Statistique et analyse des données

Michel Bruynooghe

Méthodes nouvelles en classification automatique de données taxinomiques nombreuses

Statistique et analyse des données, tome 2, n° 3 (1977), p. 24-42.

<http://www.numdam.org/item?id=SAD_1977__2_3_24_0>

Statistique et Analyse des Données
3 - 1977

méthodes nouvelles en classification automatique de données taxinomiques nombreuses

B R U Y N O O G H E Michel

Université Aix-Marseille II, Centre de Recherche en Economie des Transports
Avenue Gaston Berger, I3IOO AIX EN PROVENCE

CLUSTERING METHODS USING THE CONCEPT OF SPACE CONTRACTION

Abstract :

The aim of cluster analysis is to structure a large number of entities which are characterized by values of several variables, so that the entities are hierarchically classified.

A number of the usual hierarchical algorithms are quite hopeless to apply to more than a comparatively small number of objects.

If the sample size is too small, the clusters observed may have little meaning and cluster analysis methods should be able to handle a large multivariate data set.

The graph theoretical clustering methods based on the concept of "space-contraction", which are presented, are optimal methods for the known sorting strategies : single linkage, complete linkage, mean linkage, variance...

The computational performance of these methods is such that they are applicable to a wide class of practical problems involving large sample size and high dimensionality.

Résumé :

L'un des principaux objectifs de la classification est de condenser, de structurer l'information contenue dans des données nombreuses, caractérisées par de multiples descripteurs.

Les nouvelles méthodes de classification ascendante hiérarchique présentées dans cet article permettent d'édifier une hiérarchie exacte sur un vaste ensemble de données selon une stratégie d'agrégation quelconque : lien minimum, diamètre minimum, distance moyenne, variance...

Ces nouvelles méthodes permettent d'obtenir rapidement des partitions emboîtées d'un grand ensemble de données et peuvent trouver des applications dans de nombreux domaines : météorologie, phytosociologie, géologie, microbiologie, médecine, économie, géographie...

INTRODUCTION

La description fine de la réalité en géographie, en économie, en géologie.. nécessite le recueil de données nombreuses et riches caractérisées par de multiples descripteurs.

La méthode classique de classification ascendante hiérarchique permet d'édifier une hiérarchie exacte selon un critère d'agrégation quelconque, mais il est exclu d'en faire usage sur un tableau de plus d'un millier d'objets [3], [5], [10], [11], [12], [13].

Par contre, d'autres méthodes capables de prendre en compte un plus grand nombre d'objets présentent certaines contraintes : choix a priori du nombre de classes [6], obligation d'utiliser la stratégie d'agrégation selon le lien minimal [7], [9].

Les nouvelles méthodes de classification ascendante hiérarchique, présentées dans cet article, combinent les avantages des deux types et permettent de dépasser les limites des algorithmes de classification usuels ; il est désormais possible d'envisager l'analyse hiérarchique de données nombreuses en un temps de calcul très faible [4].

Ces méthodes permettent d'obtenir rapidement un certain nombre de partitions emboîtées par coupure de la hiérarchie totale et peuvent être considérées comme des méthodes de classification non hiérarchique d'un grand ensemble de données taxinomiques.

I - CLASSIFICATION HEURISTIQUE PAR LA "METHODE DES GRAPHES DE CONTIGUITE"

1 - Le principe de la procédure heuristique de classification.

La "méthode des graphes de contiguïté", ici présentée, permet le traitement d'ensembles de données composés de quelques milliers d'objets et construit une hiérarchie sur l'ensemble des objets à classifier par l'une quelconque des méthodes d'agrégation selon le saut minimal, le diamètre minimal, la distance moyenne, la distance entre centres de gravité, la variance...

Cette méthode recherche initialement les voisins les plus proches de chaque objet à classer et construit un graphe de similarité, mis à jour étape par étape, après agrégation des deux sommets les plus proches entre lesquels il existe au moins un lien dans le graphe de similarité initial.

La méthode d'agrégation, ainsi définie, peut éventuellement conduire à des inversions dans la croissance de l'indice de stratification dont est munie la hiérarchie, néanmoins une première expérimentation sur un "grand" ensemble de données a montré que la hiérarchie engendrée selon le critère de la variance, ne présentait pas d'inversions [4].

2 - Description de la méthode de classification heuristique.

Cette méthode détermine une hiérarchie totale binaire A de parties sur l'ensemble I des objets à classer, selon une stratégie d'agrégation quelconque, en construisant une suite d'arbres binaires A_0, A_1 --- A_h --- $A_{|I|-1}$ et une suite de graphes $G_0 = (X_0, U_0)$, $G_1 = (X_1, U_1)$ --- $G_h = (X_h, U_h)$ --- $G_{|I|-1} = (X_{|I|-1}, U_{|I|-1})$

tels que :

- $A_{|I|-1} = A$
- X_h est l'ensemble des sommets supérieurs de l'arbre A_h : $X_h = Som(A_h)$. ($Som(A_h)$ est l'ensemble des sommets supérieurs de l'arbre A_h).
- U_h est l'ensemble des paires de sommets de $Som(A_h)$ susceptibles de s'agréger à l'étape h.

