La classification automatique de données quantitatives

1 Introduction

Parmi les méthodes de statistique exploratoire multidimensionnelle, dont l'objectif est d'extraire d'une masse de données des " informations utiles ", on distingue les méthodes d'analyse factorielle des méthodes de classification automatique. L'objectif des méthodes d'AF est entre autre la visualisation des données, la réduction du nombre de variables. L'objectif de la classification automatique est de former des groupes d'individus ou de variables afin de structurer un ensemble de données. On cherche souvent des groupes homogènes c'est à dire que les objets sont ressemblant à l'intérieur d'un même groupe. Les méthodes de classification se distinguent entre autre par la structure de classification obtenue (partition, recouvrement, hiérarchie, pyramide).
Ces techniques peuvent être utilisées dans de nombreux domaines comme :

- le domaine médicale : regrouper des patients afin de définir une thérapeutique adaptée à un type particulier de malades
- le domaine du marketing, afin de définir un groupe cible d'individus pour une campagne publicitaire. On parle alors souvent de segmentation à ne pas confondre avec la segmentation qui est une méthode de discrimination par arbre.

En anglais, le terme désignant les méthodes de classification automatique est Clustering. En biologie on parle souvent de Taxonomie et en Intelligence Artificielle on parle d'Apprentissage non supervisé. Il ne faut pas confondre les méthodes de classification avec les méthodes explicatives de discrimination dont l'objectif est d'expliquer une partition connue à priori, c'est à dire d'expliquer une variable qualitative dont chaque modalité décrit une classe de la partition, par un ensemble de variables de type quelconque (et non pas d'expliquer une variable qualitative comme en régression). En anglais le terme désignant les méthodes de discrimination est Classification et en I.A. on parle d'Apprentissage supervisé ou encore de Reconnaissance des formes (Pattern Recognition). Aujourd'hui le Data Mining (synonymes : Fouille de données, extraction de connaissance, KDD) et un champs d'application à l'interface de la statistique et des technologies de l'information (bases de données, I.A., apprentissage). On définie parfois le Data Mining comme l'extraction de connaissances de grandes bases de données. Le Data Mining utilise donc souvent les méthodes d'A.D. comme les méthodes de classification, d'analyse discriminante.
Avant de pouvoir appliquer une méthode d'A.D. et en particulier l'analyse factorielle ou la classification, il faut généralement définir :

- le tableau de données et en particulier l'homogénéiser lorsque les variables sont de type différent. Les méthodes classiques s'appliquent en général sur des données toutes quantitatives ou toutes qualitatives.

– une mesure de ressemblance (distance, similarité, dissimilarité) entre individus (lignes) ou variables (colonnes) du tableau

Dans ce cours, on présente des méthodes classiques de classification automatiques d'individus (objets) décrits dans un tableau de données quantitatives. On reprend donc les notations et le vocabulaire qui ont été déjà utilisés dans le cours sur l'ACP.

1.1 Tableau de données quantitatives

On considère un ensemble $\Omega = \{1, ..., i, ..., n\}$ de n individus décrits par p variables $X^1, \ldots, X^p$ dans une matrice $\mathbf{X}$ de n lignes et p colonnes :

$$\mathbf{X} = (x_{ij})_{n\times p} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} 1 \cdots j \cdots p \\ \left[\begin{array}{c} \cdot \\ \vdots \\ \cdots \quad x_{ij} \in \mathbf{R} \quad \cdots \\ \vdots \\ \cdot \end{array}\right. \end{array}.$$

Un individu $i \in \Omega$ est donc décrit par un vecteur $x_i \in \mathbf{R}^p$ (ligne i de X). Un poids w_i est associé à chaque individu i. Pour des données résultant d'un tirage aléatoire avec probabilité uniforme, on a généralement $w_i = 1/n$ pour tout i. Mais il peut être utile pour certaines applications de travailler avec des poids non uniformes (échantillons redressés, données aggrégées).

On dispose donc, comme en ACP, d'un nuage pondérés de n points-individus de $\mathbf{R}^p$.

1.2 Mesure de ressemblance

Une mesure de ressemblance entre individus peut être une similarité, une dissimilarité ou une distance.

Définition 1 *Un indice de similarité $s : \Omega \times \Omega \to \Re^+$ est tel que $\forall i, i' \in \Omega$ on a :*

$$\begin{aligned} s(i,i') &\geq 0 \\ s(i,i') &= s(i',i) \\ s(i,i) &= s(i',i') = smax \geq s(i,i') \end{aligned} \tag{1}$$

Si smax=1, alors s est une similarité normalisée.

