841

Les Cahiers de l'Analyse des Données
Vol. V - 1980 - n° 2 - p. 135-144

LA CLASSIFICATION HIÉRARCHIQUE ASCENDANTE SELON LA MÉTHODE DES VOISINS RÉCIPROQUES

[CAH. VOIS. RECIP.]

par C. de Rham (1)

1 *Introduction* : En classification hiérarchique ascendante, le procédé consiste à grouper les observations individuelles en classes par agrégation successive jusqu'à ce que toutes les observations fassent partie de la même classe. Les méthodes se distinguent par le choix de la distance entre les observations et la définition de la stratégie d'agrégation.

Dans l'algorithme de base, le calcul des distances (il s'agit plus exactement d'une quantité critère que l'on appelle distance par abus de langage) entre classes se fait par récurrence à partir de la matrice des distances entre observations.

Dès que l'utilisateur de l'algorithme de base veut traiter des ensembles de plus de quelques centaines d'observations, il se trouve confronté à un problème de place nécessaire en mémoire centrale et à un problème de temps de calcul.

Nous montrons dans cet article que ces deux problèmes peuvent être résolus de manière élégante par la méthode des voisins réciproques . Cette méthode permet d'édifier une hiérarchie totale exacte suivant les principaux critères ne comportant pas d'inversions (agrégation suivant le saut minimum, le diamètre, la distance moyenne et la variance).

On trouve les bases théoriques concernant les voisins réciproques chez Bock et quelques caractéristiques sont également mentionnées chez Bruynooghe (cf bibliographie). On ne reprendra ici que les notions nécessaires à l'élaboration des programmes.

La méthode des voisins réciproques comprend deux phases distinctes. La première consiste à chercher tous les couples d'éléments tels que chaque élément du couple soit le plus proche élément de l'autre élément. Ces couples sont appelés voisins réciproques. La deuxième phase agglomère tous les couples de voisins réciproques et forme une nouvelle classe pour chacun des couples. Les deux phases sont exécutées alternativement jusqu'à ce que tous les objets soient réunis en une seule classe.

Dans ce travail, nous développerons deux réalisations de la méthode des voisins réciproques. La première, basée sur la matrice des distances est très proche de l'algorithme de base. Il y a toutefois une différence importante : comme l'algorithme forme plusieurs classes par itération, il faudra étendre la formule de récurrence au cas suivant :
$d(i \cup j, k \cup m) = f(d(i,j),d(i,k),d(i,m),d(j,k),d(j,m),d(k,m),ni,nj,nk,nm)$

La seconde méthode est basée sur la matrice des données. Cette application est la plus intéressante pour le praticien . Car , avec les

(1) Diplômé de l'école polytechnique de Zürich, MBA INSEAD de Fontainebleau, Dr ès s.ces. Cet article est extrait de la thèse de doctorat de l'auteur présentée à l'univ. de Neuchâtel, Suisse. Directeur de thèse : Professeur A. Strohmeier.

mêmes moyens en place et en temps nécessaires à l'algorithme de base pour quelques centaines d'observations, celui-ci peut en traiter quelques milliers.

Comme les calculs de distance se font à partir de la matrice des données, cette seconde méthode n'est applicable efficacement qu'aux stratégies de centre de gravité et de variance. Ce petit inconvénient est largement compensé par le fait que l'on peut éviter de mettre en mémoire la matrice des distances.

2 *Définitions*

2.1 *La formule de récurrence* : La distance entre une nouvelle classe $i \cup j$ et une classe quelconque k se calcule par une formule de récurrence de la forme (1) :

$$d(i \cup j,k) = \alpha_i \; d(i,k) + \alpha_j \; d(j,k) + \beta \; d(i,j) + \gamma |d(i,k) - d(j,k)|$$

Le tableau suivant donne les coefficients α_i , α_j , β et γ pour les 7 stratégies d'agrégation.

