COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 26 au 30 AVRIL 1977

ANALYSE DE CLUSTERS : méthodes non paramétriques d'analyse statistique multidimensionnelle

André BERTHON

Société d'Etudes et Conseils AERO - 3 avenue de l'Opéra - 75001 PARIS

RESUME

On passe brièvement en revue les méthodes courantes de recherche de clusters, considérant particulièrement le cas d'ensembles statistiques importants et de dimension élevée dont on cherche à étudier la loi de probabilité.

SUMMARY

A short review of current methods in cluster analysis is presented, with emphasis on the case of large statistical samples having high dimensionality whose probability density function is to be investigated.

ANALYSE DE CLUSTERS : méthodes non paramétriques d'analyse statistique multidimensionnelle

1. INTRODUCTION

L'analyse statistique multidimensionnelle s'applique à des problèmes aussi variés que la taxinomie, l'estimation paramétrique, la reconnaissance de formes, l'identification de processus aléatoires quelconques. Elle se définit par le fait que chaque échantillon ou événement aléatoire est décrit par plusieurs variables et se caractérise, dès que le nombre des variables dépasse 2 ou 3, par la quasi impossibilité pour l'analyste de se faire une représentation intuitive des données.

La nature de ces données est très variable. Chaque composante peut représenter, pour ne prendre que trois exemples :

- une caractéristique des individus d'une population; elle peut ne prendre que des valeurs discrètes,
- un coefficient de la décomposition d'un signal suivant une base de fonctions données,
- une variable cinématique affectée à une particule de l'état final dans une réaction entre particules élémentaires [1].

Dans tous les cas l'ensemble statistique se présente comme une famille de N points dans l'espace $\mathbb{R}^n (n>1)$. Nous nous plaçons dans le cas où le "nuage" de points est de nature statistique, le problème posé étant d'en extraire le plus d'informations possible sur la fonction de distribution $p(\vec{x})$ à laquelle il obéit. On suppose que l'information a priori sur cette distribution est soit inexistante, soit de nature essentiellement qualitative (par exemple, que $p(\vec{x})$ représente la superposition d'un nombre indéterminé de processus indépendants dont la loi n'est pas donnée analytiquement); en d'autres termes l'étude à mener n'est pas paramétrique.

L'estimation directe de la densité de probabilité multidimensionnelle est généralement impraticable. Remarquons que le nuage est extrêmement peu dense. Si l'on divise l'intervalle de variation de chacune des variables en 10 le nombre de points dans chaque hypercube élémentaire est en moyenne $N/10^n$. Or N dépasse rarement quelques milliers.

De plus dès que la distribution présente quelque structure cette densité varie énormément d'une région de l'espace à l'autre. Aussi est-il exclu d'estimer $p(\vec{x})$ à l'aide d'un histogramme multidimensionnel. Parmi les estimateurs non paramétriques qui s'adaptent à la densité locale le plus populaire est celui des k-plus proches voisins [2] :

$$\hat{p}(\vec{x}) = k/N \; V_k(\vec{x})$$

où $V_k(\vec{x})$ est le volume de la sphère qui contient les k points du nuage les plus proches de $\vec{x}$. On montre [3] que le biais et la variance tendent asymptotiquement vers 0 pourvu que k dépende de la taille de l'échantillon selon une loi vérifiant :

$$\lim_{N\to\infty} k(N) = \infty \qquad \lim_{N\to\infty} \frac{k(N)}{N} = 0$$

Mais pour un échantillon à n dimensions le biais est donné par :

$$E\left[\hat{p}(\vec{x})\right] = p(\vec{x}) + \frac{\Gamma^{2/n}\left(\frac{n+2}{2}\right)}{2\pi(n+2)} \frac{Tr\left(p''(\vec{x})\right)}{p^{2/n}(\vec{x})} \left(\frac{k}{N}\right)^2$$

où p'' est la matrice des dérivées secondes de p. La densité p étant faible et sa variation rapide il faut s'attendre pour les statistiques finies, à un biais très important presque partout.