Définition 2 *Un indice de dissimilarité (une dissimilarité) $d : \Omega \times \Omega \to \Re^+$ est tel que $\forall i, i' \in \Omega$ on a :*

$$\begin{aligned} d(i,i') &\geq 0 \\ d(i,i') &= s(i',i) \\ d(i,i) &= 0 \end{aligned} \tag{2}$$

Définition 3 *Une distance est une dissimilarité qui vérifie en plus l'inégalité triangulaire : $\forall i, i', k \in \Omega$ on a $d(i,i) \leq d(i,k) + d(k,i')$.*

Remarque : Il est facile de transformer un indice de similarité s en un indice de dissimilarité d. Il suffit de poser :

$$d(i,i') = smax - s(i,i')$$

La mesure de ressemblance utilisée varie en fonction du type des données c'est à dire du type des variables : tableau quantitative, qualitatif, ou mixte (qui doit alors être recodé).
Lorsque les données sont quantitatives les distances classiques sont :
• Les distances définies par :

$$d^2(x_i, x_{i'}) = (x_i - x_{i'})^t M (x_i - x_{i'})$$

et
– si $M = I$, d est la distance euclidienne simple,
– si $M = D_{1/s^2}$ (matrice diagonale des inverses des variances empiriques des p variables) on se ramène à une distance euclidienne simple entre variables réduites ($j$ième colonne divisée par l'écart-type empirique s^j). On parle de distance euclidienne normalisée par l'inverse de la variance,
– si $M = V^{-1}$, d est la distance de Mahalanobis.
• La distance de city-block ou de Manhattan :

$$d(i,i') = \sum_{j=1,\ldots,p} \mid x_{ij} - x_{i'j} \mid$$

• La distance de Chebychev, ou distance du max :

$$d(i,i') = \max_{j=1,\ldots,p} \mid x_{ij} - x_{i'j} \mid$$

En général, on utilise la distance euclidienne lorsque tous les paramètres ont une variance équivalente. En effet, si une variable a une variance bien plus forte, la distance euclidienne simple va accorder beaucoup plus d'importance à la différence entre les deux individus sur cette variable qu'à la différence entre les deux individus sur les autres variables. Il est préférable dans ce cas d'utiliser la distance euclidienne normalisée par l'inverse de la variance, afin de donner la même importance à toutes les variables. Cela revient à réduire tous les variables (les diviser par leur écart-type) et à calculer ensuite la distance euclidienne simple.

1.3 Les espaces de classification

Les espaces de classification classique sont les partitions et les hiérarchies.

Définition 4 *Une partition P en K classes de Ω est un ensemble $(C_1, \ldots, C_k, \ldots, C_K)$ de classes non vides, d'intersections vides deux à deux et dont le réunion forme Ω :*
– $\forall k \in \{1, \ldots K\},\ C_k \neq \emptyset$
– $\forall k, k' \in \{1, \ldots K\}\ C_k \cap C_{k'} = \emptyset$
– $\cup_{k=1,\ldots,K}, C_k = \Omega$

Par exemple, si $\Omega = \{1, \ldots, 7\}$ est un ensemble de 7 points du plan, $P_3 = (C_1, C_2, C_3)$avec $C1 = \{7\}$, $C_2 = \{5, 4, 6\}$ et $C_3 = \{1, 2, 3\}$ est une partition en trois classes de Ω.

Définition 5 *Une hiérarchie H de Ω est un ensemble de classes non vides (appellés paliers) qui vérifient :*
- $\Omega \in H$
- $\forall i \in \Omega,\ \{i\} \in H$ *(la hiérarchie contient tous les singletons)*
- $\forall A, B \in H, A \cap B \in \{A, B, \emptyset\}$ *(deux classes de la hiérarchie sont soit disjointes soit contenues l'une dans l'autre)*

Par exemple, $H = \{\{1\}, \ldots, \{7\}, \{4,5\}, \{2,3\}, \{4,5,6\}, \{1,2,3\}, \{4,5,6,7\}, \Omega\}$ est une hiérarchie de $\Omega = \{1, \ldots, 7\}$. Les classes de la hiérarchie une fois indicée sont représentées dans un arbre appellé arbre hiérarchique. Si l'on coupe un arbre hiérarchique par une suite de lignes horizontales, on obtient une suite de partitions emboîtées.