N°	Nom	α_i	α_j	β	γ	monotone	conservation de l'espace
1	saut minimum (2)	1/2	1/2	0	-1/2	oui	contractant
2	diamètre (3)	1/2	1/2	0	+1/2	oui	dilatant
3	dist. moy. non pond .	$\frac{ni}{ni+nj}$	$\frac{nj}{ni+nj}$	0	0	oui	conservant
4	dist. moy. pondérée	1/2	1/2	0	0	oui	conservant
5	c. de gr . non pond .	$\frac{ni}{ni+nj}$	$\frac{nj}{ni+nj}$	$-\alpha_i\alpha_j$	0	non	conservant
6	c. de gr . pondéré	1/2	1/2	-1/4	0	non	conservant
7	variance inter-cl. N=ni+nj+nk	$\frac{ni+nk}{N}$	$\frac{nj+nk}{N}$	$\frac{-nk}{N}$	0	oui	dilatant

ni, nj, nk : poids des classes i, j et k

On trouve le calcul détaillé des coefficients pour chaque stratégie chez Jambu.

2.2 *L'algorithme de base de la classification hiérarchique ascendante*:

Soit l'ensemble I des observations

pas a) calculer les distances d(i,j) pour toute paire (i,j) d'éléments de I

pas b) trouver i et j $\in$ I tels que pour tout i', j' $\in$ I : $d(i,j) \leq d(i',j')$.

pas c) pour le couple (i,j)

-former une nouvelle classe c = i $\cup$ j

-faire I:=I - {i}-{j} et C ={c}

---

(1) cf Lance et Williams, 1973 ; in l'Analyse des Données, TI B n° 4 ; C.A.H.; 1973 ; ou Jambu 1974 ; des commentaires importants sont dans Bruynooghe [CLASS. RAP.] *§ 2.2.*

(2) en anglais : single linkage.

(3) en anglais : multiple linkage.

pas d) si |I| = 0 : STOP

pas e) calcul des distances entre les classes existantes

i ∈ I et la nouvelle classe c à l'aide de la formule de récurrence choisie : (d(i,c) pour tout i ∈ I

pas f) faire I := I ∪ C, aller à b)

2.3 *Les voisins réciproques* : Soit I une partition de l'ensemble I_0 des objets en classes disjointes, $I \in P(I_0)$. Soit i et j deux classes de I, i, j ∈ I.

On note : ∀ i ∈ I. : dd(i) = inf{d(i,j) | j ∈ I, j ≠ i} ;

et v(i) = {j | j ∈ I, j ≠ i, d(i,j) = dd(i)}.

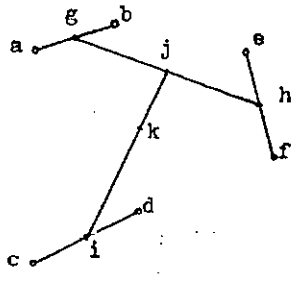
Par définition, i et j sont deux classes voisines réciproques si on a :

i ∈ v(j) et j ∈ v(i)

On notera M l'ensemble des paires voisines réciproques de la partition I

M = {(i,j) | i, j ∈ I, i ≠ j, i ∈ v(i), j ∈ v(i)}
= {(i,j) | i, j ∈ I, i ≠ j, dd(i) = dd(j)} .

2.4 *Exemple numérique pour l'algorithme de base* : Soit un ensemble I = {a,b,c,d,e,f} de 6 observations représentées par des points dans un espace à deux dimensions avec la condition ∀ i ∈ I : ni = 1.



	coordonnées	
a	0	40
b	15	45
c	0	0
d	20	10
e	40	40
f	45	20

pas a) calcul des (6-1)(6/2) = 15 distances d(i,j) pour la stratégie n° 7 (critère de variance) :

$$d^2(i,j) = ((ni.nj)/(ni+nj)) . \|\bar{X}i - \bar{X}j\|^2$$

i \ j	a	b	c	d	e	f
a	0	125	800	650	800	1213
b	125	0	1125	625	325	763
c	800	1125	0	250	1600	1213
d	650	625	250	0	650	363
e	800	325	1600	650	0	213
f	1213	763	1213	363	213	0

