Chapitre 5

Arbre des classifications

I - INTRODUCTION

De manière intuitive, un algorithme est un procédé de calcul mécaniquement réalisable. Nous avons vu au chapitre 1 l'exemple de deux algorithmes (des "Transferts" et des "Nuées dynamiques") de recherche d'une partition. Ici, le but du calcul est de déterminer une chaîne de partitions, représentable par un arbre de classifications, qui respecte de la manière la plus satisfaisante les ressemblances entre éléments de l'ensemble à classifier qui peut être l'ensemble des objets ou, d'abord l'ensemble des variables. Dans ce derniers cas, la technique offre une approche non-linéaire, sensiblement plus souple que celle de l'Analyse Factorielle, pour dégager les "dimensions" sous-jacentes au comportement de la population étudiée.

Nous avons déjà souligné que dans la définition du couple (Critère, Algorithme), on pouvait faire prévaloir deux attitudes. La première consiste dans l'établissement d'un bon critère et la conception d'un algorithme qui optimise au mieux le critère défini ; dans ce cas, généralement, chaque pas de l'algorithme optimise la part locale du critère (global) qui se trouve affectée. Dans la seconde attitude on voit apparaître un algorithme "naturel", lequel pouvant d'ailleurs faire intervenir un critère essentiellement local et on cherche s'il existe un critère global que cet algorithme optimise ; ou bien,on soumet les résultats de l'algorithme à un critère sûr.

Dans la première partie de nos recherches, nous avons eu le souci d'établir une technique de recherche d'une hiérarchie de classifications qui soit très liée à cette précieuse donnée de base qu'est la préordonnance. Nous avons ainsi abouti à un algorithme dont le caractère naturel est exprimé au moyen d'une propriété d'optimalité relativement au critère d'ordre lexicographique vu au chapitre précédent. Nous nous sommes vite rendus compte que notre algorithme était analogue à celui de Sneath (1957) ; il restait néanmoins plus systématique en ne tenant compte que de la préordonnance de départ. A également le même caractère systématique l'algorithme de M. Roux (cf. [8]) dit de l'"Ultramatrique Inférieure Maxima", dont les résultatssont équivalents, mais conçus différemment à la suite d'un travail de N. Jardine et R. Sibson (cf. [3]) dans un cadre, à notre goût, par trop métrique.

D'ailleurs, nous avons tenu à désigner notre premier algorithme au moyen du critère qu'il optimise, d'où le nom d'algorithme "lexicographique", pour insister sur le caractère ordinal et non métrique du critère optimisé.

Nous verrons ; c'est trivial, mais ce n'est pas apparu immédiatement, que cet algorithme revient à réunir à chaque pas les deux classes les plus proches au sens de la plus grande valeur de la similarité observée entre deux éléments appartenant respectivement aux deux classes.

Dans ces conditions et dans le cas important où l'ensemble D à classifier est l'ensemble V des variables descriptives, il est naturel, pour définir la proximité entre deux classes de variables d'adopter le même principe qui a prévalu à la définition de la proximité entre deux variables où, rappelons-le, partant d'un indice brut de proximité, on considère, pour juger de sa grandeur relative, sa distribution dans une hypothèse adéquate N d'absence de lien (cf. Chap. 2). Deux indices bruts de proximité entre deux classes B et C apparaissent ; le premier, mis en évidence dans l'algorithme "lexicographique", est la plus grande proximité observée entre un élément de B et un élément de C. Le second est la somme étendue sur B x C des proximités. Une fois associés à chacun de ces deux indices de base celui, statistiquement "significatif", on considérera l'algorithme, définissant un arbre détaillé des classifications, qui, à chaque pas réunit les deux classes les plus proches. Cette question sera développée au paragraphe III, alors que le paragraphe II sera consacré à l'étude de l'algorithme "lexicographique" dont l'intérêt, compte tenu de l'état actuel de développement de la méthode, est surtout théorique.

S'il s'agit d'établir un arbre de classifications sur l'ensemble E des individus ou bien sur l'ensemble des lignes (resp. colonnes) d'une juxtaposition de tableaux de contingence, le problème se ramène au précédent compte tenu de l'extension de notre indice de proximité entre variables (cf. Chap. 2).

Dans le cas où les variables descriptives sont numériques ou ordinales, pour définir la proximité entre deux classes d'objets, on peut considérer la partie locale qui se trouve affectée, par la réunion des deux classes, d'un critère global d'adéquation de la partition. Ce critère global peut être l'inertie expliquée s'il s'agit de variables numériques (cf. § IV Chap.6) ou bien, quelque soit le type de variable, le critère général basé sur l'ordonnance que nous avons établie. Cette approche sera considérée au paragraphe IV.

Dans le cas particulier, mais important, où le tableau de données est un tableau de contingence croisant deux partitions ayant chacune un grand nombre de modalités, on optimisera également, dans la réunion de deux classes, la part locale d'un critère global ; lequel peut être l'inertie expliquée au sens de la métrique de χ^2 ou bien la quantité d'Information (Benzecri, Orloci, 1968) (cf. Chap.6 § III et IV). C'est également au paragraphe IV qu'on explicitera les quantités critères utilisés.

Au paragraphe V nous montrerons comment notre critère basé sur l'ordonnance permet son interprétation dynamique de l'arbre des classifications et la reconnaissance de ses noeuds statistiquement significatifs ; ce qui permet la condensation de l'arbre à ses niveaux les plus pertinents.

II - ALGORITHME "LEXICOGRAPHIQUE"

Nous désignerons ici par E l'ensemble à classifier sur lequel est supposé donnée la préordonnance ω. Compte tenu de la représentation d'une chaîne de partitions au moyen d'une préordonnance ultramétrique (cf. Chap. 0 § IV.3) ; nous commencerons par définir une telle préordonnance ultramétrique ω_u sur E, liée à ω au moyen d'une fonction ordinale H sur F pour laquelle[u]:

$$H(x,y)\leqslant r \quad \text{et} \quad H(y,z)\leqslant r \Longrightarrow H(x,z)\leqslant r$$

pour tout x,y et z de E. Nous exposerons ensuite l'algorithme, qui détermine ce préordre total sur F par ses sections commençantes (qui sont des parties saturées de F) et démontrerons son caractère optimal. Nous terminerons en nous intéressant à l'étude de l'algorithme dans le cas où la donnée est une ordonnance sur E (ordre total sur F).

1- DEFINITION D'UNE PREORDONNANCE ULTRAMETRIQUE ω_u LIEE A ω.

ω est la préordonnance donnée sur E et $\rho(p)$ désigne le rang pour ω d'un élément p quelconque de F ; c'est-à-dire, rappelons-le,

$$\text{card}\{q\in F / q\leqslant p \text{ pour } \omega\}$$

c(x,y) indiquera une chaîne finie d'éléments de E d'origine x et d'extrémité y telle que $(z_1, z_2, \ldots, z_k)$; $z_1=x$, $z_k=y$ et $z_i\neq z_j$ pour tout $i\neq j$

$\mathcal{C}(x,y)$ sera l'ensemble de toutes les chaînes d'origine x et d'extrémité y.

Enfin H(x,y) désignera une fonction ordinale sur F compatible avec la préordonnance ultramétrique ω_u à déterminer, c'est-à-dire telle que :

$$H(p)\leqslant H(q) \Longleftrightarrow p\leqslant q$$

pour ω_u ; pour tout p et q de F.

1.1. Définition de H(x,y)

Les opérations de maximum et de minimum dans N étant représentées par V et Λ, nous définissons

$$\phi(c(x,y)) = \rho(z_1,z_2)V\rho(z_2,z_3)V\ldots V\rho(z_{(k-1)},z_k)$$

où $c(x,y)=(z_1,z_2,\ldots,z_k)$.

On posera

$$(1)\quad H(x,y) = \bigwedge_{c\in\mathcal{C}(x,y)} (c(x,y))$$

$H(x,y)=\rho(x,y)$ si le minimum de $\phi(c(x,y))$ est atteint pour la chaîne réduite à (x,y).

Nous allons montrer que H(x,y) définit bien une préordonnance ultramétrique, c'est-à-dire

1.2. PROPOSITION

$H(x,y)\leqslant r$ et $H(y,z)\leqslant r \Longrightarrow H(x,z)\leqslant r$ *pour tout* x, y *et* z *de* E

$H(x,y)\leqslant r$ signifie qu'il existe une chaîne c(x,y) pour laquelle $\phi(c(x,y))\leqslant r$ soit par exemple

$$c(x,y)=(u_1,u_2,\ldots,u_k)$$

$$u_1=x, u_k=y \text{ et } u_i \neq u_j \text{ pour } i \neq j$$

$H(y,z) \leqslant r$ signifie qu'il existe une chaîne $c(y,z)$ pour laquelle $\phi(c(y,z)) \leqslant r$ soit par exemple

$$c(y,z)=(v_1,v_2,\ldots,v_h) \quad v_1=y, v_h=z \text{ et } v_i \neq v_j \text{ pour } i \neq j$$

Désignons par u_ℓ le premier élément de la suite $c(x,y)$ qui se retrouve dans la suite $c(y,z)$; cet élément existe nécessairement puisque les deux suites ont en commun y.

Soit $u_\ell = v_m$. Dans ces conditions la suite :

$$(u_1,u_2,\ldots,u_\ell,v_{m+1},v_{m+2},\ldots,v_h)$$

est un élément $c(x,z)$ de $\mathcal{C}(x,z)$. En vertu de la définition de ϕ et des hypothèses, on a pour ce $c(x,z)$, $\phi(c(x,z)) \leqslant r$.

Donc à fortiori $H(x,z) \leqslant r$.

2- ALGORITHME PERMETTANT D'OBTENIR LA PREORDONNANCE ULTRAMETRIQUE ω_u DEFINIE PAR H.

Soient $B_1, B_2, \ldots, B_p$, les différentes classes du préordre total ω, avec $B_1 < B_2 < \ldots < B_p$ pour l'ordre total quotient et soit $C_1, C_2, \ldots, C_p$, la suite obtenue de la façon suivante :

$$C_1=B_1,\ C_2=B_1 \cup B_2, \ldots, C_p=B_1 \cup B_2 \ \ldots \cup B_p$$

Pour tout i, C_i est une section commençante de F pour ω

Notons $\overline{C}_i$ la saturée de C_i, (rappelons que si A est une partie de F, la saturée $\overline{A}$ est l'ensemble des paires, $\{x,y\}$, telles que : $\{x,y\} \in \overline{A}$ si $\{x,y\} \in A$ ou $(\exists z)$: $\{x,z\} \in A$ et $\{y,z\} \in A$).

