Les Cahiers de l'Analyse des Données
Vol. XXII - 1997 - n°2 - pp.191-198.

RAPPEL: CONSTRUCTION D'UNE CLASSIFICATION ASCENDANTE HIÉRARCHIQUE PAR LA RECHERCHE EN CHAÎNE DES VOISINS RÉCIPROQUES

[RAPPEL CAH CHAÎNE RÉCIP.]

J.-P. BENZÉCRI

La méthode de construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques a été d'abord exposée dans CAD, *Vol. VII, n°2, (1982). Puis une version élaborée de l'exposé a paru dans la 4-ème édition du Tome I du Traité de l'* Analyse des Données. *Présentement, ces exposés sont difficilement accessibles; cependant qu'à notre connaissance, la méthode n'a pas été publiée ailleurs. Nous publions donc le présent rappel, à l'intention des lecteurs qui étudieront, d'après [SOURCES PASCAL IIB], le programme de CAH en langage Pascal, fondé sur cet algorithme.*

1 Principe de la recherche en chaîne

L'algorithme de base de la classification hiérarchique ("base") constrit les nœuds un par un, en cherchant, sur l'ensemble des sommets, la paire (sa, sb) réalisant le minimum de la quantité d(sa, sb), prise pour critère. Avec cet algorithme, pour un ensemble I de cardI = n individus, la classification ne s'achève que par le calcul (ou, au moins, la consultation) de $n^3/6$ distances.

L'algorithme des voisinages réductibles ("réduc"), dû à M. BRUYNOOGHE, et utilisé, notamment par M. JAMBU, vise à diminuer ce nombre, en bornant les comparaisons de distances à une liste définie par un seuil supérieur, et convenablement tenue à jour à chaque création de de nœud: "réduc" accélère en effet généralement la construction d'une CAH; mais l'entretien de la liste est fatidieux, et le choix du seuil aléatoire. En fait, ce choix se renouvelle, car, en bref, chaque fois que la liste se trouve vide, on doit élever le seuil. De la sorte, on ne peut fixer une échelle générale de temps – e.g.: en n^2, ou: $n^2.\text{Log } n$ – au lieu de n^3 pour "base".

Les cahiers de l'analyse des données 0339-3097/97/02/$ 7.00/© Gauthier-Villars

L'algorithme des voisins réciproques ("récip"), dont le principe remonte à Mac QUITTY (1966), et dont des réalisations furent programmées par BRICE et GRANJOUAN (1977), et de RHAM (1980), introduit, sans compliquer l'algorithme de base, un principe d'accélération très efficace: au lieu de construire des nœuds un par un, on repère, à chaque parcours de la matrice des distances, un esnsemble de paires de sommets (sa, sb), (sa', sb'), (sa", sb"),... qu'on peut agréger parce que, même s'il ne s'agit pas de la paire (sa, sb) de réalisant le minimum du critère d(sa, sb), chaque membre d'une paire est, relativement à l'autre, le sommet qui réalise ce minimum: i.e., quel que soit s, autre que sa" et sb", le critère d(sa", sb") est inférieur, à la fois, à d(s, sa") et à d(s, sb").

Utilisée comme un principe d'agrégation simultanée, la notion de voisin réciproque est déjà visiblement efficace, des calculs sur un modèle de densité uniforme montrant qu'en moyenne la moitié (voire le tiers) des individus d'un ensemble forment des paires de voisins réciproques; en sorte qu'en un parcours de la matrice des distances on décide simultanément d'un si grand nombre d'agrégations de paires que le nombre, CardSom, des sommets en cours de traitement se trouve réduit à (3/4).CardSom (voire (5/6).CardSom).

Il existe cependant un modèle d'ensemble de points en chaîne sur un axe pour lequel, à chaque lecture du tableau des distances, il n'apparaît qu'une seule paire de voisins réciproques: il suffit de poser:

$$I = \{3^{-i} \mid i = 1, 2, \ldots, n\} \quad .$$

On voit, en bref, que chaque i a pour voisin le plus proche: i-1; et est le plus proche voisin de (i-1); en sorte que seule la paire ultime, (n, n-1), satisfait au critère de réciprocité (cf. [CHAÎNE COMP. CAH]; in *CAD*, Vol. VII, pp.189-208; 1982).

