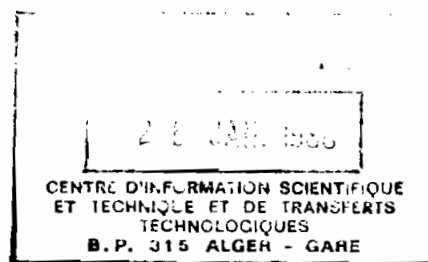


optimisation en classification automatique et reconnaissance des formes

E. DIDAY



RÉSUMÉ

Les algorithmes actuellement opérationnels, consistant à fournir une «bonne partition» d'un ensemble fini, produisent des solutions dont rien ne permet d'affirmer qu'elles soient optimales. Le principal but de ce texte est une étude synthétique de propriétés d'optimalité dans des espaces formés de partitions d'un ensemble fini. On formalise et on prend pour modèle de cette étude, une famille de techniques particulièrement efficaces du type «nuées dynamiques». Après avoir développé l'aspect programmation, on illustre les différents résultats par un exemple artificiel et surtout par deux applications concrètes; l'une, en géologie minière pour la recherche de familles géographiques de sondages verticaux, l'autre en médecine pour le dépistage de profils biologiques permettant une aide au diagnostic.

ABSTRACT

Algorithms which are operationally efficient and which give a good partition of a finite set, produce solutions that are not necessarily optimum. The main aim of this paper is a synthetical study of properties of optimality in spaces formed by partitions of a finite set. We formalize and take for a model of that study a family of particularly efficient technics of «clusters centers» type. The proposed algorithm operates on groups of points or «cores» (called «noyau»); these cores adapt and evolve into interesting clusters. Compared with other clustering algorithms, this algorithm requires less machine time and storage. After having developed the notion of «strong» and «weak» patterns, and the computer aspects, we illustrate the different results by an artificial example and by two applications; one in mineral geology, the other in medicine to determine biological profiles.

sommaire

1 - INTRODUCTION	4
1 - 1 Le problème	
1 - 2 La méthode des nuées dynamiques	
1 - 3 Etude synthétique des solutions obtenues	
2 - QUELQUES NOTATIONS ET DÉFINITIONS	7
3 - CONSTRUCTION DE TRIPLETS (f, g, W) PAR LES NUÉES DYNAMIQUES .	9
3 - 1 Formulation générale	
3 - 2 Les différentes variantes et intérêt comparé	
3 - 3 Construction de triplets rendant la suite v_n décroissante	
4 - STRUCTURATION DE L_k, P_k et V_k, PROPRIÉTÉS D'OPTIMALITÉ	13
5 - RECHERCHE D'INVARIANTS	16
5 - 1 Mesure des arborescences	
5 - 2 Formes fortes, Fuzzy sets et information	
5 - 3 Optimum global de V_k	
5 - 4 Approche de l'optimum global par changement d'arborescence	
6 - PROGRAMMATION DU TABLEAU DES FORMES FORTES ET INTERPRÉTATION HEURISTIQUE	21
7 - EXEMPLES D'APPLICATIONS	23
7 - 1 L'exemple artificiel de Ruspini	
7 - 2 Classement de sondages d'un gisement minier	
7 - 3 Recherche de profils biologiques	
CONCLUSION	27
BIBLIOGRAPHIE	28
ANNEXES	36



introduction

1.1 – LE PROBLEME

Dans différents domaines scientifiques (médecine, biologie, archéologie, économie etc.), il apparaît fréquemment de vastes ensembles d'objets représentés par un nombre fini de paramètres. Pour le spécialiste, l'obtention des groupements " naturels et homogènes ", ainsi que des éléments les " plus représentatifs " d'un tel ensemble constitue une étape importante dans la compréhension de ses données.

Une bonne approche de la solution de ce problème est fournie par les techniques de classification automatique qui consistent à trouver une partition d'un ensemble fini E telle que chaque objet ressemble plus aux objets intérieurs à son groupe, qu'aux objets extérieurs. En termes mathématiques, le problème peut s'énoncer sous l'une des deux formes suivantes ; étant donné un certain critère W :

A – Trouver la partition de E qui optimise W .

B -- Trouver la partition de E qui optimise W parmi toutes les partitions en K classes.

La famille de méthodes dont il sera question concerne surtout le problème B, mais elle pourra également aider le praticien dans la résolution du problème C suivant :

C – Chercher parmi toutes les partitions en K classes celles dont chaque classe aura le noyau le plus représentatif. (Un noyau est un groupe de points de la population à classifier)*.

Dans le paragraphe 1.2 nous donnerons sommairement les principales propriétés des méthodes des nuées dynamiques**. Cette famille de méthodes servira de modèle au but véritable de cette étude qui sera développé dans le paragraphe 1.3.