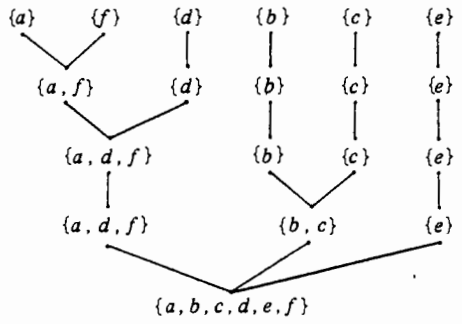
Soit q le plus grand indice pour lequel la saturée $\overline{C}_q$ est identique à $F(\overline{C}_q=F)$; la suite des saturées : $\overline{C}_1, \overline{C}_2, \ldots, \overline{C}_q$ est une suite croissante au sens de l'inclusion, puisque $C_i \subset C_j$ pour $i<j$. Cette suite est, précisément, celle des différentes sections commençantes de la préordonnance ultramétrique ω_u définie par $H(x,y)$.

Exemples. Pour illustrer le fonctionnement de l'algorithme ; nous nous proposons de l'appliquer pour le tableau d'incidence (ch.2 § III.1) sur la base de chacune des préordonnances $\underline{\omega} = \omega_s$, $\omega = \omega_{(s+t)}$ (ch.2 § III.2), ω^1 (la préordonnance "la plus fine", ω^2 (la préordonnance "la moins fine" (ch.2 § III.3).

1' Application de l'algorithme à $\underline{\omega}$

af < *ad* = *df* < *bc* < *ac* = *ef* < *ab* = *bd* = *cd* = *be* = *ce* = *de* = *bf* < *cf* = *ae*														
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	3	4	4	4	4	4	4	4	4	4	4	4	4

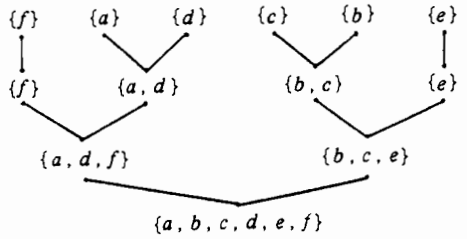
Le vecteur de composantes 1 ou 0 du niveau i représente la fonction caractéristique de $\overline{C}_i$. le vecteur de la dernière ligne représente la fonction ordinale qui définit la préordonnance ultramétrique ω_u associée à $\underline{\omega}$. Une quelconque des composantes de ce vecteur est obtenue en retranchant à 5 (nombre maximum de niveaux - 1) le nombre de 1 de la colonne correspondante. La chaîne de partitions associée est la suivante :



2' Application de l'algorithme à $\overline{\omega}$.

ad = bc < be = af = df < de < bd = ce = ef < ab = ac = cd < ae = bf < cf														
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	1	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	3	3	3	4	4	3	4	4	4	4	4	4	4

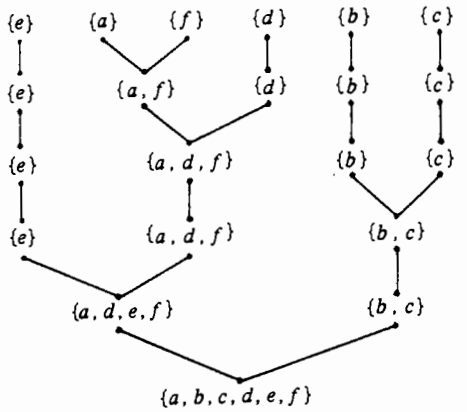
La chaîne de partitions associée est la suivante :



3' Application de l'algorithme à la préordonnance "la plus fine"

af < ad < df < bc < ef < ac < be < de < ce < bd < ab < cd < bf < ae < cf														
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	1	0	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	2	3	4	4	3	4	4	4	4	4	3	4

La chaine de partitions associée est la suivante :

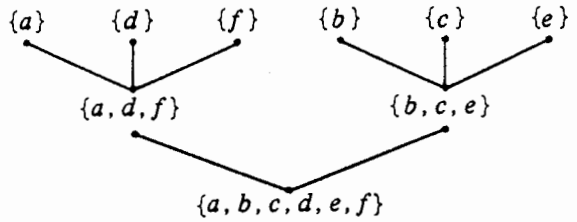


La préordonnance "la plus fine" étant une ordonnance, la chaine de partitions associée est élémentaire [cf. § ci-dessus].

4' Application de l'algorithme à la préordonnance "la moins fine".

$ad = af = bc = be = df < ab = ac = ae = bd = bf = cd = ce = de = ef < cf$

1	1	1	1	1	0	0	0	0	0	0	1	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	3	3	3	3	4	4	4	4	4	4	3	4	4	4



2.3. THEOREME *(Propriété optimale de la préordonnance ultramétrique* ω_u*)*

Parmi les préordonnances ultramétriques de même type que ω_u, ω_u *est la plus voisine de* ω *au sens du critère d'ordre lexicographique,* [cf. Ch.4 § II A.4.2]. *De plus toute préordonnance ultramétrique dont la suite des sections commençantes,*

$$R_1, R_2, \ldots, R_i, \ldots,$$

satisfait la condition:

$$C_1 \subset R_1, C_2 \subset R_2, \ldots, C_i \subset R_i, \ldots,$$

est telle que :

$$\overline{C}_1 \subset R_1, \overline{C}_2 \subset R_2, \ldots, \overline{C}_i \subset R_i, \ldots,$$

Preuve

Soit $t = (t_1, t_2, \ldots, t_q)$ le type de ω_u ; on a :

$$\sum_{i=1}^{j} t_i = |\overline{C}_j| > |C_j|$$

Désignons par $(T_1, T_2, \ldots, T_q)$ la suite des sections commençantes d'un préordre ultramétrique, quelconque, sur F, de type t. Parmi les préordonnances ultramétriques de type t, toute celle pour laquelle $C_1 \subset T_1$ est plus proche de ω que de toute celle pour laquelle ce n'est pas le cas. D'autre part si $C_1 \subset T_1$, alors $\overline{C}_1 \subset T_1$ donc $\overline{C}_1 = T_1$ puisque $|\overline{C}_1| = |T_1|$.

De nouveau, parmi les préordonnances ultramétriques de type t, pour lesquelles $T_1 = \overline{C}_1$, toute celle pour laquelle $C_2 \subset T_2$ est plus proche de ω que toute celle pour laquelle ce n'est pas le cas. D'autre part si $C_2 \subset T_2$, alors $\overline{C}_2 \subset T_2$, donc $\overline{C}_2 = T_2$ puisque $|\overline{C}_2| = |T_2|$ et ainsi de suite... d'où le résultat annoncé.

La deuxième partie de la proposition est évidente, elle exprime une propriété de "finesse" de ω_u. $\overline{C}_1$ est la première section commençante, la plus petite (au sens de la cardinalité), pour une préordonnance ultramétrique dont la première section commençante contient C_1 ; $\overline{C}_2$ est la deuxième section commençante la plus petite pour une préordonnance ultramétrique dont la première section est $\overline{C}_1$ et dont la seconde contient C_2 ; etc.

Il résulte de la définition de l'algorithme la

2.4. PROPOSITION

La condition nécessaire et suffisante pour que la préordonnance ultramétrique ω_u, associée à ω par l'algorithme, soit comparable ((1)) *avec ω est que la saturée de toute section strictement commençante (pour ω), C_i, soit une section commençante, au sens large, (pour ω).*

Nous voulons dire que pour tout i, il existe h, tel que :

$$C_i = C_{i+h} \cup A_{i+h+1}$$

où A_{i+h+1} est une partie de B_{i+h+1}.

Nous allons à présent dégager certaines propriétés de l'algorithme dans le cas où la donnée ω est une ordonnance sur E, (ordre total sur F).

3- HIERARCHIE DE CLASSIFICATIONS ASSOCIEE A UNE ORDONNANCE

3.1. PROPOSITION

Dans le cas où la donnée ω est une ordonnance, la chaîne de partitions, définie par l'algorithme, est élémentaire [cf. Ch. 0 § 1.2].

Ce cas est caractérisé par la propriété : chacune des classes B_i, (ci-dessus définies), se réduit à un élément de F ; on a dans ces conditions $f = n(n-1)/2$ classes Comparons $\overline{C}_{i+1}$ et $\overline{C}_i$.

On a

$$\overline{C_{i+1}} = \overline{\overline{C}_i B_{i+1}}$$

(1) Deux préordres totaux ω et ω' sur un ensemble donné sont comparables si le graphe de l'un est contenu dans celui de l'autre.

soit, par exemple, {x,y} l'élément unique de la classe B_{i+1}. Deux cas sont possibles :

a) $\{x,y\} \in \overline{C}_i$; et alors, $\overline{C}_{i+1} = \overline{C}_i$,

b) $\{x,y\} \notin \overline{C}_i$; dans ces conditions, la partition définie par $\overline{C}_{i+1}$ se déduit de celle définie par $\overline{C}_i$, par réunion d'exactement deux classes, celle de x et celle de y.

D'après la proposition 2.4. ci-dessus, ω est une ordonnance pour laquelle la préordonnance ultramétrique, ω_u, associée par l'algorithme est comparable avec ω, si et seulement si la saturée de toute section, C, commençante pour ω, est une section commençante pour ω.

Une question se pose que nous allons exprimer de manière intuitive : "la propriété, qu'on vient d'énoncer, est-elle rare ?" Nous sommes ainsi amenés à étudier, dans l'ensemble, Ω, de toutes les ordonnances sur E, la proportion de celles, ω_u, pour lesquelles l'algorithme définit une préordonnance ultramétrique, ω_u, comparable avec ω ; nous nous rendrons compte, en particulier, du comportement de cette proportion lorsque n augmente. ($n = |E|$).

Le cardinal de l'ensemble, Ω, de toutes les ordonnances est f ! soit

$$(n(n-1)/2)!,\ f = |F| = n(n-1)/2.$$

Désignons par $\Omega_u(E)$ l'ensemble des ordonnances pour lesquelles la propriété ci-dessus est satisfaite.

Le cardinal de $\Omega_u(E)$ dépend seulement de l'entier n, cardinal de E, posons, dans ces conditions, $\gamma(n) = |\Omega_u(E)|$. Le rapport qui nous intéresse est $\gamma(n)/f$!

3.2. Dénombrement de Ω_u.

Une ordonnance ω est un élément de Ω_u si et seulement si ω est compatible avec une préordonnance ultramétrique ω_u, associée à une chaîne élémentaire de partitions, [cf.Prop. 3.1].

Si ω_u est une telle préordonance, qu'on dira élémentaire, ω_u comporte (n-1) classes ; en désignant, respectivement par $t_1, t_2, \ldots, t_{(n-1)}$ les cardinaux des différentes classes (du préordre), prises de gauche à droite ($t_1, t_2, \ldots, t_{(n-1)}$ sera le type de ω_u. Il y a $t_1 ! t_2 ! \ldots t_{(n-1)}!$ ordonnances de Ω_u compatibles avec ω. Un calcul direct de $\gamma(n)$ suppose la détermination de l'ensemble des types des préordonnances ultramétriques élémentaires. Plutôt que de procéder directement, nous nous contenterons d'établir une relation de récurrence pour $\gamma(n)$.