Initialisation faire h = o

$A_o = \{ \{i\} | i \in I\}$, $X_o = Som(A_o) = I$

Phase 0 : Détermination du graphe de similarité $G_o = (X_o, U_o)$

On note $V_0(s,\rho)$ et $V_F(s,\rho)$ le voisinage ouvert et le voisinage fermé de centre s et de rayon ρ.

Déterminer une suite de nombres $\{\rho(s) \mid s \in X_o\}$ telle que si :
$U_o = \{(s,t) \mid s \in X_o,\ t \in V(s,\rho(s))\}$
avec $\forall s \in X_o$, $V_0(s,\rho(s)) \subset V(s,\rho(s)) \subset V_F(s,\rho(s))$
alors $G_o = (X_o, U_o)$ est connexe.

Le choix des rayons $\rho(s)$ peut être tel que le voisinage $V(s,\rho(s))$ de chaque élément s de l'ensemble I, soit constitué des k voisins les plus proches de s ; le rayon $\rho(s)$ dépend alors de la densité des éléments situés à proximité de s.

Phase 1 : Agrégation des deux parties disjointes voisines les moins dissemblables entre elles.

Faire h = h + 1

$\delta(s_h, s'_h) = \inf \{\delta(s,s') \mid s,s' \in X_{h-1},\ (s,s') \in U_{h-1}\}$

Phase 2 : Construction de l'arbre A_h à partir de l'arbre A_{h-1}

Noter : $a_h = s_h \cup s'_h$ $\alpha(a_h) = s_h$ $\beta(a_h) = s'_h$ $\tau(a_h) = \delta(s_h, s'_h)$

$A_h = A_{h-1} \cup \{a_h\}$ $Som(A_h) = Som(A_{h-1}) \cup \{a_h\} - \{s_h, s'_h\}$

Phase 3 : Test d'arrêt

Si $h = |I| - 1$, alors $Som(A_h) = I$; $A_{|I|-1} = A$; $a_{|I|-1} = I$ FIN.

Phase 4 : Construction du graphe G_h à partir du graphe G_{h-1}

Déterminer $V(s_h) = \{t \mid t \in X_{h-1},\ (s_h,t) \in U_{h-1}\}$

$V(s'_h) = \{t \mid t \in X_{h-1},\ (s'_h,t) \in U_{h-1}\}$

$$V(a_h) = V(s_h) \cup V(s'_h) - \{s_h, s'_h\}$$

$X_h = \text{Som}\ (A_h)$

$U_h = U_{h-1} - \{(s_h,t) \mid t \in V(s_h)\} - \{(s'_h,t) \mid t \in V(s'_h)\} + ---$

$+ \{(a_h,t) \mid t \in V(a_h)\}$

Aller ensuite à la Phase 1.

REMARQUE :

Cette méthode heuristique de classification ascendante hiérarchique est applicable à l'analyse arborescente de données géographiques avec contraintes de contiguïté spatiale.

II - LES STRATEGIES D'AGREGATION CONTRACTANTES

La notion de stratégie d'agrégation contractante définie ci-après, sera utilisée par les nouvelles méthodes de classification d'un grand ensemble de données pour édifier une hiérarchie exhaustive exacte selon l'une quelconque des stratégies d'agrégation suivantes : lien minimum, diamètre minimum, distance moyenne ou variance.

Soit ρ un seuil de stratification, Q une partition de l'ensemble I des objets en classes disjointes, a une classe de Q ;
On note $\delta(t,a)$ l'indice de dissimilarité entre les classes t et a, dont la définition dépend de la stratégie d'agrégation utilisée, et $V(a,\rho)$ le voisinage de la classe a :

$$V(a,\rho) = \{t \mid t \in Q\ ;\ \delta(t,a) \leqslant \rho,\ t \neq a\}$$

Par définition, une stratégie d'agrégation contractante est telle que :

$$V(a \cup b, \rho) \subset V(a,\rho) \cup V(b,\rho) \quad \text{pour } \forall(a,b) \text{ tel que } \delta(a,b) \leqslant \rho$$

On montre que les méthodes d'agrégation selon le saut minimum, le diamètre minimum, la distance moyenne ou la variance sont contractantes ; par contre, les stratégies d'agrégation selon la distance entre centres de gravité des classes ou la distance angulaire dans un espace euclidien ne le sont pas [4-c]

III - CLASSIFICATION ASCENDANTE HIERARCHIQUE EXACTE PAR LA "METHODE DES GRAPHES REDUCTIBLES"

1 - Description de la méthode de classification.

La "méthode des graphes réductibles" détermine une hiérarchie totale exacte A sur l'ensemble I des objets à classifier en construisant une suite d'arbres binaires A_o, A_1 --- A_h --- $A_{|I|-1}$ et une suite de graphes de similarité $G_o = (X_o,U_o)$, $G_1 = (X_1,U_1)$ ---

--- $G_h = (X_h,U_h)$ --- $G_{|I|-1} = (X_{|I|-1},U_{|I|-1})$

tels que :

- $A_{|I|-1} = A$
- X_h est l'ensemble des sommets supérieurs non isolés de l'arbre A_h.
- U_h est l'ensemble des paires de sommets de X_h dont la dissimilarité est inférieure ou égale à un seuil de stratification ρ_h fixé à priori.

