# Approximation rapide et interpretation d'une partition centrale pour les algorithmes de partitionnement

Gilles Celeux

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 301

# APPROXIMATION RAPIDE ET INTERPRÉTATION D'UNE PARTITION CENTRALE POUR LES ALGORITHMES DE PARTITIONNEMENT

Gilles CELEUX

Mai 1984

# APPROXIMATION RAPIDE ET INTERPRETATION D'UNE PARTITION CENTRALE POUR LES ALGORITHMES DE PARTITIONNEMENT

Gilles CELEUX

Résumé

Les partitions obtenues par des algorithmes de classification de type nuées dynamiques dépendent de l'initialisation de la méthode.

Les stratégies les plus répandues pour s'affranchir de ce problème (sélection de la partition fournissant la meilleure valeur du critère, construction de la partition des formes fortes, ...) sont peu satisfaisantes.

A partir des diverses partitions obtenues par un algorithme de partitionnement, nous proposons une méthode très rapide pour construire une partition centrale.

Cette partition est une approximation de la partition qui optimise le critère majoritaire de Concordet. Vu le caractère particulier des similarités considérées, l'approximation obtenue est excellente.

Nous nous intéressons ensuite à l'utilisation de cette partition centrale, consensus entre les partitions obtenues.

On montre, en particulier, que cette partition centrale est très peu sensible au nombre de classes demandé lors de l'initialisation de l'algorithme de partitionnement.

Dans les cas où il existe un "bon" nombre (inconnu) de classes, pour l'algorithme considéré, la partition centrale permet de le retrouver ainsi que la partition associée sous des conditions très larges d'initialisation de l'algorithme de partitionnement.

Abstract

Clusters obtained by dynamic clustering algorithms depend on starting states.

The most used strategies to rid of this problem are unsatisfactory.

We propose a fast construction of a middle partition from those obtained by a clustering algorithm.

This middle partition approximates the middle partition which optimizes Condorcet's majority rule.

Applications of this strategy show that this middle partition is not very sensitive to starting states and to the asked number of clusters.

In particular, if data are well structured in k clusters, this middle partition gives back the right partition in k clusters under very large conditions on starting states of the clustering algorithm.

# APPROXIMATION RAPIDE ET INTERPRETATION D'UNE PARTITION CENTRALE POUR LES ALGORITHMES DE PARTITIONNEMENT

## 1 - Le problème :

Généralement les algorithmes de classification par partitionnement donnent des solutions qui dépendent de l'initialisation de la méthode.

C'est le cas, en particulier, pour les méthodes de partitionnement de la bibliothèque MODULAD (cf. [M083]) : méthodes des nuées dynamiques, classification croisée d'un tableau de contingence, d'un questionnaire, méthodes des boules optimisées, méthodes k-means, Isodata, méthode des transferts.

Dans cet article, nous proposons la construction d'une partition concensus entre les m partitions obtenues après m applications d'un algorithme quelconque de partitionnement.

Auparavant nous présentons les approches les plus répandues pour s'affranchir du problème d'initialisation en classification.

## 2 - Les stratégies actuelles :

La stratégie la plus répandue, car la plus simple, consiste à retenir la partition qui fournit la meilleure valeur du critère à optimiser par l'algorithme utilisé.

La pratique montre que la partition obtenue n'est pas très stable. C'est en particulier le cas si le nombre de classes demandé par l'utilisateur est inadequat avec la structure des données étudiées.

Une stratégie plus élaborée consiste à construire la partition des formes fortes [Di 72]. Cette partition est la partition intersection des m partitions obtenues. Elle correspond à un consensus unanime [BaleMo 83].

En pratique, cette partition s'avère trop fine pour être interprétée globalement. Seules les classes de cardinal élevé de cette partition permettent de dégager des regroupements de points très typés [Di 79, ch 23].

Une autre approche consiste à rechercher, par diverses heuristiques, une position initiale de l'algorithme de classification qui conduit à des résultats meilleurs que ceux obtenus après initialisation au hasard.