2 Méthodes de partitionnement

Etant donné un ensemble Ω d'individus, la structure classificatoire recherchée est la partition. Si on définit un critère de qualité W sur une partition $P = (C_1, \ldots, C_k, \ldots, C_K)$ mesurant généralement l'homogénéité des classes, le problème de classification est un problème d'optimisation parfaitement définit :
Trouver, parmi l'ensemble $\mathcal{P}_K(\Omega)$ de toutes les partitions possibles en K classes de Ω, la partition qui optimise le critère $W : \mathcal{P}_K(\Omega) \to \mathbf{R}^+$
Comme Ω est fini,$\mathcal{P}_K(\Omega)$ est fini et le problème d'optimisation est soluble par énumération compète. En fait, dès que le nombre d'individus n de Ω est assez grand, on a approximativement $card(\mathcal{P}_K(\Omega)) \sim \frac{K^n}{K!}$. L'énumération complète de toutes les solutions possibles est donc irréalisable en pratique dès que n devient grand.
En pratique, on applique souvent un algorithme itératif du type :
- On part d'une solution réalisable c'est à dire d'une partition P^0
- A l'étape m+1, on cherche une partition $P^{m+1} = g(P^m)$ telle que $W(P^{m+1}) \leq W(P^m)$

Le critère W mesurant l'homogénéité des classes (la qualité de la partition) est souvent l'inertie intra-classe de la partition et l'algorithme itératif utilisé pour minimiser (localement) ce critère est l'algorithme des centres mobiles (k-means).

2.1 Inerties d'un nuage de points pondérés

On considère le nuage des n points de $\Omega = \{1, ..., i, ..., n\}$ pondérés par w_i et décrit par un vecteur $x_i \in \mathbf{R}^p$. On note :

$$\begin{aligned} \mu_k &= \sum_{i \in C_k} w_i \text{ poids de } C_k \\ g_k &= \frac{1}{\mu_k} \sum_{i \in C_k} w_i x_i \text{ centre de gravité de } C_k \\ d_M(x_i, x_{i'}) &= {}^t(x_i - x_{i'}) M (x_i - x_{i'}) \text{ où M est une matrice symétrique définie positive} \end{aligned}$$

Définition 6 *L'inertie I_a du nuage des n individus par rapport à un point $a \in \mathbf{R}^p$ est*

$$I_a = \sum_{i=1}^{n} w_i d_M^2(x_i, a) \tag{3}$$

Définition 7 *L'inertie totale T du nuage des n individus est*

$$T = \sum_{i=1}^{n} w_i d_M^2(x_i, g) \tag{4}$$

et g est le centre de gravité du nuage. L'inertie totale est indépendante de la partition

Définition 8 *L'inertie inter-classe B de la partition P est*

$$B = \sum_{k=1}^{K} \mu_k d_M^2(g_k, g) \tag{5}$$

B est donc l'inertie du nuage des centres de gravité des K classes, munis des moids μ_k

Définition 9 *L'inertie intra-classe W de la partition P est*

$$W = \sum_{k=1}^{K} I(C_k) \tag{6}$$

et

$$I(C_k) = \sum_{i \in C_k} w_i d_M^2(x_i, g_k) \tag{7}$$

Remarque 1 *En général avec $M = I$, on choisit :*
- $w_i = \frac{1}{n}$, on parle alors de variance intra et inter-classe
- $w_i = 1$,on parle alors de somme des carrés intra et inter-classe

Remarque 2 *Lorsque $w_i = \frac{1}{n}$ et $M = I$, l'inertie totale T est égale à la somme des variances des p variables. Si le tableau est réduit, alors $T = p$*

Remarque 3 *Les inerties calculées sur le tableau centrée-réduit sont égales aux inerties calculées avec la distance euclidienne simple sur le tableau des p composantes factorielles de l'ACP.*

Définition 10 *Le pourcentage d'inertie expliquée d'une partition P est :*

$$(1 - \frac{W}{T}) \times 100 \tag{8}$$

Ce critère varie entre zero et cent. Il vaut cent pour la partition en n classes des singletons, et zero pour la partition réduite à une classe Ω. Le pourcentage d'inertie expliquée augmentant avec le nombre de classe, il ne permet donc que de comparer deux partitions ayant le *même nombre de classes*. Si le pourcentage d'inertie expliquée d'une partition est supérieur au pourcentage d'inertie expliquée d'une autre partition ayant le même nombre de classe, on considérera que la première partition est meilleure que la seconde, au sens du critère d'inertie.