Soit $\{\rho_k^* \mid k = 0, 1, 2 \text{ ---}\}$ une suite quelconque de seuils de stratification.

Initialisation faire h = o, k = o $A_o = \{\{i\} \mid i \in I\}$

Phase 0 : Construction du graphe des liaisons de similarité entre les sommets supérieurs de l'arbre binaire.

Faire $\rho_h = \rho_k^*$

Soit $U_h = \{(s,s') \mid s,s' \in Som(A_h),\ s \neq s',\ \delta(s,s') \leqslant \rho_h\}$

$X_h = \{s \mid s \in Som(A_h), \exists\, s' \in Som(A_h)$ tel que $(s,s') \in U_h\}$

$G_h = (X_h,U_h)$

Si $X_h = \emptyset$ et $U_h = \emptyset$, faire k = k + 1 et aller à la phase 0. Sinon, aller à la phase 1.

Phase 1: Agrégation des deux sommets s_h et s'_h reliés par l'arête la plus courte du graphe de similarité G_{h-1}.

Phase 2 : Construction de l'arbre A_h à partir de l'arbre A_{h-1}.

Phase 3 : Test d'arrêt.

Phase 4 : Construction du graphe G_h à partir du graphe G_{h-1}

Faire $\rho_h = \rho_{h-1}$

Déterminer $V(s_h,\rho_h) = \{t \mid t \in X_{h-1},\ (t,s_h) \in U'_{h-1}\}$

$V(s'_h,\rho_h) = \{t \mid t \in X_{h-1},\ (t,s'_h) \in U_{h-1}\}$

Noter $E(a_h,\rho_h) = V(s_h,\rho_h) \cup V(s'_h,\rho_h) - \{s_h,s'_h\}$

si la stratégie d'agrégation est contractante, sinon noter

$E(a_h,\rho_h) = Som(A_{h-1}) - \{s_h,s'_h\}$

Calculer ensuite les indices de dissimilarité $\delta(t,a_h)$ entre le nouveau sommet créé a_h et les sommets supérieurs de l'arbre A_{h-1}, qui appartiennent à l'ensemble $E(a_h,\rho_h)$.

Déterminer $V(a_h,\rho_h) = \{t \mid t \in E(a_h,\rho_h),\ \delta(t,a_h) \leqslant \rho_h\}$

et $U_h = U_{h-1} - \{(t,s_h) \mid t \in V(s_h,\rho_h)\} - \{(t,s'_h) \mid t \in V(s'_h,\rho_h)\} + ---$

$+ \{(t,a_h) \mid t \in V(a_h,\rho_h)\}$

$X_h = \{s \mid s \in Som\ (A_h),\ V(s,\rho_h) \neq \emptyset\}$

Si $X_h = \emptyset$ et $U_h = \emptyset$, faire $k = k + 1$ et aller à la phase 0. Sinon, aller à la phase 1.

2 - Justification de la méthode de classification.

On peut montrer que la méthode proposée engendre une hiérarchie exacte qui ne dépend pas de la suite des seuils de stratification choisis à priori ou déterminés lors de la procédure de classification. De plus dans le cas d'une stratégie d'agrégation contractante, la hiérarchie exacte édifiée sur l'ensemble à classifier est munie d'un indice de stratification qui ne présente pas d'inversions [4-c].

L'expérimentation numérique de cette nouvelle méthode a permis de construire une hiérarchie exacte sur un ensemble de plus de 1 000 objets, en un temps de calcul très faible.

IV - CLASSIFICATION ASCENDANTE HIERARCHIQUE EXACTE PAR LA "METHODE DES SEUILS LOCALISES".

La "méthode des seuils localisés" détermine une hiérarchie totale exacte A sur l'ensemble I des objets à classifier, en construisant une suite d'arbres binaires, une suite de famille de voisinages et une suite de graphes de similarité.

A chaque étape de la construction de la hiérarchie, la "méthode des graphes réductibles" [4-c] suppose défini un seuil de stratification unique pour déterminer le voisinage de chaque sommet supérieur de l'arbre en cours de construction.

Par contre, la "méthode des seuils localisés" est fondée sur une définition locale des seuils de stratification, susceptible d'améliorer l'efficacité de la procédure de classification.

Initialisation : $X_o = I$

Phase 0 : Détermination du graphe de similarité initial $G_o = (X_o, U_o)$

Pour tout $s \in X_o$, déterminer un seuil de stratification local $\rho(s)$ tel que le voisinage associé $V(s,\rho(s))$ ne soit pas vide (en recherchant les p plus proches voisins de s par exemple).

$V(s,\rho(s)) = \{t \mid t \in X_o - \{s\},\ \delta(t,s) \leqslant \rho(s)\}$

Faire $U_o = \{(s,t) \mid s \in X_o,\ t \in V(s,\rho(s))\}$

Phase 1 : Agrégation des deux parties les moins dissemblables entre elles.

Phase 2 : Construction de l'arbre A_h à partir de l'arbre A_{h-1}.

Phase 3 : Test d'arrêt.

Phase 4 : Construction du graphe de similarité $G_h = (X_h, U_h)$

Faire $X_h = Som(A_h)$

4-1 Mise à jour des voisinages des sommets supérieurs de l'arbre A_h, autres que le nouveau sommet créé a_h.

