LES CAHIERS DE L'ANALYSE DES DONNÉES

I. KHARCHAF

Sur la recherche des plus proches voisins suivant une décomposition cellulaire de l'espace en classification ascendante hiérarchique

Les cahiers de l'analyse des données, tome 12, n° 2 (1987), p. 198-202

<http://www.numdam.org/item?id=CAD_1987__12_2_198_0>

Les Cahiers de l'Analyse des Données
Vol XII - 1987 - n° 2 - pp. 198-202

SUR LA RECHERCHE DES PLUS PROCHES VOISINS SUIVANT UNE DECOMPOSITION CELLULAIRE DE L'ESPACE EN CLASSIFICATION ASCENDANTE HIERARCHIQUE

[CAH. VOIS. CELL.]

I. Kharchaf (*)

Supposons définie sur un échantillon E d'effectif 2^p de l'ensemble I, une classification hiérarchique binaire. Sous sa forme idéale cette classification aura une profondeur exactement égale à p; en ce sens que E étant divisé au plus haut niveau en deux classes d'égal effectif A(E) et B(E), chacune de celle-ci se partage de même, ... et ainsi de suite jusqu'aux éléments de E, chacun contenu dans un système de p classes emboîtées; depuis E tout entier (de cardinal 2^p) jusqu'à la classe immédiatement supérieure à l'élément (classe de cardinal $2^1 = 2$). Cette classification sur E, permet de définir pour l'espace ambiant tout entier une partition polyédrale, *a priori* bien adaptée à la structure de I, et dans laquelle il est facile de placer tout point de l'espace; et, en particulier tout élément de I. On utilise pour cela la technique des éléments supplémentaires (cf. Jambu; thèse, et le livre).

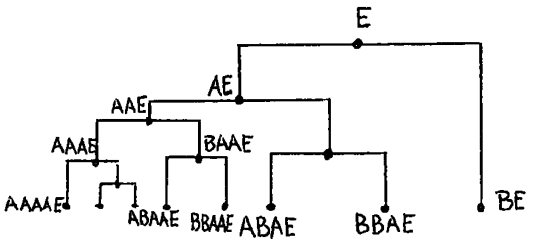
L'espace entier est d'abord partagé par l'hyperplan H(E) mené par le point G(E) (centre de gravité de l'échantillon E) perpendiculairement à la ligne G(AE) G(BE) (joignant les centres de gravité des deux descendants immédiats de E); le demi-espace "AE" (nous noterons aussi en bref, le demi-espace contenant AE et limité à l'hyperplan H(E)) est lui-même divisé par l'hyperplan H(AE) mené par G(AE) perpendiculairement à la ligne G(AAE) G(BAE); d'où deux domaines "AAE" et "ABE" : "AAE" étant défini comme le quart d'espace contenant G(AAE) et limité par les deux hyperplans H(E) et H(AE). Finalement on aboutit à une division de E en 2^p cellules polyédrales, construites par voie descendante; indicées par les éléments e de E; avec associée à chaque classe c de la hiérarchie taxinomique bâtie sur E un domaine "c" que l'hyperplan H(c) mené par G(c)

(*) Maître assistant. Faculté des Sciences. Rabat.

perpendiculairement à la ligne G(Ac) G(Bc) subdivise en deux domaines "Ac" et "Bc". Quel que soit le parti que nous espérons tirer de cette subdivision, on doit toutefois noter que bien désignés par les classes de la hiérarchie sur E, les domaines polyédraux ne contiennent pas nécessairement la classe du même nom: en particulier si e est un élément de E, le domaine "e" peut ne pas contenir le point e lui même.

En revanche il est facile de déterminer dans quel domaine "e" se trouve un point quelconque s de l'espace ambiant (et notamment un élément i de I); par n produits scalaires en effet, on descend la hiérarchie. Convenons de noter u(c) le vecteur associé à une classe c de la CAH sur E et défini comme suit: u(c) est unitaire, et égal au quotient de (G(Ac) - G(Bc)) par sa norme. Partons du sommet, le produit scalaire (s - G(E)) o u(E) est positif si s ∈ "AE" , négatif si s ∈ "BE"; de plus la valeur absolue de ce produit scalaire donne la distance de s à l'hyperplan frontière H(E). Supposons que s ∈ "AE", le produit scalaire (s - G(AE)) o u(AE) permet de décider si s ∈ "AAE" or s ∈ "BAE"; et donne cette fois encore la distance de s à l'hyperplan de séparation. Finalement s sera placé dans l'une des 2^n cellules "e" (étiquetées par les éléments e de l'échantillon E); et on connaîtra la distance de s à la frontière de la cellule; ainsi que celui des hyperplans frontières H(c) (faces de "e") desquels s est le plus proche; et cela ne nous aura coûté que le calcul de p produits scalaires.