L'idée générale de ces heuristiques est de partir d'une partition dont les classes sont assez séparées les unes des autres pour la distance considérée.

Ces heuristiques sont assez lourdes à mettre en oeuvre, et n'atteignent pas en général leur but.

La méthode des pôles d'attraction [Le 81 chap. 8] , [MO 83] est plus utile pour une bonne initialisation des algorithmes de partitionnement où chaque classe de la partition cherchée est caractérisée par un point.

Par contre, rien ne permet d'affirmer que l'on obtiendra une meilleure valeur du critère en partant des pôles d'attraction plutôt que d'une partition au hasard.

Enfin, la méthode des pôles d'attraction exige le calcul du tableau de distance des objets à classer.

## 3 - Notion de partition centrale :

Soient m partition $P^1,\ldots,P^m$ sur un ensemble E de n objets. A chaque partition $P^\ell$ de E est associée la relation d'équivalence $v^\ell$ sur E définie par :

$$\forall (i,i') \in E \times E \qquad v^{\ell}(i,i') = \begin{cases} 1 \text{ si } i \text{ et } i' \text{ appartiennent à une même classe de } P^{\ell} \\ 0 \text{ sinon.} \end{cases}$$

Munissons l'ensemble $\mathcal{S}$ des partitions sur E de la distance de la différence symétrique :

$$\forall P, Q \in \mathcal{S} \qquad d(P,Q) = \frac{1}{2} \sum_{i,i' \in E} |v(i,i') - w(i,i')|$$

v étant la relation d'équivalence associée à P, w étant celle associée à Q.

On appelle partition centrale des partitions $(P^1, \ldots, P^m)$ la partition $P^*$ dont la relation d'équivalence associée $u^*$ minimise le critère :

$$C(u) = \frac{1}{2} \sum_{\ell=1}^{m} \sum_{i,i' \in E} |u(i,i') - v^{\ell}(i,i')|$$

$u \in U$ (ensemble des relations d'équivalence sur E)

Intuitivement, la partition centrale $P^*$ minimise le nombre de désaccords moyen avec les m partitions $(P^1 \ldots, P^m)$.

Le problème de recherche de la partition centrale revient donc à trouver la relation sur E, réflexive, symétrique et transitive qui minimise le critère C(u).

Ce problème a été considéré, en classification automatique par Régnier [Re 65] pour construire une partition d'objets décrits par m variables qualitatives.

En effet, chaque variable définit une partition sur E, la partition centrale approximera les relations induites par les variables qualitatives.

Dans ce cadre, Régnier a défini l'algorithme des transferts. Partant d'une partition P, cet algorithme construit une suite de partitions par transfert, à chaque pas, d'un objet d'une classe à l'autre. A chaque pas, l'algorithme choisit le transfert qui assure le meilleur gain du critère C. La solution obtenue à la convergence dépend de la position initiale.

L'intérêt de l'algorithme des transferts dépasse le cadre de recherche d'une partition centrale et peut s'appliquer à d'autres critères que celui mentionné ci-dessus [Le 81] , [MO 83].

Récemment, des auteurs ont situé le problème de recherche d'une partition centrale dans le cadre générale d'agrégation de relations binaires par une autre relation binaire particulière. Ils ont ainsi montré les liens existants entre les problèmes d'agrégation des préférences et d'agrégation de similarités (recherche d'une partition centrale) [Ba Mo 81], [Ma Mi 79].

A partir de ce cadre, des algorithmes performants de résolution du problème de recherche d'une partition centrale ont été proposé.

Marcotorchino et Michaud [Ma Mi 82] formulent le problème en un problème de programmation linéaire en nombres entiers. Pour la résolution effective ils remplacent les contraintes d'intégrité $u(i,i') = 0$ ou 1 par les contraintes $0 \leq u(i,i') \leq 1$. Ils traitent de manière analogue le cas de l'agrégation de préférence.

Pour ce dernier problème, Arditi [Ar 83] a proposé un algorithme de relaxation Lagrangienne.