Soit $t \in X_{h-1} - \{s_h, s'_h\}$

Noter $I_{h-1}(s)$ l'ensemble des sommets x qui possèdent le sommet s dans leur voisinage : $I_{h-1}(s) = \{x \mid x \in X_{h-1},\ s \in V_{h-1}(x,\rho(x))\}$ et $V^*_{h-1}(t,\rho(t)) = V_{h-1}(t,\rho(t)) - \{\{s_h,s'_h\} \cap V_{h-1}(t,\rho(t))\}$

4-1-1 Si $t \notin I_{h-1}(s_h) \cup I_{h-1}(s'_h) - \{s_h,s'_h\}$,

faire $V_h(t,\rho(t)) = V_{h-1}(t,\rho(t))$

si la stratégie d'agrégation est contractante, sinon, faire $V_h(t,\rho(t)) = V_{h-1}(t,\rho(t))$ si $\delta(t,a_h) > \rho(t)$ et faire $V_h(t,\rho(t)) = V_{h-1}(t,\rho(t)) + \{a_h\}$ si $\delta(t,a_h) \leqslant \delta(t)$.

4-1-2 Si $t \in I_{h-1}(s_h) \cup I_{h-1}(s'_h) - \{s_h,s'_h\}$, distinguer les deux cas suivants :

i) Si : $\delta(t,a_h) \leqslant \rho(t)$,
faire $V_h(t,\rho(t)) = V^*_{h-1}(t,\rho(t)) + \{a_h\}$

ii) Sinon $\delta(t,a_h) > \rho(t)$

ii-1) si $V^*_{h-1}(t,\rho(t)) \neq \emptyset$, faire
$V_h(t,\rho(t)) = V^*_{h-1}(t,\rho(t))$

ii-2) si $V^*_{h-1}(t,\rho(t)) = \emptyset$, déterminer un seuil de stratification local $\rho(t)$ tel que le voisinage correspondant $V_h(t,\rho(t))$ ne soit pas vide.

4-2 Détermination du voisinage du nouveau sommet créé a_h.

Soit $\rho(a_h) = \inf \{\rho(s_h)\ ;\ \rho(s'_h)\}$

Noter $E(a_h,\rho(a_h)) = V_{h-1}(s_h,\rho(a_h)) \cup V_{h-1}(s'_h,\rho(a_h)) - \{s_h,s'_h\}$ si la stratégie d'agrégation est contractante, sinon noter :

$$E(a_h,\rho(a_h)) = \mathrm{Som}(A_{h-1}) - \{s_h,s'_h\}$$

Déterminer $B(a_h,\rho(a_h)) = \{t \mid t \in E(a_h,\rho(a_h)),\ \delta(t,a_h) \leqslant \rho(a_h)\}$

4-2-1 Si $B(a_h,\rho(a_h)) \neq \emptyset$, faire $V_h(a_h,\rho(a_h)) = B(a_h,\rho(a_h))$

4-2-2 Si $B(a_h,\rho(a_h)) = \emptyset$, déterminer un seuil de stratification local $\rho(a_h)$ tel que le voisinage associé $V_h(a_h,\rho(a_h))$ ne soit pas vide (en recherchant, par exemple, les p plus proches voisins de a_h).

On a alors : $V_h(a_h,\rho(a_h)) = \{t \mid t \in Som(A_{h-1}) - \{s_h,s'_h\};\ \delta(t,a_h) \leqslant \rho(a_h)\} \neq \emptyset$

et $Card(V_h(a_h,\rho(a_h)) \leqslant p$

4-3 Détermination de l'ensemble des arcs orientés du graphe de similarité G_h.

$$U_h = \{(s,t) \mid s \in X_h,\ t \in V_h(s,\rho(s))\}$$

Aller ensuite à la Phase 1.

V - CLASSIFICATION ASCENDANTE HIERARCHIQUE EXACTE PAR LA "METHODE DES VOISINS RECIPROQUES".

1 - Le principe de la procédure de classification.

La "méthode des voisins réciproques" est une méthode de décomposition d'un grand problème de classification, qui permet d'édifier une hiérarchie totale exacte sur l'ensemble des objets à classifier, selon une stratégie d'agrégation contractante quelconque (lien minimal, diamètre minimal, distance moyenne, variance)

Cette méthode recherche d'abord les couples d'éléments voisins réciproques sur l'ensemble des objets à classifier, tels que chaque élément d'un couple soit le plus proche voisin de l'autre élément. On agglomère ensuite progressivement les éléments voisins réciproques à l'intérieur des composantes connexes du graphe des liaisons de réciprocité (ou liaisons conjuguées). Après réduction des composantes connexes, on recherche les couples de composantes connexes voisines réciproques pour définir le graphe des liaisons conjuguées entre les sommets supérieurs de l'arbre en cours de construction. On poursuit alternativement le processus d'agglomération et le processus de détermination des classes voisines réciproques jusqu'à ce que tous les objets

soient réunis en une seule classe.