Mais, surtout, tant que subsiste le principe de la recherche et de la création simultanée de plusieurs nœuds, il est difficile de se former un modèle qui soit assez précis pour permettre de calculer des bornes de durée valant uniformément, quels que soient les aléas des données.

En réalité, l'efficacité du principe des voisins réciproques n'est pas liée à la création simultanée de plusieurs nœuds, mais plus exactement à ce que les paires de voisins réciproques peuvent être agrégées au fur et à mesure qu'on les découvre, sans s'astreindre à créer les nœuds dans l'ordre croissant de leurs niveaux. La seule contrainte est que, conformément au principe ascendant de la CAH, un nœud n ne peut être créé qu'après ses deux dépendants, a(n) et b(n). Dès lors, le problème se pose de concevoir au mieux la recherche des paires de voisins réciproques.

On propose ici une recherche en chaîne qui, d'une part, ainsi que l'a suggéré C. de RHAM, évite de rechercher le plus proche voisin, v(s), d'un sommet s, si un v(s) déterminé antérieurement subsiste; et, d'autre part, n'opère jamais que sur une seule chaîne: {s, v(s), v(v(s)), v(v(v(v(s))), ...}, laquelle aboutit nécessairement à une paire de voisins réciproques qu'on agrège; d'où un racourcissement de la chaîne, que l'on préserve pourtant (si sa longueur dépassait 2) pour chercher une paire de voisins réciproques dans le nouvel état de l'ensemble des sommets.

Ainsi, il apparaît que le coût de la création d'un nœud est essentiellement celui de la création de 3 maillons de chaîne, parfois seulement de 2. Or, le coût de la création d'un maillon, i.e. de la recherche du n-ème plus proche voisin, est de traiter autant de distances qu'il reste de sommets (plus précisément; de sommets, en dehors de la chaîne): on voit que ce coût est borné par k.n; où n=CardI, et k dépend de la complexité des calculs de distance; mais non de CardI. En sorte que le coût global de la création de la hiérarchie toute entière admet une borne de l'ordre de n^2, valant quels que soient les aléas du parcours; et, en particulier, pour le modèle de [CHAÎNE COMP. CAH], cité plus haut comme contrexemple à l'efficacité de "récip".

Toutes les notions en jeu ici sont simples: la difficulté principale a été de les découvrir puis de les conjuguer. Un spécialiste de la Classification Ascendante Hiérarchique n'a sans doute aucune peine à concevoir un algorithme d'après les considérations qui précèdent. Pourtant, entraîné de développement en développement assez loin de l'algorithme de base et de ses fondements théoriques désormais classiques, nous croyons utile de donner un exposé théorique qui ne laisse aucun doute quant à la validité des constructions nouvelles (si du moins le critère satisfait à l'axiome de la médiane, énoncé ci-après); validité d'ailleurs impliquée par le succès des traitements sur machine.

Des publications antérieures, donnent un algorithme général: ici, nous nous bornerons à renvoyer à [SOURCES PASCAL IIB]; où figure un programme utilisant exclusivement le critère de la variance (ou du moment d'inertie); et traitant des données soumises préalablement à l'analyse de corrspondance.

2 Construction d'une CAH par une suite croissant d'arbres compatibles compatibles avec son critère

2.1 L'axiome de la médiane

Tout repose sur le caractère local de la notion de voisin réciproque, i.e., sur le fait que cette relation n'est pas détruite par une agrégation effectuée ailleurs. Le caractère local résulte de l'axiome suivant. Si on suppose définie

une notion de "distance" ou niveau d'agrégation, d(a, b) entre parties finies, l'axiome relatif à 3 parties, {n, b, s}, deux à deux d'intersection vide, s'énonce:

$$d(a, b) \leq \inf\{d(a, s), d(b, s)\}$$
$$\Rightarrow \quad \inf\{d(a, s), d(b, s)\} \leq d(a \cup b, s) \ .$$

On peut dire que {a, b, s} sont les trois sommets d'un triangle dont (a, b) est le plus petit côté; que (a ∪ b) sert de milieu à (a, b); et que (a ∪ b, s) est une médiane. L'axiome s'énonce alors: la médiane opposée au plus petit côté est supérieure en longueur au plus court des deux autres côtés; énoncé manifestement faux en géométrie euclidienne, mais qui suggère un nom pour l'axiome.