Les résultats obtenue par ces deux méthodes semblent très bons.

Malheureusement, ils sont d'un coût prohibitif si le nombre d'objets à classer est important.

Notons enfin, que le problème de recherche d'une partition centrale est NP - complet. Il est donc impossible d'assurer, sauf cas particuliers, l'obtention d'un optimum exact.

## 4 - La partition centrale comme consensus entre partitions :

Soient m partitions $P^1, \ldots, P^m$ obtenues à l'issue de m passage d'un algorithme de partitionnement sur un ensemble E de n objets.

La partition centrale déduite de ces m partitions fournit un bon compromis entre les différents résultats obtenus.

Dans ce cadre, la recherche d'une partition centrale doit être extrêmement rapide. Il s'agit de fournir une information supplémentaire pour des algorithmes usuels de classification sans entrainer une détérioration de leur performances.

Il est donc exclu d'utiliser des algorithmes du type de ceux proposés dans [Ma Mi 82] ou [Ar 83].

On peut plus sérieusement envisager d'utiliser l'algorithme de transfert. Cela étant, l'algorithme de transfert est surtout un algorithme de classification.

Il serait assez lourd de l'intégrer systématiquement à des programmes de classification pour la recherche d'une partition centrale.

### 4.1. L'heuristique proposée :

Rappelons le problème de recherche de la partition centrale :

On cherche parmi les relations d'équivalence sur E, la relation $u^*$ qui minimise

$$c(u) = \frac{1}{2} \sum_{\ell=1}^{m} \sum_{i,i' \in E} |u(i,i') - v^{\ell}(i,i')|$$

avec $\forall \ell = 1, m$ $v^{\ell}$ relation d'équivalence associée à la partition $P^{\ell}$

On définit la matrice carrée a de dimension n par :

$$\forall i, \forall i' \in E \qquad a(i,i') = \sum_{\ell=1}^{m} v^{\ell}(i,i')$$

a(i,i') est le nombre de fois où les objets i et i' de E ont été classés ensemble dans les partitions $p^1,\ldots,p^m$

Pour simplifier l'exposé, nous allons supposer que m est impair (m = 2 p + 1, avec p entier supérieur ou égal à 1).

La généralisation ne pose pas de problème. Si l'on s'affranchit de la contrainte de transitivité, la solution optimale est obtenue par la relation $u^\circ$ associée à la règle majoritaire de Condorcet.

$$\forall i,i' \in E \quad u^\circ(i,i') = \begin{cases} 1 \text{ si } a(i,i') \geq p+1 \\ 0 \text{ sinon} \end{cases}$$

Cette relation $u^\circ$ n'est pas en général transitive . De manière analogue, on peut définir les relations $u^j$ , j variant de 1 à p par :

$$\forall i,i' \in E \quad u^j(i,i') = \begin{cases} 1 \text{ si } a(i,i') \geq p+1+j \\ 0 \text{ sinon.} \end{cases}$$

Ces relations ne sont pas en général transitives. Seule $u^p$ est transitive, la partition associée est la partition des formes fortes.

De même, on peut définir les relations $u^{-j}$, j variant de 1 à p+1, par :

$$\forall i,i' \in E \quad u^{-j}(i,i') = \begin{cases} 1 \text{ si } a(i,i') \geq p+1-j \\ 0 \text{ sinon.} \end{cases}$$

Ces relations ne sont pas en général transitives. Seule $u^{-(p+1)}$ est transitive, elle correspond à la partition grossière.

Il est facile de voir que :

$$c(u^\circ) \leq c(u^1) \leq \ldots \leq c(u^p)$$

et $$c(u^\circ) \leq c(u^{-1}) \quad \ldots \leq c(u^{-(p+1)})$$

Considérons la fermeture transitive de chacune de ces relations. A partir d'une relation $u^j$, on notera $\bar{u}^j$ sa fermeture transitive et $\Gamma^j$ la partition associée à $\bar{u}^j$.