3.2.1. PROPOSITION

On a, pour la fonction $\gamma(n)$, la relation de récurrence :

$$\gamma(n) = \sum_{1 \leqslant k \leqslant \frac{n}{2}} \gamma(k)\gamma(n-k)(k(n-k))!\binom{n}{k}\binom{n-2}{k-1}$$

Le niveau (n-1) de la chaîne élémentaire de partitions correspond à une partition de E en deux classes. Le type de partition du niveau (n-1) est donc de la forme (n-k, k) ou k peut prendre les valeurs : 1, 2, ...,[n/2] (rappelons que [n/2] est la partie entière de n/2).

Il y a $\binom{n}{k}$ partitions de E en deux classes de types (n-k,k).

Considérons la partition de Ω_u selon le type de la partition obtenue par l'algorithme au niveau (n-1) et désignons par Ω_k le sous-ensemble de Ω_u pour lequel la partition obtenue au niveau (n-1) est de type (n-k,k) en posant $\gamma(n,k)$ le cardinal de Ω_k, on a

$$\gamma(n) = \sum_{1 \leqslant k \leqslant \frac{n}{2}} \gamma(n,k) \qquad (1)$$

Considérons une partition particulière de type (n-k,k) que nous noterons $(\overline{n-k},\overline{k})$, soit $\delta(n,k)$ le nombre d'ordonnances pour lesquelles la partition obtenue par l'algorithme au niveau (n-1) est précisément $(\overline{n-k},\overline{k})$. Donc :

$$\gamma(n,k) = \binom{n}{k}\ \delta(n,k)$$

et

$$\gamma(n) = \sum_{1 \leqslant k \leqslant \frac{n}{2}} \binom{n}{k}\ \delta(n,k) \qquad (2)$$

Désignons, respectivement, par L et par M les ensembles des parties à deux éléments, de $(\overline{n-k})$ et de $\overline{k}$. Si ω est l'ordre total sur F, ω_1 et ω_2 seraient respectivement les restrictions de ω à L et à M, ω_{1u} et ω_{2u} les préordonnances ultramétriques associées à ω_1 et à ω_2 par l'algorithme.

La condition nécessaire et suffisante pour qu'une ordonnance ω sur E appartienne à Ω_u et que de plus le niveau (n-1) de la chaîne de partitions associée par l'algorithme corresponde à la partition $(\overline{n-k},\overline{k})$ est que :

a) $\omega_1 \in \Omega_u(\overline{n-k})$ et $\omega_2 \in \Omega_u(\overline{k})$,

b) LUM définisse une section de F commençante pour ω, et :

c) chaque classe de ω_{1u} ou ω_{2u} définisse un intervalle de F pour ω ; les différents intervalles, de la suite des intervalles que définit ω_{1u}, s'intercalant avec les différents intervalles de la suite des intervalles que définit ω_{2u}.

Rappelons que ω_{1u} et ω_{2u} comportent respectivement (n-k-1) classes et (k-1) classes.

$\delta(n,k)$ est le nombre de façons dont on peut constituer une ordonnance, sur E, remplissant les conditions a, b et c.

Pour définir ω, il y a lieu :

1/ de déterminer une ordonnance ω_1 sur $\overline{(n-k)}$ appartenant à $\Omega_u(\overline{n-k})$, cela peut se faire de $\gamma(n-k)$ façons différentes ;

2/ de déterminer une ordonnance ω_2 sur $\overline{k}$ appartenant à $\Omega_u(k)$, cela peut se faire de $\gamma(k)$ façons différentes ;

3/ d'intercaler les différents intervalles de la suite des intervalles définie par ω_{1u} avec les différents intervalles définis par ω_{2u}. Cela peut se faire de $\binom{n-2}{k-1}$ façons différentes, puisque ω_{1u} comporte (n-k-1) classes et ω_{2u} (k-1) classes.

4/ enfin de définir un ordre total sur les k(n-k) paires {x,y} pour lesquelles x appartient à $\overline{k}$ et y à $\overline{(n-k)}$, cela peut se faire de (k(n-k))! manières différentes.

On a donc bien

$$\delta(n,k) = \gamma(n-k)\ \gamma(k)\ \binom{n-2}{n-1}(k(n-k))!$$

$\gamma(n)$ augmente très vite lorsque n croit, le rapport $\gamma(n)/f!(f=n(n-1)/2)$ tend très rapidement vers 0 lorsque n augmente pour tendre vers l'infini.

n	3	4	5	6	7	8	9	10	11	12	13	14	15
$\gamma(n) \geqslant$	6	10^2	10^5	10^{10}	10^{14}	10^{21}	10^{30}	10^{40}	10^{51}	10^{65}	10^{81}	10^{99}	10^{118}
$\frac{\gamma(n)}{f!} \leqslant$	1	6/10	$5/10^2$	$3/10^3$	$1/10^5$	$3/10^8$	$3/10^{12}$	$2/10^{16}$	$8/10^{22}$	$2/10^{27}$	$2/10^{34}$	$9/10^{42}$	$3/10^{50}$

4- EXTENSION DE L'ALGORITHME "LEXICOGRAPHIQUE" POUR LA RECHERCHE D'UNE FAMILLE DE SUITES DE RECOUVREMENTS.

Rappelons qu'un recouvrement d'un ensemble E est défini par la donnée d'une famille de parties de E, dont la réunion est E ; une partition étant un recouvrement particulier. A un recouvrement Q est associée la relation binaire :

$(\forall x,y \in E)$: xQy $\Longleftrightarrow$ x et y appartiennent à une même partie. Un recouvrement Q sera dit plus fin qu'un recouvrement Q' si xQy $\Longrightarrow$ xQ'y pour tout x et y de E.

Soit $Q^1, Q^2, \ldots, Q^h, \ldots$ une suite de recouvrements de E, totalement ordonnée par finesse décroissante ; une telle suite respecte les ressemblances entre objets si le premier indice h, pour lequel deux objets donnés quelconque de E se trouvent réunis dans une même partie de Q^h, est d'autant plus petit que la ressemblance des deux objets est grande.

L'algorithme que nous définirons nous permettra d'obtenir, de manière naturelle, une famille finie de suites totalement ordonnées de recouvrement dont chacune respecte de manière satisfaisante, pour un degré de finesse donné, les ressemblances entre objets. Ce point sera clarifié par ce

qui suit.

Reprenons les notations du paragraphe 2 précédent ; B_1, B_2,...B_p désignent les différentes classes du préordre ω, avec $B_1<B_2<...B_p$ pour l'ordre total quotient ; C_1, C_2,...,C_p est la suite des sections strictement commençantes de F pour ω, soit

$$C_i = B_1 \cup B_2 \cup \cdots \cup B_i$$

soit d'autre part, pour i<p et pour h<p-i, l'intervalle

$$D_i^h = B_{i+1} \cup B_{i+2} \cup \cdots \cup B_{i+h}$$

soit enfin $(P_0,P_1,...,P_q)$ la suite des partitions définie par l'algorithme "lexicographique" ; P_0 est la partition pour laquelle chaque classe contient un seul objet et P_j celle déterminée par la saturée $\overline{C}_i$ de C_i, $j \leqslant i$.

A chaque partition P_i est associée l'une des suites de recouvrements (Q_i^h), h=1, 2,...,(p-i), de la façon suivante :

Notons E_ℓ une quelconque des classes de P_i. Pour h fixé la saturée $\overline{D}_{i-1}^h$ de D_{i-1}^h définit une partition sur une partie de E ; e_m désignera une classe quelconque de cette partition. Une partie de E est un élément du recouvrement Q_i^h cherché si et seulement si cette partie est une classe de type E_ℓ ou une classe de type e_m. On remarquera que la suite (Q_0^h) n'est autre que la suite

$$(P_0,P_1,P_2,...,P_q).$$

Dans la pratique on est surtout intéressé par des recouvrements faiblement empiétants de sorte que nous proposons de déterminer en même temps que la suite $(P_0,P_1,...,P_q)$, la suite

$$(Q_1^1, Q_2^1,...,Q_{q-1}^1) ;$$

Q_i^1 étant défini à partir de P_i et de B_i.

Un tel algorithme est moins vulnérable à l'effet des "enchaînements successifs" auquel nous avons fait allusion dans l'introduction.

Exemple : Reprenons l'exemple (2') du paragraphe 2 ci-dessus. La suite des recouvrements

$$(Q_1^1, Q_2^1,...,Q_{q-1}^1)$$

est ici

$$Q_1^1 = \{\{f\}, \{a,d,f\}, \{b,c\}, \{b,e\}\}$$

$$Q_2^1 = \{\{a,d,f\}, \{b,c,e\}, \{d,e\}\}$$

q étant égal à 3.

III - ALGORITHME OPTIMISANT UN CRITERE LOCAL

Comme nous l'avons annoncé dans l'introduction, la proposition 3.1 ci-dessus montre que l'algorithme "lexicographique" revient à agréger à chaque pas les deux classes les plus voisine au sens de la plus grande proximité entre deux éléments appartenant respectivement aux deux classes.

L'algorithme désormais envisagé de construction d'un arbre de classifications consiste dans la réunion à chaque pas des deux classes les plus "proches". La notion de proximité considérée acquiert ainsi une importance cruciale. Cette proximité peut ne faire intervenir que les deux classes ; il s'agira alors d'un algorithme d'optimisation d'un critère local que nous développerons ici. La proximité entre les deux classes peut être comprise comme la part locale, affectée dans leur agrégation, d'un critère global de jugement de toute la partition ; il s'agira d'un algorithme d'optimisation locale d'un critère global que nous développerons au paragraphe suivant. Il est intéressant de noter, qu'en raison du théorème 2.3 et de la proposition 3.1, l'algorithme développé au paragraphe précédent peut être regardé de chacun de ces deux points de vue.

1. ALGORITHME DE LA VRAISEMBLANCE DU LIEN (A.V.L.)

1.1. Définition

Nous avons, conformément à un principe général, élaboré au chapitre 2, un indice de proximité qui est une sorte de corrélation généralisée, entre deux lignes ou colonnes d'une table à double entrée des données, pour chaque type de structure mathématique de cette dernière. Soit D l'ensemble à organiser par proximité, selon une structure en classes et sous-classes. D peut correspondre à un ensemble de variables descriptives d'un type fixé, au bien à un ensemble d'individus décrits par de telles variables, ou enfin à l'ensemble des lignes (resp. colonnes) d'une table ou d'une juxtaposition de tables de contingences.

Désignons par $\{U(\beta,\gamma)/(\beta,\gamma)\in DXD\}$ la table définitive de l'indice de proximité sur D avant la référence à une échelle de probabilité. Rappelons que l'indice $U(\beta,\gamma)$ s'obtient en deux étapes. La premières conduit à la table des proximités $\{Q(\beta,\gamma)/\ (\beta,\gamma)\in DXD\}$ où $Q(\beta,\gamma)$ s'obtient à partir d'un indice brut $s(\beta,\gamma)$ en centrant et en réduisant par rapport à une hypothèse d'absence de lien tenant compte des caractéristiques de taille de β et de γ. La deuxième étape est une réduction globale des similarités $Q(\beta,\gamma)$ (cf. chap. 2 § IV.5).

