Revue Belge de Statistique, d'Informatique et de Recherche Opérationnelle, Vol.23, n°3.
Belgisch Tijdschrift voor Statistiek, Informatica en Operationeel Onderzoek, Vol.23, Nr 3.

TUTORIAL PAPER XV.

CLASSIFICATION AUTOMATIQUE

G.LIBERT

Faculté Polytechnique
Rue de Houdain, 9
7000 MONS.

SUMMARY : Very diverse procedures have been proposed in order to separate a set of objects into classes. This paper presents an overview of the methods. Firstly, one considers the hierarchical methods for metric or non metric data with agglomerative or divisive algorithms. Then one recalls some non hierarchical methods the resulting clusters of which form a partition or an overlapping partition or a fuzzy partition. Finally, one gives various criteria for the assessment of classifications produced by all these methods.

1. INTRODUCTION

Lorsqu'on souhaite une représentation algébrique d'un tableau de données, on essaie de lui associer une structure définie par des groupes. On utilise alors des techniques de classification automatique qui découpent l'ensemble des individus de la population en divers sous-ensembles de points homogènes.
Il convient d'insister immédiatement sur le fait que normalement on ne sait s'il existe une structure dans les données. De plus qu'appelle-t-on structure ? Quels sont les caractéristiques d'une structure ? En fait, la classification apparaît avant tout comme un problème mal posé. On recherche des groupes sans avoir défini préalablement ce qu'est un groupe.

C'est peut-être la raison pour laquelle tant de méthodes de classification ont été développées. Toutes se justifient compte tenu des données analysées et des structures recherchées. Classiquement, on les divise en deux classes : celles qui constituent une hiérarchie et celles qui recherchent une partition.

Les algorithmes hiérarchiques élaborent des classes par agglomérations successives des individus deux à deux ou par dichotomies successives de l'ensemble des individus. Les partitions quant à elles sont obtenues par des algorithmes directs optimisant un critère.

De nombreuses questions surgissent dès cette définition plus précise des méthodes.

Comment apprécie-t-on les ressemblances entre les individus ?
C'est le problème du choix de la distance.

Doit-on prendre les variables telles quelles ou doit-on les transformer ? En d'autres termes, les échelles de mesure des variables reflètent-elles bien les écarts que l'on veut voir entre individus ?

Comment choisir les individus que l'on agrège dans une hiérarchie ? Selon quel critère admet-on que deux individus différents peuvent être considérés comme semblables ?

Quel critère optimiser lorsqu'on recherche une partition ? Quelles sont les conséquences de ce choix sur la structure obtenue ?

Combien y a-t-il de groupes dans les données ?

Confronté à toutes ces questions, plusieurs indications de réponse peuvent être formulées :

- on interroge le fournisseur des données sur le sens de ses mesures

- on décrit au mieux les données par des techniques multidimensionnelles descriptives

- on choisit la méthode en fonction de ses caractéristiques et de l'objectif poursuivi

- on valide les résultats en utilisant d'autres méthodes, d'autres distances, d'autres critères d'agrégation, etc.

Ces problèmes pratiques soulevés, la description des méthodes peut être entreprise. La donnée de base est la matrice, D, de distances ou d'indices de distance (d(i,i')) entrè tous couples d'individus i et i'. Cette matrice est soit fournie directement soit calculée après avoir résolu le problème de codage lié à la transformation des variables et au choix de la distance

2. LES METHODES HIERARCHIQUES OU ARBORESCENTES

2.1. Hiérarchie indicée et ultramétrique

L'objet de la classification hiérarchique est de construire une hiérarchie des parties de l'ensemble des individus I, notée H, c'est-à-dire une famille des parties de I, vérifiant les propriétés :

$\forall\, k,k' \in H;\ k \cap k' \neq \phi \Rightarrow k \subset k' \quad ou \quad k' \subset k$

$I \in H$

$\forall\, i \in I;\ i \in H$

Cette hiérarchie peut être représentée par un dendrogramme.
Soit un indice appelé diamètre (d) présentant les propriétés suivantes :

$d(i) = 0$

$d(I)$ est maximum

$d(k) > d(k')$ si $k \supset k'$

Une hiérarchie H munie d'un diamètre d est appelée hiérarchie indicée et est notée H_d

Une distance ultramétrique est définie par

$d(i,i') = 0$ si et seulement si $i = i'$

$d(i,i') = d(i',i)$

$d(i,i') \leqslant MAX\ \{d(i,i^*), d(i',i^{**})\}$, condition plus forte que l'inégalité triangulaire $d(i,i') \leqslant d(i,i^{**}) + d(i',i^{**})$ car tous les triangles sont équilatéraux ou isocèles à petite base.
Il existe une correspondance biunivoque entre l'ensemble des hiérarchies indicées construites sur I et l'ensemble des ultramétriques définies sur l'espace produit I x I.

Le problème de la classification hiérarchique se ramène donc à la construction d'une ultramétrique qui déforme le moins possible les indices de distance donnés.
Plusieurs procédures sont proposées suivant qu'elles agrègent deux à deux les individus ou qu'elles séparent par dichotomies successives l'ensemble des individus et suivant qu'elles utilisent des indices de distance ou des distances.

2.2. Classification hiérarchique ascendante

On regroupe séquentiellement les points qui sont les plus proches et on associe au niveau correspondant de la hiérarchie la distance entre les points agrégés. On appelle points les individus initiaux ou les groupes formés.
Il reste donc à définir la distance entre points. Selon cette définition, on obtient diverses méthodes.

2.2.1. Méthodes non métriques

- Critère d'agrégation suivant le saut minimum (single linkage).

La distance entre deux points est donnée par

$$d(k,k') = \underset{\substack{i\in k \\ i'\in k'}}{\text{MIN}}\ d(i,i')$$

et on agrège chaque fois les points les plus proches. La hiérarchie créée ne dépend donc que de l'ordonnance de départ des indices et non des valeurs spécifiques.
La construction qui en résulte s'apparente à la recherche de l'arbre minimal (minimum spanning tree : MST) et l'ultramétrique est l'ultramétrique inférieure maximale. C'est l'ultramétrique la plus proche du tableau des indices de distance initiaux parmi celles qui réduisent ou conservent ces indices.
Comme il suffit d'un lien étroit entre deux sous-ensembles pour qu'ils soient agrégés, on peut s'attendre à un effet de chaînage.

- Critère d'agrégation suivant le diamètre maximum (complete linkage)

La distance est

$$d(k,k') = \underset{\substack{i\in k \\ i'\in k'}}{\text{MAX}}\ d(i,i')$$

et on agrège les points dont le plus grand lien est le plus petit. Ceci conduit à créer des agrégats "sphériques". La hiérarchie ne dépend que de l'ordonnance de départ des indices et on obtient l'ultramétrique supérieure minimale.
Des faibles variations des données peuvent entraîner des inversions et des hiérarchies très différentes.

- Critère d'agrégation de la moyenne des indices de distance (average linkage)

La distance est

$$d(k,k') = \frac{1}{n_k \, n_{k'}} \sum_{\substack{i\in k \\ i'\in k'}} d(i,i')$$

n_k : effectif de la classe k

2.2.2. Méthodes métriques

Soit l'ensemble I des individus $X_i = (x_{i1},\ldots,x_{ip})'$, $(i = 1,\ldots,n)$, points de l'espace euclidien R^p muni de la norme notée

$\|.\|$ et munis des poids p_i.