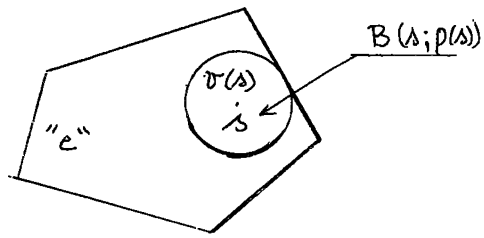
Avant de poursuivre, il importe de noter que la structure dichotomique idéale postulée pour la classification sur E ne joue pas ici un rôle indispensable; la profondeur p étant fixée *a priori*, il est toujours possible de ne retenir de la classification édifiée sur E que la partie supérieure de l'arbre définie par la condition que seules soient retenues les classes c ayant un nombre de prédécesseurs inférieur ou égal à p : l'exemple ci-dessous, du cas n = 3, Card = 8 (on pourrait éventuellement prendre Card E > 8...) montre comment procéder: les classes retenues étant marquées d'un point plein on aboutit à une partition ultime en 5 classes.



Ceci posé, pour construire une CAH sur I , on sait qu'il est commode de procéder par recherche en chaîne des voisins réciproques au sein de l'ensemble s des sommets. La construction élémentaire est celle de v(s), plus proche voisin de s. Nous proposons de chercher d'abord v(s) au sein de la cellule "e" où se trouve

s. Au départ S = I; pour avoir un ordre de grandeur des calculs à effectuer postulons ici une situation idéale: chaque cellule "e" contient exactement k points de S (i.e. de I).

Dans la pratique, il n'en sera pas ainsi; mais seul importe pour l'hypothèse que pour la plupart des cellules "e" le nombre de points de S y contenus, est de l'ordre de (CardS)/2^p. Le coût requis pour affecter les éléments de I aux cellules "e" est np (si n = Card I): p produits scalaires pour chacun des n points. La recherche de v(s) au sein de "e" $\cap$ I se fait par k calculs de distances, puisqu'il y a k points de I à considérer dans cette cellule. Mais obtient-on aussi v(s)? Il est facile de le savoir: en effet on connaît la distance $\rho(s)$ de s à la frontière de la cellule "B"; on peut donc affirmer que tout point de I situé à une distance de s inférieure à $\rho(s)$ est dans "e"; si donc la boule de centre s et de rayon $\rho(s)$, $B(s;\rho(s))$ contient le ppv de s dans "e", ce point est aussi le ppv de s dans I. La question est donc : $B(s;\rho(s)) \cap$ "e" $\cap$ I est-il non vide?



Dans l'affirmative on aura obtenu v(s) pour un coût de l'ordre de k. Sinon, il faut remonter dans la hiérarchie des cellules; jusqu'à attendre un prédécesseur de "e" où soit incluse une boule de centre s dont le rayon soit supérieur à $\rho(s)$: cela est aisé si on a noté les distances de s à tous les hyperplans H(c) qui par dichotomies successives, à partir de "E" (qui n'est autre que l'espace entier) définissent "e", mais chaque fois qu'on s'élève d'un degré, le coût est doublé: il atteint n = Card I, si la recherche de v(s) doit s'étendre à I tout entier.

Ainsi on rencontre un nouveau problème: majorer la probabilité de l'événement défavorable: $B(s;\rho(s)) \cap$ "e" $\cap$ I = $\varnothing$; le problème peut être transformé comme suit: les individus i à classer sont issus d'une source potentiellement infinie caractérisée par une distribution de probabilité sur l'espace ambiant "E"; sur "e" il y a une loi induite; et "e" $\cap$ I est un échantillon d'effectif k de cette loi; le point s lui-même est aussi issu de cette loi. On aboutit à un énoncé précis: soit "e" un domaine polyédral, muni d'une loi de probabilité prob, s un point issu de la loi prob; I(e) un échantillon d'effectif k muni de la loi prob (par tirages indépendants); notons $\rho(s)$ la distance de s à la frontière de "e"; quelle est la probabilité A pour que: $B(s;\rho(s)) \cap I(e)$ soit vide? Il est clair que pour s fixé, on a

une probabilité conditionnelle égale à la puissance k de l'intégrale de prob étendue à "e" - B(s;ρ(s)); ce qu'on doit calculer est donc:

$$Esp(1 - prob(B(\rho;\rho(s)))^k) = A ;$$

l'espérance étant à prendre pour s issu de la loi prob.Calculer A sans fixer "e" et la loi prob, est impossible; en général on doit attendre que le résultat dépende grandement de la dimension de l'espace ambiant: une valeur typique intéressante étant fournie par le cas où "e" est un hyperbole de dimension r (l'hypothèse "e" sphérique, certes peu réaliste, ayant l'avantage d'offrir les calculs les plus faciles) et où la loi prob est la loi uniforme. Sans faire de calcul, on peut adopter pour A l'estimation $(1 - \varepsilon(r))^k$; ε étant petit; et d'autant plus petit que r est plus élevé.