Dans l'hypothèse N retenue d'absence de lien entre β et γ ; $u = U(\beta,\gamma)$ est la réalisation d'une variable aléatoire normale centrée réduite N(0,1) ; d'où

$$P(\beta,\gamma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-x^2/2}\, dx$$

est une réalisation d'une variable aléatoire uniformément répartie entre 0 et 1.

Nous allons établir la notion de proximité entre classes que suppose A.V.L. en appliquant à l'indice brut de proximité apparu dans l'algorithme "lexicographique" le principe, dégagé au chapitre 2, de la définition d'une proximité entre structures finies de même type.

Soient B et C deux parties disjointes de D définissant deux classes de D de cardinaux respectifs ℓ et m. Considérons la suite des valeurs de la mesure de proximité sur l'ensemble des couples (β,γ) où β décrit B et γ, C. Soit

$$\{U(\beta,\gamma) \ / \ (\beta,\gamma)\in B\times C\} \qquad (1)$$

ou bien, en se référant à une échelle de probabilité

$$\{P(\beta,\gamma)/(\beta,\gamma)\in B\times C\} \qquad (2)$$

Pour A.V.L., la base de la constitution de la mesure de proximité entre B et C, est

$$P'(B,C) = \max_{(\beta,\gamma)\in B\times C} p(\beta,\gamma). \qquad (3)$$

qu'on peut écrire, $P'(B,C) = \max_{\beta\in B} P'(\beta,C)$ où $P'(\beta,C) = \max_{\gamma\in C} P(\beta,\gamma)$;

Or l'ensemble des valeurs $\{P(\beta,\gamma)/\ \gamma\in C\}$ constitue dans l'hypothèse N, un échantillon de m points indépendants d'une variable aléatoire uniformément répartie entre 0 et 1 ; d'où

$$P_r^N\{P'(B,C)<t\} = t^m \quad \text{avec } 0<t<1.$$

D'autre part, l'ensemble des valeurs $P'(\beta,C)/\beta\in B\}$ constitue dans l'hypothèse N, un échantillon de ℓ points indépendants d'une variable aléatoire dont la fonction de répartition vient d'être établie ; par conséquent

$$P_r^N\{P'(B,C)<t\}=(t^m)^{\ell} = t^{\ell m} ;$$

et si t_0 est la valeur observée de P'(B,C) on retiendra comme mesure de la proximité entre les deux classes

$$P(B,C) = t_0^{\ell m} \qquad (4)$$

Signalons, que pour des raisons de précision du calcul, on travaille au niveau du programme avec l'indice, fonction strictement croissante de P(B,C) :

$$-\text{Log } \ell - \text{Log } m - \text{Log } [-\text{Log } P'(B,C)] \qquad (4')$$

L'algorithme a été programmé avec élégance par Mme M.H. Nicolaü dans le cadre d'une thèse de 3ème cycle (cf. [6]). La structure du programme permet de façon très générale de passer d'un indice de proximité entre parties disjointes de l'ensemble à classifier à la représentation polonaise de l'arbre détaillé puis condensé des classifications. Cette thèse, qui portait sur l'analyse expérimentale de l'indice de comparaison entre classes que nous avons introduit dans [14], a montré, sur la base de nombreux exemples concrets, l'indiscutable progrès que représente l'algorithme de la vraisemblance du lien sur celui, "lexicographique" : Une classe de faible cohésion, formée notamment de variables relativement neutres (cf. Chap.3), à laquelle correspond une interprétation claire bien que faiblement apparente, se dégage nettement avec la nouvelle notion de proximité alors qu'elle se trouvait éparpillée dans le précédent algorithme ; ses éléments étant attirés par enchaînements successifs par des classes dont le noyau présente une forte cohésion. Néanmoins, l'algorithme "lexicographique"

garde, en raison précisément de son effet de chaînage, un certain intérêt en faisant apparaitre un lien faible "limite" de quelques attributs à certaines classes qui donne de ces dernières une vision parfois insoupçonnée.

Une adaptation de A.V.L. a été mise au point par A. Prod'homme (thèse de 3ème cycle, cf [17]) pour tenir compte d'une connexité spatiale ou statistique quant à la structure des classes. Le problème se pose surtout dans le cadre de la classification d'unités géographiques où l'utilisateur peut imposer l'obtention de classes connexes respectant autant que se peut la proximité de comportement statistique. Dans ce cas, en partant de la valeur de l'indice la plus faible, on retiendra la plus grande proximité statistique des paires de classes qui se "touchent".

Nous avons également utilisé ce dernier algorithme (A.V.L. sous contrainte de contiguïté) dans le cas de tableaux d'incidence "creux", où, pour la classification des lignes ou bien des colonnes, on impose une contrainte statistique de la forme suivante : on ne retiendra une proximité entre deux classes que s'il existe au moins un couple d'éléments (de lignes (resp. colonnes) s'il s'agit de la classification des lignes (resp. colonnes)), appartenant respectivement aux deux classes et ayant au moins k associations positives (cf. Chap.2, §III). Une telle contrainte se justifie d'ailleurs par des considérations statistiques liées au passage à une échelle de probabilité pour l'évaluation de la proximité.

1.2. Exemple.

Nous allons illustrer le fonctionnement de l'algorithme sur le tableau d'incidence T suivant, résultant de la description d'un ensemble $E = \{x_1, x_2, \ldots, x_9\}$ de neuf objets par un ensemble $A = \{a_1, a_2, \ldots, a_6\}$, formé de six attributs descriptifs qu'on désire classifier par A.V.L. en vertu de leurs proximités respectives établies à partir de la description de E et relativement à l'hypothèse N_1 d'absence de lien cf.Chap.2 §IV.1).

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
a_1	1	1	1	0	1	0	1	0	0
a_2	1	0	1	0	0	1	0	1	1
a_3	1	0	1	0	0	1	0	1	1
a_4	0	0	1	1	1	1	1	0	0
a_5	1	1	0	1	0	0	1	0	1
a_6	0	1	0	0	1	1	0	1	1

Tableau T

A partir de ce tableau on établit celui, symétrique, des nombres $s(a_i,a_j)$ où $s(a_i,a_j)$ est l'indice brut de proximité entre les deux attributs a_i et a_j ; il s'agit du nombre d'objets possédant les deux attributs $s(a_i,a_j) = \text{card}\,(E_{a_i} \cap E_{a_j})$.

	a_1	a_2	a_3	a_4	a_5	a_6
a_1	5					
a_2	2	5				
a_3	2	5	5			
a_4	3	2	2	5		
a_5	3	2	2	2	5	
a_6	2	3	3	2	2	5

Pour des raisons de simplicité, nous avons considéré, comme on peut le remarquer, des attributs également fréquents dans E : card $(E_{a_i}) = 5$ pour tout i, $1 \leqslant i \leqslant 6$. Nous allons appliquer A.V.L. à A en considérant un tableau attaché à la formation de l'arbre des classifications qui représente à une étape donnée les proximités entre couples de classes déjà formées. L'état initial de ce tableau est celui des proximités $P(a_i,a_j)$ où $P(a_i,a_j)$ est l'indice : proportion de parties Y de même cardinal que E_{a_j} pour lesquelles card $(E_{a_i} \cap Y)$ est inférieur ou égal à $s(a_i,a_j)$. Introduisant la variable aléatoire S_{a_i} = card$(E_{a_i} \cap Y)$ où Y est un élément aléatoire dans l'ensemble, muni d'une probabilité uniformément répartie, $P_5(E)$ de toutes les parties de E de même cardinal 5 ;

$$P(a_i,a_j) = Pr^N \{S_{a_i} \leqslant s(a_i,a_j)\}.$$

La proportion de parties Y pour lesquelles $S_{a_i} = k$ est donnée par la formule $\binom{5}{k}\binom{4}{5-k}/\binom{9}{5}$ où $1 \leqslant k \leqslant 5$; de sorte que S_a est susceptible de prendre la suite des valeurs (1,2,3,4,5) avec respectivement les probabilités $(\frac{5}{126}, \frac{40}{126}, \frac{60}{126}, \frac{20}{126}, \frac{1}{126})$ que nous garderons sous cette forme.

Le tableau des nombres $P(a_i,a_j)$ est donc le tableau P suivant où l'attribut a_i a été représenté par le code i.

L'agrégation des attributs a_2 et a_3 donnant lieu à une classe notée 23 transforme le tableau initial en celui $P^{(1)}$ suivant.

L'examen des proximités nous conduit à la formation de la classe $\{a_1,a_4,a_5\}$ après laquelle le tableau des proximités devient celui $P^{(2)}$ suivant.

D'où la réunion des classes $\{a_2,a_3\}$ et $\{a_6\}$ à laquelle succède le tableau des proximités $P^{(3)}$.

	1	2	3	4	5	6
1	1	$\frac{45}{126}$	$\frac{45}{126}$	$\frac{105}{126}$	$\frac{105}{126}$	$\frac{105}{126}$
2	$\frac{45}{126}$	1	1	$\frac{45}{126}$	$\frac{45}{126}$	$\frac{105}{126}$
3	$\frac{45}{126}$	1	1	$\frac{45}{126}$	$\frac{45}{126}$	$\frac{105}{126}$
4	$\frac{105}{126}$	$\frac{45}{126}$	$\frac{45}{126}$	1	$\frac{45}{126}$	$\frac{45}{126}$
5	$\frac{105}{126}$	$\frac{45}{126}$	$\frac{45}{126}$	$\frac{45}{126}$	1	$\frac{45}{126}$
6	$\frac{45}{126}$	$\frac{105}{126}$	$\frac{105}{126}$	$\frac{45}{126}$	$\frac{45}{126}$	1

Tableau P

	1	2,3	4	5	6
1	1	$\left(\frac{45}{126}\right)^2$	$\frac{105}{126}$	$\frac{105}{126}$	$\frac{45}{126}$
2,3	$\left(\frac{45}{126}\right)^2$	1	$\left(\frac{45}{126}\right)^2$	$\left(\frac{45}{126}\right)^2$	$\left(\frac{105}{126}\right)^2$
4	$\frac{105}{126}$	$\left(\frac{45}{126}\right)^2$	1	$\frac{45}{126}$	$\frac{45}{126}$
5	$\frac{105}{126}$	$\left(\frac{45}{126}\right)^2$	$\frac{45}{126}$	1	$\frac{45}{126}$
6	$\frac{45}{126}$	$\left(\frac{105}{126}\right)^2$	$\frac{45}{126}$	$\frac{45}{126}$	1

Tableau $P^{(1)}$

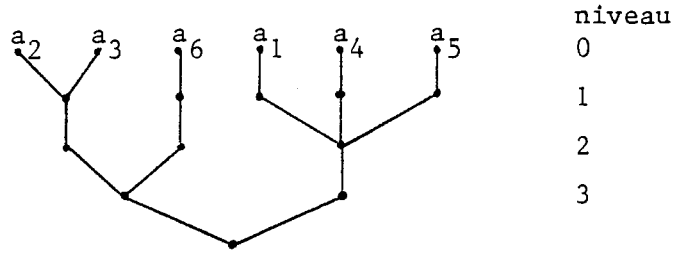
	145	23	6
145	1	$\left(\frac{45}{126}\right)^6$	$\left(\frac{45}{126}\right)^3$
23	$\left(\frac{45}{126}\right)^6$	1	$\left(\frac{105}{126}\right)^2$
6	$\left(\frac{45}{126}\right)^3$	$\left(\frac{105}{126}\right)^2$	1

Tableau $P^{(2)}$

	145	236
145	1	$\left(\frac{45}{126}\right)^9$
236	$\left(\frac{45}{126}\right)^9$	1

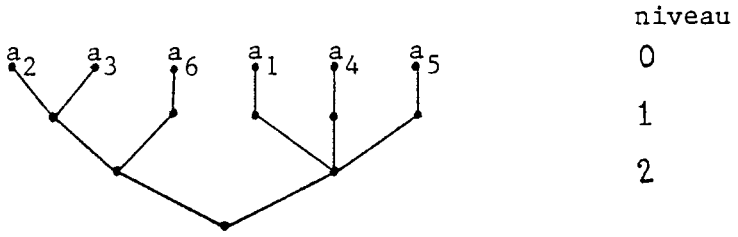
Tableau $P^{(3)}$

L'arbre des classifications associé à la suite des agrégations est



La préordonnance associée par ressemblance décroissante est $23<14\sim15\sim26\sim36<12\sim13\sim16\sim24\sim25\sim34\sim35\sim45\sim46\sim56$ où on a noté ij la paire $\{a_i,a_j\}$ avec $i<j$.