Si l'estimation directe est impossible, l'idée sous-jacente, de caractériser la densité locale en un point du nuage par la distribution de ses plus proches voisins, n'en est pas moins à la base de la plupart des méthodes d'analyse, avec des variantes diverses [4].

Ce qui vient d'être dit ne s'applique pas, du moins pour N grand, au cas de deux dimensions; ce cas est très différent des autres pour une autre raison, qui est la remarquable aptitude de l'oeil à extraire l'information contenue dans une image, en effectuant très rapidement des opérations élémentaires telles que :

- estimation de la densité locale de points,
- estimation du gradient de cette densité, et des directions principales,
- partage du plan en régions disjointes séparées par des zones de faible densité,
- reconnaissance des structures à une dimension (points groupés le long d'un arc de courbe).

Selon certains auteurs [5] ces performances reposent sur la construction instinctive d'un squelette du graphe dont les sommets sont les points du nuage (arbre minimal).

2. POURQUOI CHERCHER DES "CLUSTERS" ?

D'un point de vue purement probabiliste, l'analyse d'un ensemble statistique multidimensionnel peut se concevoir comme un effort pour se ramener à une situation idéale dans laquelle la distribution des points est uniforme, ce qui implique que la densité de probabilité peut être complètement factorisée. En général ce programme n'est pas réalisable globalement, mais on peut chercher à structurer l'information à plusieurs niveaux en faisant intervenir plusieurs lois de distribution dont le mélange rende compte de la distribution observée. L'idée est que chaque point du nuage résulte d'un double tirage au sort, d'abord celui d'un type de processus auquel correspond une fonction de distribution $p_\alpha(\vec{x})$, ensuite celui du point selon la distribution de fréquences correspondante. La fonction de distribution totale est alors une combinaison linéaire convexe des fonctions qui rendent compte des différents processus :

$$p(\vec{x}) = \sum_\alpha \lambda_\alpha \, p_\alpha(\vec{x})$$

Cette démarche est naturelle lorsqu'on s'attend à ce que certaines des n variables soient fortement corrélées entre elles mais avec une corrélation différente selon le domaine occupé par un autre groupe de variables (par exemple, les variables d'environnement s'il s'agit de mesures de propagation).

Si les fonctions $p_\alpha(\vec{x})$ sont connues a priori, fût-ce à la valeur de certains paramètres près, on est évidemment ramené à un problème d'estimation paramétrique. Dans le cas contraire il faut les construire à partir de l'échantillon lui-même en imposant certaines conditions dont la plus naturelle est qu'elles aient des domaines disjoints. Ainsi l'approche statistique initiale conduit à considérer pour commencer un problème de classification consistant à partitionner l'espace et à distinguer dans le nuage de points des amas (clusters) dont chacun constituera au moins de

./...

manière approchée un ensemble statistique homogène au regard du processus aléatoire sous-jacent. Une autre idée et un autre espoir sont ici implicites, c'est que chaque agrégat a des chances d'obéir à une loi plus simple, notamment par la réduction du nombre des variables significatives, c'est-à-dire de la dimension intrinsèque de l'ensemble.

Depuis les débuts de l'analyse statistique une assez grande variété de méthodes de regroupement et de séparation ont été développées; chacune a ses mérites et son domaine d'application privilégié, car il est clair que le problème n'admet pas de solution universelle. Elles n'ont pu connaître leur plein développement qu'avec l'apparition d'ordinateurs rapides et à grande capacité de mémoire et, ce qui n'est pas moins important, de leur utilisation en mode interactif. C'est en effet l'oeil qui demeure le meilleur juge de la réussite et la visualisation des résultats, qui suppose une projection à deux dimensions, est le meilleur moyen de contrôler les algorithmes et d'en ajuster les paramètres.