La "méthode des voisins réciproques" fait donc alterner successivement la définition d'un graphe de similarité et la construction de l'arbre binaire par agglomérations successives, comme dans la "méthode des graphes réductibles". Le principe de la "méthode des voisins réciproques" est aussi analogue à celui de la "méthode des seuils localisés" ; en effet, à chaque élément, objet ou classe, peut être associé un seuil de stratification local, égal à la distance de cet élément à son plus proche voisin.

2 - Les éléments voisins réciproques.

Soit Q une partition de l'ensemble I des objets en classes disjointes, s et t deux classes de Q.

On note : $\forall s \in Q$: $d(s) = \inf \{\delta(s,t) \mid t \in Q, t \neq s\}$

et $B(s) = \{t \mid t \in Q, t \neq s, \delta(t,s) = d(s)\}$

Par définition, s et t sont deux classes voisines réciproques si on a :

$$t \in B(s) \quad \text{et} \quad s \in B(t)$$

On dit alors que (s,t) et (t,s) sont des liaisons conjuguées. Le couple (s,t) peut alors être représenté par une arête non orientée d'un graphe de similarité construit sur l'ensemble des classes de Q :

$G = (Q,U)$

avec $U = \{(s,t) \mid s,t \in Q, s \neq t, t \in B(s), s \in B(t)\}$

$= \{(s,t) \mid s,t \in Q, s \neq t, d(s) = d(t)\}$

3 - Description de la méthode de classification.

Initialisation $X_o = I$

Phase 0 : Détermination du graphe des liaisons conjuguées entre les sommets supérieurs de l'arbre A_h.

$\forall s \in Som(A_h)$, noter :

$d_h(s) = \inf \{\delta(s,t) \mid t \in Som(A_h), t \neq s\}$

$$B_h(s) = \{t \mid t \in Som(A_h),\ t \neq s,\ \delta(s,t) = d_h(s)\}$$

et $V_h(s) = \{t \mid t \in B_h(s),\ d_h(s) = d_h(t)\}$

Déterminer :

$X_h = \{s \mid s \in Som(A_h),\ V_h(s) \neq \emptyset\}$

$U_h = \{(s,t) \mid s,t \in X_h\ ;\ s \neq t\ ;\ t \in V_h(s),\ s \in V_h(t)\}$

$G_h = (X_h, U_h)$

Phase 1 : Agrégation des deux sommets voisins réciproques reliés par l'arête la plus courte du graphe de similarité G_{h-1}

Phase 2 : Construction de l'arbre A_h à partir de l'arbre A_{h-1}

Phase 3 : Test d'arrêt

Phase 4 : Construction du graphe G_h à partir du graphe G_{h-1}

Déterminer : $V_h(a_h) = V_{h-1}(s_h) \cap V_{h-1}(s'_h)$

$\forall\, s \in V_h(a_h) : V_h(s) = V_{h-1}(s) - \{s_h, s'_h\} + \{a_h\}$

$\forall\, s \in X_{h-1} - V_{h-1}(s_h) \cup V_{h-1}(s'_h) : V_h(s) = V_{h-1}(s)$

$\forall\, s \in V_{h-1}(s_h) \cup V_{h-1}(s'_h) - V(a_h) : V_h(s) = \emptyset$

$U_h = U_{h-1} - \{(t,s_h) \mid t \in V_{h-1}(s_h)\} - \{(t,s'_h) \mid$ ----

---- $\mid t \in V_{h-1}(s'_h)\} + \{(t,a_h) \mid t \in V_h(a_h)\}$

$X_h = \{s \mid s \in Som(A_h),\ V_h(s) \neq \emptyset\}$

$= X_{h-1} - V_{h-1}(s_h) \cup V_{h-1}(s'_h) + V(a_h)$

Si $V(a_h) \neq \emptyset$, faire $\delta(t,a_h) = \delta(s_h,s'_h)$ pour $\forall\, t \in V(a_h)$

(Remarque : Si $V(s_h) = \{s'_h\}$ et si $V(s'_h) = \{s_h\}$, s_h et s'_h forment une composante connexe isolée du graphe G_{h-1} et on a :
$V_h(a_h) = \emptyset$, $U_h = U_{h-1} - \{(s_h,s'_h)\}$ et $X_h = X_{h-1} - \{s_h,s'_h\}$)

Si $U_h = \emptyset$, aller ensuite à la Phase 1, sinon aller à la Phase 0.

4 - Justification de la procédure de classification et aspects informatiques.

On peut montrer que la "méthode des voisins réciproques" engendre une hiérarchie exacte selon une stratégie d'agrégation contractante. Par contre si la stratégie d'agrégation n'est pas contractante (stratégie des centres de gravité, par exemple), la méthode proposée n'est plus une procédure exacte mais une procédure heuristique de classification de données nombreuses.

De plus, si les composantes connexes des différents graphes de similarité, sont des sous-graphes complets, on peut démontrer l'unicité de la hiérarchie exacte édifiée selon une stratégie d'agrégation contractante. On retrouve la propriété classique d'unicité de la hiérarchie lorsque les composantes connexes sont toutes des couples disjoints de voisins réciproques.

Du point de vue informatique, l'existence éventuelle de composantes connexes de cardinal élevé, nécessite l'introduction de seuils de stratification globaux, comme dans la "méthode des graphes réductibles", pour tenir compte de la limitation de l'espace mémoire disponible et pour permettre la classification de données nombreuses sur un mini-ordinateur.