L'axiome, dont l'mportance apparaît chez M. BRUYNOOGHE (cf. *CAD*, Vol. III, pp.9-10; 1978), se vérifie pour les principaux critères d'agrégation, dont on rappelle ici les formules (cf. [ALG. & ALG. CAH]):

critère du saut : $d(a, b) = \inf\{d(i, j) \mid i \in a, j \in b\}$;

c. du diamètre : $d(a, b) = \sup\{d(i, j) \mid i \in a, j \in b\}$;

distance moyenne : $d(a, b) = \sum\{d(i, j).\ mi.mj\ /\ (ma.mn) \mid i \in a, j \in b\}$,

où on a noté mi, mj les masses des éléments i et j ; et ma, mb les masses totales des parties a et b ;

moment d'inertie : $d(a, b) = ((ma.mb)\ /\ (ma+mb))\ .\ d(ga, gb)^2$,

où on a noté ga, gb les centres de gravité des parties a et b; [ce critère, encore appelé: *critère d'agrégation suivant la variance*, est celui quasi exclusivement utilisé par nous; et il est le seul calculé par le programme CLH; dont le listage est expliqué dans [SOURCES PASCAL IIB]; cf. §2.1.2].

L'axiome vaut aussi pour le critère proposé par D. DOMENGÈS (Mme. Chr. MULLON) dans l'article [CAH FLUX], in *CAD*, Vol. VII, n°2, pp.169-172; 1982.

2.2 Caractère local de la relation de plus proches voisins réciproques

Soit un ensemble de parties de l'ensemble initial I; pouvant être l'ensemble des sommets d'un arbre non connexe A qui est une étape de la construction de la CAH sur I. Notons v(s) un plus proche voisin, dans S, d'un sommet s de S: i.e. :

$$\forall\ s' \in S - \{s\} \ :\ d(s,\ s') \geq d(s, v(s)) \ ;$$

alors, v(s) reste un plus proche voisin de s si on modifie S par l'agrégation d'une paire (a, b) de plus proches voisins réciproques. De façon précise:

Soit : a, b, s, v $\in$ S, quatres sommets distincts ;

$v(a) = b$; $v(b) = a$; $v(s) = v$, i.e. :

$\forall\, s' \in S$: $d(s, v(s)) \leq d(s, s')$;

$S' = (S - \{a, b\}) \cup (a \cup b)$ (i.e.: on supprime a et b; on crée $a \cup b$);

alors : $\forall\, s' \in S'$: $d(s', v(s)) \leq d(s, s')$.

La vérification est immédiate: elle porte seulement sur $s' = a \cup b$; et s'identifie à l'axiome de la médiane pour le triangle {s, a, b}.

Si, en particulier, s et v sont voisins réciproques, (i.e. : v=v(s), et: s=v(v)), ils le restent après agrégation de a avec b.

2.3 Arbre compatible avec un critère

Par l'algorithme de base, et aussi avec "réduc" et avec "récip" sous sa forme la plus simple, on construit des états successifs de l'arbre, lesquels, relativement à la hiérarchie binaire complète (au sommet de laquelle est l'ensemble I lui-même, considéré comme le nœud supérieur) peuvent être définis par la suppression des nœuds dont le niveau dépasse un seuil x. Présentement, on désire jouir d'une liberté plus grande dans l'ordre de création des nœuds. Nous définissons donc la notion d'arbre binaire sur I compatible avec le critère d ; la construction s'effectuera par une suite croissante d'arbres compatibles aboutissant à un arbre connexe (i.e. à un seul sommet) qui réalise la CAH cherchée.

Définition : Soit A un arbre binaire sur I admettant comme ensemble de terminaux toutes les parties à un élément: i.e. $Ter(A) = \{\{i\} \mid i \in I\}$; on définit, par récurrence, la notion d'arbre compatible avec d par les deux conditions suivantes:

1°) $A = \{\{i\} \mid i \in I\}$ est un arbre compatible (cet arbre constitue l'état initial de la CAH)

2°) A est compatible s'il existe un sommet s de A, $s \in Som(A)$, tel que:

2°a) $A - \{s\}$ est lui-même un arbre compatible ;

2°b) dans A, le sommet *s* a pour successeurs immédiats deux parties a(s) et b(s) qui sont plus proches voisins réciproques au sein de $Som(A - \{s\})$. [on peut donc considérer A comme provenat d'un état antérieur $(A-\{s\})$ par agrégation de deux sommets qui sont plus proches voisins réciproques).