Il est possible de classer les méthodes de recherches de groupements locaux suivant quelques caractéristiques en distinguant :

. D'une part, des méthodes de recherche guidée et des méthodes globales. Les premières consistent à réduire la masse des informations pour faire apparaître plus facilement des structures, généralement par une application sur un espace de dimension plus faible. Cette application peut être linéaire (projection) ou non. Les méthodes globales au contraire utilisent la donnée des n composantes et cherchent à classer directement les points, c'est-à-dire à construire une application de l'ensemble des points dans un ensemble fini obéissant à certaines conditions.

. D'autre part, des méthodes reposant sur la métrique de l'espace, et des méthodes qui ne font intervenir que la notion de proximité entre les points, c'est-à-dire une fonction définie sur les couples de points et vérifiant :

$$d(\vec{x_i},\vec{x_j}) = d(\vec{x_j},\vec{x_i}) \geqslant 0 \qquad d(\vec{x_i},\vec{x_j}) = 0 \Rightarrow i = j$$

mais non l'inégalité triangulaire.

3. METHODES DE REDUCTION

3.1 Préliminaires

Le fait de représenter les données dans $\mathbb{R}^n$ introduit implicitement une notion de proximité, et même une mesure de cette proximité, par exemple à l'aide de la distance euclidienne. Toutefois cette mesure ne fait pas nécessairement partie de l'information contenue dans les données; il est fréquent que les grandeurs portées sur les axes soient de nature différente; le choix des échelles est alors arbitraire; certaines variables peuvent être discrètes. Souvent le choix des variables est lui-même largement arbitraire. Il y a donc une étape préalable à la recherche des clusters, qui est de définir un premier système de coordonnées tel que des points "voisins" au sens intuitif (compte tenu des plages de variation des paramètres, des erreurs de mesure dont ils peuvent être affectés, etc.) soient également voisins dans l'espace et inversement.

3.2 Méthode des composantes principales

Le premier usage possible de la métrique est l'analyse en composantes principales. Soit V un sous-ensemble du nuage de points et C la matrice $n \times n$ de coefficients :

$$C_{ij} = \sum_{\vec{x} \in V} (x_i - \bar{x}_i)(x_j - \bar{x}_j)$$

où $\bar{x}_i$ est la valeur moyenne sur V de la coordonnée x_i. La diagonalisation de cette matrice fournit n combinaisons linéaires $y_1, \ldots, y_n$ orthogonales des x_i, dont les variances (relativement à V) sont les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$. On en tire les informations suivantes :

- la direction y_1 qui correspond à la plus grande valeur propre est de toutes ces combinaisons linéaires des variables originales, celle qui est la plus dispersée, et ainsi de suite dans l'ordre décroissant des valeurs propres;
- le nombre de valeurs propres non nulles ou non négligeables est la dimension effective de l'ensemble de points, ou plutôt du plus petit espace linéaire dans lequel il peut être plongé.

Appliquée à l'ensemble des données cette méthode n'apporte en général rien de plus que la détection éventuelle de dépendances linéaires entre les variables, et ne permet pas de déterminer la dimension intrinsèque du nuage : des points répartis uniformément sur une hypersphère donneront une matrice C égale à l'unité. Quant aux axes principaux, ils fournissent les variables, linéaires dans les x_i, dont les histogrammes ont les largeurs les plus grandes. Ce ne sont pas nécessairement ceux qui sont les plus structurés au sens de la formation d'amas.

En revanche la méthode des composantes principales appliquée au voisinage d'un point, défini soit par une sphère de rayon fixe centrée en ce point, soit par les k plus proches voisins du point ($k > n$) renseigne sur la dimension locale du nuage. A condition que la statistique soit assez fournie on peut ainsi détecter les zones où la dimension réelle est inférieure à n, l'analyse en composantes principales fournissant la variété linéaire tangente au lieu des points [6].

3.3 Méthode "projection pursuit"

Si l'on veut trouver la projection linéaire à une ou deux dimensions qui fait apparaître le maximum de structure il faut se donner une mesure de cette structure. On peut définir un indice de projection à une dimension [7] de la forme :

$$I(u) = s(u)\, d(u)$$

où $s(u)$ est la variance de la famille de points obtenus par projection sur la direction u, et d une fonction qui mesure le degré d'agglutination de ces points, par exemple :

$$(1) \quad d(u) = \sum_{i=1}^{N} \sum_{j=1}^{N} f(r_{ij}),$$

r_{ij} étant la distance entre les projections des points i et j, et f une fonction décroissante de r, nulle pour les valeurs supérieures à un seuil R. Ce paramètre R définit la taille (en projection) de la structure cherchée.

./...

./...

On cherche alors la direction de projection qui rend maximum l'indice I, c'est un problème non linéaire de maximisation à $N-1$ inconnues. La généralisation à deux dimensions consiste à chercher deux directions orthogonales u et v telles que l'indice :

$$I(u,v) = s(u)\, s(v)\, d(u,v)$$

soit maximum, la fonction d étant encore définie par la formule (1).

Cette méthode est souvent très efficace pour déterminer les directions pour lesquelles l'histogramme à une ou deux dimensions présente des concentrations importantes. On peut répéter son application en masquant les régions de haute densité déjà trouvées, et faire apparaître ainsi des pics dans d'autres projections linéaires; alors que la méthode des composantes principales ne permet pas de conclure lorsqu'on a affaire à plusieurs amas dont les axes principaux ont des orientations différentes.

3.4 Applications non linéaires

L'idée de déformer le nuage de points, c'est-à-dire de changer les distances tout en respectant une notion qualitative de proximité est ancienne. Elle s'applique, par exemple, à la recherche de la dimension intrinsèque [8]. En ce qui concerne la recherche d'agrégats la méthode de SAMMON [9] consiste à trouver une configuration de N points dans un espace de dimension plus faible, généralement 2, dont les distances mutuelles D_{ij} soient aussi proches que possible des distances euclidiennes d_{ij} dans R^n. On cherche le minimum de la fonction :

$$E(\vec{y}_1,\ldots,\vec{y}_n) = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\frac{D_{ij} - d_{ij}}{d_{ij}} \right)^2$$

par rapport aux $2N$ coordonnées des vecteurs $\vec{y}_i$.

L'avantage de cet algorithme est d'être indépendant de toute hypothèse sur les données et de tout paramètre de contrôle, en laissant le soin de déterminer les groupements sur la carte ainsi obtenue à l'oeil de l'observateur. Le prix à payer est une minimisation où le nombre de paramètres augmente comme N et le temps de calcul de la fonction comme N^2, ce qui limite le domaine d'application de la méthode aux ensembles n'ayant pas plus de quelques centaines de points.

Remarquons qu'il n'est fait usage que des distances mutuelles des points dans l'espace d'origine, qui peuvent être en fait de simples fonctions de proximité au sens défini plus haut.

4. METHODES GLOBALES

Il y a deux grandes manières de rechercher la décomposition du nuage, l'une est d'ordonner l'ensemble globalement suivant un ou plusieurs graphes, l'autre de procéder par agrégation locale selon des critères de proximité; elles peuvent se combiner. Les divers algorithmes ont en commun soit de comporter un ou plusieurs paramètres de contrôle, soit de conduire à un optimum local, dépendant d'une classification de départ.

./...

4.1 L'arbre minimal (Minimum Spanning Tree)

C'est une méthode de représentation des ensembles de points multidimensionnels qui n'est pas limitée à la recherche d'une décomposition. Là encore seule la notion de proximité des couples intervient. On considère le graphe dont les sommets sont les points du nuage, un arc reliant chaque couple de points (ij) avec un poids égal à d_{ij}. L'arbre minimal est le sous-graphe complet sans cycles dont le poids total est minimum. Etant donné la matrice des d_{ij} il peut se construire des plus simplement, par exemple en partant d'un couple dont la fonction de proximité est minimum, et en ajoutant à chaque pas le sommet relié à l'un des points déjà inclus par un arc minimal, ainsi que l'arc correspondant [10]. On montre que dans le graphe résultant chaque point est relié à l'un de ses plus proches voisins. Il est facile de voir que si le nuage est composé de groupes tels que :

- la distance entre deux points d'un même groupe est inférieure à R,
- la distance entre points appartenant à des groupes différents est supérieure à R.

il suffira de supprimer les arcs de longueur supérieure à R pour obtenir un graphe non connexe dont chaque composante connexe correspondra à l'un des groupes.

Diverses stratégies moins simplistes permettent d'identifier les amas disjoints dans lesquels la densité des points est variable, et même les zones de contact entre deux amas incomplètement séparés [5]. Parmi les autres applications de l'arbre minimal signalons la recherche de la dimension intrinsèque du nuage par simplification progressive de sa structure [11].

Cette méthode a l'avantage de ne pas être limitée par la dimension de l'espace. En revanche lorsque N est très grand le nombre des informations à conserver en mémoire peut devenir prohibitif.

4.2 Méthode des plus proches voisins

Cet algorithme est d'une simplicité surprenante [12]. On se donne a priori le nombre k de groupements cherché, et une classification arbitraire des points en k catégories, que l'on modifie pas à pas en associant à chaque point la classe à laquelle appartenait à l'étape précédente le plus grand nombre de ses voisins, dans un voisinage de dimension fixe. L'algorithme s'arrête lorsqu'aucun point n'a changé de catégorie d'une étape à l'autre.

La réussite dépend de trois facteurs :

- La taille des voisinages : s'ils sont trop grands une classe absorbe les autres; s'ils sont trop petits chaque classe reste scindée en groupes disjoints.
- Le caractère aléatoire de la classification initiale. S'il n'est pas respecté il y a formation de "noyaux durs" et la séparation "naturelle" entre classes ne se fait pas.
- L'adaptation de la métrique aux données.

Lorsque ces facteurs sont bien réglés il suffit que k soit plus grand que le nombre d'amas effectivement présent, les catégories en surnombre disparaissent d'elles-mêmes.

./...

4.3 Méthode des nuées dynamiques

L'idée /13/ est d'ajuster à chaque itération les centres de gravité des amas, dont on se donne le nombre a priori. Partant d'un ensemble de centres aléatoire, on attribue chaque point au centre dont il est le plus proche. Puis les centres sont redéfinis comme étant les centres de gravité des amas ainsi constitués. On montre /13/ que la somme des variances intra-classe décroît à chaque itération. Il en résulte que l'algorithme devient stationnaire après un nombre fini d'étapes, fournissant une partition de l'ensemble (qui dépend du choix initial des centres).

4.4 Une méthode de gradient non paramétrique

Cet algorithme publié récemment /14/ utilise à la fois l'idée de graphe et celle d'agrégation. Il repose sur un estimateur de la densité locale qui peut être soit du type de PARZEN (dont le plus simple consiste à compter les points situés dans un voisinage fixe du point considéré), soit celui des k plus proches voisins, et comporte donc un paramètre de contrôle.

Soit V_i le voisinage du point $\vec{x}_i$ utilisé par l'estimateur, p_i la densité locale estimée. On la compare aux densités p_j obtenues pour les points $x_j \in V_i$. L'algorithme procède par construction d'arbres orientés dont les origines sont les points où p_i a un maximum local, une branche reliant $\vec{x}_i$ à $\vec{x}_j$ si l'estimateur du gradient de densité :

$$g_{ij} = \frac{p_j - p_i}{d_{ij}}$$

est maximal sur l'ensemble des points $x_j \in V_i$.

Le graphe complet est la réunion des arbres obtenus, dont chacun correspond à un agrégat, l'arborescense partant de son point de densité maximum.

Le fait remarquable est que l'algorithme ne suppose pas de classification initiale, détermine automatiquement le nombre des amas, et ne comporte pas d'itération. En contrepartie le résultat peut dépendre de l'ordre dans lequel les points sont traités.

