vecteur de mesures. Ce vecteur correspond aux réponses de cet objet à un ensemble de paramètres ou variables définis a priori.

La nature multidimensionnelle de la description de ces objets présente l'une des difficultés les plus importantes dans la résolution d'un problème de classification.

En général l'information initiale s'exprime sous la forme d'un système d'hypothèses probabilistes ou sous la forme d'un critère objectif qui doit être optimisé.

Hypothèses initiales

On considère qu'un objet ou individu à classer est une entité appartenant à une population théorique constituant l'ensemble des individus susceptibles être classés.

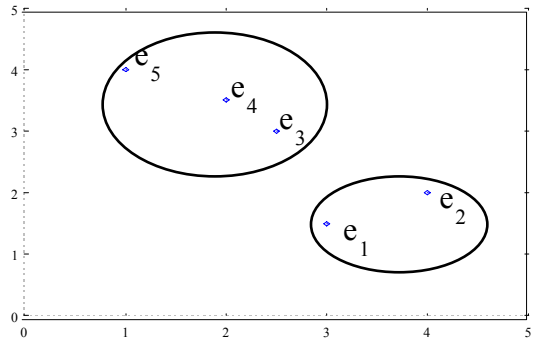
• il existe sur cette population une *structure classificatoire*;

• il existe un tableau de données,

•il existe une *distance sur ce tableau de données*. Cette «distance» mesure la proximité ces individus.

Structures classificatoires

Partition



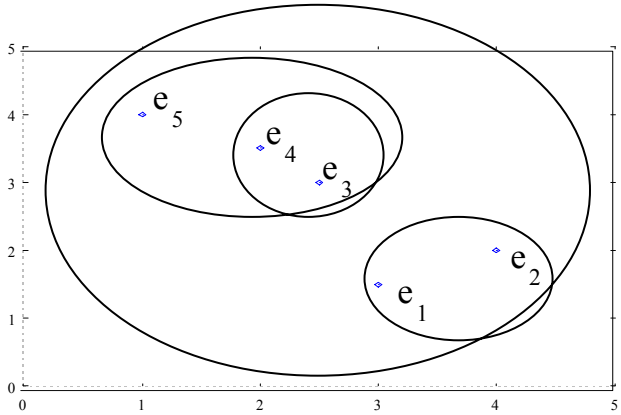
1) $\forall \ell = 1, \cdots, K$ on a $P_\ell \neq \varnothing$

2) $\bigcup_{\ell=1}^{K} P_l = E$

3) $\forall \ell, m = 1, \cdots, K$ *et* $\ell \neq m$

alors $P_\ell \cap P_m = \varnothing$

Hiérarchie



1) $E \in H$

2) $\forall e \in E$ alors $\{e\} \in H$

3) $\forall h, h' \in H$ on a :

$h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

Classe « homogène »

Classe P_k	Critère de classification	
Approche géométrique		
d distance	$w(P_k) = \sum_{e_i \in P_k} \sum_{e_\ell \in P_k} d^2(\mathbf{z_i}, \mathbf{z}_\ell)$	Graph theory
Modèle probabiliste		
$p(\mathbf{z}/\theta) = \sum_{j=1}^{K} p(\mathbf{z}/\theta_j).\pi_j$	$L(P_k/\theta_k) = \prod_{e_i \in P_k} p(\mathbf{z}_i/\theta_k)$	Mixture density
Prototype		
L_k prototype	$w(P_k, L_k) = \sum_{e_i \in P_k} D(\mathbf{z_i}, L_k)$	Iterative optimization

Méthodes de partitionnement

La structure classificatoire recherchée est la *partition*. En définissant un critère d'homogénéité sur les classes ou sur une partition le problème de classification devient un problème parfaitement défini en optimisation discrète.

Trouver, parmi l'ensemble de toutes les partitions possibles, une partition qui optimise un critère d'homogénéité défini a priori.

Ω est fini donc il y a un ensemble fini de partitions possibles alors le problème est toujours soluble par l'énumération complète. Cependant, en pratique, cette approche est irréalisable car nous avons approximativement avec un ensemble de N objets en K classes $K^N / K!$ solutions possibles.

Problème d'optimisation combinatoire

Soit un critère W, défini de $\wp_K(\Omega) \to \Re^+$, où est l'ensemble $\wp_K(\Omega)$ de toutes les partitions en K classes non vides de Ω alors le problème d'optimisation combinatoire se pose sous la forme:

$$W(P) = \underset{Q \in \wp_K(\Omega)}{Min} W(Q) = \sum_{k=1}^{K} w(Q_k)$$

où $w(Q_k)$ mesure l'homogénéité de la classe Q_k.