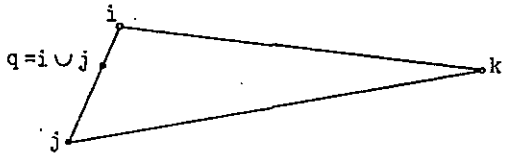
Itérations

pas: itération	b) couple d_{min}	c) nouvelle classe	d) |I|	e) calcul des distances
1	(a,b)	g = a ∪ b	4	d(g,c) = 1242 d(g,d) = 808 d(g,e) = 708 d(g,f) = 1275
2	(e,f)	h = e ∪ f	3	d(h,g) = 1381 d(h,c) = 1804 d(h,d) = 604
3	(c,d)	i = c ∪ d	2	d(i,g) = 1413 d(i,h) = 1681
4	(h,g)	j = h ∪ g	1	d(i,j) = 1602
5	(i,j)	k = i ∪ j	0 STOP	

3 *L'application de l'algorithme selon les voisins réciproques avec la matrice des distances*

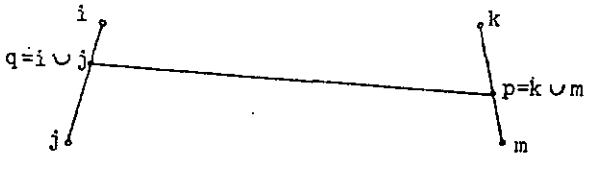
3.1 *Extension de la formule de récurrence pour la calcul de la distance entre deux classes nouvellement formées* :

La formule de récurrence vaut pour le cas suivant :



$$d(q,k) = d(i\cup j,k) = f(d(i,j),\ d(i,k),\ d(j,k),\ ni,\ nj,\ nk)$$

L'algorithme selon la méthode des voisins réciproques forme plusieurs classes par itération. Il faut donc étendre la formule de récurrence au cas nouveau :



$$d(q,p) = d(i\cup j,\ k\cup m) = f(d(i,j),\ d(i,k),\ d(i,m),\ d(k,m),\ d(j,k),\ d(j,m),\ ni,\ nj,\ nk,\ nm)$$

Afin de simplifier les calculs, on traitera séparément les stratégies pour lesquelles $\gamma \neq 0$ ou $\gamma = 0$.

Agrégation suivant le saut minimum et le diamètre ($\gamma \neq 0$)

saut minimum :

$$d(i \cup j,k) = \inf(d(i,k),\ d(j,k))$$
$$d(i \cup j,m) = \inf(d(i,m),\ d(j,m))$$
$$d(i \cup j,k \cup m) = \inf(d(i \cup j,k), d(i \cup j,\ m))$$
$$\Rightarrow d(i \cup j,k \cup m) = \inf(d(i,k),\ d(j,k),\ d(i,m),\ d(j,m))$$

diamètre :

$$d(i \cup j,k) = \sup(d(i,k),\ d(j,k))$$
$$d(i \cup j,m) = \sup(d(i,m),\ d(j,m))$$
$$d(i \cup j,k \cup m) = \sup(d(i \cup j,k),\ d(i \cup j,m))$$
$$\Rightarrow d(i \cup j,k \cup m) = \sup(d(i,k),\ d(j,k),\ d(i,m),\ d(j,m))$$

Ces deux expressions très simples seront utilisées pour l'application informatique.

Moyenne, centre de gravité et variance ($\gamma = 0$)

Il faut noter que la stratégie du centre de gravité est incluse ici tout en rappelant que n'étant pas monotone, elle ne peut donner que des solutions heuristiques. Selon la formule de récurrence :

$$d(i \cup j,k) = \alpha 1.d(i,k) + \alpha 2.d(j,k) + \beta 1.d(i,j)$$
$$d(i \cup j,m) = \alpha 3.d(i,m) + \alpha 4.d(j,m) + \beta 2.d(i,j)$$
$$d(i \cup j,k \cup m) = \alpha 5.d(i \cup j,k) + \alpha 6.d(i \cup j,m) + \beta 3.d(k,m)$$
$$d(i \cup j,k \cup m) = \alpha 5.(\alpha 1.d(i,k) + \alpha 2.d(j,k) + \beta 1.d(i,j))$$
$$+ \alpha 6.(\alpha 3.d(i,m) + \alpha 4.d(j,m) + \beta 2.d(i,j))$$
$$+ \beta 3.d(k,m)$$
$$d(i \cup j,k \cup m) = d(i,k).\alpha 1.\alpha 5 + d(j,k).\alpha 2.\alpha 5 + d(i,m).\alpha 3\alpha 6$$
$$+ d(j,m).\alpha 4.\alpha 6 + d(i,j).(\beta 1.\alpha 5 + \beta 2.\alpha 6) + d(k,m).\beta 3$$