L'algorithme "lexicographique" aurait donné l'arbre suivant



qui diffère peu de l'arbre précédent ; compte tenu de l'exiguïté de l'exemple.

1.3. Contractance de A.V.L.

La propriété de contractance a été introduite en classification automatique par M. Bruynoogh (thèse de 3ème cycle, Université Paris VI, 1976) (cf. dans [13] où l'approche est reprise). Relativement à un ensemble muni d'un indice de distance ou de proximité, cette propriété s'exprime intuitivement par : " le voisinage de la réunion de deux parties disjointes de l'ensemble est inclus dans la réunion des voisinages des deux parties".

Considérons dans notre cadre une partition $\mathcal{P}(D)$ de D. Si B est une classe de cette partition, nous indiquerons par $\mathcal{B}(B,\pi)$ la boule de centre B et de rayon $(1-\pi)$; soit

$$\mathcal{B}(B,\pi) = \{H/P(B,H) \geq \pi\} \qquad (1)$$

où P est l'indice de proximité entre classes qu'on vient de définir ci-dessus.

Relativement à deux classes B et C de $\mathcal{P}(D)$, nous allons montrer qu'on a bien

$$\mathcal{B}(B\cup C,\ \pi) \subset \mathcal{B}(B,\pi)\bigcup\mathcal{B}(C,\pi) \qquad (2)$$

En d'autre termes si G est un élément de $\mathcal{B}(BUC,\pi)$, alors nécessairement

$$(\max\{P(e,g)/(e,g)\in B\times G\})^{\ell\times r} \geq \pi \qquad (3)$$

ou

$$(\max\{P(e,g)/(e,g)\in C\times G\})^{m\times r} \geq \pi$$

où on a noté ℓ = card(B), m = card (C) et r = card (G).

Sinon, on aurait

$$(\max\{P(e,g)/(e,g)\in C\times G\})^{r} < \pi^{1/\ell} \qquad (4)$$

et

$$(\max\{P(e,g)/(e,g)\in C\times G\})^{r} < \pi^{1/m}$$

Supposons, sans restreindre la généralité, que $\ell \geq m$.

On en déduit (à partir de (4)),

$$(\max\{P(e,g)/(e,g)\in (B\cup C)\times G\})^{r} < \pi^{1/\ell}, \qquad (5)$$

ce qui entraînerait

$$P(B\cup C,G) < \pi \qquad (6)$$

et qui est contraire à l'hypothèse de départ. Il en résulte la propriété suivante.

PROPRIETE

Le critère de formation des classes de l'algorithme de la vraisemblance du lien est contractant .

Cette propriété a d'abord été démontrée pour un certain nombre de critères classiques basés sur les distances au carré, tel que celui de l'inertie expliquée (cf. paragraphe IV suivant). Nous avons pu, ainsi envisager l'adaptation pour nos critères d'algorithmes accélérés de classification hiérarchique dont le principe général s'exprime comme suit (cf.[13]).

On commence par se restreindre aux fusions de paires de classes qui, respectivement, correspondent à une valeur de l'indice de proximité entre classes supérieure à un premier seuil π_1. Pour cela on forme autour de chaque élément (ou classe) une boule dont le rayon correspond à ce seuil.

On ne compare entre elles que les paires dont les deux composantes appartiennent à une même boule.

Après chaque réunion, on reforme une boule autour de la classe fusion ce qui, compte tenu de la contractance, ne nécessite que l'examen de la réunion des deux boules entourant respectivement les deux classes composantes.

Lorsqu'aucune réunion n'est plus possible, on passe d'un seuil π_1 à un seuil π_2 correspondant à un indice de proximité plus petit pour reprendre la procédure ... et ainsi de suite jusqu'à aboutir à une seule classe.

Cet algorithme a été programmé pour le critère de la vraisemblance du lien par J.P. Chapron (stage en Informatique de 5ème année de l'INSA de Rennes). L'arbitraire du choix des seuils définissant les rayons des boules se trouve considérablement amoindri pour notre approche ; en outre, ce choix peut être effectué de façon à utiliser au mieux la place de mé-

moire centrale de l'ordinateur (mémoire de travail) (cf. [18]).

2. ALGORITHME DE LA MOYENNE DES PROXIMITES (A.M.P.).

Un autre indice brut de proximité entre les deux classes d'éléments de D, B et C, qui semble "naturel" est la somme des proximités $U(\beta,\gamma)$, (cf. expression (1), § 1 ci-dessus).

Dans l'hypothèse N, l'ensemble suivant des valeurs de $U(\zeta,\theta)$;

$$\{U(\beta,\gamma)/\gamma\in C\}$$

est un échantillon de m points indépendants d'une variable aléatoire normale entrée réduite $\mathcal{N}(0,1)$; en notant

$$U(\beta,C) = \sum_{\gamma\in C} U(\beta,\gamma)$$

$\frac{1}{\sqrt{m}}\,U(\beta,C)$ est la réalisation d'une v.a. $\mathcal{N}(0,1)$.

Dans ces conditions l'ensemble des valeurs

$$\{\frac{1}{\sqrt{m}}\ (\beta,C)/\ \beta\in B\}$$

est, dans l'hypothèse N, un échantillon de ℓ points indépendants d'une v.a. $\mathcal{N}(0,1)$; d'où

$$\frac{1}{\sqrt{\ell}} \sum_{\beta\in B} \frac{1}{\sqrt{m}}\ (\beta,C)$$

est, dans l'hypothèse N, une réalisation d'une v.a. $\mathcal{N}(0,1)$. Il en résulte que la première statistique de proximité à adopter entre les deux classes B et C, est

$$U(B,C) = \frac{1}{\sqrt{\ell m}} \sum_{(\beta,\gamma)\in C\times C} U(\beta,\gamma) \qquad (1)$$

Compte tenu de l'échelle de référence définie dans l'hypothèse d'absence de liaison par la loi $\mathcal{N}(0,1)$, cette statistique permet de comparer sans biais les proximités U(B,C) et U(B',C') attachées à deux couples de classes.

Si nous avons proposé ce dernier critère (1) pour la formation des classes c'est pour répondre de façon claire à ceux qui considèrent qu'"il faut tenir compte de toutes les distances entre points des deux classes comme peut le faire le critère la "moyenne des distances"". Cependant, l'expérience montre de façon claire et indiscutable que si A.M.P. va le plus directement vers la formation des plus grosses tendances, c'est A.V.L. qui produit les résultats les plus raffinés et les plus cohérents dans leurs nuances, distinguant dans chacune des tendances, les sous tendances composantes. Une réflexion peut être menée sur la raison d'une telle différence de nature entre les résultats de A.V.L. et de A.M.P.

Cette différence se conserve curieusement lorsqu'on prend comme le fait F. Nicolaü dans sa thèse (cf. [16]) pour A.M.P. un indice basé non pas sur la loi de probabilité dans l'hypothèse N de

$$\frac{1}{\ell m} \sum \{ (\beta,\gamma)/(\beta,\gamma)\in B\times C\}$$

mais sur celle de

$$\frac{1}{\ell m} \sum \{P(\beta,\gamma)/(\beta,\gamma)\in B\times C\} \qquad (2)$$

où P est l'indice défini qui se réfère à une échelle [0,1] de probabilité. On peut d'autre part s'étonner de constater que c'est ce dernier indice qui permet les plus grandes valeurs de la statistique globale $\sum$ des niveaux (cf. § V ci-dessous) (cf. [16]). On trouvera d'ailleurs dans la thèse précitée une analyse comparative très intéressante de différents critères de formation hiérarchique des classes, basés sur la vraisemblance des liens.

Signalons pour terminer que J.L. Buard a démontré dans sa thèse (cf. [10]). que A.M.P. (au sens de F. Nicolaü) n'est pas contractant.

IV - ALGORITHME OPTIMISANT LOCALEMENT UN CRITERE GLOBAL

1. CRITERE DE L'INERTIE

Il s'agit plutôt ici de la classification de l'ensemble E des objets représentable au moyen d'un nuage $\mathcal{N} = \{(M_i, \mu_i)/i\in I\}$ dans un espace euclidien muni d'une métrique q. Avec des notations que nous reprendrons au chapitre 6, on a la formule de décomposition du moment total d'inertie du nuage :

$$\sum_{i\in I} \mu_i || M_i-G||^2 = \sum_{1\leq h\leq k} \{ \sum_{i\in I_h} \mu_i || M_i-G_h||^2 + \sum_{1\leq h\leq k} \nu_h ||G_h-G||^2 \qquad (1)$$

formule dite de l'analyse de la variance ; $\{I_h/1\leq h\leq k\}$ désigne une partition en k classes de I indiquant une partition de l'ensemble E des objets. A chacune des classes I_h se trouvent associées d'une part sa masse $\nu_h = \sum_{i\in I_h} \mu_i$ et d'autre part, son centre de gravité $G_h = \frac{1}{\nu_h} \sum_{i\in I_h} \mu_i M_i$

Le moment d'inertie de la classe I_h

$$\mathcal{W}_h = \sum_{i\in I_h} \mu_i || M_i-G_h||^2 \qquad (2)$$

définit la <u>cohésion</u> de la h-ème classe ; cette dernière est d'autant plus grande que $\mathcal{W}_h$ est plus petit.

Le critère global de jugement de la partition est défini par la petitesse de $\sum_{1\leq h\leq k} \mathcal{W}_h$, premier terme du second membre de la formule (1).