L'intérêt de cette définition est que tout arbre compatible A peut avoir été obtenu en créant les nœuds par ordre de niv eau croissant, comme le fait

l'algorithme de base. C'est ce que montre la proposition suivante.

Proposition : Soit A compatible; soit s un sommet de A de niveau maximum [i.e. s est un nœud; et, parmi les sommets, s', qui sont des nœuds, s réalise le maximum de d(a(s'), b(s'))] ; alors, A-{s} est compatible et {a(s), b(s)} sont plus proches voisins réciproques au sein de Som(A-{s}).

On démontre la proposition par récurrence sur le nombre N de nœuds de A. Si N=1, la proposition résulte immédiatement de la définition même: l'unique sommet, s, est réunion de deux individus, a(s) et b(s), qui sont plus proches voisins réciproques dans I.

Soit N quelconque; et supposons la proposition démontrée jusqu'au rang N-1. Considérons un arbre compatible A, qui a N nœuds, celui de plus haut niveau étant noté s, comme dans l'énoncé de la proposition. Il existe, par définition, un sommet s' tel que A-s' est compatible (nous omettons désormais les accolades autour de s, s',...) et que a(s') et b(s') sont v.r. au sein de Som(A-s'). Si s=s', il n'y a rien à démontrer.

Supposons que s est distinct de s'. Du fait du niveau de s, a(s) et b(s) sont, en vertu de l'hypothèse de récurrence, des voisins réciproques au sein de Som(A-{a(s'), b(s')}): car A-s', a (N-1) nœuds. Dès lors, de par l'axiome de la médiane,, a(s) et b(s) restent voisins réciproques si on agrège a(s') et b(s'); c'est à dire qu'ils sont voisins réciproques au sein de Som(A-s); ce qu'il fallait démontere.

2.4 Recherche des sommets par une chaîne de plus proches voisins

La proposition du §2.3 permet de montrer qu'un arbre obtenu à partir de A par CardI-1 agrégations successives (de deux plus proches voisins réciproques d'entre les sommets de l'état antérieur de l'arbre) aboutit au résultat désiré, comme si les nœuds avaient été créés dans l'ordre croissant des niveaux. Reste à guider la recherche des paires de voisins réciproques. On utilisera pour cela les chaînes de plus proches voisin.

Définition : Soit A un arbre binaire sur I compatible avec le critère d : une suite ordonnée {s(1), ..., s(p)} de sommets distincts de A est appelée une chaîne de plus proches voisins de longueur p si on a:

$$\forall\ q \in S\{2,\ \ldots, p\} : s(q) = v(s(q-1))\ ;$$

i.e. si, au sein de Som(A), chaque sommet de la chaîne est un plus proche voisin de son prédécesseur, soit:

$$\forall\ s \in \mathrm{Som}(A) - s(q-1) : d(s(q-1), s(q)) \leq d(s(q-1), s)\ ;$$

nous disons, en toute rigueur, "un" plus proche voisin, et non "le" plus proche voisin, car on peut ici faire abstraction de l'unicité du plus proche voisin;

quitte à admettre plusieurs solutions pour la CAH elle-même que l'on construit.

Proposition : Toute chaîne de plus proches voisins, convenablement prolongée, aboutit à une paire de plus proches voisins réciproques.

Pour prolonger la chaîne {s(1), ..., s(p)} il faut déterminer s(p+1)=v(s(p)), le plus proche voisin de s(p) au sein de Som(A) : il est clair, d'abord, que v(s(p)) n'est pas à chercher parmi les p-2 premiers s(q), car on a:

$$d(s(p), s(p-1)) \leq d(s(q), s(q+1)) \leq d(s(q), s(p)) \quad ;$$

la deuxième de ces inégalités résulte de ce que s(q+1) est un plus proche voisin pour s(q) ; et la première, de ce que, dans une chaîne de plus proches voisins, la longueur des maillons successifs ne peut que décroître, i.e.:

$$d(s(q+1), s(q+2)) \leq d(s(q), s(q+1)) \quad ;$$

parce que s(q+2) = v(s(q+1)). Ceci posé, on voit qu'en épuisant éventuellement Som(A) par une chaîne de longueur p = Card(Som(A)), on aboutit à une paire {s(p), s(p-1)} de plus proches voisins réciproques.

