LES CAHIERS DE L'ANALYSE DES DONNÉES

M. ROUX

Sur la valeur maxima des indices de niveau en classification ascendante hiérarchique dans le cas de l'agrégation par la distance moyenne

Les cahiers de l'analyse des données, tome 7, n° 2 (1982), p. 155-161

<http://www.numdam.org/item?id=CAD_1982__7_2_155_0>

Les Cahiers de l'Analyse des Données
Vol. VII - 1982 - n° 2 - p. 155-161

SUR LA VALEUR MAXIMA DES INDICES DE NIVEAU EN CLASSIFICATION ASCENDANTE HIÉRARCHIQUE DANS LE CAS DE L'AGRÉGATION PAR LA DISTANCE MOYENNE {BORNES NIV. C.A.H.]

par M. Roux (1)

1 *Introduction*

En classification ascendante hiérarchique, l'un des principaux obstacles à la mise en oeuvre de l'algorithme des voisinages réductibles (cf. Bruynooghe, Jambu) est le choix *a priori* d'un seuil de stratification, en particulier si l'on utilise comme point de départ un indice de distance, comme celui de Jaccard, couramment employé sur les données booléennes. C'est pourquoi nous avons cherché s'il n'était pas possible de calculer, en même temps que la matrice des distances, les limites entre lesquelles le plus grand indice de diamètre peut varier. On pourrait alors choisir le seuil de stratification avec plus de précision et améliorer du même coup l'efficacité de cet algorithme.

Le cas de l'agrégation par la distance moyenne nous a paru important à résoudre à cause des bons résultats que fournit en général cette stratégie. La difficulté du problème provient de ce qu'à chaque pas de l'algorithme, l'indice de niveau ne représente pas la moyenne des distances à l'intérieur du sous-ensemble créé mais la moyenne des distances entre les deux groupes qui l'ont formé. Le plus grand indice de niveau n'échappe pas à cette règle et dépend donc de l'histoire des agglomérations précédentes.

Nous exposerons d'abord l'aspect mathématique du problème, on jugera ensuite, sur des exemples réels, l'intérêt des bornes trouvées pour le plus grand indice de niveau.

Dans les démonstrations qui suivent, on supposera que les points i de l'ensemble I (qui doit être muni d'une classification) sont tous de même masse, cette restriction n'est toutefois pas requise pour la validité de nos résultats.

2 *Quelques remarques théoriques sur l'agrégation par la distance moyenne*

Dans tout ce qui suit on ne considérera que des distances entre des points *distincts*, celles-ci pouvant cependant être nulles (on exclut la diagonale principale de la matrice des distances).

2.1 *Relation entre indice de niveau et moyenne des distances* : On appelle I l'ensemble des objets, ou terminaux, de l'arbre hiérarchique, d(i,i') la distance initiale (ou l'indice de distance) entre les points i et i' de I : on considère la classification obtenue grâce à l'agglomération par la distance moyenne; soit q et q' des parties

(1) *Chargé de recherches CNRS.*

de la partition Q_h correspondant à une étape quelconque h de l'algorithme) on appelle $\delta(q,q')$ l'indice de niveau associé à une éventuelle fusion des classes q et q'. Cet indice est celui qui provient des modifications successives de la matrice d des distances. Enfin on désigne par $N(q) = \text{Card } q \times (\text{Card } q - 1)/2$ le nombre de paires d'éléments distincts de q.

Proposition 1

La moyenne des distances entre objets appartenant à la réunion de deux classes est inférieure ou égale à l'indice de niveau associé à cette réunion.

$$\forall\, q \in Q,\ \forall\, q' \in Q : \Sigma\{d(i,i') \mid i \in q \cup q' \ ;\ i' \in q \cup q'\}/N(q \cup q') \leq \delta(q,q')$$

Pour la démonstration nous procéderons par récurrence. Au début de l'algorithme toutes les classes ne contiennent qu'un seul objet. La relation à démontrer se réduit donc à :

$$\forall\, i \in I,\ \forall\, i' \in I : d(i,i') \leq \delta(\{i\},\ \{i'\}) = d(i,i')$$