Proposition 1 *On a la relation fondamentale suivante :*

$$T = W + B \tag{9}$$

On en déduit que minimiser l'inertie intra-classe c'est à dire l'homogénéité des classes est équivalent à maximiser l'inertie inter-classe, c'est à dire la séparation entre les classes. Cette relation se déduit du théorème de Huygens (voir la démonstration section 4.1)

Proposition 2 *L'inertie d'une classe C_k s'écrit également indépendamment du centre de gravité, en ne faisant intervenir que les distances des individus deux à deux :*

$$I(C_k) = \sum_{i \in C_k} p_i d^2(x_i, g_k) = \sum_{i \in C_k} \sum_{i' \in C_k} \frac{p_i p_{i'}}{2\mu_k} d^2(x_i, x_{i'}) \tag{10}$$

Cette égalité se déduit du théorème de Koenig-Huygens et de la définition de l'inertie par rapport à un point (voir la démonstration section 4.2)

Proposition 3 *Une relation importante est :*

$$I(C_1 \cup C_2) = I(C_1) + I(C_2) + \frac{(\mu_1 \mu_2)}{(\mu_1 + \mu_2)} d^2(g_1, g_2) \tag{11}$$

(voir la démonstration section 4.3)

2.2 Algorithme des centres mobiles

L'algorithme des centres mobiles est un cas particulier de l'algorithme des Nuées dynamiques qui part d'une partition intitiale et itère :

- une **étape de représentation** qui consiste à définir un centroide pour chaque classe, ici le centre de gravité
- une **étape d'affectation** où l'on affecte chaque individu à la classe dont le centroide est le plus proche, ici au sens de la distance euclidienne.

Plus précisément, l'algorithme des centres mobiles est le suivant :

(a) Initialisation

On se donne une partition $P = (C_1, \ldots, C_k, \ldots, C_K)$ et on calcul $g_1, \ldots, g_k$

(b) Etape d'affectation

$test \leftarrow 0$

Pour tout i de 1 à n faire

déterminer la classe $k*$ telle que

$$k* = arg \min_{k=1,\ldots,K} d(i, g_k)$$

déterminer la classe l de i

si $k* \neq l$

$test \leftarrow 1$

$C_k \leftarrow C_k \cup \{i\}$

$C_l \leftarrow C_l \backslash \{i\}$

(c) Etape de représentation

Pour tout k de 1 à K calculer le centre de gravité de la nouvelle classe C_k

(d) Si $test = 0$ FIN, sinon aller en (b)

Proposition 4 *Cette algorithme converge vers une partition réalisant un minimum local de l'inertie intra-classe (voir la démonstration section 4.4).*

Remarque 4 *Souvent, l'étape d'initialisation consiste à prendre K individus comme centres initiaux, la partition initiale étant obtenue par affectation des individus au centre le plus proche. Ces K individus peuvent être (entre autre) :*

– *les K premiers individus dans le tableaux,*

– *K individus tirés au hasard,*

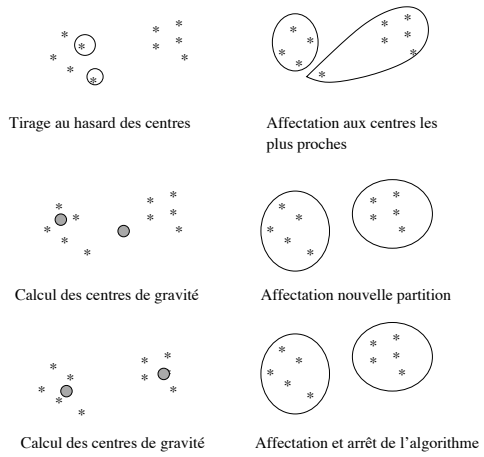

FIG. 1 – Exemple de déroulement de l'algorithme des centres mobiles

– *K individus éloignés (en terme de distance) les uns des autres.*

Bien sûre, la partition finale obtenue après convergence de l'algorithme (ou après un nombre fini d'itérations) dépend de la partition initiale et donc des centres initiaux. Le minimum local de W trouvé dépend en effet du "point" de départ et donc de la partition initiale. En pratique, on peut répéter N fois l'algorithme en tirant à chaque fois K individus au hasard comme centres initiaux. On retient alors la partition finale qui a la pourcentage d'inertie expliqué le plus grand (qui donne la plus petite valeur de l'inertia intra-classe W). C'est possible car la complexité de l'algorithme des centres mobiles est o(KpnT) où T est le nombre d'itérations. L'algorithme va donc "fonctionner" avec beaucoup d'individus et peut donc être répété sans que cela "coûte trop cher".

3 Les méthodes de classification hiérarchiques

La structure classificatoire recherchée est maintenant la hiérarchie. On a vu qu'une hiérarchie H de Ω est un ensemble de classes de Ω appellés paliers comprenant les singletons (classes réduites à un élément), l'ensemble Ω, et des classes dont l'intersection est soit vide,soit l'une ou l'autre des classes.

Définition 11 *Une hiérarchie binaire est une hiérarchie dont chaque palier est la réunion de deux paliers. Le nombre de paliers non singletons d'une hiérarchie binaire vaut $n-1$.*

Cette définition d'une hiérarchie est ensembliste. Maintenant, pour pourvoir représenter une hiérarchie par un graphique, il faut pouvoir valuer ses paliers c'est à dire leur attribuer une hauteur.

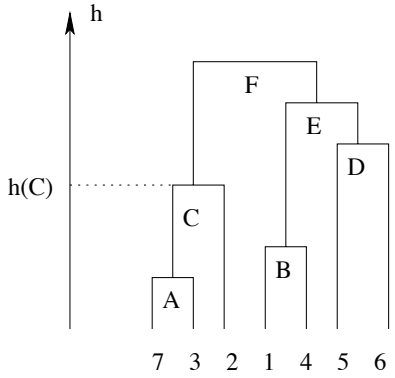

FIG. 2 – Exemple de dendogramme d'une hiérarchie indicée

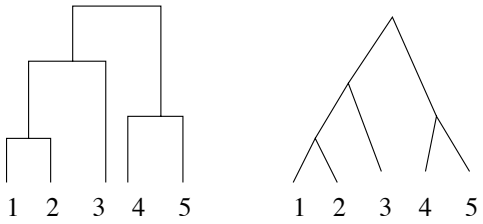

FIG. 3 – Deux formats possibles d'un dendogramme

Définition 12 *Une hiérarchie indicée est un couple (H, h) où H est une hiérarchie et h est une application de H dans $\Re^+$ telle que :*

$$\begin{aligned} \forall A \in H, h(A) = 0 &\Leftrightarrow A \text{ est un singleton} \\ \forall A, B \in H, A \neq B, A \subset B &\Rightarrow h(A) \leq h(B) \end{aligned} \tag{12}$$

Le graphique représentant une hiérarchie indicée est appelé un dendogramme ou arbre hiérarchique.

Remarque 5 *Une hiérarchie indicée définie une suite de partitions emboîtées de 2 à n classes. Elles sont obtenues en coupant l'arbre hiérarchique suivant une suite de lignes horizontales. De plus, la condition 12 vérifiée par l'indice h assure qu'il n'y a pas d'inversions dans la représentation de la hiérarchie, c'est à dire que si $C = A \cup B$ le palier C est bien représenté plus haut que les paliers A et B.*

Il existe plusieurs formats possibles pour représenter un dendogramme (figure 3) et surtout de nombreuses représentations équivalentes d'une même hiérarchie indicée. En effet on peut facilement permuter l'ordre dans lequel sont représentés les individus (voir figure 4). Si n est le nombre d'individus de Ω, il y a 2^{n-1} représentations possibles équivalentes d'une hiérarchie indicée par un arbre binaire.

Remarque 6 *Il existe d'autres modes de représentation d'une hiérarchie que le dendogramme permettant, lorsque le nombre d'individu est grand, et le dendogramme difficilement lisible, de trouver rapidement la classe d'un individu. Par exemple le stalactite (icicle plot) du tableau 1 correspond à la hiérarchie des figures 3 et 4. A chaque niveau,*

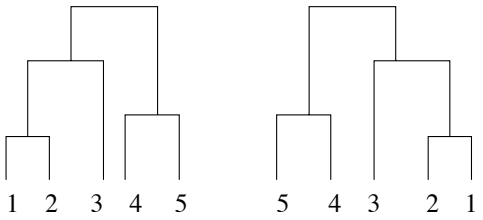

FIG. 4 – Deux représentations équivalentes d'une hiérarchie indicée

les objets dans une même classe sont liés par le signe égal. On lit ainsi que la première classe crée est la classe $\{1, 2\}$ *puis la classe* $\{4, 5\}$ *etc...*