VI - UNE METHODE RAPIDE DE DETERMINATION D'UNE FAMILLE DE VOISINAGES

1 - Performances des méthodes de classification utilisant le concept de voisinage .

Les nouvelles méthodes de classification, présentées dans cet article, font alterner la détermination d'une famille de voisinages et l'édification de la hiérarchie binaire par agglomérations successives.

Le temps de calcul nécessaire à la construction de la hiérarchie par agglomérations successives des classes les moins dissemblables varie proportionnellement avec le nombre d'objets à classifier, dans le cas d'une stratégie d'agrégation contractante.

Par contre, la détermination d'une famille de voisinages sur un ensemble E, par énumération séquentielle exhaustive des couples d'éléments de E, fait appel au calcul de N(N-1) /2 indices de dissimilarité sur E, N étant le nombre d'éléments de E. Le temps de calcul correspondant, proportionnel au carré du nombre d'éléments, augmente rapidement avec la taille du problème de classification.

Il est donc exclu de faire usage d'une telle procédure de classification sur un tableau de plusieurs dizaines de milliers d'objets.

La nouvelle méthode de détermination d'une famille de voisinages, fondée sur la notion d'indice de dissimilarité minorant, permet de dépasser la limite précédente, en construisant un graphe de similarité en un temps de calcul beaucoup plus réduit , lorsque les données sont représentées dans un espace euclidien multidimensionnel. Il est ainsi possible d'envisager la classification d'un très grand ensemble de données en un temps de calcul acceptable.

2 - Définition d'une famille de voisinages sur une partition de l'ensemble à classifier.

Soit E une partition de l'ensemble I des objets à classifier. Au début de la construction de la hiérarchie sur I, E est l'ensemble des objets de I puis E est l'ensemble $Som(A_h)$ des sommets supérieurs de la hiérarchie A_h, lors de chaque définition d'un graphe de similarité.

On note $\delta(s,t)$ l'indice de dissimilarité entre les éléments s et t de E et $V(s,\rho;E)$ le voisinage de rayon ρ, de l'élément s, dans l'ensemble E.

$$V(s,\rho;E) = \{t \mid t \in E \ ; \ \delta(s,t) \leqslant \rho\}$$

On se propose de rechercher une famille de voisinages

$$F = \{V(s,\rho(s);E) \mid s \in E\}$$

pour construire un graphe de similarité G = (E,U) tel que

$$U = \{(s,t) \mid s, t \in E \ ; \ t \in V(s,\rho(s);E)\}$$

3 - Evaluation minorante d'un indice de dissimilarité.

Soit Q une partition de E, t et t' deux classes de Q.

L'indice de dissimilarité δ^*, défini sur la partition Q, est une évaluation minorante de l'indice de dissimilarité δ entre deux classes quelconques de la partition E, si on a :

$\forall t,t' \in Q \ ; \ s,s' \in E \ ; \quad s \subset t, \quad s' \subset t' \ : \ \delta(s,s') \geq \delta^*(t,t')$

La connaisance d'un indice de dissimilarité minorant permet d'éliminer implicitement un nombre élevé de paires d'éléments à examiner pour déterminer une famille de voisinages sur E.

En effet, soit ρ un seuil de stratification quelconque, on a :

$$\delta^*(t,t') > \rho \Longrightarrow \forall s \in t,\ s' \in t' \ :\ \delta(s,s') > \rho$$

Cette élimination implicite de paires d'éléments dissemblables permet d'accélérer la procédure de détermination du voisinage d'un élément s de E, puisque l'on a :

$$V(s,\rho;E) = \{s' \mid s' \in V_Q(s,\rho)\ ;\ \delta(s,s') \leqq \rho\}$$

$$\text{avec } V_Q(s,\rho) = \cup \{t \mid t \in Q,\ \delta^*(s,t) \leqq \rho\}$$

4 - Classification dans un espace euclidien multidimensionnel

Dans le cas d'une stratégie d'agrégation selon le saut minimum, le diamètre minimum, la distance moyenne, la distance entre centres de gravité, l'indice de dissimilarité $\delta(s,s')$ entre deux éléments s et s' d'un ensemble E, qui appartiennent respectivement aux classes t et t' d'une partition Q sur E, est minoré par l'indice de dissimilarité $\delta^*(t,t')$ défini ci-après :

$$\delta^*(t,t') = ||g(t) - g(t')|| - [r(t) + r(t')]$$

où $g(t)$ représente le centre de gravité de la classe t

et où $r(t) = \sup \{||x - g(t)|| \mid x \in t\}$

Dans le cas de la stratégie d'agrégation selon la variance, on a :

$$\delta^*_{var}(t,t') = \frac{m(t).\ m(t')}{M(t) + M(t')} \left[||g(t) - g(t')|| - [r(t) + r(t')]\right]^2$$

où $m(t) = \inf \{m_s \mid s \in t\}$ (m_s est la masse de l'élément s de E)

$M(t) = \sup \{m_s \mid s \in t\}$

5 - Arbre de longueur minimum et indice de dissimilarité minorant.