avec
$$\alpha ik = \alpha 1.\alpha 5$$
$$\alpha jk = \alpha 2.\alpha 5$$
$$\alpha im = \alpha 3.\alpha 6$$
$$\alpha jm = \alpha 4.\alpha 6$$
$$\beta ij = \beta 1.\alpha 5 + \beta 2.\alpha 6$$
$$\beta km = \beta 3$$

on peut écrire l'expression générale :

$$d(i \cup j,k \cup m) = \alpha ik.d(i,k) + \alpha jk.d(j,k) + \alpha im.d(i,m) + \alpha jm.d(j,m)$$
$$+ \beta ij.d(i,j) + \beta km.d(k,m)$$

Le tableau de la page suivante donne les valeurs de α et β pour les différentes stratégies.

Tableau des valeurs de αik, αjk, αim, αjm, βij, βkm, pour les stratégies 3 à 7

N°	Nom	αik	αjk	αim	αjm	βij	βkm
3	dist. moy. non pond.	ni.nk / nij.nkm	nj.nk / nij.nkm	ni.nm / nij.nkm	nj.nm / nij.nkm	0	0
4	dist. moy. pondérée	.25	.25	.25	.25	0	0
5	c. de gr. non pond.	ni.nk / nij.nkm	nj.nk / nij.nkm	ni.nm / nij.nkm	nj.nm / nij.nkm	$-\frac{ni.nj}{(ni+nj)^2}$	$-\frac{nk.nm}{(nk+nm)^2}$
6	c. de gr. pondéré	.25	.25	.25	.25	-.25	-.25
7	variance	$\frac{ni+nk}{nijkm}$	$\frac{nj+nk}{nijkm}$	$\frac{ni+nm}{nijkm}$	$\frac{nj+nm}{nijkm}$	$-\frac{(nk+nm)}{nijkm}$	$-\frac{(ni+nj)}{nijkm}$

nij = ni + nj, nkm = nk + nm ; nijkm = ni + nj + nk + nm

3.2 *L'algorithme* : Soit l'ensemble I des observations

pas a) calcul des distances d(i,j) pour tout i, j ∈ I, i ≠ j et définition de v(i) pour tout i ∈ I

pas b) trouver l'ensemble M des couples de voisins réciproques

M = {(i,j) i, j ∈ I, i ≠ j, i ∈ v(j), j ∈ v(i)}

définir C := ∅

pas c) faire pour chaque couple (i,j) ∈ M

-former une nouvelle classe c = i ∪ j

-faire I := I -{i}-{j} et C := C + {c}

pas d) si |I| + |C| = 1 : STOP

pas e) calcul des distances

-entre les classes existantes i ∈ I et les nouvelles classes c ∈C à l'aide de la formule de récurrence :

d(i,c) pour tout i dans I et pour tout c dans C

- entre les nouvelles classes c ∈ C à l'aide de l'extension de la formule de récurrence (pour tout c, c' dans C)

pas f) faire I := I∪C, aller à b)

3.3 *Exemple numérique pour l'algorithme selon les voisins réciproques avec matrice des distances* : Soit le même ensemble de 6 observations I_0 = {a,b,c,d,e,f} que sous 2.4.

pas a) calcul des 15 distances entre observations

Exemple : calcul détaillé de d(g,h) selon l'extension de la formule de récurrence.

$$d(g,h) = \alpha ae.d(a,e) + \alpha af.d(a,f) + \alpha be.d(b,e) + \alpha bf.d(b,f) + \beta ab.d(a,b) + \beta ef.d(e,f)$$

$$d(g,h) = 1/2.\ 800 + 1/2.\ 1213 + 1.2.\ 325 + 1/2.\ 763 - 1/2.\ 125 - 1/2.\ 213 = 1381.$$

L'algorithme de base avait fait 5 itérations et calculé 10 distances. Celui-ci a fait 3 itérations et calculé 4 distances.

pas : itération	b) ensemble M	c) nouvelle classe	d) |I| |C|	e) calcul des distances
1	(a,b) (e,f) (c,d)	g = a∪b h = e∪f i = c∪d	0 3	d(g,h)=1381 d(g,i)=1413 d(h,i)=1681
2	(g,h)	j = g∪h	1 1	d(i,j)=1602
3	(i,j)	k = i∪j	0 1 STOP	

4 *L'application de l'algorithme selon les voisins réciproques avec la matrice des données*

4.1 *Introduction* : cette application a l'avantage de ne pas dépendre de la matrice des distances. Par contre on ne peut l'appliquer efficacement qu'aux stratégies basées sur le centre de gravité des classes (centre de gravité pondéré et non pondéré, variance interclasses).

La place nécessaire à l'algorithme de base est de

$(N - 1)(N/2) + 2N$ mots

La solution proposée ici a besoin de

$N * (K + 5)$ mots

(N = nombre d'observations, K = nombre de variables)

L'algorithme basé sur la matrice des données est plus économique dès que

$N > 2K + 7$

ce qui est presque toujours le cas dans la pratique.

4.2 *L'algorithme* : En ce qui concerne l'algorithme proprement dit, il ne diffère du précédent que dans la mesure où les distances ne sont plus calculées sur la base de la matrice des distances entre observations mais sur la base des centres de gravité des classes.

Soit l'ensemble I des observations

pas a) calcul des distances d(i,j) pour toute paire (i,j) d'éléments de I et définition de v(i) et dd(i) pour tout $i \in I$.

pas b) trouver l'ensemble M des couples de voisins réciproques

$M = \{(i,j) \mid i, j \in I, i \neq j, i \in v(j), j \in v(i)\}$

définir $C := \emptyset$

pas c) pour chaque couple $(i,j) \in M$

- -former une nouvelle classe $c = i \cup j$
- -calculer le centre de gravité de c
- -faire $I := I - \{i\} - \{j\}$ et $C := C + \{c\}$

pas d) si $|I| + |C| = 1$: STOP

pas e) calcul des distances sur la base des centres de gravité

-entre les classes existantes $i \in I$ et les nouvelles classes $c \in C$: $d(i,c)$ pour tout $i \in I$ et pour tout $c \in C$

-entre les nouvelles classes $c \in C$:

$d(c,c')$ pour tout c, $c' \in C$, $c < c'$

-définition de $v(i)$ et $dd(i)$ pour tout $i \in I$ et pour tout $i \in C$

pas f) Faire $I := I \cup C$, aller à b)

4.3 *Exemple numérique pour l'algorithme selon les voisins réciproques avec matrice des données* : Soit le même ensemble de 6 observations $I = \{a,b,c,d,e,f\}$ que sous 2.4

pas a) calcul des 15 distances entre observations et définition de $v(i)$ et $dd(i)$ $\forall\, i \in I$:

i:	a	b	c	d	e	f
v(i):	b	a	d	c	f	e
dd(i):	125	125	250	250	213	213

pas: itération	b) ensemble M	c) nouvelle classe	d) |I| |C|	e) calcul des distances	i	v(i)	dd(i)
1	(a,b) (e,f) (c,d)	g = a ∪ b h = e ∪ f i = c ∪ d	0 3	d(g,h)=1381 d(g,i)=1413 d(h,i)=1681	g h i	h g g	1381 1381 1413
2	(g,h)	j = g ∪ h	1 1	d(i,j)=1602	i j	j i	1602 1602
3	(i,j)	k = i ∪ j	0 1 STOP				

5 *Comparaison des performances* : Les performances ont été comparées sur trois ensembles de données construits à l'aide d'un programme de génération de hiérarchies artificielles.

Les comparaisons concernent le nombre d'itérations, le nombre de calculs de distances, la place en mémoire centrale et le temps d'exécution.