Déterminons la variation Δ de cette quantité critère lorsqu'on réunit les deux classes I_j et $I_{j'}$.

$$\sum_{i\in I_j\cup I_{j'}} \mu_i ||M_i-G_{juj'}||^2 - \sum_{i\in I_j} \mu_i ||M_i-G_j||^2 - \sum_{i\in I_{j'}} \mu_i ||M_i-G_{j'}||^2 \qquad (3)$$

où nous avons noté $G_{juj'}$ le centre de gravité de la classe $I_j \cup I_{j'}$;

soit
$$G_{juj'} = \frac{1}{\nu_j+\nu_{j'}} \sum_{i\in I_j\cup I_{j'}} \mu_i M_i$$
$$= \frac{1}{\nu_j+\nu_{j'}} \{\nu_j G_j + \nu_{j'} G_{j'}\} \qquad (4)$$

Or, en vertu de la formule de décomposition de l'analyse de la variance, le premier terme du second membre de (3) définissant Δ , se met sous la forme

$$\sum_{i\in I_j} \mu_i ||M_i-G_j||^2 + \sum_{i\in I_{j'}} \mu_i ||M_i-G_{j'}||^2 + \nu_j ||G_j-G_{juj'}||^2 + \nu_{j'} ||G_{j'}-G_{juj'}||^2 ;$$

d'où
$$\Delta = \nu_j ||G_j-G_{juj'}||^2 + \nu_{j'} ||G_{j'}-G_{juj'}||^2$$

et compte tenu de (4), on obtient finalement
$$\Delta(j,j') = \frac{\nu_j \nu_{j'}}{\nu_j+\nu_{j'}} || G_{j'}-G_j ||^2 \qquad (5)$$

Par conséquent, l'algorithme consistera à réunir à chaque pas la paire $\{ I_j, I_{j'} \}$ de classes pour laquelle $\Delta(j,j')$ est le plus petit.

Il est intéressant de remarquer que la part locale (5) du critère global, affecté par la réunion des deux classes I_j et $I_{j'}$, ne fait intervenir que ces deux classes ; on aurait pu par conséquent, présenter l'algorithme d'agrégation au paragraphe précédent mais alors on n'aurait pas pu justifier a priori le choix du critère local (5), (cf. [19]).

2. CAS D'UN TABLEAU DE CONTINGENCE

J.P. Benzécri a particulièrement développé l'analyse métrique de la dépendance entre deux variables définissant chacune une partition, ayant un grand nombre de classes, sur la population étudiée. Soit I (resp.J) l'ensemble des modalités de l'une (resp. de l'autre) variable ; posons $n = \text{card}(I)$ et $m = \text{card}(J)$. Le support de l'information est le tableau de contingence de croisement des deux partitions indéxé par IxJ. A l'intersection de la ligne i et de la colonne j est le nombre k_{ij} des sujets (d'un échantillon de la population étudiée, en général) qui possèdent la i-ème modalité du premier caractère et la j-ème du second. On substitue à ce tableau celui des fréquences relatives $f_{ij} = k_{ij}/k$, où $k = \sum_{(i,j)} k_{ij}$ est l'effectif total de l'échantillon concerné. On complète ce tableau par deux marges ; une marge colonne qui contiendra les proportions $p_{i\bullet} = \sum_{j\in J} f_{ij}$, pour $i = 1,2,\ldots,n$; et une marge ligne qui contiendra celle $p_{\bullet j} = \sum_{i\in I} f_{ij}$, pour $j = 1,2,\ldots,m$. Ces fréquences relatives sont

dites marginales ; $p_{i.}$ (resp.$p_{.j}$) est une mesure de l'"importance" (numérique) de la i-ème (resp. de la j-ème) modalité du premier caractère (resp. du second).

La perception de la ressemblance entre deux modalités d'un même caractère à travers l'autre caractère (où il s'agit de ne pas tenir compte de l'importance relative de présence des deux modalités) conduit à représenter géométriquement I par le nuage (cf. chap. 6)

$$\mathcal{N}(I) = \{(f_J^i, p_{i.})/i\in I\} \qquad (1)$$

où on associe à $i\in I$ le point de R^m

$$f_J^i = (f_1^i,\ldots,f_j^i,\ldots,f_m^i) \qquad (2)$$

suite des proportions conditionnelles

$$f_j^i = f_{ij}/p_{i.} \quad ;$$

proportion, dans la i-ème catégorie, des individus possédant la j-ème modalité du second caractère. On retient ainsi pour la description de i son "profil" à travers J ; c'est-à-dire, la suite des parts des différents $j\in J$ qui entrent dans la composition de la catégorie i. On préserve l'importance de présence $p_{i.}$ de i en affectant le sommet f_J^i du poids $p_{i.}$. $\mathcal{N}(I)$ est un système de points de simplexe $\mathcal{S}_J$ des lois de probabilités sur J, chacun des points $i\in I$ étant affectée de la masse $p_{i.}$.

On introduit de façon tout-à-fait symétrique le nuage de R^n associé à J ; avec des notations qu'on comprend aisément

$$\mathcal{N}(J) = \{(f_J^i, p_{.j})/j\in J\} \qquad (3)$$

Relativement à l'analyse du nuage $\mathcal{N}(I)$ on munit l'espace ambiant R^m, dont la base canonique est notée $\{e_j/1\leq j\leq m\}$, de la métrique q suivante

$$q(e_j,e_k) = \begin{cases} 0 & \text{si } j\neq k \\ 1/p_{.j} & \text{si } j=k \end{cases} \qquad (4)$$

L'introduction de cette métrique, dite du χ^2, se justifie pour des raisons algébrique et statistique. La raison algébrique est définie par la condition de l'équivalence distributionnelle qui signifie que les distances entre éléments de I ainsi que celles entre éléments de J restent invariants lorsqu'on remplace deux éléments i_1 et i_2 de I de masses respectives $p_{i._1}$ et $p_{i._2}$ qui ont la même représentation dans R^m par un seul point i_0 de masse

$$p_{i_0.} = p_{i_1.} + p_{i_2.}$$

Avec une telle métrique l'expression de la distance entre i et i' de I devient

$$d^2(i,i') = \sum_{1 \leq j \leq m} \frac{1}{p_{.j}} (f_j^i - f_j^{i'})^2 \qquad (5)$$

qui est exactement la distance du χ^2 associée à la loi de probabilité $\{p_{.j}/j \in J\}$ entre les deux distributions $\{f_j^i/j \in J\}$ et $\{f_j^{i'}/j \in J\}$.

D'autre part, le moment total d'inertie du nuage est exactement la statistique du χ^2 attachée au tableau de contingence. L'analyse des correspondances que nous présenterons rapidement au chapitre suivant se propose précisément une décomposition axiale de cette inertie ; laquelle se met, relativement à l'étude du nuage $\mathcal{N}(I)$, sous la forme

$$\sum_{i \in I} p_{i.} ||f_J^i - g_J||^2 = \sum_{i \in I} p_{i.} \sum_{j \in J} \frac{1}{p_{.j}} (f_j^i - p_{.j})^2 \qquad (6)$$

où nous avons noté $g_J = \sum_{i \in I} p_{i.} f_J^i = (p_{.1}, \ldots, p_{.j}, \ldots, p_{.m})$, le centre de gravité du nuage.

Relativement à une classification $\{I_h/1 \leq h \leq k\}$ de I, la formule (1), du paragraphe précédent, de décomposition de l'inertie, devient

$$\sum_{i \in I} p_{i.} ||f_J^i - g_J||^2 = \sum_{1 \leq h \leq k} \{ \sum_{i \in I_h} ||f_J^i - g_J^h||^2 + \sum_{1 \leq h \leq k} p(I_h) ||g_J^h - g_J||^2 \qquad (7)$$

où, à la classe I_h se trouvent associés d'une part son poids

$$p(I_h) = \sum_{i \in I_h} p_{i.}$$

et d'autre part, son centre de gravité

$$g_J^h = \frac{1}{p(I_h)} \sum_{i \in I_h} p_{i.} f_J^i = (f_1^{I_h}, \ldots, f_j^{I_h}, \ldots, f_m^{I_h})$$

où nous notons

$$f_j^{I_h} = \frac{f(I_h,j)}{f(I_h)} = \sum_{i \in I_h} f(i,j) / \sum_{i \in I_h} p_{i.} \qquad (8)$$

qui est une proportion conditionnelle : il s'agit de la proportion dans la classe I_h des individus possédant la j-ème modalité du second caractère.

Conformément au résultat du paragraphe précédent (cf. formule (5)), la construction d'un arbre de classifications sur I par agrégations successives conduit à réunir à chaque pas la paire de classes $\{I_h, I_h\}$ pour laquelle est minimum

$$\eta(h,h') = \frac{p(I_h)p(I_{h'})}{p(I_h)+ p(I_{h'})} \; || \; g_J^h - g_J^{h'} \; ||^2 \qquad (9)$$

où $$||g_J^h - g_J^{h'}||^2 = \sum_{j\in J} \frac{1}{p_{.j}} \left(f_j^{I_h} - f_j^{I_{h'}}\right)^2 \qquad (10)$$

qui est la distance, au sens de la métrique du χ^2, entre les deux points du simplexe $\mathcal{R}_J$ représentant respectivement les deux classes I_h et $I_{h'}$; la j-ème composante du premier (resp. du second) est la proportion dans I_h (resp. $I_{h'}$) des sujets possédant la j-ème modalité de l'autre caractère.

Compte tenu de la parfaite symétrie du tableau de contingence ; il est tout à fait clair que le critère local à utiliser pour la définition d'un arbre de classifications sur J, est le suivant

$$\eta'(h,h') = \frac{p(J_h)\; p(J_{h'})}{p(J_h) + p(J_{h'})} \; || \; g_I^h - g_I^{h'} \; ||^2 \qquad (11)$$

avec des notations que l'on comprend immédiatement et que nous laissons le soin au lecteur de préciser.

Exercice.

Soit $\mathcal{N}(I) = \{(M_i,\mu_i)/i\in I\}$ un nuage de points dans un espace euclidien et soit $\{I_j/1\leq j\leq m\}$ une partition de I définissant une classification en m classes non vides de $\mathcal{N}(I)$. On désigne par $\Delta(I_j,I_h)$ la perte de l'inertie expliquée par la classification résultant de la fusion des deux classes respectivement indexées par I_j et I_h.

1- Exprimer $\Delta(I_j \cup I_h, I_k)$ en fonction de $\Delta(I_j,I_h)$, $\Delta(I_j,I_k)$, $\Delta(I_h,I_k)$ et des masses des trois classes.

2- On considère l'algorithme de formation ascendante d'un arbre des classifications sur I où, à chaque niveau on réunit les paires de classes qui optimisent le critère Δ présenté ci-dessus.

A une même classification $\{I_j/1\leq j\leq m\}$, on associe

$$\min \{\Delta(I_j,I_h)/1\leq j\neq h\leq m\} \; ;$$

montrer que cette quantité est strictement croissante le long de la suite des niveaux de l'arbre formé des classifications.