Dans le cas d'une stratégie d'agrégation contractante, la connaissance de l'arbre de longueur minimum du graphe de similarité complet, défini sur l'ensemble des objets à classifier, permet de définir une évaluation minorante "optimale" de l'indice de dissimilarité, supérieure à l'indice de dissimilarité minorant fondé sur l'inégalité triangulaire.

En effet, on peut montrer que :

$$\forall t,t' \in Q,\ \forall s \in t,\ s' \in t' \ :\ \delta(s,s') \geqq \delta_{UIM}(t,t')$$

où δ_{UIM} est l'ultramétrique inférieure maxima à l'indice de dissimilarité $\delta(\{i\}, \{i'\})$ dont la définition sur les parties $\{i\}$ et $\{i'\}$ de I, réduites à un élément, dépend de la stratégie d'agrégation [3].

De plus la connaissance de l'arbre de longueur minimum permet de définir des seuils de stratification locaux et de séparer le problème de classification en sous-problèmes disjoints.

En effet, la suppression de l'arête la plus longue de l'arbre de longueur minimum fait apparaître deux composantes connexes disjointes auxquelles on peut associer un seuil de stratification local égal à la longueur de l'arête supprimée. Il est ainsi possible de séparer le problème de classification initial en deux sous-problèmes disjoints. La séparation ultérieure de chacun de ces sous-problèmes, permet d'envisager la classification d'un vaste ensemble de données par décompositions successives.

6 - Classification heuristique des données et détermination d'une famille de voisinages.

Connaissant des seuils de stratification locaux et une évaluation minorante précise de l'indice de dissimilarité, il est possible d'envisager la détermination rapide d'une famille de voisinages sur un grand ensemble de données, après avoir défini une partition Q, sur l'ensemble à classifier, par une méthode heuristique quelconque.

Il est possible de structurer rapidement un grand ensemble de données, en un temps de calcul proportionnel à $N\text{Log}_2 N$, soit par classification descendante hiérarchique, chaque classe étant subdivisée en deux sous-classes par une variante de la "méthode des nuées dynamiques" par exemple [3], soit par classification ascendante hiérarchique d'un échantillon de taille réduite, de 1 500 à 2 000 objets par exemple, et détermination d'une partition munie d'une hiérarchie par la "méthode des éléments supplémentaires".

La partition définie sur l'ensemble a classifier étant munie d'une structure hiérarchique, il est possible d'accélérer la détermination d'une famille de voisinages, par une procédure de séparations et éliminations implicites séquentielles, fondée sur la connaissance d'un indice de dissimilarité minorant.

7 - Description de la méthode exacte de détermination du voisinage $V(s,\rho;E)$

Soit $H_o(Q)$ une hiérarchie binaire édifiée sur les classes de la partition Q, par une méthode heuristique quelconque. La méthode suivante de détermination du voisinage $V(s,\rho;E)$, par séparations et éliminations séquentielles, est une méthode exacte, dont le résultat ne dépend pas du choix de la hiérarchie $H_o(Q)$.

Initialisation Faire n = o

Soit $L_o = \{a_{|Q|-1}\}$ et $V_o(s,\rho) = \emptyset$

Phase 0 : Sélection d'un sommet pendant a_n dans la liste L_{n-1}

Faire n = n + 1

Soit $\phi(s)$ une mesure quantitative de l'importance de la classe s ; $\phi(s)$ peut être le cardinal, le diamètre, la variance ou le moment d'ordre deux de la classe s, par exemple.

Déterminer $\phi(a_n) = \sup \{\phi(s) \mid s \in L_{n-1}\}$

Soit $Sui(a_n)$ l'ensemble des successeurs immédiats de a_n dans la hiérarchie binaire $H_o(Q)$.

Si $Sui(a_n) \neq \emptyset$, aller à la Phase 1

Sinon, $Sui(a_n) = \emptyset$, aller alors à la Phase 2.

Phase 1 : Séparation de la classe a_n et évaluation des deux successeurs de a_n.

Faire $V_n(s,\rho) = V_{n-1}(s,\rho)$

Soient $\alpha(a_n)$ et $\beta(a_n)$ les deux successeurs immédiats de a_n dans la hiérarchie binaire $H_o(Q)$: $Sui(a_n) = \{\alpha(a_n), \beta(a_n)\}$

. Si $\inf \{\overset{*}{\delta}(s,\alpha(a_n)) ; \overset{*}{\delta}(s,\beta(a_n))\} > \rho$, faire $L_n = L_{n-1} - \{a_n\}$

. Si $\sup \{\overset{*}{\delta}(s,\alpha(a_n)) ; \overset{*}{\delta}(s,\beta(a_n))\} \leqq \rho$, faire $L_n = L_{n-1} - \{a_n\} + \{\alpha(a_n), \beta(a_n)\}$

. Si $\overset{*}{\delta}(s,\alpha(a_n)) \leqq \rho$ et $\overset{*}{\delta}(s,\beta(a_n)) > \rho$, faire $L_n = L_{n-1} - \{a_n\} + \{\alpha(a_n)\}$

. Si $\overset{*}{\delta}(s,\alpha(a_n)) > \rho$ et $\overset{*}{\delta}(s,\beta(a_n)) \leqq \rho$, faire $L_n = L_{n-1} - \{a_n\} + \{\beta(a_n)\}$

Aller à la Phase 3.