$\|.\|^2$ représente la carré de la distance pour la norme $\|.\|$

Si k est une classe (ou partie de I), on définit

$$p_k = \sum_{i\in k} p_i$$

$$\bar{X}_k = \frac{1}{p_k} \sum_{i\in k} p_i \, X_i \quad : \text{barycentre de k}$$

$$M_k^2 = \sum_{i\in k} p_i \, \| X_i - \bar{X}_k \|^2 \quad : \text{moment centré d'ordre 2 de k}$$

= variabilité du groupe k

$$V_k = \frac{1}{p_k} M_k^2 \quad : \text{variance de k}$$

On dispose de la relation de KOENIG-HUYGENS

$$\sum_{i\in k} p_i \, \| X_i - X_{i'} \|^2 = M_k^2 + p_k \, \| \bar{X}_k - X_{i'} \|^2 \quad \forall i' \in R^p$$

On considère alors 1,2,...,k,...,q avec $k \cap k' = \emptyset \quad k \neq k'$

$Uk = I$

Il vient

$$\sum_{i \in k} p_i \, \| X_i - \bar{X}_I \|^2 = M_k^2 + p_k \, \| \bar{X}_k - \bar{X}_I \|^2$$

et

$$\sum_{k} \sum_{i \in k} p_i \, \| X_i - \bar{X}_I \|^2 = \sum_{k} M_k^2 + \sum_{k} p_k \, \| \bar{X}_k - \bar{X}_I \|^2$$

$$M_I^2 = \sum_{k=1}^{q} M_k^2 + M_{1,\ldots,q}^2$$

La variabilité totale se décompose en une somme de variabilités intra-groupes et une variabilité inter-groupes.

Soient 3 classes k,k',k" disjointes. On a

$M_{kUk'}^2 = M_k^2 + M_{k'}^2 + M_{k,k'}^2$ et on obtient aisément

$$M_{k,k'}^2 = \frac{p_k \, p_{k'}}{p_k + p_{k'}} \, \| \bar{X}_k - \bar{X}_{k'} \|^2$$

$$M_{k",kUk'}^2 = \frac{(p_{k"} + p_k) M_{k",k}^2 + (p_{k"} + p_{k'}) M_{k",k'}^2 - p_{k"} M_{k,k'}^2}{p_k + p_{k'} + p_{k"}}$$

- Critère d'agrégation du moment centré d'ordre 2 d'une partition maximum.

A l'étape (t-1), on a la partition $1^{(t-1)},\ldots,q_{t-1}^{(t-1)}$. On se propose de regrouper à l'étape t, $k'^{(t-1)} = k$ et $k"^{(t-1)} = k'$.

On a

$$M_I^2 = \sum_{k\neq k',k''} M^2_{k(t-1)} + M_k^2 + M_{k'}^2 + M^2_{l(t-1),\ldots,q_{t-1}^{(t-1)}}$$

$$= \sum_{k\neq k',k''} M^2_{k(t-1)} + M^2_{kUk'} + M^2_{l(t),\ldots,q_t^{(t)}}$$

On réunit les classes k et k' de manière à maximiser le moment centré d'ordre 2 de la partition, à savoir :

$$M^2_{l(t),\ldots,q_t^{(t)}}$$

Comme $M^2_{l(t-1),\ldots,q_{t-1}^{(t-1)}}$ ne peut être modifié (pas de rétroaction dans l'algorithme), maximiser le moment centré d'ordre 2 de la partition revient à minimiser

$$M^2_{kUk'} - M_k^2 - M_{k'}^2 = M^2_{k,k'}$$

On choisit donc, en une des étapes de l'algorithme, de réunir deux classes k et k' telle

$$d(k,k') = M^2_{k,k'} \quad \text{soit minimum}$$

Les classes k et k' étant agrégées, on met à jour les distances entre classes par la formule $M^2_{k'',kUk'}$

- Critère d'agrégation de la variance d'une partition maximum.

  Ce critère se déduit du précédent puisque $V_k = \frac{1}{p_k} M_k^2$

- Critère d'agrégation du moment centré d'ordre 2 de la réunion de deux classes minimun.

La distance entre 2 classes k et k' est fournie par

$$d(k,k') = M^2_{kUk'}$$

On agrège les classes dont le moment centré d'ordre 2 de leur réunion est minimum.

- Critère d'agrégation de la variance de la réunion de deux classes minimum.

La distance est

$$d(k,k') = V_{kUk'} = \frac{1}{P_{kUk'}} M^2_{kUk'}$$

- Critère d'agrégation de la distance minimum entre centres de gravité des classes (centroïd method).

La distance est donnée par

$$d(k,k') = \|\bar{X}_k - \bar{X}_{k'}\|^2$$

On regroupe les classes dont les centres de gravité sont les plus proches au sens de $\|.\|$. Cette procédure peut conduire à des inversions.

- Critère d'agrégation de la méthode minimum ou critère de la médiane.

La distance est définie par

$$d(k,k') = \|\bar{Y}_k - \bar{Y}_{k'}\|^2 \quad \text{avec} \quad \bar{Y}_k = \frac{1}{n_k} \sum_{i \in k} X_i$$: moyenne arithmétique des coordonnées.

D'autres critères existent comme ceux basés sur la théorie de l'information ou ceux conservant une préordonnance. Il est de peu d'intérêt de tous les présenter d'autant plus que ceux signalés précédemment peuvent se multiplier en fonction de la norme choisie. On peut ainsi utiliser la norme euclidienne usuelle, la norme euclidienne sur variables réduites, la norme de MAHALANOBIS, la norme du χ^2.

Pour un tableau de contingence, on choisit évidemment la norme du χ^2 et le critère de maximisation du moment centré d'ordre 2 afin de pouvoir comparer les résultats avec ceux d'une analyse des correspondances. Si on dispose d'un tableau de mesures, on travaille avec la norme euclidienne sur variables réduites et le critère du moment centré d'ordre 2 pour permettre le lien avec l'analyse en composantes principales normée.
La norme de MAHALANOBIS est coûteuse en temps de calcul car elle nécessite des inversions de matrices.

2.3. Classification hiérarchique descendante

On sépare par dichotomies successives les classes existantes en commençant par l'ensemble des individus et on associe au niveau correspondant de la hiérarchie une mesure de l'étalement de la classe avant son éclatement.

2.3.1. Méthode non métrique de L.HUBERT

La classe k" est éclatée en les classes k et k' si elle contient le plus grand diamètre. Le diamètre d'une classe k est donné par $\underset{i,i' \in k}{MAX} d(i,i')$

Les classes k et k' sont formées de la façon suivante

$$\forall i \in k; \exists\, i' \in k';\ d(i,i') \geqslant d(i,i'') \quad \forall i'' \in k$$

$$\forall i' \in k'; \exists\, i \in k\ ;\ d(i',i) \geqslant d(i'',i') \quad \forall i'' \in k'$$

On associe au niveau de séparation de la hiérarchie, le diamètre maximum de la classe k".

2.3.2. Méthodes métriques

- Méthode de MAC QUEEN

A chaque étape, on éclate le groupe k" ayant le plus grand moment centré d'ordre 2 $M^2_{k"}$. On associe au niveau de la hiérarchie cette valeur.
Les groupes k et k' sont constitués de façon à minimiser M^2_k et $M^2_{k'}$. Cependant, comme il existe $2^{n_{k"}} - 1$ partitions en 2 classes, il n'est pas possible de les énumérer toutes et on utilise un algorithme approché (KMEANS avec q = 2, voir paragraphe 3.5.1).
On stoppe les dichotomies successives lorsque tous les individus sont séparés ou lorsque le nombre de groupes q choisi est atteint.

- Méthode de EDWARDS et CAVALLI-SFORZA

A chaque étape, on éclate tous les groupes en deux groupes jusqu'à ce qu'un groupe ne contienne plus qu'un individu. Pour chaque éclatement, on procède comme ci-dessus.