La suite des $\Gamma^j$ constituent une suite de partitions emboitées et constitue une hiérarchie sur l'ensemble des partitions de E. (cf. [Di 72])

$$\Gamma^p \subseteq \Gamma^{p-1} \subseteq \ldots \Gamma^1 \subseteq \Gamma^\circ \subseteq \Gamma^{-1} \subseteq \ldots \subseteq \Gamma^{-p} \subseteq \Gamma^{-(p+1)}$$

On a noté la relation "plus fine que" par $\subseteq$ .

L'heuristique est la suivante :

Pour j variant de -(p+1) à p , détermination de $\bar{u}^j$ et calcul de $c(\bar{u}^j)$

La procédure ci-dessus est arrêtée dès que la relation déterminée est associée à la partition grossière sur E.

La partition central $\Gamma^*$ construite par cette heuristique est associée à la relation $\bar{u}^*$ qui vérifie :

$$c(\bar{u}^*) = \min\ (c(\bar{u}^j),\ j = -(p+1),p)$$

5 - Applications à des problèmes de classification sur variables qualitatives :

Insistons sur le fait que l'heuristique proposée ci-dessus n'a pas la prétention de rentrer en compétition avec des méthodes plus élaborées de recherche d'une partition centrale.

Elle a pour but de fournir rapidement un consensus entre partitions obtenues à l'issue d'un algorithme de classification. Dans ce cadre, sa simplicité n'est pas un handicap car les partitions considérées ont tendance à se ressembler de par la nature de leur obtention. C'est en particulier le cas, lorsque l'ensemble E des objets est "bien classifiable".

Il est toutefois important de voir comment cette heuristique se comporte pour des problèmes de classification sur variables qualitatives.

Exemple 1 : classification de félidés.

Cet exemple est tiré de [Ma Mi 82].

Il s'agit de 30 félins décrits par 14 variables qualitatives concernant des caractéristiques morphologiques et de comportement.

Par notre heuristique, la partition optimale $\Gamma^*$ est associé à la fermeture transitive de la relation $u_1$ définie par :

$$\forall i,i' \in E \quad u_1(i,i) = \begin{cases} 1 \text{ si } a(i,i') \geq 10 \\ 0 \text{ sinon} \end{cases}$$

La valeur du critère associé à $\Gamma^*$ est : $c(u_1) = 2124$

La borne inférieure, définie par $u^*$, du critère est : $c(u^*) = 1994$

La partition $\Gamma^*$ comporte 4 classes :

classe 1 : 1 à 4
classe 2 : 5,8,11
classe 3 : 6
classe 4 : 7,9,10,12 à 30.

Par leur méthode, Marcotorchino et Michaud obtiennent une partion $P^*$ en 4 classes :

classe 1 : 1,2
classe 2 : 3,4,5,7,8,11
classe 3 : 6
classe 4 : 9,10,12 à 30

La valeur du critère associé à $P^*$ est : $c(u_2) = 2094$

Commentaires :

Les deux partitions $\Gamma^*$ et $P^*$ sont très analogues :

- elles comportent toutes les deux le même nombre de classes.
- les classes 3 sont identiques, les classes 4 ne différent que par un élément.
- la différence la plus sensible concerne la classe 1. Nous avons regroupé les gros félins (lion (1), tigre (2)) avec le jaguar (3) et le léopard (4).

Malheureusement, nous n'avons pas d'élément pour faire une interprétation zoologique de la partition $\Gamma^*$.

- la détérioration du critère de $\Gamma^*$ par rapport à $P^*$ est faible.

Exemple 2 : classification de cétacés

Il s'agit d'un ensemble de 36 cétacés décrits par 15 variables qualitatives concernant des caractéristiques morphologiques et de comportement.

Cet ensemble de données avait été proposé, dans un but de confrontation de méthodes, au workshop d'analyse de données de Bruxelle (Juin 83).