De façon précise, s est inclus dans un emboitement de p+1 cellules depuis "e" jusqu'à l'espace entier "E".

$$s \in \text{"e"}\ 1\ \text{"c1(s)"}\ 1...1\ \text{"ch(s)"}\ 1...1\ \text{"cp(s)"} = \text{"E"} ;$$

sous l'hypothèse simplificatrice adoptée des dichotomies parfaites, on a:

$$\text{Card "ch(s)"} \cap I = k.2^h = n.2^{(h-p)}.$$

Notons ρh(s) la distance de s à la frontière de la cellule "ch(s)"; en particulier, si on pose "e" = "c0(s)", ce qu'on a noté ρ(s) ci-dessus doit être noté ρ0(s). Nous désirons estimer la probabilité que la recherche de v(s) doive être faite non dans "c0(s)" ∩ I; mais plus haut dans la hiérarchie, donc dans "ch(s)" ∩ I. Estimons par exemple la probabilité que v(s) ne soit pas dans "ch(s)" ∩ I; cela équivaut à dire que:

$$B(s; \rho h(s)) \cap I = \varnothing$$

estimée au niveau de "ch(s)" comme plus haut au niveau de "c0(s))" = "e", cette probabilité est: $((1 - e) \uparrow (k.2^h))$ = reste(h). La probabilité que la recherche de v(s) doive être poussée jusqu'au niveau h+1 est majorée par reste(h): ainsi reste(0), probabilité que v(s) ne soit pas dans B(s;ρ0(s)), majore la probabilité que v(s) soit trouvé précisément dans la cellule "c1(s))" immédiatement supérieure à "e" = "c0(s)", plus précisément dans B(s;ρ1(s)); reste(1), majore la probabilité que v(s) doive être cherché dans "c2(s)", et trouvé dans B(s;ρr(s)); etc; reste (p-1) majore la probabilité que v(s) doive être cherché en parcourant I tout entier. Or le coût d'une exploration de "ch(s)" n'est autre que Card("ch(s)" ∩ I) = $k.2^h$. Ainsi on a une estimation du coût moyen de la recherche de v(s) (compte tenu du niveau hiérarchique variable auquel il faut remonter) par la formule:

$$k+2k((1-\varepsilon)\uparrow k)+2^2k((1-\varepsilon)\uparrow(2.k))+..+2^{h+1}k((1-\varepsilon)\uparrow(2^h k))+..$$

la sommation doit être poursuivie jusqu'à h = p-1; mais en fait il apparaît que peut être majorée la somme infinie, (poursuivie pour $p \to \infty$). En effet notons $a=(1-\varepsilon)\uparrow k$; on a pour estimation du coût

$$\text{coût} \leq k(1+2a+2^2 a^2 +...+ 2^{h+1}(a\uparrow 2^h) +...)$$

Si on suppose $a \leq 1/2$, on voit aisément que coût $< 4k$.

Finalement la construction d'une CAH sur I (d'effectif CardI=n) sera décomposée comme suit:

1°) Choix du nombre k: selon nos hypothèses, ε est estimé d'après la dimension de l'espace ambiant; k est choisi tel que $(1-\varepsilon)^k < 1/2$.

2°) Tirage d'un échantillon E de I d'effectif n/k; et construction d'une CAH sur E: nous noterons CO(n/k) le coût de cette construction.

3°) Partition de I suivant les cellules associées à la CAH sur E: la profondeur de cette CAH étant log(n/k): le coût de cette partition sera n.log(n/k) (chaque élément i étant placé en descendant l'arbre, par une suite de produits scalaires); de même les sommets créés devant au fur et à mesure être placés dans la partition: au total puisque le nombre des noeuds successivement créés n'excède pas n, le coût global de cette maintenance restera en n log(n/k).

4°) Recherche en chaîne des voisins réciproques: on sait qu'il faut construire un nombre de maillons de l'ordre de n (de 2n à 3n): le coût d'un maillon étant de l'ordre de k (4k selon ce qu'on a estimé) on aboutit à un coût global en kn.

En récapitulant les majorations faites il vient:

$$CO(n) < CO(n/k) + K\, n \log n$$

(où K désigne une constante) ; par récurrence on a:

$$CO(n/k) < CO(n/k^2 + K(n/k) \log(n/k)$$

d'où pour CO(n) la majoration:

$$CO(n) < Kn \log n(1+k^{-1} + k^{-2} +...);$$

c'est-à-dire, puisque k est un entier (peut-être de l'ordre de 50 voire de 1000....) une majoration pour CO(n) en n log n.

Malgré la fragilité des estimations auxquelles nous avons dû avoir recours, nous pensons que ce résultat doit encourager à écrire des algorithmes de CAH, fondés sur la recherche des voisins suivant une décomposition en cellules, elle-même issue d'une CAH effectuée sur un échantillon.