Elle est donc vraie dans ce cas. Supposons qu'elle le soit encore au pas h, et que le pas h+1 consiste à fusionner les deux classes q' et q". Il faut alors montrer que la relation tient encore avec la nouvelle classe $q' \cup q"$:

$$\forall\, q \neq q', q" : \Sigma\{d(i,i') \mid i,i' \in q \cup q' \cup q"\}/N(q \cup q' \cup q") \leq \delta(q, q' \cup q")$$

Les classes q, q' et q" étant d'intersections vides, on peut calculer la moyenne M, qui constitue la partie gauche de cette inégalité, à l'aide de 3 moyennes partielles.

$$M = \frac{\alpha\,\Sigma\{d(i,i') \mid i,i' \in q\}}{N(q)} + \frac{\beta\Sigma\{d(i,i') \ i,i' \in q' \cup q"\}}{N(q' \cup q")} + \frac{\gamma\Sigma\{d(i,i' \mid i \in q, i' \in q' \cup q"\}}{\text{Card}q\ \text{Card}(q' \cup q")}$$

où les coefficients α, β et γ sont de somme égale à l'unité. Considérons le premier terme de cette somme. Si la classe q est réduite à un seul objet ce terme n'existe pas, mais cela n'entache pas la suite de la démonstration. Si la classe q a au moins deux objets c'est qu'elle provient de la fusion de deux classes q_1 et q_2 à une étape antérieure. Par l'hypothèse de récurrence on a donc :

$$\Sigma\{d(i,i') \mid i,i' \in q\}/N(q) \leq \delta(q_1, q_2)$$

d'autre part, le principe même de la construction hiérarchique fait que $\delta(q_1, q_2) \leq \delta(q, q' \cup q")$ puisque q_1 et q_2 ont déjà été fusionnées au pas h alors que q ne l'est pas encore avec $q' \cup q"$.

Considérons maintenant le deuxième terme de la décomposition de M. Par l'hypothèse de récurrence on a :

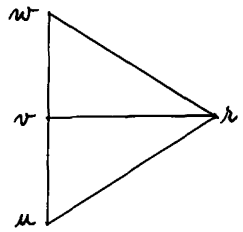
$$\Sigma\{d(i,i') \mid i,i' \in q' \cup q"\}/N(q' \cup q") \leq \delta(q', q")$$

Et $\delta(q',q") \leq \delta(q, q' \cup q")$ car on sait qu'il ne peut y avoir inversion de niveaux dans l'agrégation par la distance moyenne.

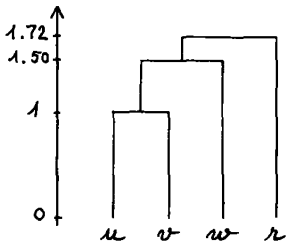
Enfin le troisième terme de la décomposition de M n'est autre que $\delta(q, q' \cup q")$ lui-même comme les coefficients α, β et γ sont tels que $\alpha + \beta + \gamma = 1$ il en résulte que $M \leq \delta(q, q' \cup q")$ ce qui achève la démonstration.

Remarque 1

Il peut arriver que certaines distances intra-classes soient supérieures à une distance interclasse (indice de diamètre) ainsi dans l'exemple suivant : I = {r, u , v, w}



	r	u	v	w
r	0			
u	1.89	0		
v	1.6	1	0	
w	1.89	2	1	0



$d(u,v) = d(v,w)$; on choisit d'agglomérer d'abord u et v ; si l'on prend la classe $q = \{u, v, w\}$, $d(u,w) = 2$ est plus grand que $\delta(q, \{r\}) = 1.79$; cependant la moyenne des distances internes à la classe q n'est que de 1.33 ; mais ce que dit la proposition précédente est plus contraignant : la moyenne des distances internes à $q \cup \{r\} = I$ est inférieure à $\delta(q,\{r\})$. Cette moyenne est en effet égale à 1.56.

2.2 *Borne supérieure de l'indice de niveau* : Pour tout i on pose : $t(i) = \max\{d(i,i') \mid i' \in I - \{i\}\}$; q et q' désignent maintenant les deux dernières classes de la hiérarchie à fusionner au dernier pas de l'algorithme de la distance moyenne : $q \cup q' = I$; $\delta(q,q')$ représente donc la plus grande valeur de l'indice de niveau, nous l'écrirons PGiN en abrégé.