Par notre heuristique, la partition optimale $\Gamma^*$ est associée à la fermeture transitive de la relation $u_1$ définie par :

$$\forall i,i' \in E \quad u_1(i,i') = \begin{cases} 1 \text{ si } a(i,i') \geq 11 \\ 0 \text{ sinon} \end{cases}$$

La valeur du critère associé à $\Gamma^*$ est $c(u_1) = 2715$

La borne inférieure, définie par $u^*$ , du critère est $c(u^*) = 2666$

La partition $\Gamma^*$ comporte 7 classes :

| | | |
|---|---|---|
| classe 1 : | 1,9,21 | (Baleen Whales) |
| classe 2 : | 2,3,8,18 | (Grey Whale, Finback Whales) |
| classe 3 : | 4,12,19,34,36 | (Beaked Whales) |
| classe 4 : | 5,7,10,11,15,17,22 à 25,28 à 32,35 | (Dolphins Porpoises) |
| classe 5 : | 13,16,27,33 | (River Dolphins) |
| classe 6 : | 6,20 | (White Whales) |
| classe 7 : | 14,26 | (Sperm Whales) |

Commentaires :

Cette classification est exactement la même que celle obtenue par la méthode de Marcotorchino et Michaud [Ve 83].

Un autre auteur, Hinde [Hi 83] a proposé une méthode, pour ces données, basée sur un critère de rapport de vraisemblance, conduisant à une suite de partitions

emboitées. Pour son critère le meilleur résultat conduit à une partition en 9 classes. Mais pour 7 classes, son résultat est exactement le même que le nôtre.

Par ailleurs, la partition est très satisfaisante du point de vue zoologique (cf. l'intitulé zoologique des classes obtenues).

Nous pensons que cet accord parfait entre différentes méthodes et l'interprétation zoologique aisée montre qu'il s'agit d'un problème facile.

Cela tend à prouver que dans de tels cas simples à traiter (en particulier celui de la recherche de consensus de classification) l'heuristique proposée est satisfaisante.

Pour terminer ce paragraphe, nous citerons la remarque suivante dûe à Lerman [Le 81] :

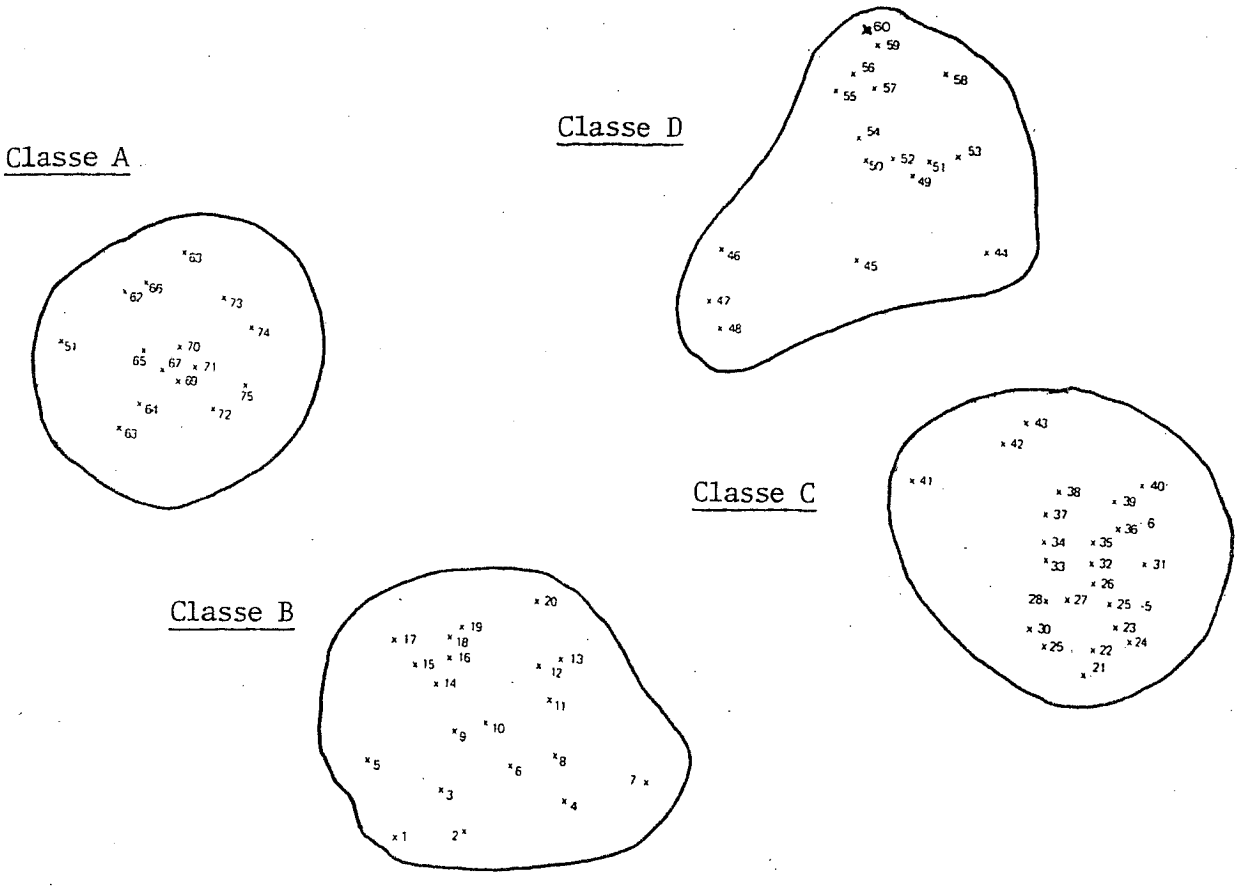
"Le point le plus faible de cette méthode (recherche d'une partition centrale) provient de ce que les différentes partitions définies par les différents caractères qualitatifs sont considérés, a priori, comme également discriminantes d'un point de vue statistique".