1.2 – LES METHODES DES NUÉES DYNAMIQUES

On se donne une fonction g permettant de transformer une partition de E en un ensemble fini de noyaux et une fonction f permettant le passage de plusieurs noyaux à une partition. Le principe de ces méthodes est simple, il consiste à appliquer de manière alternative les fonctions f et g à partir d'un choix initial de noyaux. Moyennant certaines hypothèses qui seront données, la décroissance du critère W est assurée jusqu'à la convergence. Le formalisme que nous donnons permet d'obtenir de nombreuses variantes de cette technique et notamment, comme cas particuliers, la méthode de HALL et BALL (1965) de FREEMAN (1969) et de DIDAY (1970). Nous avons pris cette famille de méthodes comme modèle de notre étude pour de multiples raisons.

a) Elles permettent d'éviter la mise en mémoire du tableau $\frac{N \cdot (N-1)}{2}$ (où $N = \text{card}(E)$)

des similarités des objets deux à deux. Cela permet le traitement de populations beaucoup plus importantes que par d'autres techniques plus classiques (SOKAL et SNEATH (1963), JOHNSON (1967), ROUX (1968), LERMAN (1970)).

* Le problème C est formalisé et un exemple simple est donné en 2.1.

** Voir également DIDAY E. (1970), DIDAY E. (1971).

- b) Ces techniques sont très rapides, par exemple la variante étudiée dans DIDAY (1970), permet le traitement sur IBM 360/91 d'une population de 900 objets caractérisés chacun par 35 paramètres en $3 \frac{1}{2}$ minutes.
- c) Ces techniques ne souffrent pas de l'effet de chaîne (voir JOHNSON (1967), ZAHN (1971)). Autrement dit, elles n'ont pas tendance à rapprocher deux points éloignés si ces deux points sont liés par une file serrée de points.
- d) Il n'est pas nécessaire de définir des seuils arbitraires pour la détermination des classes ni pour l'arrêt du processus, (voir SEBESTIEN (1966), BONNER (1964), HILL (1967) etc.).

L'utilisation des noyaux introduits de plus les avantages suivants :

- a) La possibilité de prendre des noyaux de plusieurs éléments de la population permet une reconnaissance plus efficace des formes. On pourra voir, par exemple, dans le chapitre VI de DIDAY (1971), comment un bon choix de f et g revient à construire des noyaux épousant la structure des formes de E, ce qui ne serait évidemment pas possible si ces noyaux étaient réduits à un seul point. L'utilisation de noyaux permet également un vaste choix des fonctions f et g, par exemple l'utilisation de la distance de Mahalanobis (voir ROMEDER (1969)) qui n'aurait pas de sens si chaque noyau était réduit à un point unique.
- b) En prenant pour noyaux des éléments de la population elle-même, plutôt que des centres de gravités, on évite l'effet artificiel que ces centres peuvent créer (cf. fig. 11). D'un autre côté, pour certains types de données, la notion de centre de gravité peut ne pas avoir de sens.
- c) L'utilisation de noyaux permet la réalisation de partitions autour des agglomérations à forte densité en atténuant l'effet des points marginaux (cf. fig. 14 et fig. 15).
- d) Signalons enfin, que l'utilisation des noyaux permet de donner des " optimum locaux " au problème C, permettant ainsi d'aider l'utilisateur intéressé par de bons échantillonnages.

1.3 – ETUDE SYNTHÉTIQUE DES SOLUTIONS OBTENUES

Toutes les techniques réalisables dont le but est de minimiser le critère W, fournissent des solutions dont rien ne prouve qu'elles soient optimales. Or, les différentes études faites récemment sur l'état actuel des recherches en " clustering " (voir BOLSHEV (1969), FISHER et VAN NESS (1971), BALL (1970), WATTANABÉ (1971), CORMACK (1971)) font ressortir l'inexistence d'étude synthétique des solutions obtenues pour un algorithme donné. C'est à cette étude que ce texte est consacré. Nous nous sommes restreint à un type particulier d'algorithme mais, évidemment, cette analyse pourrait s'étendre à d'autres techniques.

On appellera V_k l'ensemble des solutions possibles. Chaque solution* obtenue par un algorithme des nuées dynamiques est optimale vis-à-vis d'une certaine partie de V_k qui est une arborescence particulière. Cela conduit à donner une structuration à l'espace V_k . On montre en particulier que, sous certaines hypothèses, cet espace peut être partitionné en un nombre fini d'arborescences qui ont pour racine une solution stable dite "non-biaisée" et pour sommets pendants des éléments d'un certain type, appelés " éléments impasses ". On applique les différents résultats obtenus de la manière suivante :

- a) On construit une variable aléatoire permettant de se faire une idée réelle de la structure de V_k . On obtient ainsi un invariant intéressant pour de multiples raisons, notamment pour les données évoluant dans le temps et pour comparer l'efficacité des différentes techniques.

b) On définit différents types de "fuzzy-sets"* dans W_k : les formes "fortes" et "faibles" ainsi que les points "charnières". **Bien mieux que l'optimum global, ce sont à notre avis ces "fuzzy-sets" et les optimums locaux obtenus qui fourniront véritablement à l'utilisateur les différentes facettes de la réalité qu'il désire saisir.**

c) Nous donnons un nouveau type de techniques permettant, par passage d'une arborescence à l'autre, une approche de l'optimum global.

Les exemples d'application qui seront donnés font notamment ressortir l'intérêt des "formes fortes" qui sont un outil d'une grande utilité pour le praticien, en lui permettant d'extraire de sa population les groupes de points les plus significatifs.

Signalons enfin, que nous avons évité les développements théoriques, en nous restreignant aux résultats intéressants pour la compréhension et l'utilisation **informatique** des méthodes.

* Voir ZADEH (1965) et RUSPINI (1970).