Proposition 2

La moyenne sur l'ensemble I des longueurs maxima t(i), des segments issus d'un même point i , est supérieure ou égale au plus haut niveau de la hiérarchie de la distance moyenne (PGiN)

$$\Sigma\{t(i) \mid i \in I\}/\text{Card } I \geq \delta(q,q')$$

En effet :

$$\delta(q,q') = \Sigma\{d(i,i') \; i \in q, \mid i' \in q'\}/(\text{Card } q \;\; \text{Card } q')$$

En remplaçant les d(i,i') par les t(i) qui leur sont supérieurs :

$$\delta(q,q') \leq \Sigma\{t(i) \mid i \in q, \; i' \in q'\}/\,\text{Card } q \;\; \text{Card } q'$$

Les termes à additionner ne dépendant pas de i' cela s'écrit :

$$\delta(q,q') \leq \text{Card } q' \;\; \Sigma\{t(i) \mid i \in q\}/\,\text{Card } q \;\; \text{Card } q'$$
$$\delta(q,q') \leq \Sigma\{t(i) \mid i \in q\}/\text{Card } q$$

d'une façon analogue on aurait pu obtenir

$$\delta(q,q') \leq \Sigma\{t(i') \mid i' \in q'\}/\text{Card } q'$$

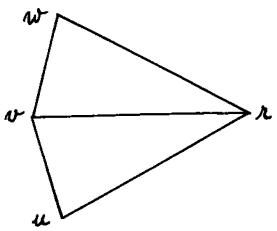
En multipliant la première de ces deux inégalités par Card q , la seconde par Card q' , et en ajoutant membre à membre :

$$(\text{Card } q + \text{Card } q')\;\delta(q,q') \leq \Sigma\{t(i) \mid i \in q \cup q'\}$$

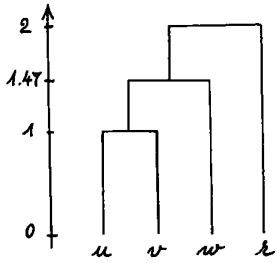
soit : $\delta(q,q') \leq \Sigma\{t(i) \mid i \in I\}/\text{Card } I$

Remarque 2

On peut trouver des exemples où cette borne supérieure est atteinte, comme dans la situation plane suivante où r se trouve au centre d'un cercle de rayon 2 passant par les points u, v, w qui sont entre eux à des distances plus petites que 2.



	r	u	v	w
r	0			
u	2	0		
v	2	1	0	
w	2	1.94	1	0



$t(r) = t(u) = t(v) = t(w) = 2$; la moyenne des $t(i)$ vaut donc aussi 2... comme le plus haut niveau de la hiérarchie.

On appellera désormais h cette borne supérieure du PGiN.

Remarque 3

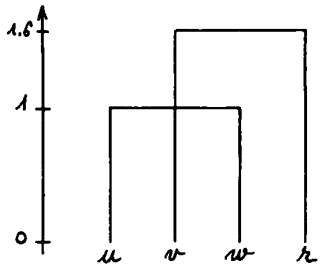
Si l'on considère la hiérarchie obtenue par le saut minimum q et q' étant toujours les deux dernières classes agglomérées on a :

$$\delta(q,q') = \min\{d(i,i') \mid i \in q,\ i' \in q'\}$$

En remplaçant les d(i,i') par les t(i)

$$\delta(q,q') \leq \min\{t(i) \mid i \in q\} = \min_i \max_i, d(i,i')$$

Là encore cette borne supérieure peut être atteinte. Reprenons l'exemple de la remarque 1 ; on prend encore q = {u, v, w}



$\delta(q, \{r\}) = 1.6$

$t(u) = 2$; $t(v) = 1.6$; $t(w) = 2$;
$t(r) = 1.89$

$\min_i t(i) = 1.6$

Enfin dans le cas de la hiérarchie de l'agglomération par le diamètre on a évidemment dans tous les cas :

$$\delta(q,q') = \max\{d(i,i') \mid i \in q,\ i' \in q'\}$$

$$= \max\{d(i,i') \mid i,\ i' \in I\}$$

car à l'intérieur d'une partie q de la hiérarchie toutes les distances sont inférieures à celles qui impliquent des points de q.