De fait, il existe bien d'autres méthodes intéressantes pour classer des objets décrits par des variables qualitatives (cf. [Le 81] , [Go 83],[Mo 83])

## 6 - Applications à la recherche de consensus en classification

Exemple 1 : Données de Ruspini.

Il s'agit de 75 points répartis dans le plan. Leurs images ci-dessous montre qu'il s'agit d'un ensemble de points bien classifiable, le bon nombre de classes étant quatre.




Nous avons utilisé notre heuristique à l'issue de 15 tirages de l'algorithme des nuées dynamiques, la distance étant la distance euclidienne usuelle (cf. [Di 79]).

Nous avons fait des essais en faisant varier le nombre de classes de 3 à 9, l'initialisation étant faite au hasard.

Les résultats sont résumés dans le tableau suivant :

| Nb de classes des partitions étudiées | Borne inférieure du critère | Critère pour la partition centrale | Nb de classes de la partition centrale |
|---|---|---|---|
| 3 | 4 837 | 4 837 | 3 |
| 4 | 5 617 | 5 617 | 4 |
| 5 | 7 070 | 7 070 | 4 |
| 6 | 4 399 | 4 399 | 4 |
| 7 | 5 720 | 6 092 | 4 |
| 8 | 4 274 | 4 830 | 4 |
| 9 | 4 058 | 4 744 | 10 |

Commentaires :

Lorsque la partition centrale a 4 classes, il s'agit de la partition "naturelle" visible sur la figure.

Lorsque la partition centrale a 3 classes, elle regroupe ensemble les classes A et B de la partition naturelle.

La partition centrale à 10 classes est inintéressante. Cet exemple, montre que la partition centrale est très peu sensible au nombre de classes demandé lors de l'initialisation de l'algorithme des nuées dynamiques.

Du moins, cette constatation est vraie pour un ensemble d'objets "bien classifiable".

Cet exemple semble montrer qui'il vaut mieux partir d'une sur-estimation du bon nombre de classes que d'une sous-estimation.

Il est remarquable que pour 3,4,5 et 6 classes la partition centrale est associée à la relation optimale $u^*$. De plus, lorsqu'on s'éloigne du bon nombre de classes l'écart entre le critère pour la partition centrale et la borne inférieure du critère augmente.