Une autre mesure du lien, que le χ^2, entre deux variables partitions, nous est fourni par la théorie de l'Information au sens de Hartley et Shannon ; il s'agit de l'information mutuelle entre les deux variables que nous allons rapidement présenter en nous référant à [2] (voir aussi [7]).

Soit I l'ensemble des modalités d'un caractère descriptif dont la distribution sur la population étudiée E est définie par

$$p_I = \{p_i / i \in I\} \qquad (1')$$

Relativement à l'expérience aléatoire : extraction au hasard d'un individu x dans E muni d'une probabilité uniformément répartie ; on définit l'entropie de Shannon

$$H(p_I) = - \sum_{i \in I} p_i \log_2 p_i \qquad (2')$$

qui est l'espérance mathématique de la quantité d'information $- \log_2 p_i$ apportée par l'événement : x possède la i-ème modalité du caractère.

De nombreuses axiomatiques permettent de caractériser au moyen de conditions "naturelles" l'entropie de Shannon (cf. [1]) ; nous ne nous y étendrons pas.

I et J désignant les deux ensembles de modalités respectivement associées à deux caractères descriptifs ; considérons l'entropie $H(f_{IXJ})$ attachée au tableau de contingence de croisement des deux partitions définies par les deux caractères

$$H(f_{IXJ}) = - \sum_{(i,j)} f_{ij} \log_2 f_{ij} \qquad (3')$$

On a la relation très importante

$$H(f_{IXJ}) \leqslant H(p_I) + H(p_J) \; ; \qquad (4')$$

l'égalité caractérisant l'indépendance entre les deux partitions qui s'exprime par les relations

$$(\forall\ (i,j) \in IXJ),\ f_{ij} = p_{i.} p_{.j}$$

L'inégalité (4') peut être démontrée en utilisant les propriétés de concavité de la fonction $x \log_2 x$ pour $x \geqslant 0$

La différence positive

$$H(p_I) + H(p_J) - H(f_{IXJ}) \qquad (6')$$

peut être mise sous l'une des deux formes

$$H(p_I) - H(I/J) \text{ ou bien } H(p_J) - H(J/I) \qquad (7')$$

avec

$$H(J/I) = \sum_{i \in I} p_{i.} H(f_J^i) \qquad (8')$$

où, bien entendu,

$$H(f_J^i) = - \sum_{j \in J} f_j^i \log_2 f_j^i$$

La quantité (6') qui mesure le lien entre les deux variables partitions est l'information mutuelle entre ces deux variables que nous noterons conformément à [2], $H(f_{IJ} ; p_I\, p_J)$ où $p_I\, p_J = \{p_{i.}p_{.j}/(i,j) \in I\times J\}$ désigne la distribution de probabilité sur IxJ dans le cas de l'indépendance. On a

$$H(f_{IJ} ; p_I p_J) = \sum_{(i,j)\in I\times J} f_{ij} \log_2(f_{ij}/p_{i.}p_{.j}) = \sum_{(i,j)\in I\times J} p_{i.}p_{.j}\,\phi(f_{ij}/p_{i.}p_{.j})$$

où $\phi(x) = x\log_2 x$.

L'information mutuelle apparait ainsi comme l'espérance mathématique pour la loi $p_I \times p_J$ produit des lois marginales de la fonction ϕ de la densité $(f_{IJ}/p_I \times p_J)$. la stricte concavité de la fonction $\phi(x)$ sur $(0,\infty)$ permet de montrer la relation (4') ; en effet, la valeur d'une telle fonction au centre de gravité d'une suite de points est inférieure à la moyenne pondérée de ses valeurs sur cette suite de points, l'égalité n'ayant lieu que si tous les points sont confondus.

Le lien entre I et J défini par la statistique du χ^2 peut se mettre sous une forme analogue à la seconde expression (9') ; il s'agit de

$$\sum_{(i,j)\in I\times J} p_{i.}p_{.j}\ \psi\ (f_{ij}/p_{i.}p_{.j}) \qquad (10')$$

où $\psi(x) = x^2-1$

J.P. Benzecri note avec intérêt que $\phi(f_{ij}/p_{i.}p_{.j})$ et $\psi(f_{ij}/p_{i.}p_{.j})$ se comportent également au voisinage de l'indépendance $(f_{IJ}\#p_I \times p_J)$; en effet les deux fonctions $\phi(x)$ et $\frac{1}{2\log 2}\ \psi(x)$ s'annullent pour x=1 et ont les mêmes dérivées premières et secondes.

La fusion des deux classes de l'ensemble E des sujets indexés par les deux modalités i_1 et i_2 de I conduit à la baisse de l'information mutuelle de la quantité, positive en raison de la concavité de ϕ ,

$$\sum_{j\in J} \{-p_{i_0}.p_{.j}\ \phi(f_{i_0j}/p_{i_0}.p_{.j}) + p_{i_1}.p_{.j}\ \phi[p_{i_1}.p_{.j}\phi(f_{i_1j}/p_{i_1}.p_{.j}) + p_{i_2}.p_{.j}\ \phi(f_{i_2j}/p_{i_2}.p_{.j})]\}$$

où $p_{i_0}. = p_{i_1}. + p_{i_2}.$ et $f_{i_0j} = f_{i_1j} + f_{i_2j}$

Le premier pas de l'algorithme de construction d'un arbre de classifications sur I consiste à réunir la paire $\{i_1,i_2\}$ de modalités pour lesquelles la quantité (10') est minimum ; de la sorte, dans le remplacement de I par $I'=I-\{i_1,i_2\}+\{i_0\}$; le lien, mesuré par l'information mutuelle, entre I' et J, reste le plus voisin de celui entre I et J.

A une partition $\{I_1,\ldots,I_h,\ldots,I_k\}$ de I, on peut naturellement associer

une variable partition sur E dont l'ensemble des modalités peut être noté $K = \{1,2,\ldots,h,\ldots,k\}$. Un même pas de l'algorithme consiste en la réunion, de la paire de classes $\{I_h, I_{h'}\}$, qui minimise la perte dans l'information mutuelle entre K et J ; soit

$$H(f_{KJ} ; p_K \times p_J) - H(f_{K'J} ; p_{K'} \times p_J) \qquad (12')$$

qui se calcule par une formule analogue à (11') et où K' se déduit de K par la concaténation de h et de h'.

Ici encore, il est intéressant de noter que la part locale, du critère global, affectée par la réunion des deux classes, ne fait intervenir que ces deux dernières.

En remplaçant dans l'expression (11') la fonction ϕ par celle ψ ; on a une autre forme de la quantité critère basée sur le χ^2 (cf. formule (9) ci-dessus).

Il est enfin tout-à-fait clair que le problème de la classification de J sur la base du critère de l'information mutuelle est identique à celui de I. Le problème de la détermination d'une "bonne" classification simultanée de I et de J est toutefois différent.

3. CAS OU LA DONNEE EST UNE ORDONNANCE

C'est le cas très général où, pour la définition d'un critère, on retient comme information relative aux ressemblances de l'ensemble D à classifier, un ordre total sur l'ensemble F des paires d'éléments distincts de D. Nous supposons que l'ordre total a été établi de telle sorte que le rang d'une paire soit une fonction décroissante de la ressemblance entre ses composants. Nous avons dans ces conditions établi au chapitre 4 le critère suivant d'adéquation d'une partition $\pi = \{D_1, D_2, \ldots, D_j, \ldots, D_k\}$ à l'ordonnance ω

$$\{\text{card}\,(\text{gr}(\omega) \cap (R(\pi) \times S(\pi)) - r.s/2\} / \sqrt{r.s(f+1)/12} \qquad (1)$$

où, rappelons le, $R(\pi)$ (resp. $S(\pi)$) est l'ensemble des paires réunies (resp. séparées) par π. $r = \text{card}\,(R(\pi))$, $s = \text{card}\,(S(\pi))$ et $f = r+s$.

Nous avons

$$R(\pi) = \sum_{1 \leq j \leq k} P_2(D_j) \quad \text{et} \quad S(\pi) = \sum_{(j,h)} D_j * D_h \qquad (2)$$

(sommes ensemblistes)

où $P_2(D_j)$ est l'ensemble des parties à deux éléments de D_j et où $D_j * D_h$ est l'ensemble des paires dont l'une des composantes appartient à D_j et l'autre à D_h.

La statistique (1) mesurant la cohésion des classes formées ; l'algorithme de construction d'un arbre de classifications retiendra à chaque pas l'agrégation de la paire $\{D_j, D_{j'}\}$ de classes, qui maximise l'accroissement de (1). Dans ces conditions, déterminons la variation des deux quantités intervenant dans l'expression (1) : $\text{card}(\text{gr}(\omega) \cap (R(\pi) \times S(\pi)))$ et r.s.

Désignons par π' la partition couvrant π, résultant de l'agrégation des deux classes D_j et $D_{j'}$; on a

$$R(\pi') = R(\pi) + D_j * D_{j'},$$
$$\text{et} \qquad S(\pi') = S(\pi) - D_j * D_{j'}, \tag{3}$$

(somme et différence ensemblistes) ; par conséquent,

$$R(\pi') \times S(\pi') = R(\pi) \times S(\pi) - R(\pi) \times (D_j * D_{j'}) + D_j * D_{j'} \times S(\pi) - (D_j * D_{j'}) \times (D_j * D_{j'}). \tag{4}$$

d'où, l'accroissement de card $(gr(\omega) \bigcap (R(\pi) \times S(\pi)))$, dans le remplacement de π par π' :

$$\text{card}\ \{gr(\omega) \bigcap (D_j * D_{j'}) \times S(\pi) - R(\pi) \times (D_j * D_{j'}) - (D_j * D_{j'}) \times (D_j * D_{j'}))\}\ ; \tag{5}$$

d'autre part, il est immédiat que la nouvelle valeur de r.s est

$$(r + n_j n_{j'})\ (s - n_j n_{j'})$$

où n_j = card (D_j) et $n_{j'}$ = card $(D_{j'})$

Exercice :

Déterminer la variation du critère (1) pour un pas de l'algorithme des transferts (cf. Chap. 2 (α)).

Il est important de noter que, contrairement aux cas précédents (cf. § 1 et 2 ci-dessus) la part locale affectée dans la réunion des deux classes, du critère global (1), fait intervenir toute l'information quant aux ressemblances définie par ω ainsi que chacune des deux partitions π et π'.

On peut envisager un critère plus local que le taux d'accroissement de la stratistique (1) ; et ce, en ne considérant que l'ensemble des paires restant séparées à un niveau donné. Supposons que ce niveau définisse la partition π et posons $K = S(\pi)$, $L = D_j * D_{j'}$, et M la partie complémentaire dans K de L. Pour juger de la signification de l'agrégation des deux classes D_j et $D_{j'}$, nous commencerons par considérer la statistique de base card $\{gr(\omega_K) \bigcap (L \times M)\}$ où ω_K est la restriction de ω à K. Cette statistique prend en compte l'ensemble des paires à réunir par rapport à celles qu'on laissera séparées.

Des considérations analogues à celles qui ont prévalu à la constitution de la statistique (1) de proximité ci-dessus (voir chapitre précédent) montrent que la distribution de l'indice qu'on vient d'envisager, lorsque ω_K décrit uniformément l'ensemble de tous les ordres totaux sur K ; respectivement, lorsque L décrit de façon uniforme l'ensemble de toutes les parties de K de cardinal ℓ, ℓ = card(L), est approximativement normale de moyenne $\ell \times m/2$ (m=card(M)) et de variance $\ell \times m(\ell+m+1)/12$. D'où le critère local.

$$\text{card}\{gr(\omega_K \bigcap (L \times M)) - \ell \times m/2\} / \sqrt{\ell \times m(\ell+m+1)/12}, \tag{6}$$

qu'on peut chercher à maximiser à chaque pas de la construction d'un arbre

de classifications.

V - NOEUDS SIGNIFICATIFS D'UN ARBRE DE CLASSIFICATIONS, CONCLUSION.

En raison même du degré de généralité des critères d'agrégation, basés sur l'ordonnance, que nous venons d'établir ; nous préférons commencer par utiliser, dans la construction d'un arbre détaillé de classifications, un critère local tenant étroitement compte de la structure des données. Le rôle très important du critère basé sur l'ordonnance consistera alors à interpréter de façon dynamique l'arbre des classifications et à en reconnaître les noeuds significatifs ; et ce, quel que soit le type de données.

Nous avons vu aux numéros 1 et 2 du paragraphe IV précédent que la part locale d'un critère global se réduisait à un critère essentiellement local ne faisant intervenir que les deux classes à réunir. D'où la justification a postériori des critères locaux que nous avons introduit au paragraphe III pour comparer deux classes de variables d'un même type. Ce qu'il y a de nouveau par rapport à d'autres méthodes de classification hiérarchique c'est que la distribution de la proximité entre deux classes de variables est appréhendée dans l'hypothèse N d'absence de liaison ; ce qui permet la référence à une échelle claire pour juger de la grandeur observée d'une mesure de proximité et pour comparer sans biais les mesures attachées à deux couples de classes. Nous accordons la plus grande importance à la proximité entre classes que nous avons étudiée au paragraphe III.1 et qui a donné lieu à l'algorithme de la vraisemblance du lien (A.V.L.), lequel a donné des résultats très fins et très nuancés pour n'importe quelle structure mathématique du tableau des données et quel que soit l'ensemble à classifier (celui indexant les lignes ou bien les colonnes).

Cependant dans les cas suivants :

(a) Classification d'un ensemble E d'objets ou de sujets décrits au moyen d'un ensemble V de variables numériques.

(b) Classification de l'ensemble des lignes (resp. colonnes) d'une table de contingence ou d'une juxtaposition de tables de contingences ;

On peut considérer la comparaison des résultats de A.V.L. couplé avec l'indice adéquat de proximité (cf. Chap. 2) avec les résultats du critère basé sur l'inertie expliquée (cf. § IV.1). Une telle étude comparative a été réalisée par B. Tallur dans le cas (b) ci-dessus sur la base de données en économie rurale (cf. partie II et 15).

Quelle que soit la nature de la donnée ; supposons établis l'arbre des classifications ainsi d'ailleurs que l'ordonnance associée à l'indice de similarité entre éléments de l'ensemble à classifier ; indice que comprend la définition du critère local ayant permis la construction de l'arbre. Le plus souvent un même pas de l'algorithme consiste dans la réunion d'une seule paire de classes qui réalise le seul maximum du critère local ; de sorte que le nombre de niveaux de l'arbre est presque aussi grand que le cardinal de l'ensemble à classifier. Il en résulte d'abord une certaine peine du spécialiste à s'y reconnaître ; d'autre part, et c'est le plus important, pour un degré de "finesse" de la partition produite à un niveau donné les distinctions entre classes ne sont pas aussi significatives les unes que les autres. D'où la nécessité d'un guidage statistique accompagnant la formation des classes et la condensation de l'arbre aux niveaux correspondants à ses noeuds significatifs. A cette fin, on attachera à chacun des niveaux de l'arbre détaillé des classifications, deux statistiques ; la

première, globale, est fournie par la formule (1) du paragraphe IV.3 précédent, sa valeur pour la partition du niveau i, notée $\sum_i$, mesure globalement l'adéquation de la partition à l'information quant aux ressemblances qu'exprime l'ordonnance. la deuxième statistique est locale, elle peut être fournie par la formule (6) ; il peut également s'agir de taux d'accroissement θ de la statistique globale $\sum$ entre deux niveaux successifs. La valeur de θ attachée au niveau i : $\theta_i = \sum_i - \sum_{(i-1)}$ mesure la contribution du noeud formé au i-ème niveau. Une très riche expérience a montré que la distribution de la statistique locale le long de la suite des niveaux est telle que sa valeur augmente lorsqu'une classe en cours de formation se confirme et décroît sensiblement devant l'arrêt de constitution d'une classe ayant quelque consistance au profit de la naissance de l'embryon d'une autre classe. Les niveaux associés aux maximums locaux de cette distribution correspondent par conséquent à des niveaux d'achèvement de classes. Nous retiendrons donc comme les plus significatifs les noeuds définis à ces niveaux qui seront seuls représentés dans l'arbre condensé.

Du point de vue que nous venons de présenter, le comportement de l'une ou de l'autre des deux statistiques locales envisagées, est quasiment le même. Toutefois les dernières expériences semblent montrer, dans des cas délicats, un plus grand accord avec l'interprétation du comportement de θ. Il est important d'insister sur le sens d'un noeud significatif ; ce dernier exprime, pour un certain degré de "finesse" dans l'interprétation, l'achèvement d'une classe. La statistique θ peut même accuser un minimum local devant la réunion de deux classes formées alors que ce lien peut particulièrement intéresser le spécialiste car, faiblement apparent, il était insoupçonné à ce niveau de la formation des classes.

On peut déduire la distribution de la statistique globale $\sum$: $\{\sum_i / 1 \leqslant i \leqslant \nu\}$, à partir de celle de : $\{\theta_i / 1 \leqslant i \leqslant \nu\}$, où ν est le nombre de niveaux de l'arbre ; toutefois, l'allure de la distribution de $\sum$, sur la suite des niveaux, a un intérêt propre. La fonction $i \to \sum_i$ a tendance à être d'abord croissante jusqu'à atteindre, avec une pente proche de l'horizontale, son maximum, d'où, après un petit palier, la fonction décroît brutalement. Tout se passe comme si, dans la suite des agrégations, on cherchait à atteindre la "meilleure" classification : celle, qui en un faible nombre de classes constitue le meilleur résumé global ; à partir de ce dernier, toute agrégation ne peut être que "contre nature" et se trouve par conséquent taxée par une forte baisse $\sum$.

Il s'agit là de la tendance générale de la distribution de $\sum$ sur la suite des niveaux ; il faut signaler que durant son intervalle de croissance générale, certaines réunions de classes ont pu être accompagnées par une certaine diminution de $\sum$; ce qui exprime que, considérée globalement, la partition du niveau résultant est moins en accord avec l'ordonnance initiale que celle qui la précède, laquelle sera retenue s'il s'agit de choisir entre les deux. Nous préférons dans ce cas ne pas trop insister sur de tels groupements laissant le soin au spécialiste d'en dégager la signification faiblement apparente pour le degré de finesse de la classification ; nous rejoignons ici ce que nous disions relativement à l'ap-

parition d'un minimum local de $\sum$ pouvant accompagner l'agrégation de deux sous classes d'une classe plus générale.

Pour rendre plus concrètes les considérations ci-dessus exprimées, on se reportera à la partie II traitant des exemples réels

BIBLIOGRAPHIE

[1] J. ACZEL, B. FORTE et C.T. Ng, *"Why the Shannon and Hartley entropies are "natural""*, Adv. Appl. Prob. 6, 131-146, Printed in Israel : © Applied Probability Trust, 1974.

[2] J.P. BENZECRI, *"Théorie de l'Information et Classification d'après un tableau de contingence"*, in *"l'Analyse des Données"*, tome I : *"La Taxinomie"*, Dunod, Paris, 1973.

[3] N. JARDINE et R. SIBSON, *"Mathematical Taxonomy"*, Wiley, London, 1971.

[4] I.C. LERMAN, *"Les bases de la classification automatique"*, Gauthier Villars, "collection Programmation", Paris, 1970.

[5] I.C. LERMAN, *"Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique"*, Cahiers du B.U.R.O. n°19, Paris, 1979.

[6] Mme M.H. NICOLAU, *"Analyse d'un algorithme de classification"*, Thèse de 3ème cycle, Univ. Paris VI (I.S.U.P.), Nov. 1972.

[7] L. ORLOCI, *"Information theory models for hierarchic and non-hierarchic classifications"*, in A.J. Cole (ed.), Numerical Taxonomy, Proceedings of the Colloquium in Numerical Taxonomy, Held in the University of St. Andrews, Sept. 1968, pp. 148-164, Academic Press, London.

[8] M. ROUX, *"Un algorithme de construction d'une hiérarchie de classifications"*, thèse de 3ème cycle, Univ. Paris VI (I.S.U.P.), 1968-69.

[9] P.H. A. SNEATH et R.R. SOKAL, *"Numerical Taxonomy"* W.H. Freeman and Company, San Francisco, 1971.

[10] J.L. BUARD, *"Gestion statistique des demandes d'actes biologiques. Typologie des unités fonctionnelles"*. Thèse de 3ème cycle, Université de Rennes I, Déc. 1980.

[11] J.L. CHANDON, S. PINSON, *"Analyse Typologique"* Masson, Paris, 1981.

[12] J.A. HARTIGAN, *"Clustering Algorithms"*, John Wiley, New-York, 1975.

[13] M. JAMBU, *"Classification automatique pour l'analyse des données"* tome 1, Dunod, Paris, 1978.

[14] I.C. LERMAN, *"Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité)"*, Rev. Math et Sc. Num., 8è année, n° 32, 1970.

[15] I.C. LERMAN, B. TALLUR, *"Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence"*, Publ. IRISA n° 127, et Rev. Stat. Appl., 1980, n° 28, 3.

[16] F. NICOLAU, *"Criterios de análise classificatória hierarquica baseados na funçao de distribuiçao"*, Faculté des Sciences de Lisbonne, Laboratoire de Statistique, Lisbonne 1980, Thèse de doctorat sou-

tenue en février 1981.

[17] A. PROD'HOMME, *"Indice d'explication des classes obtenues par une méthode de classification hiérarchique respectant la contrainte de contiguïté spatiale. Application à la viticulture girondine et à la construction de logements dans les Bouches du Rhône"*. Thèse de 3è cycle, Université de Rennes I, Déc. 1980.

[18] PH. VILLOING, *"Classification ascendante hiérarchique et indices de similarité sur données qualitatives nominales selon l'algorithme de la vraisemblance du lien"*. Thèse de 3ème cycle, Université de Rennes I, Déc. 1980.

[19] J.H., Jr. WARD, *"Hierarchical grouping to optimise an objective function"*, JASA, 58, 236-244.