2.5 Schéma et coût de l'algorithme de recherche en chaîne

Initialement, A0 = I (plus exactement, A0 est l'ensemble des parties de I réduites à un seul élément); et Som(A0) = A0. On part d'un élément quelconque, i1, pris pour s(1) et on construit une chaîne; qui aboutit à une paire de v.r. {s(p), s(p-1)} qu'on agrège; d'où un arbre A1, tel que, si on note: n1 = s(p) ∪ s(p-1), on a:

$$A1 = A0 \cup \{n1\} \quad ; \quad Som(A1) = (A0 + \{s(p), s(p-1)\}) \cup \{n1\} \quad .$$

Si p=2, (i.e. si l'élément i1 dont on est parti admet un voisin réciproque), la chaîne est détruite par l'agrégation de ses deux éléments: on doit donc choisir dans A1 un sommet quelconque s(1) (qui peut être n1), à partir duquel on construit une nouvelle chaîne, aboutissant nécessairement à une paire de v. r. dont l'agrégation fournira le nœud n2.

Si p>2, l'agrégation de s(p) et s(p-1) laisse subsister une chaîne de longueur (p-2) :

$$s(1), s(2), s(3), \ldots, s(p-3), s(p-2) \quad ;$$

il s'agit bien d'une chaîne au sens défini au §2.4, car, d'après le §2.2, l'agrégation de s(p-1) et s(p) la relation s(q) = v(s(q-1)) pour q = 2, ..., p-2. (Le cas p=3 fait toutefois exception, et ne se distingue pas véritablement du cas n=2). On reprendra donc la recherche d'une paire de v.r. en prolongeant la chaîne restante: éventuellement, s(p-3) et s(p-2) peuvent être devenus v.r. du

fait de l'absorption de s(p-1) au sein dù nouveau sommet: s(p-1) ∪ s(p). C' est, en particulier ce qui se produit nécessairement si la chaîne créée a épuisé l'ensemble des sommets, comme c'est le cas pour l'exemble de progression géométrique considéré au §1.

Après création du second nœud, n2, l'algorithme se poursuit en n'opérant jamais que sur une chaîne, éventuellement recréée à partir d'un sommet quelconque, si elle a été épuisée.

On voit que le coût de la création d'un nœud s(p-1) ∪ s(p), pour p>2, équivaut à trois recherches de plus proche voisin: celles de v(s(p-2))=s(p-1); de v(s(p-1))=s(p) ; et de v(s(p)), qui nous ramène à s(p-1). Si p=2, on a seulement deux recherches: v(s(1)) = s(2), et v(s(2)) = s(1). Quant à la recherche de v(s(p)), son coût est proportionnel à Card(Som(A))-p, nombre des sommets à envisager. Pour la création du r-ème nœud, si CardI=n, le coût est:

$C(r) \leq 3.(n-r).k$ (k dépendant non de n, mais de la formule de distance).

En tout état de cause, le coût total admet une majoration en $K.n^2$, le coefficient K ne dépendant pas de n.

Références bibliographiques

[CAH CHAÎNE RÉCIP] : J.-P. BENZÉCRI : "Construction d'une classification ascendante hiérarchique par la recherche en chapine des voisisns réciproques"; in *CAD*; Vol.VII, n°2; pp. 209-218; (1982); repris, avec des démonstrations complètes, dans:

J.-P. BENZÉCRI et collab.: *L'Analyse des Données, T.I: La Taxinomie*; 4-ème édition; Dunod, Paris; (1984);

[CHAÎNE COMP. CAH] : Ch. BASTIN, J.-P. BENZÉCRI : "Exemple d'effet de chaînage complet autour d'un centre en Classification scendante Hiérarchique"; in *CAD*; Vol.VII, n°2; pp. 189-208; (1982);

ce même cahier contient d'autres mémoires consacrés à la CAH;

[ALG. & ALG. CAH] : J.-P. & F. BENZÉCRI : "Démonstration de l'équivalence des résultats des algorithmes accélérés à ceux de l'algorithme de base en CAH"; in *CAD*; Vol.X, n°3; pp. 257-271; (1985).