En résumé, pour les trois stratégies élémentaires d'agrégation - saut minimum, distance moyenne, diamètre - la borne supérieure des indices de niveau s'obtient en prenant respectivement le minimum, la moyenne et le maximum, des longueurs maxima des segments issus d'un même point.

2.3 *Borne inférieure du plus grand indice de niveau (PGiN)* : Nous ne traiterons ici que de l'agglomération par la distance moyenne. La démonstration de la proposition 2 du § précédent peut se conduire de façon analogue, en inversant le sens de l'inégalité, pour obtenir un minorant de PGiN.

Posons $s(i) = \min\{d(i,i') \mid i' \in I - \{i\}\}$

on obtient : $\delta(q,q') \geq \Sigma\{s(i) \mid i \in I\}/\text{Card } I$

Mais on peut obtenir une borne inférieure du PGiN.

Proposition 3

Le PGiN est supérieur ou égal à la moyenne de toutes les distances deux à deux entre points de I.

q et q' étant comme ci-dessus les deux dernières classes à réunir :

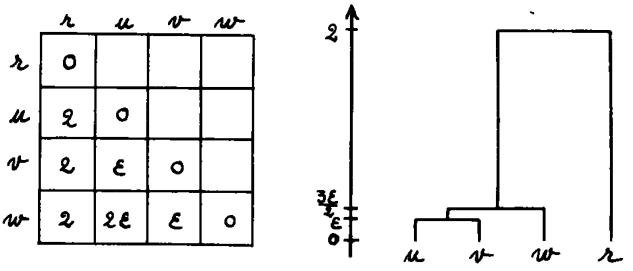
$$\delta(q,q') \geq \Sigma\{d(i,i') \mid i \in I,\ i' \in I\}/(\text{Card } I.(\text{Card } I - 1)/2)$$

c'est une conséquence immédiate de la Proposition 1.

Comme pour la borne supérieure h, la borne inférieure du PGiN, que nous appellerons désormais ℓ peut être atteinte dans des cas très particuliers : 3 points aux sommets d'un triangle équilatéral, ou 4 points aux sommets d'un tétraèdre dont toutes les arêtes sont égales, etc. .

Remarque 4

Ces cas particuliers correspondent à une distribution sphérique donc parfaitement non classifiable. Reprenons alors l'exemple de la remarque 2 ; si l'on y fait $d(u,v) = d(v,w) = \varepsilon$, $d(u,w)$ est de l'ordre de 2ε. Lorsque ε tend vers zéro les 3 points u, v, w tendent à se confondre et l'arbre hiérarchique tend vers un arbre à deux niveaux



où les 3 points u, v, w seraient agrégés ensemble au niveau 0, le PGiN restant égal à 2. On voit bien qu'il s'agit là d'une situation extrême opposée qu'on peut qualifier de parfaitement classifiable.

Quoi qu'il en soit, la diversité des situations classifiables ne se résume pas à ce seul cas aussi n'irons-nous pas plus loin dans cette direction.

3 *Applications*

En vue de l'utilisation de l'algorithme des voisinages réductibles il est tout à fait possible de calculer les bornes du PGiN au fur et à mesure du calcul des distances initiales (ce qui implique de les calculer toutes), qui est la phase préliminaire à la plupart des programmes de classification ascendante hiérarchique. (On notera toutefois que l'agrégation préalable des individus à classer en boules de rayon égal, permet d'éviter de calculer toutes les distances initiales).

Nous nous sommes limités à 4 exemples de données booléennes pour lesquelles on a calculé l'indice de distance de Jaccard qui s'est avéré utile notamment pour les données phytosociologiques. Pour chacun de ces exemples on donne les bornes inférieure et supérieure du PGiN, la vraie valeur du PGiN obtenue dans l'agglomération par la distance moyenne, le nombre et le pourcentage de distances supérieures à la borne supérieure (qui ne seraient pas prises en compte dans l'algorithme des voisinages réductibles). Voici une description succinte de ces exemples.

1) Exemple "NVZD"

L'étude porte sur la Nouvelle-Zélande et quelques îlots avoisinants, soient 11 îles au total. Les données ont été publiées par Tobler, Mielke et Detwyler (cf. bibliographie) sous forme d'une matrice carrée donnant pour chaque paire i, i' d'îles, le nombre d'espèces végétales communes à i et à i' ; la diagonale comprenant le nombre total d'espèces de chacune des îles, cela permet de calculer aisément l'indice de Jaccard.