Phase 2 : Mise à jour du voisinage du sommet s.

Faire $V_n(s,\rho) = V_{n-1}(s,\rho) \cup V(s,\rho;a_n)$

$L_n = L_{n-1} - \{a_n\}$

Aller à la Phase 3.

Phase 3 : Test d'arrêt.

Si $L_n \neq \emptyset$, aller à la Phase 0.

Sinon, faire $V(s,\rho;E) = V_n(s,\rho)$ FIN.

VII - APPLICATION A L'ANALYSE DE LA DISTANCE A LA VILLE EN LANGUEDOC-ROUSSILLON

L'application géographique traitée se situe en aval des travaux réalisés au sein du groupe DUPONT sur la distance à la ville, où à l'un de ses stades, avaient été analysées les distances kilométriques des communes péri-montpelliéraines et des communes péri-grenobloises à différents types d'équipements urbains.

Le but poursuivi n'a pas d'autre prétention que d'expérimenter les méthodes nouvelles de classification et d'en juger l'efficacité du point de vue de l'utilisateur géographe.

L'analyse factorielle des milieux urbains et ruraux de l'aire languedocienne a permis de définir des distances synthétiques à la ville, de dégager la structure qui schématise la distance de "l'urbain au rural", et l'application de la "méthode des graphes réductibles" à la classification hiérarchique des 1561 communes en Languedoc-Roussillon a permis ensuite un découpage spatial qui différencie les communes selon les distances à divers équipements urbains, distances qui se combinent souvent contradictoirement selon les lieux.

Les résultats obtenus ont présenté un intérêt certain pour les géographes qui, habitués à voir dans l'espace languedocien une organisation urbaine longitudinale, n'étaient pas prêts à voir surgir avec autant de netteté la complexité du fait urbain gardois ou bien la disposition urbaine audoise profondément différente de celle du département de l'Hérault [2].

VIII - CONCLUSION

Les nouvelles méthodes de classification présentées dans cet article, permettent de dépasser les limites des méthodes de classification usuelles et il est désormais possible d'envisager l'analyse hiérarchique d'un vaste ensemble de données, en un temps de calcul très faible.

BIBLIOGRAPHIE

[1] AURIAC F., BERNARD M.C., FERRAS F., VIGOUROUX M. (Groupe DUPONT) (1975). La distance à la ville : essais d'analyses factorielles appliquées aux cas de Grenoble et Montpellier. *L'Espace Géographique*, t. IV, n° 4, 225-238.

[2] AURIAC F., BRUYNOOGHE M. (1976). Une classification sur plus de 1500 communes : la distance à la ville en Languedoc-Roussillon. Actes du 4è Colloque sur l'analyse des données en géographie. *Cahiers de Géographie de Besançon* (à paraître).

[3] BENZECRI J.P. (1964-1973). Leçons sur les classifications. Cours 3è cycle, *Institut de Statistique de l'Université de Paris (I.S.U.P.)*, Paris.

(1973). L'analyse des données. T. 1, La taxinomie numérique ; T. 2, L'analyse des correspondances, Paris, *Dunod*.

(1976). Histoire et préhistoire de l'analyse des données. *Les Cahiers de l'analyse des données*, Vol. I., n°1, 2, 3, 4.

[4] BRUYNOOGHE M. (1976-a) Un algorithme de classification ascendante hiérarchique d'un grand ensemble de données. Communication au Congrès européen des statisticiens, Grenoble, 6-10 sept. 1976.

(1977-b) Classification ascendante hiérarchique d'un grand ensemble de données utilisant la notion de graphe de contiguïté. Application à l'analyse de la distance à la ville en Languedoc-Roussillon. Thèse doctorat 3è cycle. Laboratoire de Statistique Mathématique de l'Université Paris VI.

(1977-c) Classification automatique d'un grand ensemble de données par la "méthode des graphes réductibles". Actes des premières journées internationales. Analyse des données et Informatique. p. 85-91 7-9 septembre 1977 Versailles.

[5] CORMACK R.M. (1971). A review of classification. *J. R. Statist. Soc., A, 134*, Part 3, 321-267.

[6] DIDAY E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes. *Revue de Statistique Appliquée*, Vol. XIX, n° 2.

[7] GOWER J.C. and ROSS G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statist*, 18, 54-64.

[8] JAMBU M. (1972). Techniques de classification automatique appliquées à des données "Sciences Humaines", Thèse de Doctorat de 3è cycle, *Laboratoire de Statistique Mathématique de l'Université Paris VI.*

[9] JARVIS R.A. and PATRICK E.A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, Vol. C-22, n° 11, 1025-1034.

[10] ROUX M. (1968). Un algorithme pour construire une hiérarchie particulière. Thèse de doctorat de 3è cycle. *Laboratoire de Statistique Mathématique de l'Université de Paris VI.*

[11] SOKAL R.R. and SNEATH P.H.A. (1963). Principles of Numerical taxonomy. London *Freeman*.

[12] WARD J.H. (1963). Hierarchical grouping to optimise an objective function. *J. Am. Statist. Ass., 59, 236-244.*

[13] WISHART D. (1969). An algorithm for hierarchical classifications. *Biometrics* 25, 165-170.