L'algorithme de base (BASE) sert de référence aux deux versions de l'algorithme selon les voisins réciproques. DIST est la version avec matrice des distances, DONN celle avec matrice des données.

Les mesures ont été faites sur le CDC 6500 de Fides à Zurich. La place est mesurée en mots, le temps en ms. Le temps (nécessaire au calcul de la hiérarchie) est une moyenne des stratégies 1 à 7 pour BASE et DIST et des stratégies 5 à 7 pour DONN.

Remarques à propos du tableau comparatif des performances . Le nombre d'itérations de DIST et DDNN est très inférieur à celui de BASE. Ceci aura des répercussions très favorables sur le temps de calcul, car la recherche des voisins réciproques et la mise à jour de la matrice des distances ne sont exécutées qu'une fois par itération.

Pour *le nombre de calculs* de distance, les deux algorithmes proposés sont nettement meilleurs que BASE.

Tableau comparatif de performances des trois algorithmes

grandeur comparée	algor.	dimension de l'ensemble (obs x var)					
		(32 x 2)		(64 x 2)		(128 x 2)	
		abs	%	abs	%	abs	%
nombre d'itera-tions	BASE	31	100	63	100	127	100
	DIST	8	26	9	14	10	8
	DONN	8	26	9	14	10	8
nombre de calc. de dist.	BASE	465	100	1953	100	8001	100
	DIST	222	48	718	37	2734	34
	DONN	275	59	771	39	2878	35
place en mémoire centrale	BASE	1088	100	4224	100	16640	100
	DIST	1152	106	4352	103	16896	102
	DONN	224	21	448	11	896	5
temps d'exécu-tion	BASE	201	100	1197	100	8191	100
	DIST	97	48	294	25	987	12
	DONN	74	37	171	14	484	6

La place en mémoire centrale est à l'avantage de DONN, car pour cet algorithme, elle croît comme I au lieu de I^2 pour BASE et DIST.

Les deux algorithmes proposés obtiennent d'excellents résultats pour *le temps calcul*. L'implantation de DIST est intéressante lorsque l'on veut appliquer les stratégies du saut minimum, du diamètre et de la distance moyenne. Si cela n'est pas le cas, DONN est alors préférable.

6 Conclusions : En classification ascendante, même les méthodes d'agrégation fondées sur les critères les plus connus et mentionnés dans toutes les références bibliographiques peuvent être améliorées sensiblement grâce à de nouveaux algorithmes.

Ainsi ces méthodes sont désormais applicables à des ensembles environ dix à vingt fois plus importants que ceux traités jusqu'ici. Ce qui permet d'élargir considérablement le champ d'application de ces méthodes et nous nous en réjouissons.

Nous pensons que dans cette voie le présent travail représente un progrés notable, non seulement sur l'algorithme de base (qui requiert $O(n^3)$ opérations pour traiter n individus) ; mais aussi sur des algorithmes plus récents (cf Bruynooghe [CLASS. RAP.] ; et sa mise en oeuvre par Jambu & Lebeaux : CAH2) dans la mesure où ceux-ci requièrent le choix souvent délicat d'un seuil ; toutefois des comparaisons précises restent à faire.

De plus apparaissent désormais possibles des programmes de CAH ne requérant pas le calcul de toutes les distances entre paires (ce que fait $O(n^2)$) ; mais fondées sur l'organisation de l'ensemble I en boules

(cf [BOULES OPTIMISEES] ; § 3.4 *in Cahiers* Vol IV n° 3 pp 375 sqq ; cf

Références

Bock H.H. : *Automatische Klassifikation Vandenhoeck & Rupprecht, Göttingen,* 1974.

Bruynooghe M. : Méthodes nouvelles en classification de données taxinomiques nombreuses. Statistique et analyse des données, 3/1978[CLASS. RAP.]: Cahiers de l'A. des D. : Vol III n° 1 pp 7 sqq (1978).

Jambu M. : Sur les indices de distances en vue de la construction d'une classification hiérarchique. Consommation 2/1974.

Jambu & Lebeaux : Classification automatique (2 vol.) Dunod 1979.

Lance & Williams : *Hierarchical Classificatory Methods Statistical methods for Digital Computers,* Wiley 1973.