Optimisation itérative

- On part d'une solution réalisable $Q^{(0)} \in \wp_K(\Omega)$ ← Choix

- A l'étape *t+1*, on a une solution réalisable $Q^{(t)}$

  on cherche une solution réalisable $Q^{(t+1)} = g(Q^{(t)})$ ← Choix

  vérifiant $W(Q^{(t+1)}) < W(Q^{(t)})$

- L'algorithme s'arrête dès qu'il est impossible de trouver une solution réalisable

- Autres approches: séparation et évaluation, programmation dynamique, recuit simulé, méthode Tabou, algorithmes génétiques

Algorithme de voisinage

Une des stratégies la plus utilisée pour construire la fonction g est :

•d'associer à toute solution réalisable Q un ensemble fini de solutions réalisables $V(Q)$, appelé *voisinage* de Q,

•puis de sélectionner la solution optimale pour ce critère W dans ce voisinage, ce qui est couramment appelé *solution localement optimale*.

Par exemple on peut prendre comme voisinage de Q toutes les partitions obtenues à partir de la partition Q

en changeant un seul individu de classe (algorithme de transfert)

en permutant deux individus de classe (algorithme d'échange)

.*Le principe est simple mais peu efficace et de complexité élevée.*

Critère d'inertie intra-classe

Le critère W associé à la partition $P=(C_1,..,C_k,\ldots,C_K)$ est la somme des inerties de chacune des classes c'est-à-dire *l'inertie intra-classe*:

$$W(P)=\sum_{k=1}^{K} w(C_k)=\sum_{k=1}^{K}\sum_{i\in C_k} d^2(\mathbf{x}_i,\mathbf{w}_k)=\sum_{k=1}^{K}\sum_{i\in C_k}\left\|\mathbf{x}_i-\mathbf{w}_k\right\|^2$$

d est la distance euclidienne et $\mathbf{w}_k$ est le centre de gravité ou la moyenne de la classe C_k.

Théorème de Huygens $$\sum_{i\in C_k} d^2(\mathbf{x}_i,a)=\sum_{i\in C_k} d^2(\mathbf{x}_i,\mathbf{w}_k)+n_k d^2(\mathbf{x}_i,a)$$

Moyenne, $$\mathbf{w}_k=\arg\min_{y\in R^p}\sum_{i\in C_k} d^2(\mathbf{x}_i,\mathbf{y})=\arg\min_{y\in R^p}\sum_{i\in C_k}\left\|\mathbf{x}_i-\mathbf{y}\right\|^2$$

représentation de la classe

Algorithme des *k-means* *(MacQueen, 1967)*

Avec théorème de Huygens il suffit de trouver une solution meilleure que la solution en cours.

Aussi l'individu *i* est mis dans C_ℓ si $\ell = \arg\min_{k=1,\cdots,K} d(\mathbf{x}_i, \mathbf{w}_k)$

et les moyennes des deux classes sont mise à jour par

$$\mathbf{w}_\ell \leftarrow \frac{n_\ell . \mathbf{w}_l + \mathbf{z}_i}{n_\ell + 1} \qquad \mathbf{w}_s \leftarrow \frac{n_s . \mathbf{w}_s - \mathbf{z}_i}{n_s - 1}$$

Algorithme des centres mobiles, K-means Batch (Forgy,1965)

***(a)* initialisation**

On se donne au départ une partition P ou un sous-ensemble de K éléments de Ω.

***(b)* Étape d'affectation**

test←*0* *Pour tout i* de *1* à N faire

déterminer *l* tel que $l = \arg\min_{k=1,\cdots,K} d^2(\mathbf{x}_i, \mathbf{w}_k)$

si $\ell \neq s$ alors faire *test*←*1* $C_\ell \leftarrow C_\ell \cup \{i\}$ et $C_s \leftarrow C_s - \{i\}$

***(c)* Étape de représentation**

Pour tout k de 1 à K faire

calculer le centre de gravité et l'effectif de la nouvelle classe C_k

(d) si *test #0* (partition inchangée) aller en *(b) sinon stop*

Affectation d'un nouvel individu

Une *fonction d'affectation* ϕ de D dans $\{1,..,K\}$ définit une partition $F_\Phi = \{F_1, \cdots, F_K\}$ de l'espace de représentation avec

$$F_j = \{\mathbf{x} \in D / \Phi(\mathbf{x}) = j\}$$

A la convergence de ces algorithmes, la fonction ϕ est construite de la manière suivante :

$$\forall \mathbf{x} \in D \; \Phi(\mathbf{x}) = j \text{ si } d(\mathbf{x}, \mathbf{w}_j) = \underset{k=1,K}{Min}\{d(\mathbf{x}, \mathbf{w}_k)\}$$

Remarquons qu'il est impossible de démontrer que l'une des stratégies donne systématiquement une meilleure solution.

Ensemble flou

- Une classe C d'un ensemble Ω est usuellement associée à sa fonction caractéristique. Celle-ci s'applique sur les individus x de Ω. Elle prend la valeur 0 si x n'appartient pas à C et 1 si x appartient à C.
- Une classe floue C est caractérisée par une fonction d'appartenance, notée $\mu_{C,,}$, qui représente le degré de validité de la proposition « *x* appartient à C ». C'est une application de Ω dans [0,1].
- Pour un individu *x* donné, la valeur de la fonction d'appartenance $\mu_C(x)$ est appelée degré d'appartenance de l'individu *x* à la classe C.
- Si $\mu_C(x) = 1$, l'objet *x* appartient totalement à C, et si $\mu_C(x) = 0$, il ne lui appartient pas du tout.

.

Fuzzy c-means (Bezdek,1981)

Le critère W_F associé à la partition P est maintenant

$$W_F(P) = \sum_{k=1}^{K} w(C_k) = \sum_{k=1}^{K}\sum_{i=1}^{N} \mu_{C_k}(\mathbf{x}_i)^q d^2(\mathbf{x}_i, \mathbf{w}_k) = \sum_{k=1}^{K}\sum_{i=1}^{N} \mu_{C_k}(\mathbf{x}_i)^q \left\|\mathbf{x}_i - \mathbf{w}_k\right\|^2$$

$$\text{avec } q > 1 \text{ et } \sum_{k=1}^{K} \mu_{C_k}(\mathbf{x}_i) = 1 \quad \mu_{C_k}(\mathbf{x}_i) \in [0,1] \; \forall i \in \Omega$$

Hard

$$\mu_{C_k}(\mathbf{x}_i) \in \{0,1\}$$

$$\begin{pmatrix} \mu_{C_1}(\mathbf{x}_1) & \cdots & \mu_{C_k}(\mathbf{x}_1) & \cdots & \mu_{C_K}(\mathbf{x}_1) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_{C_1}(\mathbf{x}_i) & \cdots & \mu_{C_k}(\mathbf{x}_i) & \cdots & \mu_{C_K}(\mathbf{x}_i) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_{C_1}(\mathbf{x}_N) & \cdots & \mu_{C_k}(\mathbf{x}_N) & \cdots & \mu_{C_K}(\mathbf{x}_N) \end{pmatrix}$$

$$\sum_{k=1}^{K} \mu_{C_k}(\mathbf{x}_i) = 1 \quad \forall i \in \Omega$$

Fuzzy

$$\mu_{C_k}(\mathbf{x}_i) \in [0,1]$$

Algorithme de Fuzzy c-means (Bezdek,1981)

***(a)* Initialisation**

On se donne au départ un sous-ensemble de K éléments de Ω.

***(b)* Étape d'affectation floue**

Pour i de 1 à N et Pour k de 1 à K

calculer les degrés d'appartenance $\mu_{C_k}(\mathbf{x}_i) = \dfrac{(1/d^2(\mathbf{x}_i,\mathbf{w}_k))^{1/(q-1)}}{\sum_{k=1}^{K}(1/d^2(\mathbf{x}_i,\mathbf{w}_k))^{1/(q-1)}}$

(c) Étape de représentation

Pour k de 1 à K

calculer la moyenne pondérée $\mathbf{w}_k = \dfrac{\sum_{i=1}^{N}\mu_{C_k}(\mathbf{x}_i)^q \mathbf{x}_i}{\sum_{i=1}^{N}\mu_{C_k}(\mathbf{x}_i)^q}$

(d) si *les moyennes varient beaucoup* aller en *(b) sinon stop*

Qualité de la partition

Le coefficient de Dunn permet de mesurer la proximité de la partition floue par rapport à la partition dure.

$$F_K = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} \mu_{C_k}(\mathbf{x}_i)^2$$

Si ce coefficient est égal à 1 alors cette partition floue est une partition dure. Plus ce coefficient est petit plus cette partition est floue. Le minimum est obtenu avec

$$\mu_{C_k}(\mathbf{x}_i) = 1/K \text{ et } F_K = 1/K$$

Le coefficient de Dunn peut être utilisé comme critère d'arrêt.

Algorithme EM (Dempster et al,1977).

Initialisation $\Theta=(\pi_1,\ldots,\pi_K,\theta_1,\ldots,\theta_K)$ et K fixés. $\theta_1,\ldots,\theta_K$ dépendent des hypothèses sur les lois

Étape d'Estimation : calcul des probabilités a posteriori et a priori

Pour i=1,..,N et k=1,..,K on calcule

$$c_{ik} = \Pr[y = k / x_i; \Theta] = \pi_k f_k(x_i / \theta_k) / f(x_i)$$

$$\pi_k = \sum_{i=1}^{N} \Pr[y = k / x_i; \Theta] \Big/ N$$

Étape de Maximisation : calcul de Θ qui maximise

$$\Theta \leftarrow \arg\max_{\Gamma} Q(\Gamma, \Theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} c_{ik} . \log[\pi_k . f_k(x_i / \Gamma)]$$

Distributions Gaussiennes

$$c_{ik} = \Pr[y = k / x_i; \Theta] = \pi_k f_k(x_i / \theta_k) / f(x_i)$$

avec $\theta_k = (\mu_k, \Sigma_k)$

moyenne

$$\mu_k = \frac{\sum_{i=1}^{N} c_{ik} . x_i}{\sum_{i=1}^{N} c_{ik}}$$

Matrice de variance-covariance

$$\Sigma_k = \frac{\sum_{i=1}^{N} c_{ik} . (x_i - \mu_k)(x_i - \mu_k)}{\sum_{i=1}^{N} c_{ik}}$$

Tableau de données complexes

	Pulse Rate	Systolic pressure	Diastolic pressure
1	[60, 72]	[90,130]	[70,90]
2	[70,112]	[110,142]	[80,108]
3	[54,72]	[90,100]	[50,70]
4	[70,100]	[130,160]	[80,110]
5	[63,75]	[60,100]	[140,150]
6	[44,68]	[90,100]	[50,70]

Chaque objet *i* est décrit par un vecteur d' intervalles

$$\mathbf{x}_1 = ([60,72],...,[70,90])$$

Méthode des Nuées Dynamiques (Diday, 1971)

Cet algorithme possède deux étapes d'optimisation

•La première étape est l'étape de *représentation*, elle consiste à définir un **représentant** ou **prototype** pour chacune des classes. L'idée est d'associer à chaque classe un représentant, par exemple, centre de gravité, un individu, une droite, une loi de probabilité

•La seconde étape est l'étape d'*affectation*, elle va modifier la classe d'affectation de chacun des individus de Ω.

Moyenne, $\mathbf{w}_k = \arg\min_{y \in R^p} \sum_{i \in C_k} d^2(\mathbf{x}_i, \mathbf{y}) = \arg\min_{y \in R^p} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{y}\|^2$

représentation de la classe

Méthode des Nuées Dynamiques (Diday, 1971)

(a) Initialisation

Choisir K prototypes $L_1,...,L_K$ distincts de Λ espace des prototypes.

(b) Étape d'affectation

Pour chaque objet i de Ω calculer l'indice l de la classe d'affectation qui vérifie $l = \arg\min_{k=1,...,K} d(\mathbf{x}_i, L_k)$

(c) Étape de représentation

Pour chaque classe k rechercher le prototype L_k de Λ qui minimise $L_k = \arg\min_{L \in \Lambda} \sum_{i \in C_k} d^q(\mathbf{x}_i, L)$

Répéter (b) et (c) jusqu'à la convergence

Distances entre deux vecteurs d'intervalles

Nous proposons deux approches:

- La distance City-Block
- La distance Euclidienne
- La distance de Hausdorff

Comment construire les prototypes pour un ensemble de vecteurs d'intervalles ?

Distance entre deux vecteurs d'intervalles

Λ *l'ensemble des intervalles*

$$\mathbf{x}_1 = (x_1^1,..., x_1^j,..., x_1^p) = ([a_1^1, b_1^1],..., [a_1^p, b_1^p]) \in \Lambda^p$$

La distance d est définie par:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} d_j(x_1^j, x_2^j)$$

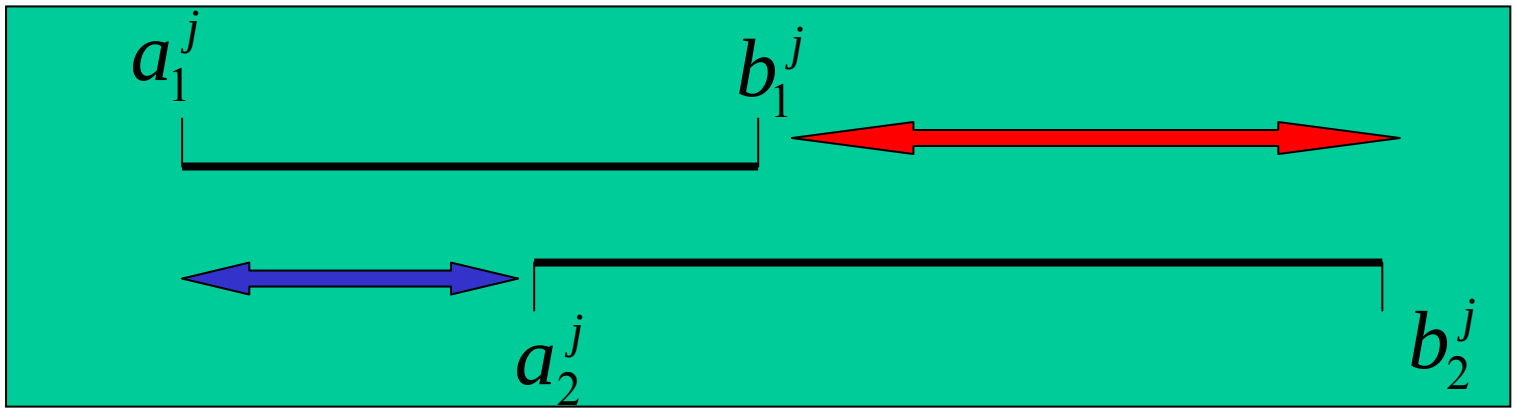
- *d est une distance si et seulement si toutes les fonctions* d_j *sont des distances*

Distance euclidienne entre deux intervalles

Pour chaque variable j la distance d entre deux intervalles

$x_1^j = [a_1^j, b_1^j]$ ***et*** $x_2^j = [a_2^j, b_2^j]$ ***est égale à***

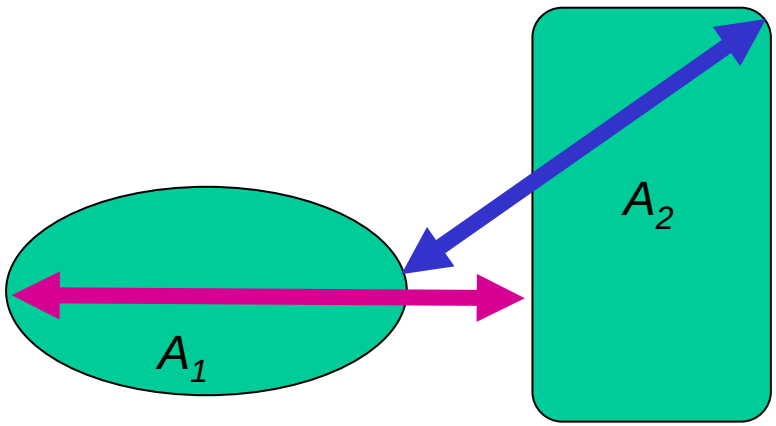
$$d_j^q(x_1^j, x_2^j) = \left|a_1^j - a_2^j\right|^q + \left|b_1^j - b_2^j\right|^q$$



Distance de Hausdorff entre deux ensembles

La distance de Hausdorff d_H entre deux ensembles A_1 et A_2 of $\Re^p$ est:

$$d_H(A_1, A_2) = max\left\{ \sup_{x \in A_1} \inf_{y \in A_2} d(x,y),\ \sup_{y \in A_2} \inf_{x \in A_1} d(x,y) \right\}$$



d_H vérifie les trois 3 propriétés d'une distance

(1) $d(\mathbf{x},\mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$

(2) $\forall \mathbf{x},\mathbf{y}\ d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$

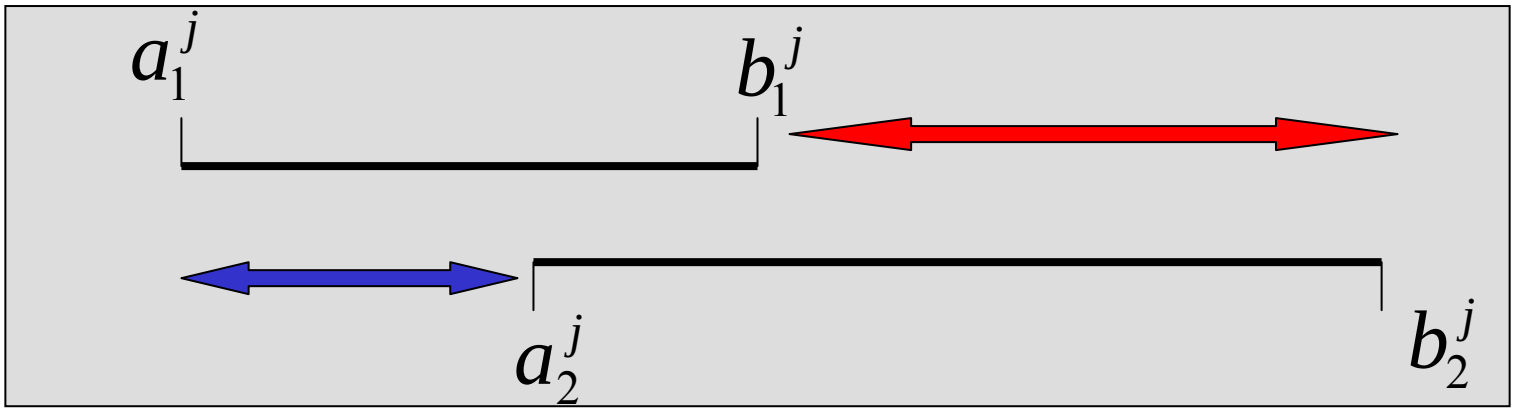
(3) $\forall \mathbf{x},\mathbf{y},\mathbf{z}\ d(\mathbf{x},\mathbf{y}) \le d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$

Distance de Hausdorff entre deux intervalles

$$x_1^j = [a_1^j, b_1^j] \qquad x_2^j = [a_2^j, b_2^j]$$

La distance de Hausdorff est définie par :

$$d_H(x_1^j, x_2^j) = \max(\left| a_1^j - a_2^j \right|, \left| b_1^j - b_2^j \right|)$$



Les prototypes

Le vecteur des prototypes $\boldsymbol{L}_k = (L_k^1,\ldots,L_k^p)$ où $L_k^j = [\alpha_k^j, \beta_k^j]$

Les bornes α_k^j, β_k^j sont définie par :

- La distance City-block :
  $\alpha_k^j = médiane\ \{a_i^j : i \in C_k\}$ et $\beta_k^j = médiane\ \{b_i^j : i \in C_k\}$
- La distance Euclidienne:
  $\alpha_k^j = moyenne\ \{a_i^j : i \in C_k\}$ and $\beta_k^j = moyenne\ \{b_i^j : i \in C_k\}$
- La distance de Hausdorff : $\alpha_k^j = \mu_k^j - \rho_k^j$ and $\beta_k^j = \mu_k^j + \rho_k^j$ où
  $\mu_k^j = médiane\ \{a_i^j + b_i^j) / 2 : i \in C_k\}$ et
  $\rho_k^j = médiane\ \{b_i^j - a_i^j) / 2 : i \in C_k\}$

Classification floue et données intervalles

Le critère W_F associé à la partition Q est maintenant

$$W_F(P) = \sum_{k=1}^{K} w(C_k) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mu_{C_k}(\mathbf{x}_i)^q d^2(\mathbf{x}_i, \mathbf{w}_k)$$

$$\text{avec } q > 1 \text{ et } \sum_{k=1}^{K} \mu_{C_k}(\mathbf{x}_i) = 1 \quad \mu_{C_k}(\mathbf{x}_i) \in [0,1] \forall i \in \Omega$$

Ce n'est pas toujours possible car le calcul du prototype est lié aux *degrés d'appartenance* $\mu_{C_k}(\mathbf{x}_i)$

Possible pour la moyenne, impossible pour la médiane.

Une solution est de prendre comme ensemble des prototype Λ l'ensemble Ω.

Classification d'un tableau de distances

(a) Initialisation

Choisir K objets $L_1,...,L_K$ distincts de Ω espace des prototypes.

(b) Étape d'affectation

Pour chaque objet i de Ω calculer l'indice l de la classe d'affectation qui vérifie $l = \arg\min_{k=1,...,K} d(\mathbf{x}_i, L_k)$

(c) Étape de représentation

Pour chaque classe k rechercher l'objet L_k de Ω qui minimise

$$L_k = \arg\min_{L \in \Omega} \sum_{i \in C_k} d^q(\mathbf{x}_i, L)$$

Répéter (b) et (c) jusqu'à la convergence

Approche MNDD ou PAM (Kaufman et Roosseuw, 1990)

FANNY (Kaufman et Roosseuw, 1990)

Le critère W_F associé à la partition P est maintenant

$$W_F(P) = \sum_{k=1}^{K} w(C_k) = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_{C_k}(\mathbf{x}_i)^2 \mu_{C_k}(\mathbf{x}_j)^2 d(\mathbf{x}_i, \mathbf{x}_j)}{2\sum_{i=1}^{N} \mu_{C_k}(\mathbf{x}_i)^2}$$

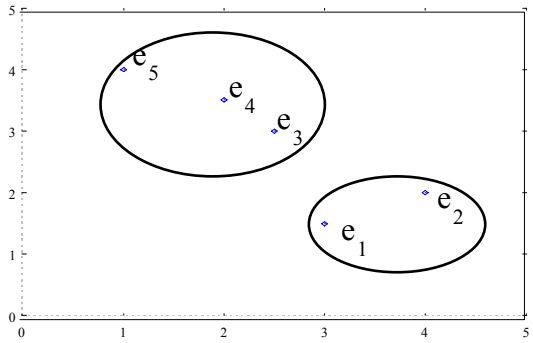
$$\text{avec } q > 1 \text{ et } \sum_{k=1}^{K} \mu_{C_k}(\mathbf{x}_i) = 1 \quad \mu_{C_k}(\mathbf{x}_i) \in [0,1] \; \forall i \in \Omega$$

3 : les structures de classification

- **Les structures de classification**
  - Recouvrement
  - Pyramide

Structures classificatoires

Partition



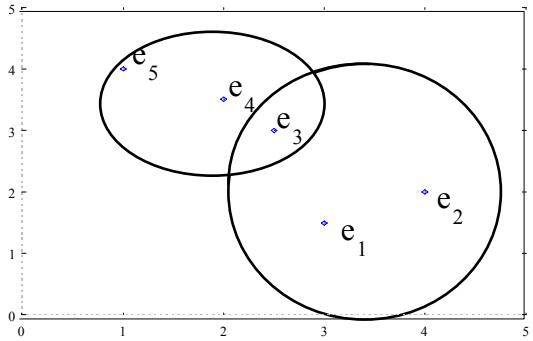
1) $\forall \ell = 1, \cdots, K$ on a $P_\ell \neq \varnothing$

2) $\bigcup_{\ell=1}^{K} P_l = E$

3) $\forall \ell, m = 1, \cdots, K$ *et* $\ell \neq m$

alors $P_\ell \cap P_m = \varnothing$

Recouvrement



1) $\forall \ell = 1, \cdots, K$ on a $P_\ell \neq \varnothing$

2) $\bigcup_{\ell=1}^{K} P_l = E$

Retour aux sources

Cuvier :

Pour qu'une partition soit bonne il faut que deux objets pris dans la même classe se ressemblent plus que deux objets pris dans deux classes différentes

Construire une partition *P* qui vérifie les deux conditions :

- si $d(\mathbf{x},\mathbf{y}) < \alpha$ alors **x** et **y** doivent être mis dans la même classe
- si $d(\mathbf{x},\mathbf{y}) > \alpha$ alors **x** et **y** doivent être mis dans deux classes différentes

Exemple : $d(A,B)=2$, $d(B,C)=3$ et $d(A,C)=4$

Pour $\alpha = 3$ il est impossible de construire une partition

Solution

Si d est une distance ultramétrique

$$(4)\ \forall \mathbf{x},\mathbf{y},\mathbf{z}\ d(\mathbf{x},\mathbf{y}) \le \mathrm{Max}\{d(\mathbf{x},\mathbf{z}), d(\mathbf{z},\mathbf{y})\}$$

alors $\mathbf{x}R\mathbf{y} \Leftrightarrow d(\mathbf{x},\mathbf{y}) \le \alpha$ est une relation d'équivalence

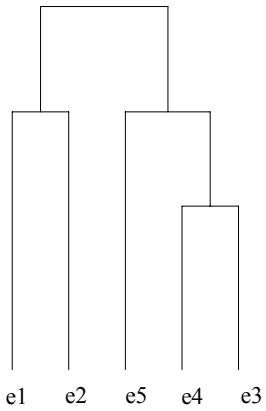
Solution :

Construire un espace ultramétrique

(méthodes hiérarchiques)

Structures classificatoires

Hiérarchie



1) $E \in H$

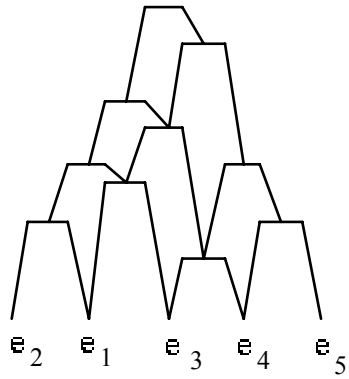
2) $\forall e \in E$ alors $\{e\} \in H$

3) $\forall h, h' \in H$ on a :

$h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

Pyramide

Hiérarchie faible

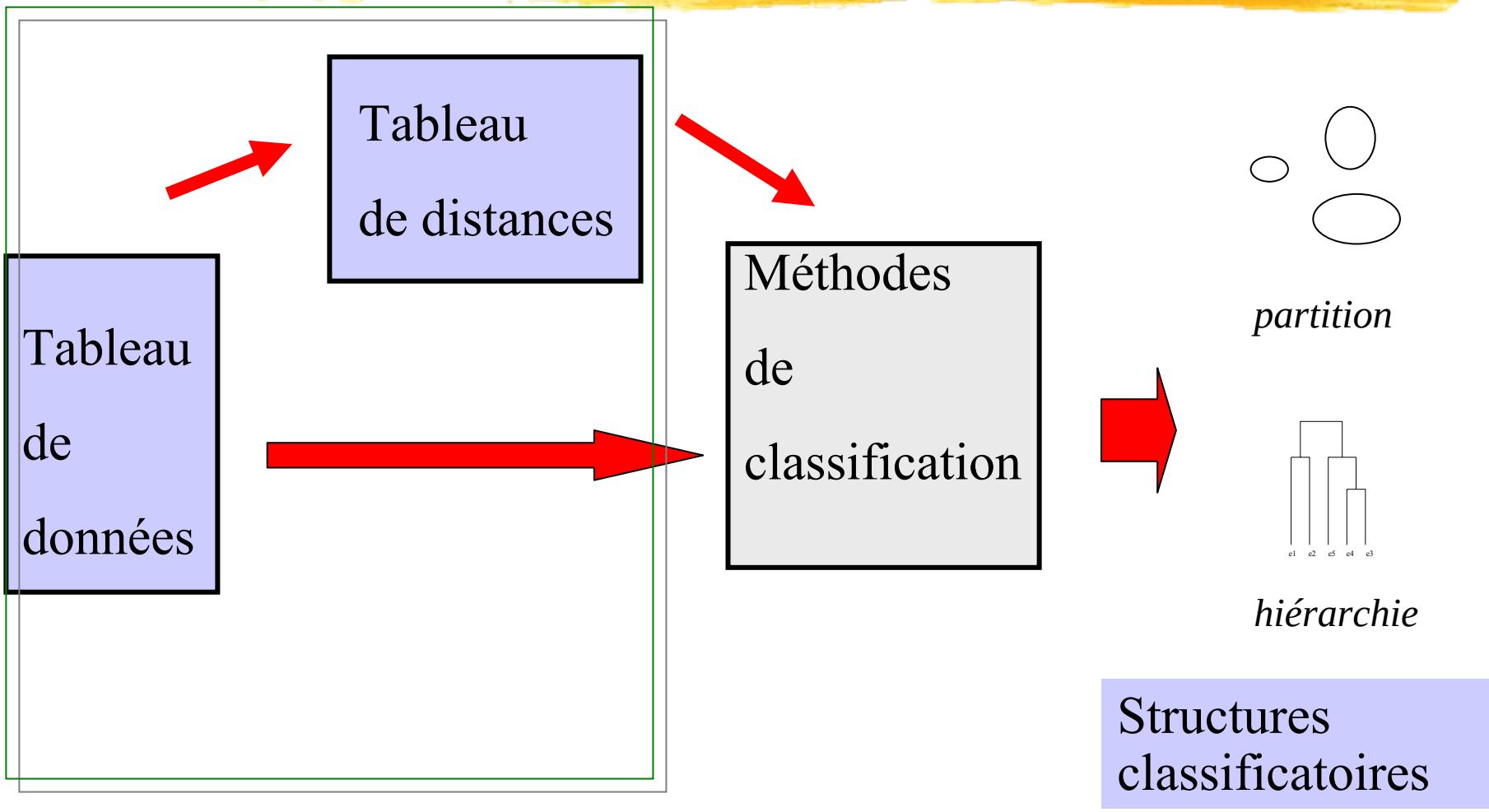


3) $\forall h, h' \in H$ on a $h \cap h' = \varnothing$ ou $h \cap h' \in H$

4) Il existe un ordre θ tel que

$\forall h \in H$, h est un intervalle de θ

La classification automatique



Yves Lechevallier
INRIA-Rocquencourt
78153 Le Chesnay Cedex
E_mail : Yves.Lechevallier@inria.fr

MERCI