5. CONCLUSION

Les méthodes qui viennent d'être décrites se prêtent à un grand nombre de variantes, et leurs performances dépendent évidemment beaucoup du problème considéré. Au reste l'analyse en termes d'amas est rarement une fin en soi, mais plutôt un moyen de faire apparaître des structures, qui seront l'objet d'une étude ultérieure. Mais c'est un moyen puissant pour distinguer les divers processus aléatoires qui interviennent concurremment dans la formation des données, non seulement lorsqu'ils sont indépendants comme l'exprime la formule (1) mais aussi lorsqu'il y a interférence entre eux comme c'est le cas des phénomènes quantiques. Même des algorithmes apparemment simples permettent d'arriver à des résultats remarquables; toutefois la masse des calculs à effectuer est en général une fonction rapidement croissante du nombre de points N. La tendance étant, en de nombreux domaines, de traiter des ensembles de points toujours plus nombreux, on peut penser que la recherche de nouvelles méthodes s'orientera surtout vers des algorithmes qui utilisent au maximum l'information locale, et ne faisant pas appel à la minimisation d'une fonction de N variables ni à des processus itératifs.

./...

REFERENCES BIBLIOGRAPHIQUES

/1/ FRIEDMAN (J). Data Analysis Techniques for High Energy Physics.
CERN Summer School on Computers, Godoy Sund, 1974.

/2/ LOFTSGAARDEN (D), QUESENBERRY (C). A Nonparametric Density Function.
Ann. Math. Statis. V 36, pp. 1049-1051 (1965).

/3/ FUKUNAGA (K), HOSTETLER (L). Optimization of K-nearest Neighbor Density Estimates;
IEEE Trans. Inf. Theory, V IT-19 pp. 320-326 (1973).

/4/ O'CALLAGHAN (J. F). An Alternative Definition for "Neighborhood of a Point".
IEEE Trans. on Computers, pp. 1121-1125 (Nov. 1975).

/5/ ZAHN (C. T). Graph-Theoritical Methods for Detecting and Describing Gestalt Clusters.
IEEE Trans. on Computers, V C-20, N° 1, pp. 68-86 (Janu. 1971).

/6/ FUKUNAGA (K), OLSEN (D. R). An Algorithm for Finding Intrinsic Dimensionality of Data.
IEEE Trans. on Computers, V C-20, N° 2, pp. 176-183 (Feb. 1971).

/7/ FRIEDMAN (J), TUKEY (J). A Projection pursuit Algorithm for Exploratory Data Analysis.
IEEE Trans. Comp. V C-24 (1974).

/8/ BENNETT (R). The Intrinsic Dimensionality of Signal Collections.
IEEE Trans. On Inf. Theory, V IT-15, N°5 (Sept. 1969).

/9/ SAMMON Jr (J. W). A Nonlinear Mapping for Data Structure Analysis.
IEEE Trans. on Computers, V C-18, N° 5 (May 1969).

/10/ PRIM (R. C). Shortest Connection Networks and some Generalizations.
BSTJ, pp. 1389-1401 (Nov. 1957).

/11/ SCHWARTZMANN (D. H), VIDAL (J. J). An Algorithm for Determining the Topological Dimensionality of Point Clusters.
IEEE Trans. on Computers, V C-24, N° 12, pp. 1175-1182 (Dec. 1975).

/12/ KOONTZ (W. L. G), FUKUNAGA (K). A Nonparametric Valley-Seeking Technique for Cluster Analysis.
IEEE Trans. on Computers, V C-21, N° 2, pp. 171-178 (Feb. 1972).

/13/ BENZECRI (J). L'analyse des données, T 1, pp. 293-303. DUNOD 1973.

/14/ KOONTZ (W. L. G), NARENDRA (P), FUKUNAGA (K). A Graph-Theoretic-Approach to Nonparametric Cluster Analysis.
IEEE Trans. on Computers, V C-25, N°9 pp. 936-944 (Sept. 1976).