LES CAHIERS DE L'ANALYSE DES DONNÉES

I. KHARCHAF

Sur la complexité des algorithmes de classification ascendante hiérarchique

Les cahiers de l'analyse des données, tome 12, n° 2 (1987), p. 195-197

<http://www.numdam.org/item?id=CAD_1987__12_2_195_0>

Les Cahiers de l'Analyse des Données
Vol XII 1987 - n° 2 - pp. 195-197

SUR LA COMPLEXITE DES ALGORITHMES DE CLASSIFICATION ASCENDANTE HIERARCHIQUE

[COMPLEXITE CAH]

I. Kharchaf (*)

On sait que le temps d'exécution de l'algorithme de base de CAH, pour un ensemble I d'individus situés dans un espace de dimension déterminé (où entre lesquels les distances, sont fournies par une procédure déterminée) croît comme la puissance troisième du cardinal n = Card I.

Le recours à l'algorithme des voisins réciproques permet d'abaisser cet ordre à n^2 car d'une part un modèle aléatoire suggère un taux de paires de voisins réciproques cf. [PROP. RECIP.] Vol VII 1982, n° 2 pp.185-188) qui aboutit à une estimation en n^2, d'autre part la recherche en chaîne des voisins réciproques conduit à un algorithme dont on a démontré (cf. [C.A.H. CHAINE RECIP.] Vol VII n° 2, pp. 209-218) que le temps d'exécution est majoré par n^2 (et il est facile de donner de ce même temps une borne inférieure, elle-aussi en n^2.

De n^3 à n^2, le progrès est considérable; mais, notamment en reconnaissance de la parole, il semble qu'on doive prochainement traiter des ensembles de données où $n \approx 10^5$ pour de tels ensembles un coût en n^2 apparaît quasi prohibitif.

Des contre-exemples ont été proposés (cf. [CHAINE COMP. C.A.H.] Vol VII 1982, pp. 189-208 et [CHAINAGE DROITE CAH] Vol IX 1984, pp. 231-238) pour montrer que la recherche des voisins réciproques sur toute l'étendue de I, ou de l'ensemble S des sommets, (nous parlerons de la recherche

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

"en surface") peut aboutir en un temps à n^3, la recherchant en chaîne permettant seule de rester toujours en n^2. Ces mêmes contre-exemples nous paraissent suggérer fortement qu'il peut être dans certains cas de figure, impossible de descendre en dessous de n^2, quel que soit l'algorithme, dans la mesure où les distances (plus exactement les écarts) entre sommets pris deux à deux peuvent demeurer quasi égaux entre eux tout au long de la construction ascendante; ce qui implique que la création d'un nouveau noeud (par agrégation de deux sommets) requiert au moins le calcul des distances d'un sommet à tous les autres (ce en vue d'allonger d'un maillon la chaîne)

Ces contre-exemples cependant ne sont aucunement le schéma d'une structure ordinaire, susceptible d'apparaître fréquemment dans la pratique: à telle enseigne que l'algorithme en chaîne conçu pour être protégé contre de telles structures semble être généralement un peu moins rapide que l'algorithme en surface, non protégé.

En effet une remarque élémentaire suggère que l'algorithme de recherche en chaîne peut être amélioré. Partant d'un sommet s on cherche le sommet v(s) qui en est le plus proche en calculant les n distances de s à tous les sommets non rangés dans la chaîne; en cherche ensuite v(v(s)) en calculant les distances de v(s) aux n-1 sommets subsistants; or il est clair que si la distance de s à s' est relativement grande, la distance à s' de v(s) (plus proche voisin de s) ne pourra être minima, en sorte que s' n'est pas candidat à être le plus proche voisin de v(s); (cette remarque n'est sans objet que si, comme dans le contre-exemple, les écarts deux à deux des sommets sont à peu près égaux entre eux). D'ailleurs la recherche des plus proches voisins ne sert pas seulement en CAH: elle est à la base de la régression par "boule" (cf. [POUBEL] Vol II 1977 n° 4 pp. 467-481): et il apparaît qu'on accélère notablement cette recherche si, l'ensemble des individus ayant été partagé en classes par une procédure d'agrégation rapide (e.g. en boules de centre fixe et de rayon variable) on borne la recherche des p.p.v. de is à la classe (ou à quelques classes) dont le centre est le plus proche de is (cf. Gleizes, thèse 3° cycle, 1980, Paris VI; et [VOIS. BOUL.] Vol IX 1984, n° 1, pp. 119-122).

Ceci revient à dire que l'ensemble des individus étant partagé entre quelques centres (dont les classes sont comme les domaines d'influence), la recherche des p.p.v. de is se fait en passant par les centres.

Un tel système de centres étendu à I tout entier est indispensable en CAH si la recherche des paires de voisins réciproques se fait en surface. Sous sa forme la plus perfectionnée, le système des centres (ou des boules recouvrant I) est organisé hiérarchiquement (cf. [VOIS. BOUL. HIER.] Vol IX, 1984, n° 1, pp. 123-124);. Afin d'assurer aux classes ultimes (boules de l'ordre inférieur) un effectif donné, la profondeur de cette hiérarchie doit être en log n: car en bref, si chaque boule du niveau x+1 se partage en p boules du niveau x, (p étant par exemple de l'ordre de la dimension de l'espace ambiant) avec h niveaux on a une hiérarchie de p^h boules. Pour trouver le p.p.v. d'un individu i (ou d'un sommet ss), on descend la hiérarchie des centres jusqu'à trouver le centre du niveau inférieur le plus proche de e en restreignant la recherche du voisinage (la boule) dépendant de ce centre. Dans un tel cadre, le coût de la recherche des v(i) apparaît

en log n, et non plus en n; d'où pour la construction de la CAH un coût en n log n. Le cadre lui-même pourraît être construit avec un coût en n log n: en effet la construction d'un système fini de centres suivant un algorithme itératif tel que celui de E. Diday; ou la variante de Flamenbaum; cf. [ALG. AGR. RAY.] Vol IV 1979, n° 3, pp. 365-375 se fait en un nombre de parcours du fichier des individus qu'on peut borner par K; d'où pour une hiérarchie de profondeur log n, Klog n lectures avec un coût en log n. Tel serait globalement l'ordre de grandeur du coût de la CAH.