L'analyse de cet écart doit donc permettre de cerner le bon nombre de classes lorsqu'il existe.

Exemple 2 : Données médicales

Il s'agit d'un ensemble de 211 patients décrits par 11 paramètres quantitatifs, résultats de diverses analyses.

Ces patients sont atteints soit de bronchite, soit d'emphisème. Mais il y a une évolution entre ces deux diagnostics principaux et les médecins distinguent quatre groupes : Bronchite (B), bronchite dominante et emphisème (Be), emphisème dominant et bronchite (Eb), emphisème (E).

Notons que les frontières entre ces quatre groupes sont floues et qu'ils sont assez mélangés.

Ici nous ne nous intéressons pas au problème de discrimination (traité dans [Ce Le 82]).

La considération des groupes a priori nous permettra juste d'interpréter les classes des partitions centrales.

Nous avons construit une partition centrale à l'issue de 20 tirages de l'algorithme des nuées dynamiques, la distance étant la distance euclidienne usuelle.

Les essais ont été réalisés en faisant varier le nombre de classes de 3 à 7 classes, l'initialisation étant faite au hasard.

Les résultats sont résumés dans le tableau suivant :

| Nb. de classes des partitions étudiées | Borne inférieure du critère | Critère pour la partition centrale | Nb. de classes de la partition centrale |
|---|---|---|---|
| 3 | 80 304 | 106 594 | 10 |
| 4 | 49 351 | 52 011 | 4 |
| 5 | 53 571 | 60 395 | 14 |
| 6 | 35 030 | 36 992 | 16 |
| 7 | 41 040 | 46 254 | 29 |

Commentaires :

Contrairement à l'exemple précédent, la partition centrale obtenue ne se fixe pas sur un nombre de classes. Cela vient du fait que l'ensemble E considéré n'est pas classifiable de manière irréfutable.

L'examen du tableau ci-dessus conduit, cependant, à retenir la partition en 4 classes.

En effet, la partition centrale a également 4 classes, et la différence relative entre la valeur du critère obtenue et sa borne inférieure est le plus faible. Enfin, les autres partitions conduisent à des partition centrales ayant des nombres de classes élevées et sont difficiles à interpréter vis à vis, en particulier, des groupes a priori.

Nous donnons, ci-dessous, le tableau de contingence croisant les groupes a priori avec les 4 classes de la partition centrale :

| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | |
|---|---|---|---|---|---|
| B | 29 | 11 | 2 | 0 | 42 |
| $B_e$ | 16 | 38 | 25 | 8 | 87 |
| $E_b$ | 8 | 6 | 26 | 15 | 55 |
| E | 2 | 1 | 12 | 12 | 27 |
| | 55 | 56 | 65 | 35 | 211 |

Ce tableau montre que la partition centrale $P = (P_1, P_2, P_3, P_4)$ s'interprête aisément vis à vis des groupes a priori :

La classe $P_1$ est reliée au groupe B.
La classe $P_2$ est reliée au groupe Be.
La classe $P_3$ est reliée aux groupes Eb et Be.
La classe $P_4$ est reliée aux groupes E et Eb.

La construction d'une partition centrale à l'issue des nuées dynamiques a permis d'exhiber une partition synthètique et facile à interpréter. Ce n'est pas le cas de la partition des formes fortes qui contient 37 classes. De fait, les formes fortes sont avant tout un outil d'une grande utilité pour extraire d'une population les regroupements de points les plus significatifs (cf.[Di72]).

## 7 - Aspects informatiques

En théorie, pour pouvoir construire les relations en jeu pour obtenir une partition centrale, on doit disposer de la matrice carrée a de dimension n (nombre d'objets à classer).

En pratique, il n'est pas nécessaire de stocker cette matrice a.

Le programme travaille à partir du tableau des formes fortes (cf. [MO 83]). En effet, soit $F = (f_1, \ldots, f_q)$ $(q << n)$

Par définition, pour la recherche d'une partition centrale, les objets appartenant à une même forme forte sont équivalents car ils ont toujours été classés ensemble.