Les méthodes métriques divisives souffrent d'être beaucoup plus coûteuses en temps de calcul que les méthodes agglomératives si on désire une solution optimale. Toutefois, l'utilisation d'algorithmes non optimaux leur permet de traiter de plus grands volumes de données tout en fournissant une hiérarchie.

D'une façon générale, les méthodes hiérarchiques ne fournissent pas un nombre de groupes "optimal". Le paragraphe "Validation de la classification" montrera comment répondre à cette question à partir de la hiérarchie et de ses niveaux.

3. LES METHODES NON HIERARCHIQUES

3.1. Classification des méthodes

Lorsqu'on doit réaliser la typologie d'un nombre important (plusieurs milliers) d'objets ou d'individus, on se limite en général, dans une première phase d'analyse, à la construction d'une partition. Une classification arborescente sur un vaste ensemble de données, même si elle s'avère réalisable, serait trop lourde à manipuler. On préfère réaliser un découpage en des sous-ensembles homogènes, se réservant ensuite la faculté de déterminer des niveaux d'agrégation sur ces sous-ensembles.

Les procédures de classification non hiérarchique font partie d'un ensemble plus vaste de méthodes de reconnaissance de formes (pattern recognition) et il est difficile de sélectionner les meilleures procédures sans préciser les objectifs poursuivis.

Quelques figures précisent ces objectifs :

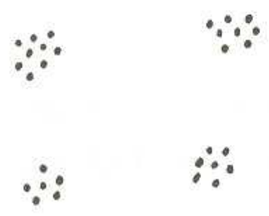

Figure 1

Figure 2

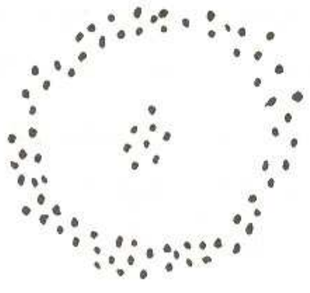

Figure 3

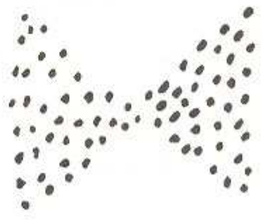

Figure 4

La reconnaissance des formes vise :

- dans la figure 1, à distinguer des groupes de points formant des sous-ensembles compacts et convexes bien distincts. Les méthodes utilisant les notions de noyau représentatif, de boule calibrée (pour les données non métriques), de centre de gravité, de variances intra- et inter-groupe (pour les données métriques) sont bien adaptées à ce type de problème relativement élémentaire;
- dans la figure 2, à relever des contours ou des tracés (examen de vues aériennes, de traces laissées par des particules dans une chambre à bulles,...). Les méthodes faisant appel à la notion de proximité (arbre minimal sous-tendu par le graphe formé des points et des arêtes valorisées par la "distance" qui les sépare) sont à recommander dans ce cas;
- dans la figure 3, à distinguer les éléments du noyau et de la couronne. Ici encore, la notion de proximité doit être retenue comme support de réflexion. Un algorithme utilisant le concept de variance conduit inévitablement à l'échec par la détermination de quartiers dans le noyau et l'anneau.
- dans la figure 4, à mettre en évidence le goulot déterminé par la rencontre de deux sous-ensembles dont certains éléments sont proches. Dans ce cas, la notion de classification valuée permet de sélectionner les points bien démarqués et les points mal identifiés alors que la détermination d'une partition conduit à placer un point du goulot dans un seul sous-ensemble.

Parmi les nombreuses méthodes proposées en matière de typologie, on peut distinguer : les méthodes heuristiques, les méthodes faisant appel à la théorie des graphes, les méthodes basées sur des structures particulières et les méthodes à critère d'optimisation.
Contrairement aux méthodes hiérarchiques, ces techniques nécessitent généralement la connaissance du nombre de classes. Comme celui-ci est presque toujours inconnu, on réalise plusieurs classifications et on choisit la meilleure.
Le problème sera discuté au paragraphe "Validation de la classification".

3.2. Les méthodes heuristiques

Un algorithme très simple est donné par une technique d'accumulation. On se fixe un nombre de classes maximum q et une valeur seuil ρ .
On considère les individus séquentiellement. Le premier constitue le centre du premier groupe. Si le second se trouve à une distance inférieure à ρ de ce centre, il appartient à ce groupe. Sinon, il devient le centre d'un deuxième groupe. On procède ainsi jusqu'au moment où tous les points ont été analysés ou lorsque le nombre de groupes cherchés est atteint. Dans ce cas, les individus restants ne sont assignés à aucune classe.

Une amélioration est obtenue en caractérisant mieux le groupe et en lui associant comme centre non plus un seul point mais son centre de gravité (à condition de disposer d'un tableau de mesures). L'individu considéré à chaque étape est celui qui est le plus éloigné de tous les autres restants à classer.

BESSON a développé la méthode IPHIGENIE fondée sur les principes de séparation et de regroupement simultanés. Elle se base sur le fait que les triangles ultramétriques sont soit isocèles, soit équilatéraux.

La méthode proposée procède comme suit :

- Considérer la matrice interdistance, regrouper les éléments les plus proches et séparer les éléments les plus éloignés.
- Eliminer les valeurs extrêmes, rechercher à nouveau les éléments les plus proches et les plus éloignés, procéder au regroupement et à la séparation.
- Itérer jusqu'à obtenir une incohérence par regroupement (i est déjà regroupé avec i', i' est déjà séparé de i"; i ne peut être regroupé avec i") ou séparation (i est déjà regroupé avec i', i' est déjà regroupé avec i"; i ne peut être séparé de i").
- Considérer les regroupements comme des arêtes et rechercher les composantes connexes.

Ces méthodes heuristiques sont rapides et fournissent une partition qui peut alors être améliorée par les procédures à critère d'optimisation.

3.3. Les méthodes faisant appel à la théorie des graphes

3.3.1. Algorithme rapide de regroupement par la recherche de l'arbre de valeur minimale

ZAHN propose de construire l'arbre de valeur minimale sous-tendu par le graphe complet construit à partir de la matrice interdistance des individus.

Chaque arête (i,i') de l'arbre optimal est testée afin d'en contrôler la "consistance". Pour cela, on supprime l'arête (i,i'); ceci conduit à déconnecter l'arbre en 2 sous-arbres. Chacun de ces sous-arbres est exploré sur une profondeur égale à c (on recherche les sommets accessibles de i, et ensuite de i', en c pas ou moins). Les moyennes des arêtes correspondant aux sommets retenus sont calculées (m_i et $m_{i'}$) ainsi que les écarts-types (σ_i et $\sigma_{i'}$).

Les critères d'inconsistance pour d(i,i') sont :

$$d(i,i') \geq \text{MAX}\{m_i + \sigma_T \sigma_i \ , \ m_{i'} + \sigma_T \sigma_{i'}\}$$

$$\text{MAX}\{\frac{d(i,i')}{m_i} \ , \ \frac{d(i,i')}{m_{i'}}\} \geq f_T$$

Les arêtes inconsistantes sont éliminées en faisant appel à un des critères ou les deux considérés simultanément. La forêt obtenue détermine une partition qui dépend du triplet (c,σ_T,f_T).

Il est intéressant, mais évident, de noter que la réunion des sommets en considérant les valeurs croissantes des arêtes de l'arbre minimal conduit à la reconstitution du dendrogramme que l'on obtient par la méthode hiérarchique "single-linkage".

3.3.2. Algorithme bi-critère de séparation par coloration

HANSEN et DELATTRE fournissent une partition qui maximise l'écart entre les classes et minimise le diamètre intra-classe. L'écart entre les classes est le plus petit saut minimum entre deux classes. Le diamètre intra-classe est le plus grand des diamètres des classes. L'écart mesure la séparation entre les groupes; le diamètre apprécie leur homogénéité.

Pour obtenir la partition qui maximise l'écart entre les classes, ces auteurs montrent qu'il suffit de supprimer les q-1 arêtes les plus grandes de l'arbre de longueur minimale (q étant le nombre de classes fixé a priori).

Pour trouver la partition dont la classe de plus grand diamètre a un diamètre minimum, ils proposent l'algorithme suivant basé sur la coloration des graphes. On considère les couples d'individus par ordre décroissant de leurs indices de distance. Au départ, tous les points à classifier (considérés comme les sommets d'un graphe) reçoivent la même couleur. Les couples sont considérés dans l'ordre retenu plus haut (s'il s'agit d'un préordre, on choisit indifféremment un des couples de même rang) et une contrainte de coloration est induite : deux sommets liés ne peuvent recevoir la même couleur.

Lorsqu'une liaison fait apparaître deux sommets présentant la même coloration, une des couleurs est modifiée tant que cela est possible. Lorsqu'une telle modification s'avère impossible (on devrait utiliser un nombre (q+1) de couleurs, ...), un algorithme de coloration permet de contrôler si le graphe construit à cette étape peut être colorié en un nombre de couleurs inférieur ou égal à q. Si oui, on reprend la procédure. Si non, la dernière liaison introduite est isolée. Elle représente le diamètre maximum de la partition en q couleurs ou moins (les sommets de même couleur appartiennent à la même classe).

Finalement, on détermine l'ensemble des solutions efficaces de ce problème de classification bi-critère. Une partition est dite efficace s'il n'est pas possible d'améliorer la valeur d'un critère sans détériorer la valeur de l'autre.

HANSEN et DELATTRE proposent également un indice permettant de juger du nombre "optimum" de classes

$$I_k = 100 \left(\frac{d_k - d_{k-1}}{d_{k-1}}\right) \text{ est calculé pour } k = 2,3,...$$

d_k est le diamètre de la partition en k classes (diamètre maximum des k classes). Le nombre de classes cherché correspond à un minimum local de I_k.

3.4. Méthodes basées sur des structures particulières

Ce paragraphe n'est pas détaillé par manque de généralité. On cite toutefois l'algorithme de GITMAN et LEVINE recherchant des groupes à concentration unimodale, celui de DUNN imposant des contraintes de continuité entre les groupes et les travaux de GORDON sur les problèmes de classification sous contraintes.

3.5. Méthodes à critère d'optimisation

Ces méthodes peuvent être scindées en trois classes : celles fournissant une partition, celles permettant aux individus d'appartenir à plusieurs classes (classes empiétantes) et celles affectant chaque individu partiellement à chaque groupe.

Les critères s'introduisent comme suit. Soient n points i formant l'espace I. Chaque point i est affecté d'un poids p_i défini a priori et tel que $\sum_{i\in I} p_i = 1$.

On se propose de déterminer une partition de I en q groupes, q étant supposé fixé. Pour réaliser cette partition, il suffit de déterminer pour tout $i\in I$, une fonction d'appartenance au groupe k, $\mu_k(i)$:

$$0 \leqslant \mu_k(i) \leqslant 1 \ , \ i\in I, \ k=1,\ldots,q$$

Si $\mu_k(i)\in \{0,1\}$, on obtient une partition (vulgaire) et si $\mu_k(i)\in [0,1]$, on a une partition valuée.

On obtient les fonctions d'appartenance en minimisant l'un des deux critères suivants (les fonctions f et g sont connues)

$$CR1 : \underset{(\mu,y_k)}{MIN} \sum_k \sum_i g\,[p_i,\mu_k(i)]\, d(i,y_k) \ !$$

sous certaines contraintes relatives à μ.

$$CR2 : \underset{(\mu,\eta)}{MIN} \sum_{k} \sum_{i} \sum_{i'} f[p_i,\eta_k(i)]\, g\,[p_{i'},\mu_k(i')]\, d(i,i') \ !$$

sous certaines contraintes relatives à μ et η

$\eta_k(i)$ représente une fonction d'appartenance auxiliaire dont le rôle est précisé ultérieurement.

$$0 \leq \eta_k(i) \leq 1, \ \forall i \in I, \ k=1,\ldots,q$$

Contrairement à CR2, CR1 doit disposer de la distance entre tout point de I et des points y_k n'appartenant pas nécessairement à I. C'est possible si on peut associer à I un espace euclidien R^p muni d'une norme de forme quadratique. On obtient alors, par dérivation de

$$CR1 = \sum_{k} \sum_{i} g[p_i,\mu_k(i)]\, \|X_i - y_k\|^2 \ ,$$

$y_k = \bar{X}_k$: le barycentre de la classe k.

$$\bar{X}_k = \frac{1}{p_k} \sum_{i \in I} g[p_i,\mu_k(i)]\, X_i$$

$$p_k = \sum_{i \in I} g[p_i,\mu_k(i)]$$

Le théorème de KOENIG-HUYGENS permet d'écrire

$$CR1 : \underset{\mu}{MIN} \sum_{k} M_k^2(g)$$

sous certaines contraintes relatives à μ

et $\quad M_k^2(g) = \sum_{i \in I} g[p_i,\mu_k(i)]\, \|X_i - \bar{X}_k\|^2$: moment centré

d'ordre 2 de k = variabilité du groupe k pour la fonction g.

On constate encore que

$$M_I^2(g) = \sum_k M_k^2(g) + M_{1,\ldots,q}^2(g)$$

avec

$$M_I^2(g) = \sum_{i\in I} \sum_k g[p_i,\mu_k(i)]\|X_i - \bar{X}_I\|^2 \quad : \text{variabilité totale pour la}$$

fonction g.

$$M_{1,\ldots,q}^2(g) = \sum_k p_k \|\bar{X}_k - \bar{X}_I\|^2 \quad : \text{variabilité entre les groupes}$$

et $\bar{X}_I = \frac{1}{\sum p_k} \sum_k p_k \bar{X}_k$: barycentre de l'ensemble I

CR2 n'impose pas la connaissance d'une distance entre des points qui n'appartiennent pas tous à I. Toutefois, si on a un espace euclidien muni d'une norme quadratique, le théorème de KOENIG-HUYGENS permet d'écrire :

$$CR2 : \min_{(\mu,\eta)} \sum_k \{p_k M_k^2(f) + p'_k M_k^2(g) + p_k p_{k'} \|\bar{X}_k - \bar{x}'_k\|^2\} !$$

sous certaines contraintes relatives à μ et η

$$p'_k = \sum_{i\in I} f[p_i,\eta_k(i)] \quad , \quad X'_k = \frac{1}{p'_k} \sum_{i\in I} f[p_i,\eta_k(i)]\, X_i$$

Si $\eta=\mu$ et $f=g$,

$$CR2 : \min_{\mu} \; 2 \sum_k p_k M_k^2(g) !$$

et si tous les p_k sont égaux, CR2 est confondu avec CR1.

3.5.1. Méthodes fournissant une partition

- Soit CR1 avec $p_i = \frac{1}{n}$ et $g = \mu_k(i)/n$. Il vient

$$\text{MIN} \sum_k \sum_i \mu_k(i) \| X_i - \bar{X}_k \|^2$$

sous les contraintes

$$0 \leqslant \mu_k(i) \leqslant 1 \qquad \forall i \in I,\ k=1,\ldots,q$$

$$\sum_k \mu_k(i) = 1 \qquad \forall i \in I$$

Chaque point i appartient partiellement à chaque groupe ($0 \leqslant \mu_k(i) \leqslant 1$) et totalement à l'ensemble $I(\sum_k \mu_k(i) = 1)$.

La méthode qui en découle est connue sous le nom de "méthode de classification des moindres carrés" (least-square clustering ou sum of squares distance criterion).

Plusieurs procédures ont été proposées. Elles conduisent à une partition vulgaire ($\mu_k(i) = 0$ ou 1).

L'algorithme HMEANS se donne une partition initiale qui permet de calculer les centres de gravité. La minimisation du critère conduit alors à affecter les individus au groupe dont le centre de gravité est le plus proche. On a ainsi une nouvelle partition et on itère jusqu'à ce qu'un minimum soit atteint. Cette procédure converge vers un optimum local qui dépend de la partition initiale. Si on connaît quelque peu les données, on peut se donner comme point de départ des centres de gravité et procéder de la même façon.

L'algorithme KMEANS se déroule presque de la même façon que HMEANS. La différence réside dans le fait que les nouveaux centres de gravité sont recalculés chaque fois qu'un individu change de groupe.
Les calculs sont menés rapidement en utilisant l'algorithme d'échange dont le principe est dû à REGNIER.

Comme $p_i = \frac{1}{n}$ et $\mu_k(i) = 0$ ou 1 pour tout i, le critère s'écrit

$$CR1 : \text{MIN} \sum_k \sum_{i \epsilon k} \| X_i - \bar{X}_k \|^2 \equiv \text{MIN} \sum_k M_k^2$$

Soit l'individu i sortant du groupe s. Le centre de gravité de ce groupe devient

$$\bar{X}'_s = \frac{p_s \bar{X}_s - X_i}{p_s - 1} \quad \text{et quelques calculs conduisent à}$$

$$M_s^{2'} = M_s^2 - \frac{p_s}{p_{s-1}} \| X_i - \bar{X}_s \|^2$$

Soit l'individu i entrant dans le groupe e. On obtient

$$\bar{X}'_e = \frac{p_e \bar{X}_e + X_i}{p_e + 1}$$

$$M_e^{2'} = M_e^2 + \frac{p_e}{p_e + 1} \| X_i - \bar{X}_e \|^2$$

Par cette opération, le critère est modifié de

$$M_e^{2'} - M_e^2 + M_s^{2'} - M_s^2$$

Si cette quantité est négative, le changement de i est profitable. Le test, basé uniquement sur des valeurs déjà calculées, s'écrit :

$$\text{i change de s à e si } \frac{p_e}{p_e + 1} \|x_i - \bar{x}_e\|^2 < \frac{p_s}{p_s - 1} \|x_i - \bar{x}_s\|^2$$

Si plusieurs e sont possibles, on choisit celui qui conduit à une valeur minimale du membre de gauche de cette expression.
KMEANS peut fournir une bonne solution très rapidement mais elle dépend de l'ordre dans lequel les individus ont été traités.

Pour essayer de déterminer le nombre de groupe en cours d'algorithme, BALL and HALL et MAC QUEEN proposent d'éclater un groupe si son moment centré d'ordre 2 est supérieur à une valeur seuil et de regrouper deux groupes si leur variabilité inter-groupe est inférieure à une autre valeur seuil. Divers paramètres peuvent ainsi être introduits. Dans leur algorithme ISODATA, BALL and HALL utilisent 7 paramètres.
Plutôt que cette approche difficile dans le choix des paramètres, certains auteurs préfèrent utiliser KMEANS pour diverses valeurs de q et diverses partitions initiales. Le nombre de groupe "optimal" est alors déduit en comparant les résultats.

- Dans le cas où on ne dispose pas de données mesurables, on considère CR2 avec

$$f = \eta_k(i) = \frac{1}{p_k} \quad , \quad g = \mu_k(i)/n \quad \text{et} \quad p_i = \frac{1}{n}$$

$$\text{CR2 : } \underset{\mu,\eta}{\text{MIN}} \sum_k \sum_i \sum_{i'} \eta_k(i)\mu_k(i')\, d(i,i')$$

avec

$$0 \leqslant \mu_k(i') \leqslant 1 \quad ; \quad \sum_k \mu_k(i') = 1 \quad ; \quad \forall i'$$

$$\eta_k(i) = \frac{1}{p_k} \quad \forall i \in k$$

$$= 0 \quad \text{sinon}$$

On obtient une partition vulgaire. Un optimum local est atteint en appliquant le principe de KMEANS.

Le critère s'écrit aussi :

$$\text{MIN} \sum_k \frac{1}{p_k} \sum_{i \in k} \sum_{i' \in k} d(i,i')$$

$$(\equiv \text{MIN} \sum_k M_k^2 \quad \text{dans le cas de valeurs mesurables})$$

On note par similitude,

$$\text{MIN} \sum_k P_k^2$$

L'algorithme d'échange fournit

$$P_s^{2'} = \frac{p_s P_s^2 - \sum_{i' \in s} d(i,i')}{p_s - 1} \quad \text{pour i sortant de s}$$

$$P_e^{2'} = \frac{p_e P_e^2 - \sum_{i' \in e} d(i,i')}{p_e + 1} \quad \text{pour i entrant dans e}$$

L'échange se fait si $P_s^{2'} - P_s^2 + P_e^{2'} - P_e^2$ est négatif.

Cet algorithme présente l'avantage par rapport à KMEANS de ne pas calculer continuellement des distances aux centres de gravité. Par contre, la matrice d'interdistance doit être conservée en mémoire centrale.

- DIDAY s'inspirant des méthodes précédentes a proposé plusieurs méthodes reposant sur la minimisation du critère CR2. On les appelle méthodes des nuées

dynamiques. L'une d'elles, connue sous le nom de méthode à pondération constante, est telle que

$$f = \eta_k(i)\ /n \quad , \quad g = \mu_k(i')$$

Il s'agit, dès lors, de

$$\text{MIN} \sum_k \sum_i \sum_{i'} \eta_k(i)\mu_k(i')d(i,i') \ !$$

avec $\quad 0 \leqslant \eta_k(i) \leqslant 1 \qquad \sum_i \eta_k(i) = n_k$ (fixé a priori)

$$0 \leqslant \mu_k(i') \leqslant 1 \qquad \sum_k \mu_k(i') = 1$$

La méthode itérative des nuées dynamiques procède en deux phases :

(i) phase de représentation : partant des valeurs μ_k obtenues dans une itération précédente, on détermine les éléments les plus représentatifs des classes obtenues. Ces éléments, en nombre égal à n_k, jouent le rôle de noyaux de la partition. On constate immédiatement que la minimisation du critère, pour μ fixé, conduit à $\eta_k(i) = 1$ ou 0.

(ii) phase d'attraction : elle consiste à regrouper les éléments de I en q classes. En effet, si η est fixé, l'optimisation conduit à choisir $\mu_k(i) = 1$ ou 0. Si $\mu_k(i) = 1$, i est associé à la classe k.

La phase initiale consiste à un choix arbitraire ou aléatoire des noyaux de départ.

DIDAY montre que la procédure converge vers un *optimum local* qui dépend du choix initial des éléments des noyaux. Pour tenter d'obtenir un optimum global, il propose une procédure heuristique qui détecte les "formes fortes", c'est-à-dire les groupements de points qui sont stables dans la partition optimale pour différents noyaux de départ.

Si $n_k = 1$, $\forall k$, les points tels que $\eta_k(i) = 1$ forment les "centres de gravité" des classes (ils appartiennent cependant à I !).

Si $f=g=p_i\ \mu_k(i)$, $\sum_k \mu_k(i)=1$, on retrouve la méthode itérative des "noyaux-partition".

- D'une façon générale, on peut émettre deux critiques à toutes les méthodes de ce paragraphe :

- on obtient une partition même si certains points sont attirés fortement par plusieurs groupes. Pour remédier à cet inconvénient, on envisage des classes empiétantes ou une classification valuée.

- par minimisation des distances à un centre de gravité réel ou fictif, on tend à créer des groupes sphériques même si les données ne se disposent pas ainsi. Des agrégats non sphériques seront obtenus par une méthode de classification valuée (MNDr).

3.5.2. Classes empiétantes

Une partition, voire une hiérarchie, est parfois insatisfaisante à cause de sa définition stricte car il arrive qu'un individu i puisse appartenir à plusieurs classes à la fois. Il existe différents types de détermination de classes empiétantes (overlapping classification) dont la plus connue est celle des B_k classes de JARDINE et SIBSON. Toutefois cette méthode est extrêmement coûteuse en temps de calcul et ses résultats sont difficilement interprétables pour de vastes ensembles de données.

Une procédure simple est proposée par JAMBU. On dispose d'une partition. On définit une distance $\delta(i,k)$ entre tout individu i et une classe k.
On considère le point $i\in k(i)$ et on calcule tous les $\delta(i,k)$, $k \neq k(i)$.
i appartient à toutes les classes k telles que $\delta(i,k) < \delta(i,k(i))$.

3.5.3. Classification valuée

Pour contourner l'aspect strict d'une partition, on peut rechercher une solution telle que tout individu appartienne plus ou moins fort à toutes les classes. On veut donc des fonctions d'appartenance $\mu_k(i)$, $k=1,\ldots,q$ prenant leurs valeurs dans tout l'intervalle $[0,1]$ (Si ensuite, on désire retrouver une partition, on peut affecter totalement tout individu à la classe pour laquelle il a le plus grand coefficient d'appartenance).

- Une des méthodes floues proposées par RUSPINI se ramène à la minimisation d'une fonction du type CR2 avec

$$\mu_k(i) = \eta_k(i) \quad \text{et}$$

$$f.g = p_i \mu_k(i) p_{i'} \mu_k(i') \frac{\sum_{i''} p_{i''} \mu_k(i'')}{\sum_{i''} p_{i''}\, d(i,i'')}$$

p_i s'interprète comme une probabilité a priori associée à i et $\mu_k(i)$ comme une probabilité conditionnelle de la classe k étant donné i.
Dès lors

$\sum_{i''} p_{i''} \mu_k(i'')$ représente la probabilité associée à la classe k

$\sum_{i''} p_{i''}\, d(i,i'')$ représente une distance moyenne, et

$\sum_{i'} P[i'|k] d(i,i')$ représente une distance moyenne conditionnelle avec

$$P[i'|k] = \frac{p_{i'} \mu_k(i')}{\sum_{i''} p_{i''} \mu_k(i'')}$$

Le programme $\text{MIN} \sum_k \sum_i \sum_{i'} f.g.d(i,i')$!

avec $0 \leqslant \mu_k(i) \leqslant 1$

$$\sum_k \mu_k(i) = 1$$

est résolu par une méthode de gradient.

Cette méthode est lourde en temps de calcul.

- Pour obtenir une solution valuée, on peut aussi reprendre les critères du § 3.5.1. et faire en sorte que $\mu_k(i)$ varie entre 0 et 1. DUNN propose pour cela une procédure qui s'inspire de la méthode de classification des moindres carrés et choisit $g = \mu_k^2(i) / n$. Une généralisation immédiate est obtenue en prenant $g = \mu_k^r(i) / n$.

Le critère de DUNN s'écrit

$$\text{MIN} \quad \sum_k \sum_i \mu_k^2(i) \, \| X_i - \bar{X}_k \|^2$$

avec $$\bar{X}_k = \frac{\sum_{i \in I} \mu_k^2(i) X_i}{\sum_{i \in I} \mu_k^2(i)}$$

Si les $\bar{X}_k$ sont connus, le programme quadratique fournit

$$\mu_k(i) = \| X_i - \bar{X}_k \|^2 \Big/ \sum_{m=1}^{q} \frac{1}{\| X_i - \bar{X}_m \|^2}$$

Cette solution est obtenue aisément par la méthode des multiplicateurs de Lagrange. On introduit un multiplicateur relatif à la contrainte $\Sigma \mu_k(i) = 1$ et on optimise par dérivation une fonction sans contrainte. On constate que la contrainte $0 < \mu_k(i) < 1$ est satisfaite.
L'algorithme procède de la façon suivante. On se fixe des fonctions d'appartenance initiales, on cherche les centres de gravité, on obtient de nouvelles fonctions d'appartenance, etc. La procédure converge vers un optimum local dépendant des conditions initiales. Elle présente le désavantage de créer des agrégats sphériques à cause de l'attraction par les centres de gravité.

- Pour remédier à cela et traiter des données non nécessairement mesurables, ROUBENS a proposé un algorithme partiellement inspiré des nuées dynamiques appelé MNDr (Méthode des Nuées Dynamiques avec exposant r).

Le critère s'écrit

$$\text{MIN} \sum_k \sum_i \sum_{i'} \mu_k^r(i)\mu_k^r(i')d(i,i') \ !$$

avec $0 \leqslant \mu_k(i) \leqslant 1$

$$\sum_k \mu_k(i) = 1$$

On note

$$D(i,k) = \sum_{i'} \mu_k^r(i')d(i,i')$$

Cette quantité mesure la distance entre i et le groupe k formé par la contribution de tous les individus. Le programme devient

$$\text{MIN} \sum_k \sum_i \mu_k^r(i)D(i,k) \text{ dont la solution est}$$

$$\mu_k(i) = \left[D(i,k) \sum_{m=1}^{q} \frac{1}{D(i,m)} \right]^{\frac{1}{r-1}}$$

L'algorithme se déroule comme suit. On choisit des fonctions d'appartenance initiales. On calcule pour le premier individu D(1,k) $\forall k$. On obtient ainsi de nouveaux $\mu_k(1)$. On passe au deuxième individu, etc. Lorsque tous les individus ont été analysés, on reprend au premier si le critère n'a pas encore atteint un minimum. La procédure converge vers un optimum local. Toutefois, la pratique montre que les conditions initiales influencent peu cette solution sauf si l'on choisit $\mu_k(i) = \frac{1}{q}$ $\forall k$ et $\forall i$. Dans ce cas, on se trouve en un point stationnaire et l'algorithme stoppe.

4. VALIDATION DE LA CLASSIFICATION

Jusqu'à présent, on a exposé des méthodes. On a présenté des hiérarchies, des partitions valuées ou non mais on n'a fourni aucun élément de réponse aux questions suivantes :

Quelle importance doit-on accorder au niveau auquel des classes se sont agrégées ? Combien de classes existent dans les données ? Quelles sont les variables responsables des séparations entre classes ? Les résultats obtenus sont-ils valables ? Comment sont-ils influencés par le codage ?

On envisage successivement dans ce paragraphe des aides à l'interprétation des résultats, la recherche du nombre "optimal" de classes et la comparaison des méthodes.

4.1. Aides à l'interprétation

On présente diverses mesures permettant de mieux décrire une classification. Certaines ne sont utilisables que si on dispose de valeurs mesurables.

4.1.1. Description de l'arborescence

- Si on dispose d'une hiérarchie, on peut tracer un histogramme des indices des niveaux d'agrégation. Si la décroissance est très forte, ceci montre qu'il n'existe que quelques séparations principales qui permettent alors de choisir le nombre de groupes. Les niveaux les plus bas de la hiérarchie peuvent être considérés comme des intermédiaires de calcul.
On peut aussi calculer la séparabilité d'un groupe qui est définie par la différence entre le niveau auquel le groupe est formé et le niveau auquel le groupe est agrégé dans un groupe plus important. Comme une hiérarchie peut

déformer sensiblement la matrice interdistance, cette mesure doit être interprétée avec prudence. Un dendrogramme "complete linkage" a vraisemblablement des valeurs de séparabilité plus grandes que celles d'un "single linkage" compte tenu d'un critère plus exigeant sur la séparation des groupes. Dès lors, une importante valeur de séparabilité dans un "single linkage" est un bon indice pour détecter un groupe isolé. La compacité d'un groupe à un certain niveau est une mesure normalisée du nombre de noeuds présents dans la représentation en graphe du groupe. Si le groupe contient n_k individus, le nombre de noeuds doit être compris entre n_k-1 et $\frac{1}{2}\, n_k(n_k-1)$. Si v noeuds sont présents, la compacité est donnée par

$$C_k = \frac{v - (n_k-1)}{\frac{1}{2}\, n_k(n_k-1) - (n_k-1)}$$

Par définition, un groupe complètement lié a une compacité égale à 1 alors que la compacité d'un groupe obtenu par "single linkage" peut atteindre la valeur minimale 0.

- A chaque valeur de la hiérarchie, on peut si possible associer le moment centré d'ordre 2 de la partition constituée des deux groupes qui vont être agrégés ($M^2_{k,k'}$). En ramenant cette inertie à l'inertie totale (M^2_I), on a la part d'inertie expliquée par la séparation à ce niveau. Cette mesure est à rapprocher du rapport des valeurs propres utilisé dans les méthodes factorielles. On peut donc aussi tracer un histogramme des taux d'inertie.

4.1.2. Histogrammes des profils par classe

Si on dispose d'un tableau variables x individus, il est toujours possible de comparer les classes en calculant leurs profils. De cette façon, on peut mettre en évidence les variables responsables des séparations et

apprécier les différences entre classes.

4.1.3. Contributions des variables à la distance des classes au centre du nuage

$$d^2(\bar{X}_k,\bar{X}_I) = \|\bar{X}_k - \bar{X}_I\|^2 = \sum_j(\bar{x}_{kj} - \bar{x}_{.j})^2 / \sigma_j^2 \qquad \text{(par exemple)}$$

représente le carré de la distance de la classe k au centre du nuage.

$\left(\frac{\bar{x}_{kj} - \bar{x}_{.j}}{\sigma_j}\right)^2$ représente la contribution de la variable j à la distance de la classe k au centre du nuage.

En divisant par $d^2(\bar{X}_k,\bar{X}_I)$, on a la contribution relative. On les affecte du signe de $\bar{x}_{kj} - \bar{x}_{.j}$ afin de positionner relativement les centres de gravité. Ces mesures permettent de connaître plus précisément l'apport d'une variable dans l'écartement d'un groupe par rapport au centre du nuage et donc de compléter la description en profils de ce groupe.

4.1.4. Contributions des variables à la distance entre deux groupes qui vont être agrégés

Le moment centré d'ordre 2 de la partition constituée des groupes k et k' qui vont être agrégés est égal à

$$M^2_{k,k'} = \frac{p_k p_{k'}}{p_k + p_{k'}} \|\bar{X}_k - \bar{X}_{k'}\|^2 = \frac{p_k p_{k'}}{p_k + p_{k'}} \sum_j(\bar{x}_{kj} - \bar{x}_{k'j})^2/\sigma_j^2 \qquad \text{(par exemple)}$$

$\frac{p_k p_{k'}}{p_k + p_{k'}} \frac{(\bar{x}_{kj} - \bar{x}_{k'j})^2}{\sigma_j^2}$ mesure donc la contribution de la variable j

dans l'écartement des groupes k et k'. En divisant par $M^2_{k,k'}$, on obtient la contribution relative.
Ces mesures permettent donc d'expliquer les différences entre groupes.

4.2. Nombre "optimal" de groupes

4.2.1. Classification non valuée

Le paragraphe précédent a fourni les éléments de réponse pour le choix du nombre de classes à partir d'une hiérarchie. En effet, en tenant compte des niveaux de la hiérarchie et des inerties inter-classes, il est possible de couper horizontalement le dendrogramme et de là statuer sur le nombre de groupes.

En présence de partitions en q groupes, q=2,3,..., on s'intéresse aux valeurs du critère. Vu la définition de celui-ci (égal ou similaire à la somme des inerties des classes), chaque fois qu'on introduit un groupe supplémentaire, le critère diminue. La valeur minimale du critère ne permet donc pas de choisir le nombre de groupes. Par contre, en construisant un histogramme des valeurs du critère, on peut retenir comme nombre de classes le plus probable celui qui conduit à une décroissance la plus rapide du critère. Si la décroissance apparaît continue, il n'existe vraisemblablement pas de groupes.
Quoiqu'il en soit, dans la mesure du possible, on s'éclairera des résultats des méthodes factorielles et on essaiera de faire la liaison entre ces deux types d'analyse.

4.2.2. Classification valuée

La qualité d'une classification valuée est liée aux valeurs des coefficients d'appartenance. Plus ceux-ci sont proches de 1 et de 0 et plus la séparation entre les groupes est nette.

$\sum_k \sum_i \mu_k(i) = n$ n'est pas un indice de validité

On considère alors $F = \sum_k \sum_i \mu_k^2(i)/n$

Si $\mu_k(i) = \frac{1}{q}$ $\forall k, \forall i$ alors $F = \frac{1}{q}$

Si $\mu_k(i) = 1$ ou 0 alors $F = 1$

Cette indice tend à croître lorsque q augmente. On choisit donc le nombre de groupes pour lequel la croissance de l'indice est la plus rapide ou on considère un indice normalisé

$$NFI = \frac{qF - 1}{q - 1}$$

Si $\mu_k(i) = \frac{1}{q}$ alors $NFI = 0$

Si $\mu_k(i) = 1$ ou 0 alors $NFI = 1$

Toutefois, ces indices qui sont fonctions du carré des coefficients d'appartenance reflètent peu les modifications de ceux-ci. D'autres indices basés sur l'entropie sont fonctions de $\mu_k \cdot \log \mu_k$ et souffrent du même inconvénient.

Pour cette raison, LIBERT a introduit l'indice $NF^{\ast}I$ défini par

$$NF^{\ast}I = \frac{qF^{\ast} - 1}{q - 1}$$

avec

$$F^{\ast} = \frac{\sum_i \frac{\mu_\cdot(i)}{n} + \underset{i}{MIN}\, \mu_\cdot(i)}{2} \quad \text{et} \quad \mu_\cdot(i) = \underset{k}{MAX}\, \mu_k(i)$$

Si $\mu_k(i) = \frac{1}{q}$ alors $F^{\ast} = \frac{1}{q}$ et $NF^{\ast}I = 0$

Si $\mu_k(i) = 1$ ou 0 alors $F^{\ast} = 1$ et $NF^{\ast}I = 1$

Cet indice ne tient compte que des coefficients maximum et n'est donc pas affecté par les autres coefficients dont la signification est souvent faible. Il est ainsi fonction linéaire des coefficients et son maximum permet de choisir le nombre de groupes.
Il est toutefois important de tenir compte du fait que plus le nombre de groupes augmente et plus le nombre de coefficients élevés à tendance à croître. Ceci est particulièrement vrai pour les procédures "à centres de gravité".
L'indice $NF^{\ast}I$ accorde un poids particulier au point le plus mal classé ($\underset{i}{MIN}\ \mu_i(i)$). On considère donc qu'une classification n'est pas satisfaisante dès qu'un individu au moins est mal défini. Dans ce cas, on recherche une nouvelle classification en éliminant ce point. En agissant de la sorte, on recherche des groupes très homogènes et on attire l'attention sur les individus mal classés.

4.3. Comparaison des méthodes

Lors de la description des méthodes, on a déjà signalé les caractéristiques essentielles de celles-ci et il apparaît peu sensé de rechercher la meilleure méthode lorsqu'on sait qu'elles ont des objectifs différents.

Leur utilisation simultanée sur un même jeu de données permet souvent de valider les résultats. On peut ainsi apprécier la stabilité des groupes en fonction des méthodes c'est-à-dire en fonction du critère d'agrégation.
De la même manière, on teste la stabilité en fonction du choix des distances.

Dans le cas d'une hiérarchie, on peut considérer l'opération associée comme une application qui transforme les distances initiales $d(i,i')$ en nouvelles distances $\hat{d}(i,i')$. Divers auteurs (SOKAL and ROHLF, KRUSKAL, JARDINE and SIBSON, GOWER) ont proposé des mesures de la distorsion créée par cette

transformation. De cette façon, on a une indication sur la validité de l'ajustement des données sur la structure imposée par la classification. Ceci peut être réalisé pour diverses méthodes et divers critères d'agrégation. Plus récemment, GORDON a proposé des algorithmes de transformation minimale de deux dendrogrammes pour qu'ils soient les plus semblables et DIDAY a construit des procédures de comparaison de dendrogrammes.

Enfin, depuis quelques années, se sont développées des recherches visant à formaliser plus précisément le problème de la classification. On essaie malgré l'absence de modèle explicite de formuler des tests d'hypothèses quant à la présence ou non de groupes et quant au nombre de groupes. De plus, on montre que sous certaines hypothèses, des méthodes de classification s'intègrent dans le cadre plus strict de la statistique inférentielle. Il en est ainsi pour la classification des moindres carrés qui est équivalente à une estimation selon le maximum de vraisemblance d'un modèle bien particularisé.

5. INDIVIDUS SUPPLEMENTAIRES

Après avoir construit une classification, on dispose parfois d'individus supplémentaires que l'on souhaite affecter à l'un ou l'autre groupe sans pour cela reprendre tous les calculs.

Dans le cas d'une hiérarchie, JAMBU propose l'algorithme suivant. On affecte un individu en parcourant le dendrogramme à partir du haut. On décide d'abord auquel des deux groupes successeurs, il appartient. On considère cette branche et on choisit un des deux groupes suivants et ainsi de suite jusqu'à rencontrer une classe terminale. Il reste à déterminer des règles d'affectation dépendant de la structure initiale de l'espace (euclidien ou non), des normes dont est muni cet espace, du type de données et de l'algorithme de classification qui a été utilisé pour construire la hiérarchie.

Dans le cas d'un tableau de mesures, il est de plus possible de mettre à jour les caractéristiques des groupes existants. Ceci permet alors de traiter progressivement de très grands tableaux de données.

Dans le cas d'une partition sur un tableau de mesures, on peut utiliser les résultats de l'analyse factorielle discriminante.

BIBLIOGRAPHIE

ANDERBERG M.R. (1973)
Cluster analysis for applications. Academic Press, New-York.

BALL G.H., HALL D.J. (1967)
A clustering technique for summarizing multivariate data.
Behavioral sciences, 12, pp.153-155.

BENZECRI J.P. (1973)
L'analyse des données. Tome 1 : La Taxinomie. Tome 2 : L'analyse des correspondances. Dunod. Paris

BESSON M. (1973)
Iphi ou un nouveau procédé de typologie. Séminaires de l'IRIA;
Classification automatique et perception par ordinateur.

BEZDEK J.C. (1981)
Pattern recognition with fuzzy objective function algorithms.
Plenum Press. New-York.

CORMACK R.M. (1971)
A review of classification. J.R.Stat.Soc., A134, part.3

DELATTRE M., HANSEN P. (1980)
Bicriterion cluster analysis. IEEE Trans. on Pattern analysis and Machine intelligence. Vol.2, 4, pp.277-291.

DIDAY E. (1971)
La méthode des nuées dynamiques. Rev.Stat.App. Vol.19,2, pp.19-34.

DIDAY E. et coll. (1980)
Optimisation en classification automatique (2 vol.)
INRIA, Le Chesnay

DUNN J.C. (1974)
A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. J.Cybernetics. 3,3,pp.32-57.

DUNN J.C. (1976)
Indices of partition fuzziness and the detection of clusters in large data sets. In Fuzzy automata and decision processes.
Ed. M.GUPTA. American Elsevier. New-York.

EDWARDS A.W.F., CAVALLI-SFORZA L.L. (1965)
A method for cluster analysis. Biometrics 21, pp.362-375.

GITMAN I., LEVINE M.D. (1970)
An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique.
IEEE Trans. on Computers, C-19, pp.583-593.

GORDON A.D. (1981)
Classification. Chapman and Hall. London.

HARTIGAN J.A. (1975)
Clustering algorithms. Wiley and Sons. New-York.

HUBERT L. (1972)
Some extensions of Johnson's hierarchical clustering algorithms.
Psychometrika. Vol.37,3.

JAMBU M. (1978)
Classification automatique pour l'analyse des données.
Tome 1 : Méthodes et algorithmes. Tome 2 : Logiciels (avec LEBEAUX M.O).
Dunod. Paris.

JARDINE N., SIBSON R. (1968)
The construction of hierarchic and non-hierarchic classification.
Computer J., 11, pp.177-184.

JARDINE N., SIBSON R. (1971)
Mathematical taxonomy. Wiley. New-York.

KRUSKAL J.B. (1964)
Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, pp.1-27.

LERMAN I.C. (1970)
Les bases de la classification automatique.
Gauthier-Villars. Paris

LERMAN I.C. (1981)
Classification et analyse ordinale des données.
Dunod. Paris

LIBERT G., ROUBENS M. (1982)
Non metric fuzzy clustering algorithms and their cluster validity.
In Approximate reasoning in decision analysis.
Eds. GUPTA M., SANCHEZ E. North-Holland. pp.417-425.

LIBERT G., ROUBENS M. (1983)
New experimental results in cluster validity of fuzzy clustering algorithms. In New trends in data analysis and applications.
Eds. JANSSEN J., MARCOTORCHINO J.F., PROTH J.M. North-Holland.
pp.205-218.

LIBERT G. (In press)
Non metric cluster analysis. In Encyclopedia of systems and control. Ed. SINGH M. Pergamon Press. Oxford.

MAC QUEEN J. (1967)
Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. 1965, pp.281-297.

REGNIER S. (1965)
Sur quelques aspects mathématiques des problèmes de classification automatique. ICC bulletin 4, pp.175-191.

ROUBENS M. (1978)
Pattern classification problems and fuzzy sets. Fuzzy sets and systems, 1, pp.239-253.

RUSPINI E.H. (1969)
A new approach to clustering. Inform. Control 15, pp.22-32.

SOKAL R.R., ROHLF F.J. (1962)
The comparison of dendrograms by objective methods. Taxon. 11, pp.33-40.

SOKAL R.R., SNEATH P.H. (1973)
Numerical Taxonomy. Freeman. San Francisco.

SPATH H. (1980)
Cluster analysis algorithms. Ellis Horwood Ltd. Chichester.

ZAHN C.T. (1971)
Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Trans.on Computers, C-20, pp.68-86.