1	=	2	=	3	=	4	=	5
1	=	2	=	3		4	=	5
1	=	2				4	=	5
1	=	2						

TAB. 1 – Exemple de stalactite horizontal

Il existe deux stratégies de construction d'une hiérarchie indicée :
- on construit la hiérarchie en partant du bas de l'arbre (des singletons) et on agrège, deux par deux les classes les plus proches, et ce jusqu'à l'obtention d'une seule classe. On parle de classification ascendante hiérarchique (C.A.H.).
- on construit la hiérarchie à partir du haut de l'arbre en procédant par divisions succéssives de l'ensemble Ω jusqu'à obtenir des classes réduites à un élément, ou des classes ne contenant que des individus indentiques. On parle de classification divisive ou classification descendante hiérarchique.

La première stratégie étant la plus utilisée et la plus présente dans la littérature, c'est celle que nous présenterons ici.

3.1 Algorithme général de C.A.H.

L'algorithme général est le suivant :

(a) Initialisation
On part de la partition la plus fine, c'est à dire la partition des singletons $P = (C_1, \ldots, C_n)$ avec $C_k = \{k\}$

(b) étape agrégative
On a part de la partition $P = (C_1, \ldots, C_K)$ en K classes obtenue à l'étape précédente et on agrège les deux classes C_k et $C_{k'}$ qui minimisent une mesure d'agrégation $D(C_k, C_{k'})$. On construit ainsi une nouvelle partition en $K-1$ classes. En cas d'égalité, on choisit la première solution rencontrée. La hiérarchie obtenue est donc toujours binaire.

(c) On recommence l'étape (b) jusqu'à obtenir la partition la plus grossière c'est à dire la partition en une seule classe Ω

Il faut donc choisir au préalable :
- la mesure d'agrégation $D : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \Re^+$ qui mesure la ressemblance entre deux classes,
- l'application h qui va indicer la hiérarchie.

3.2 Les mesures d'agrégations entre classes

Trois mesures d'agrégation classiques (il en existe d'autres) entre deux classes A et B de Ω, sont représenté figure 5. Ces mesures utilisent la distance ou la dissimilarité d choisie pour comparer deux individus.

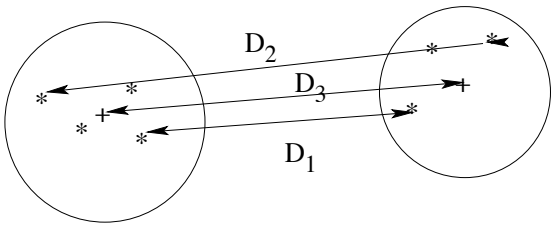

FIG. 5 – Trois mesures classiques d'agrégation

Définition 13 *La mesure d'agrégation du lien minimum entre deux classes A et B est :*

$$D_1(A,B) = \min_{i\in A, i'\in B} d(i,i') \tag{13}$$

La hiérarchie obtenue par C.A.H. avec cette mesure est la hiérarchie du lien minimum. En anglais on parle de "single linkage algorithm".

Définition 14 *La mesure d'agrégation du lien maximum entre deux classes A et B est :*

$$D_2(A,B) = \max_{i\in A, i'\in B} d(i,i') \tag{14}$$

La hiérarchie obtenue par C.A.H. avec cette mesure est la hiérarchie du lien maximum. En anglais on parle de "complete linkage algorithm".

Définition 15 *La mesure d'agrégation d'augmentation d'inertie ou mesure de Ward entre deux classes A et B est :*

$$D_3(A,B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} d^2(g_A, g_B) \tag{15}$$

où

$$\begin{aligned} \mu_A &= \sum_{i\in A} w_i \\ g_A &= \frac{1}{\mu_A} \sum_{i\in A} w_i x_i \end{aligned}$$

La hiérarchie obtenue par C.A.H. avec cette mesure est la hiérarchie de Ward.

Remarque 7 *Le hiérarchie de Ward ne peut donc être construite qu'à partir d'un tableau de données quantitative tandis que les hiérarchies du lien min et du lien max peuvent ête obtenue à partir d'un tableau de ressemblances et donc pour n'importe quel type de données.*

3.3 Définition de l'indice

La relation généralement utilisée pour définir l'indice h une hiérarchie H construite par C.A.H. est :

$$\forall A, B \in H, h(A \cup B) = D(A, B) \tag{16}$$

Cependant, certaines mesures d'agrégation comme la distance euclidienne entre les centres de gravités des classes, ne permettent pas de construire, grâce à cette relation, une hiérarchie indicée. En effet, l'indice h définie par $h(A \cup B) = d(g_A, g_B)$ ne vérifie par la condition 12, et le dendogramme de cette hiérarchie peut posséder des inversions c'est à dire que deux paliers agrégés avant deux autres peuvent être représentés plus bas que ces derniers (voir figure 6). Lorsque les données sont numériques, on lui préfère donc la mesure d'agrégation de Ward. Mais pour éviter ces inversions, on peut également utiliser la relation suivante :

$$\forall A, B \in H, h(A \cup B) = max(D(A, B), h(A), h(B)) \tag{17}$$

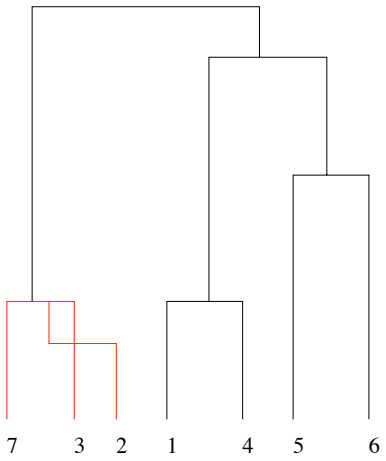

FIG. 6 – Exemple d'inversion dans le dendrogramme d'une hiérarchie

4 Annexes

4.1 Démonstration du théorème de Huygens et de la proposition 1

Théorème 1 (Huygens) *Si* $g = \frac{1}{\sum_{i=1}^{n} p_i} p_i x_i$ *est le centre de gravité du nuage des* n *individus, on a :*

$$\forall a \in \mathbf{R}^p,\ I_a = I_g + (\sum_{i=1}^{n} p_i) d_M^2(a, g)$$

Démonstration du théorème de Huygens :

D'après la définition 6 on a :

$$\begin{aligned} I_a &= \sum_{i=1}^{n} w_i \, ||x_i - a||_M^2 \\ &= \sum_{i=1}^{n} w_i \, ||x_i - g + g - a||_M^2 \\ &= \sum_{i=1}^{n} w_i \, ||x_i - g||_M^2 + 2 \sum_{i=1}^{n} w_i \, {}^t(g-a) M((x_i - g) \\ &+ \sum_{i=1}^{n} w_i \, ||g - a||_M^2 \text{ (car } M \text{ est symétrique)} \\ &= I_g + (\sum_{i=1}^{n} w_i) d_M^2(a, g) + 2 \, {}^t(g-a) M \sum_{i=1}^{n} w_i (x_i - g) \end{aligned}$$

Or $\sum_{i=1}^{n} w_i(x_i - g) = 0$ d'où le résultat.

Démonstration de la proposition 1 :
On considère g_k le centre de gravité de la classe C_k. D'après le théorème de Huygens, l'inertie du nuage des points de C_k par rapport au centre de gravité g s'écrit :

$$I_g = I_{g_k} + \overbrace{\sum_{i \in C_k} w_i}^{\mu_k} d_M^2(g, g_k)$$

soit

$$\sum_{i \in C_k} w_i d_M^2(x_i, g) = \sum_{i \in C_k} w_i d_M^2(x_i, g_k) + \mu_k d_M^2(g, g_k)$$

En sommant cette égalité pour k variant de 1 à K, on trouve $T = W + B$.

4.2 Démonstration de la proposition 2

On considère g_k le centre de gravité de la classe C_k. D'après le théorème de Huygens, l'inertie du nuage des points de C_k par rapport à un point x_i s'écrit :

$$I_{x_i} = I_{g_k} + \overbrace{\sum_{i \in C_k} w_i}^{\mu_k} d_M^2(x_i, g_k)$$

En multipliant cette égalité par p_i et en sommant pour i dans C_k ont trouve :

$$\sum_{i \in C_k} w_i I_{x_i} = \overbrace{\sum_{i \in C_k} w_i}^{\mu_k} I_{g_k} + \mu_k \overbrace{\sum_{i \in C_k} w_i d_M^2(x_i, g_k)}^{I_{g_k}}$$

En remplacant I_{x_i} par sa valeur dans cette égalité on trouve :

$$\sum_{i \in C_k} w_i p_{i'} d_M^2(x_i, x_{i'}) = \mu_k I_{g_k} + \mu_k I_{g_k}$$

et comme $I_{g_k} = I(C_k)$ on retrouve bien l'égalité de la proposition 2 :

$$I(C_k) = \sum_{i \in C_k} w_i d^2(i, g_k) = \sum_{i \in C_k} \sum_{i' \in C_k} \frac{w_i w_{i'}}{2\mu_k} d^2(x_i, x_{i'})$$

4.3 Démonstration de la proposition 3

Si on considère le nuage de points $C_1 \cup C_2$ et la partition $P = (C_1, C_2)$ de ce nuage, on a $T = I(C_1 \cup C_2)$, $W = I(C_1) + I(C_2)$ et $B = \mu_1 d_M^2(g_1, g) + \mu_2 d_M^2(g_2, g)$. En appliquant la relation fondamentale $T = B + W$ on trouve :

$$I(C_1 \cup C_2) = I(C_1) + I(C_2) + \mu_1 d_M^2(g_1, g) + \mu_2 d_M^2(g_2, g)$$

Puis on applique la proposition 2 au nuage des deux centres de gravités g_1 et g_2 munis des poids μ_1 et μ_2 et de centre de gravité g :

$$\mu_1 d_M^2(g_1, g) + \mu_2 d_M^2(g_2, g) = \frac{\mu_1\mu_2}{2(\mu_1 + \mu_2)} d_M^2(g_1, g_2) + \frac{\mu_2\mu_1}{2(\mu_2 + \mu_1)} d_M^2(g_2, g_1)$$

On retrouve donc bien :

$$I(C_1 \cup C_2) = I(C_1) + I(C_2) + \frac{\mu_1\mu_2}{\mu_1 + \mu_2} d_M^2(g_1, g_2)$$

4.4 Démonstration de la proposition 4

A l'étape m on a

$$\begin{aligned} g^m &= (g_1^m, \ldots, g_K^m) \\ P^m &= (C_1^m, \ldots, C_K^m) \end{aligned}$$

où g^m est le vecteur des centres de gravités des classes C_k^{m-1} et C_k^m est la classes des individus plus proches de g_k^m que des autres centres. L'inertie intra-classe de la partition P^m est :

$$W(P^m) = \sum_{k=1}^{K} \sum_{i \in C_k^m} d^2(i, g_k^{m+1})$$

On définit :

$$\begin{aligned} v_m &= \sum_{k=1}^{K} \sum_{i \in C_k^m} d^2(i, g_k^m) \\ v_{m+1} &= \sum_{k=1}^{K} \sum_{i \in C_k^{m+1}} d^2(i, g_k^{m+1}) \end{aligned}$$

et on montre que $\forall m$

$$v_{m+1} \leq W(P^m) \leq v_m \tag{18}$$

On montre d'abord que $W(P^m) \leq v_m$. En effet, par définition, si g_k^{m+1} est le centre de gravité de la classe C_k^m, alors

$$\sum_{i \in C_k^m} d^2(i, g_k^{m+1}) \leq \sum_{i \in C_k^m} d^2(i, g_k^m)$$

Donc en sommant sur k, on a bien

$$\overbrace{\sum_{k=1}^{K} \sum_{i \in C_k^m} d^2(i, g_k^{m+1})}^{W(P^m)} \leq \overbrace{\sum_{k=1}^{K} \sum_{i \in C_k^m} d^2(i, g_k^m)}^{v_m}$$

On montre ensuite que $v_{m+1} \leq W(P^m)$. En effet, on a toujours par construction des classes C_k^{m+1}

$$\forall i \in C_k^{m+1}, d^2(i, g_k^{m+1}) \leq d^2(i, g_k^m)$$

Donc en sommant sur i puis sur k, on a

$$\overbrace{\sum_{k=1}^{K} \sum_{i \in C_k^{m+1}} d^2(i, g_k^{m+1})}^{v_{m+1}} \leq \overbrace{\sum_{k=1}^{K} \sum_{i \in C_k^{m}} d^2(i, g_k^{m+1})}^{W(P^m)}$$

D'après 18, on aura bien

$$W(P^{m+1}) \leq v_{m+1} \leq W(P^m)$$

et donc la décroissante du critère d'inertie intra-classe.

5 Reférences

LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimentionnelle*. Dunod

GORDON A.D., *Classification*, (2de édition) Chapman.

NAKACHE J.P., CONFAIS J. (2000), *Méthodes de classification*, CISIA-CERESTA.