2) Exemple "CENCOM"

Les données publiées initialement par Berry, dans le livre de Garrison et Marble reprises par J.C. Chevailler puis par nous dans une autre étude, (cf. Roux 1976) sont constituées par 36 centres commerciaux, d'une même agglomération urbaine, pour lesquels on a relevé la présence ou l'absence de 34 types de commerces. La classification porte sur les 36 centres commerciaux.

3) Exemple "ALP.MAR"

Il s'agit de l'exemple phytosociologique déjà étudié dans Benzécri et coll., 1973 (TI C n° 3') dont on a extrait un sous-ensemble aussi représentatif que possible et comportant 28 relevés et 240 espèces végétales (cf. A. Lacoste 1971).

4) Exemple "ALPES"

C'est encore un ensemble de relevés de phytosociologie alpine, figurant également dans Benzécri et coll., 1973 (TI C n°s 2 et 3). Il comporte 55 relevés pour 174 espèces.

Voici les résultats :

	nb. d'objets	Borne inf.	Borne sup.	PGiN	dist.> b. sup.	nb. de dist.	%
NVZD	11	.906	.990	.970	7	55	13%
CENCOM	36	.786	.988	.884	55	630	9%
ALP.MAR	28	.835	.981	.938	18	378	5%
ALPES	55	.846	.984	.921	47	1485	3%

Quelques remarques s'imposent : le PGiN est dans tous les cas relativement proche de sa borne supérieure ; malgré cela le pourcentage de distances éliminées parce que supérieures à cette borne supérieure est faible et ne permet donc pas d'augurer un bon profit pour l'algorithme des voisinages réductibles.

4 *Conclusion*

Par des raisonnements assez simples nous sommes arrivés à donner une "fourchette" pour la valeur du plus grand indice de niveau associé à la hiérarchie de la distance moyenne. Sur des exemples de données booléennes, le calcul effectif de la borne supérieure, comparé à la vraie valeur de cet indice, a donné des résultats satisfaisants .

L'application de ces considérations à l'algorithme des voisinages réductibles est tout à fait possible puisqu'il est très facile de calculer, en même temps que la matrice des distances, les bornes de l'indice de niveau. Cependant si l'on fixe le seuil initial égal à la borne supérieure, bien que cela évite toute réévaluation du seuil en cours de route, le nombre des distances ainsi éliminées semble trop faible pour améliorer substantiellement les temps de calcul. La bonne solution réside certainement dans le choix d'un seuil plus petit qui ne sera réévalué qu'une fois ou deux au cours du calcul. Nos inégalités aideront alors à choisir plus précisément ce seuil.

Références

Benzécri J.P. et coll. : L'analyse des Données. Tome 1, La Taxinomie, DUNOD, Paris, 1973.

Bruynooghe M. : Classification ascendante hiérarchique de grands ensembles de données : un algorithme rapide fondé sur la construction des voisinages réductibles - *Cahiers* de l'Analyse des Données, Vol III, n° 1, pp 7-33, 1978.

Chevailler J.C. : Classification en analyse économique spatiale. Ed. Cujas, Paris 1974.

Garrison W.L. & Marble D.F. : Quantitative geography. North Western University. 1967.

Jambu M. : Classification automatique pour l'analyse des données. 1 - Méthodes et algorithmes. DUNOD, Paris 1978.

Lacoste A. : L'analyse multidimensionnelle en phytosociologie et en écologie. I - L'analyse des données floristiques. *Oecologia Plantarum*, 6, pp 353-369, 1971.

Roux M. : Etude de la dispersion des classes d'une partition. *in* classification automatique et perception par ordinateur. IRIA. B. P. 105, 78150 Le Chesnay, 1976.

Roux M. : Petites données de géographie botanique analysées dans un but didactique. Note multigraphiée. Labo. Stat. Univ. Paris 6. 5 p Paris, 1976.

Tobler W.R., Mielke H.W and Detwyler T.R. : Geobotanical distance between New-Zealand and neighbouring islands. Bioscience, Vol. 20, n° 9, 1970.