Il suffit donc de stocker la matrice carrée b de dimension q définie par :

$$\forall \ell, \ell' \in \{1,..,q\} \quad b(\ell, \ell) = \sum_{j=1}^{m} v^j(f_\ell, f_{\ell'})$$

$b(\ell,\ell')$ est le nombre de fois où les formes fortes $f_\ell$ et $f_{\ell'}$ ont été classées ensemble dans les partitions $(P^1, \ldots, P^m)$.

Le critère à minimiser prend alors la forme suivante :

$$C(u) = \frac{1}{2} \sum_{j=1}^{m} \sum_{\ell,\ell'} \text{card } f_\ell \cdot \text{card } f_{\ell'} \cdot |u(f_\ell, f_{\ell'}) - v^j(f_\ell, f_{\ell'})|$$

Si le nombre q de formes fortes n'est pas trop grand, on stocke la matrice carrée b de dimension q. Sinon, le programme calcule les accords entre les formes fortes à chaque fois qu'il en a besoin. Le surplus de coût, dans ce cas (rare en pratique) reste supportable comparé au temps mis pour calculer les m partitions par un algorithme de partitionnement.

## 8 - Conclusions :

L'heuristique présentée de recherche d'une partition centrale va être intégré à la bibliothèque MODULAD et au logiciel de classification automatique SICLA développé à l'INRIA.

Elle fournit rapidement une excellente approximation de la partition centrale (vu la nature des similarités en jeu) à partir des partitions considérées.

Du point de vue statistique, elle fournit une partition simple à interpréter qui est un bon résumé des différents tirages effectués par l'utilisateur.

On a vu d'autre part, qu'elle devait fournir, dans les cas où l'ensemble des objets à classer est bien classifiable, une bonne estimation du nombre de classes indépendamment de l'initialisation.

Incontestablement, ce point mérite d'être approfondi. On peut envisager de construire des indices statistiques, dérivés par exemple de l'écart relatif entre la valeur du critère pour la partition centrale et sa borne inférieure, permettant de juger du degré de classifiabilité d'une ensemble d'objets.

BIBLIOGRAPHIE :

[Ar 83] Arditi "Un nouvel algorithme de recherche d'un ordre induit par des comparaisons par paires" 3ème journées internationales d'analyse des données - Octobre 83.

[Ba Le Mo 83] Barthélémy, Leclerc, Monjardet "Quelques aspects du consensus en classification" 3ème journées internationales d'analyse des données - Octobre 83.

[Ba Mo 81] Barthélémy, Monjardet "The median procedure in cluster analysis and social choice theory" Math. Soc. Scien. 1 (81).

[Ce Le 82] Celeux, Lechevallier "Non parametric decision trees by bayesian approach". Compstat 82 - Springer-Verlag.

[Di 72] Diday "Optimisation en classification automatique et reconnaissance des formes" Note Scien. IRIA n° 6 (72).

[Di 79] Diday et collaborateurs "Optimisation en classification automatique" Ed. INRIA (79)

[Go 83] Govaert "Classification croisée" thèse d'état 83.

[Hi 83] Hinde "Descriptive classification of cetacae" Workshop on data analysis. Bruxelles Juin 83.

[Le 81] Lerman "Classification et analyse ordinale des données" Ed. Dunod (81).

[Ma Mi 79] Marcotorchino, Michaud "Optimisation en analyse ordinale des données" Ed. Masson (79)

[Ma Mi 82] Marcotorchino, Michaud "Agrégation de similarités en classification automatique" RSA. vol. 2 (82)

[MO 83] MODULAD "Bibliothèque FORTRAN pour l'analyse des données" Brochure de documentation version 1.0 (83)

[Re 65] Régnier "Sur quelques aspects mathématiques des problèmes de classification automatique" Mat. Sci. Hu. n° 82 - Eté 83.

[Ve 83] Vescia "Automatic classification of cetacae by similarity aggregation" Workshop on data analysis. Bruxelles-Juin 83

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique