REVUE DE STATISTIQUE APPLIQUÉE

E. DIDAY

Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques

Revue de statistique appliquée, tome 19, n° 2 (1971), p. 19-33

<http://www.numdam.org/item?id=RSA_1971__19_2_19_0>

UNE NOUVELLE MÉTHODE EN CLASSIFICATION AUTOMATIQUE ET RECONNAISSANCE DES FORMES LA MÉTHODE DES NUÉES DYNAMIQUES *

E. DIDAY

IRIA Département Informatique Appliquée

PLAN

---

* Je l'ai appelée par ailleurs "Dynamic clusters Method".

I - LE PROBLEME

Il s'agit de générer une partition à partir d'un corps de données sur lequel on ne demande pas nécessairement de faire d'hypothèse à priori. Cette partition doit réaliser au mieux les deux propriétés suivantes :

- les individus de chaque partie se ressemblent le plus possible,
- les individus de deux parties différentes se ressemblent le moins possible.

II - LES DIFFERENTES APPROCHES

Elles sont de trois types :

a) celles basées sur la recherche de la partition qui minimise un certain critère ; on se reportera par exemple à Regnier [1], Ruspini [2], Jensen [3]. Ces techniques nécessitent beaucoup de place mémoire et les calculs sont longs.

b) les techniques du type de Rocchio [4], Bonner [5], Hill [6] sont basées sur le choix d'un seuil ; contrairement aux techniques précédentes elles sont rapides et nécessitent peu de place mémoire, cependant la valeur de la partition est fortement liée au choix du seuil et au déroulement de l'algorithme qui dépend d'autres paramètres.

c) les techniques algorithmiques du type de Hall et Ball [7], Freeman [8], Diday [9] basées sur l'agrégation de groupes de points améliorée par itérations successives.

Nous ne parlons pas des méthodes de classification par hiérarchie ou d'analyse factorielle car le but direct de ces méthodes n'est pas de résoudre le problème dont il est question ici, bien qu'elles puissent lui apporter des informations enrichissantes (par exemple quant à la partition de départ de notre algorithme).

III - LA METHODE

Elle est caractérisée par des groupes de points partant par un processus itératif à la recherche des formes intéressantes, s'agrégeant en leur centre et permettant ainsi de les reconnaître.

III.1 - Quelques notations

Soit E l'ensemble des objets à classer

P l'ensemble des parties de E.

D une fonction définie sur $E \times P$ prenant ses valeurs dans $\mathbb{R}^+$

P_i l'ensemble des parties de E ayant n_i objets.

$E_1, E_2, \ldots, E_K$: K parties de E avec $E_i \in P_i$.

$L_K = (P_1 \times P_2 \times \ldots \times P_K)$

$T = \{ k \text{ entier} / 1 \leqslant k \leqslant K \}$

R : une application $E \times T \times L_K \longrightarrow \Gamma\mathbb{R}^+$. Elle sera appelée fonction d'agrégation-écartement.

III.2 - L'algorithme

Soit $L^{(n)} = (E_1^{(n)}, E_2^{(n)}, \ldots, E_K^{(n)})$ où $E_i^{(n)} \in P_i$ ainsi $L^{(n)} \in L_K$.

Connaissant $L^{(n)}$ le calcul de $L^{(n+1)}$ se fait comme suit :

1/ On calcule $D(x, E_i^{(n)})$ pour tout $x \in E$ et pour tout $i \in \{1, 2, \ldots, K\}$.

2/ Nous faisons une partition de E en K classes $C_i^{(n)}$ pour tout $i \in \{1, 2, \ldots, K\}$ comme suit :

$$C_i^{(n)} = \{x / D(x, E_i^{(n)}) < D(x, E_j^{(n)}) \ \forall j \in \{1, 2, \ldots, K\} \ j \neq i\}.$$

Nous supposerons pour simplifier que $D(x, E_i^{(n)}) \neq D(x, E_j^{(n)})$ pour $i \neq j$. (Le cas général est traité dans [9]).

3/ Nous définissons $L^{(n+1)} = (E_1^{(n+1)}, E_2^{(n+1)}, \ldots, E_K^{(n+1)})$ comme suit :

$E_i^{(n+1)}$ = les n_i objets de E qui minimisent $R(x, i, L^{(n)})$ pour

$$i = 1, 2, \ldots, K.$$

Exemples de choix de fonction d'agrégation-écartement :

Soit $L = (E_1, \ldots, E_K) \in L_K$.

Exemple 1 :

$$R(x, i, L) = \frac{D(x, E_i) \cdot D(x, C_i)}{\left[\sum_{j=1}^{K} D(x, E_j)\right]^2}$$

Appelons les éléments de E_i les étalons de la $i^{ième}$ classe.

$D(x, E_i)$ au numérateur aura pour effet d'agréger les étalons.

$D(x, C_i)$ aura pour effet de ramener les étalons vers le centre de leur classe.

$\sum_{j=1}^{K} D(x, E_j)$ au dénominateur aura pour effet d'écarter les E_j entre eux.

Exemple 2 :

$$R(x, k, L) = D(x, C_k)$$

III.3 - Etude de la convergence

Soit $E_k = \{e_{ik} \in E / i = 1, 2, \ldots, n_k\}$

$$L = (E_1, E_2, \ldots, E_K)$$

Définition d'un élément améliorant

On dira que $x \in \complement_E E_k$ est un élément améliorant de E_k dans E si $\exists\, i \in \{1, 2, \ldots, n_k\}$ tel que :

$$R(x, k, L) < R(e_{ik}, k, L)$$

e_{ik} sera appelé élément amélioré.

Définition d'un optimum local

Nous dirons que $L = (E_1, E_2, \ldots, E_K) \in L_K$ forme un optimum local si $\forall\, k \in \{1, 2, \ldots, K\}$, E_k n'a pas d'élément améliorant.

Définition de la convergence

La suite $L^{(n)}$ (définie en III.2) sera dite convergente et L sera sa limite s'il existe N : $\forall n > N \quad L^{(n)} = L$.

Définition de la suite U_p

Soit S l'application $L_K \times L_K \longrightarrow \Gamma\mathbf{R}^+$ telle que :

$$S(L, L^{(0)}) = \sum_{k=1}^{K} \sum_{x \in E_k} R(x, k, L^{(0)})$$

La suite U_p est définie comme suit à partir de la suite L_n :

$$U_0 = S(L^{(0)}, L^{(0)})$$

Si p est pair : $p = 2n$

$$U_p = S(L^{(n)}, L^{(n)})$$

Si p est impair : $p = 2n - 1$

$$U_p = S(L^{(n)}, L^{(n-1)})$$

Définition d'une fonction R carrée

Nous dirons que l'application R est carrée sur E, si $\forall$ L et M $\in L_K$ on a l'implication :

$$S(L, M) \leqslant S(M, M) \Longrightarrow S(L, L) \leqslant S(L, M) .$$

Exemple 1 :

Prenons $L_K = \Gamma\mathbf{R}^+$ et $S(L, M) = L \cdot M$ alors R est carrée car

$$S(L, M) \leqslant S(M, M) \Longrightarrow L \cdot M \leqslant M \cdot M \Longrightarrow L \leqslant M \Longrightarrow L \cdot L \leqslant L \cdot M \Longrightarrow S(L, L) \leqslant S(L, M) .$$

Exemple 2 :

Soit d une application symétrique de $E \times E \longrightarrow \Gamma\mathbf{R}^+$

$$D(x, E) = \sum_{y \in E} d(x, y)$$

et soit

$$R(x , k , L) = D(x , C_k) \quad \text{où} \quad L = (E_1 , E_2 , \ldots . E_K)$$

On montre alors (voir [9]) que R est carrée

Quelques propositions de convergences

On a démontré dans [9] successivement les trois propositions suivantes :

. Proposition 1 :

Si R est carrée la suite U_n converge en décroissant.

Proposition 2 :

Si la suite $L^{(n)}$ converge sa limite est un optimum local.

Proposition 3 :

Si la suite U_n converge alors la suite $L^{(n)}$ converge.

Théorème :

Si R est carrée la suite $L^{(n)}$ converge et sa limite est un optimum local.

Remarque :

Dans la pratique on s'aperçoit que la méthode est généralement convergente même si R n'est pas carrée.

IV - LE PROGRAMME*

IV.1 - En entrée

a) Le tableau de données.

b) Le nombre maximum de classes désirées : K.

c) Le nombre d'étalons par classe : n_i = card (E_i).

d) Optionnellement : si on dispose d'indications sur la typologie probable, les étalons peuvent être choisis expressément.

Dans le cas général où on ne dispose pas d'informations à priori les E_i sont tirés automatiquement au hasard.

IV.2 - En sortie

a) La partition obtenue : C_1, C_2, ..., C_K.

b) Le noyau E_i de chaque partie C_i.

c) Une mesure de l'homogénéité de chacune des classes obtenues.

d) Une mesure de la valeur de la partition obtenue.

e) Optionnellement : le degré de similarité de chaque individu à chaque classe.

---

* On pourra l'obtenir sur simple demande à l'IRIA - Rocquencourt - 78.

IV.3 - Test de convergence

Si p est pair on arrête les calculs quand :

$\left|1 - \frac{U_{p+1}}{U_p}\right| < \varepsilon$, si on veut être sévère on prendra ε de l'ordre de précision de la machine.

IV.4 - Distances utilisées

La méthode a été testée avec la distance du χ^2, les deux distances de Sebestien indiquées dans la thèse de Romeder [10] et avec la distance euclidienne. Il serait intéressant de l'utiliser avec des indices de similarité.

IV.5 - Performances

L'algorithme nécessite le calcul de D(x , E_k) et de R(x , i , L) pour tout $i \in \{1 , 2 , \ldots , K\}$ et $x \in E$. Un bon choix de D évite le calcul des distances deux à deux. Cela économise la mise en mémoire d'un tableau $\frac{N(N-1)}{2}$ (où N est le nombre d'objets) et aussi un gain appréciable de temps. Remarquons au passage que le programme nécessite seulement le calcul des n_i plus petites valeurs de R(x , i , L) quand x varie dans E, ainsi le programme ne calcule pas R(x , i , L) pour tout $x \in E$ quand $i = 1 , 2 , \ldots , K$.

Place mémoire nécessitée : elle est de l'ordre de $N \times M + 5N$.
Rapidité : une population de 900 individus caractérisés par 70 paramètres a donné 20 classes satisfaisantes en 3 mn 1/2 sur 360/91.
Une classification de 1800 individus caractérisés chacun par trois paramètres a été traitée en 28 secondes sur 6.600 CDC. Sur les données de Freeman [8] une analyse factorielle des correspondances en 10070 CII a mis autant de temps que 10 passages de notre méthode (en changeant les étalons de départ).

IV.6 - Informations complémentaires

La grande rapidité de la méthode permet d'obtenir en plus de la partition un certain nombre d'informations complémentaires, en changeant le choix des étalons au départ (les E_i).

Classifiabilité

Si après plusieurs tirages (10 par exemple) on obtient sensiblement les mêmes groupes on peut affirmer que les données sont classifiables.

Individus charnières

Ce sont les individus qui suivant les tirages oscillent d'une classe à l'autre.

Formes "fortes" et formes "faibles"

Un groupe de la partition sera appelé forme forte s'il apparaît bien distinctement quel que soit le tirage de départ ; les groupes appelés "forme faible" sont ceux qui ont tendance à se mélanger suivant les tirages.

V - DISCUSSION

Par rapport à la plupart des méthodes de "clustering" la méthode ne nécessite pas l'utilisation d'un seuil arbitraire pour la formation des classes ; cependant il faut fixer le nombre n_i d'individus capables de caractériser leur classe, en prenant par exemple $n_i = \frac{3}{4} \cdot \frac{N}{K}$, nous faisons simplement l'hypothèse qu'une population peut être caractérisée par les 3/4 de ses individus. Dans la pratique on constate que la valeur des n_i à moins d'être très petite n'influe pas beaucoup sur les résultats surtout s'il existe effectivement des formes "fortes".

Toutes les distances ne permettent pas de choisir R carrée cependant dans ce cas la méthode converge en général ;

Dans tous les cas on aura une bonne solution en prenant la partition qui donne la plus petite valeur à $S(L^{(n)}, L^{(n)})$ car cette quantité est proportionnelle au degré d'agrégation des classes et à leur écartement, si R est bien choisie. Remarquons ici que S décroit à chaque itération si R est carrée d'après la proposition 1.

Signalons enfin que quand K est trop grand par rapport au nombre de classes qui existent effectivement, des classes vides apparaissent.

VI - UN EXEMPLE ARTIFICIEL (Réalisé par Monsieur Barré)

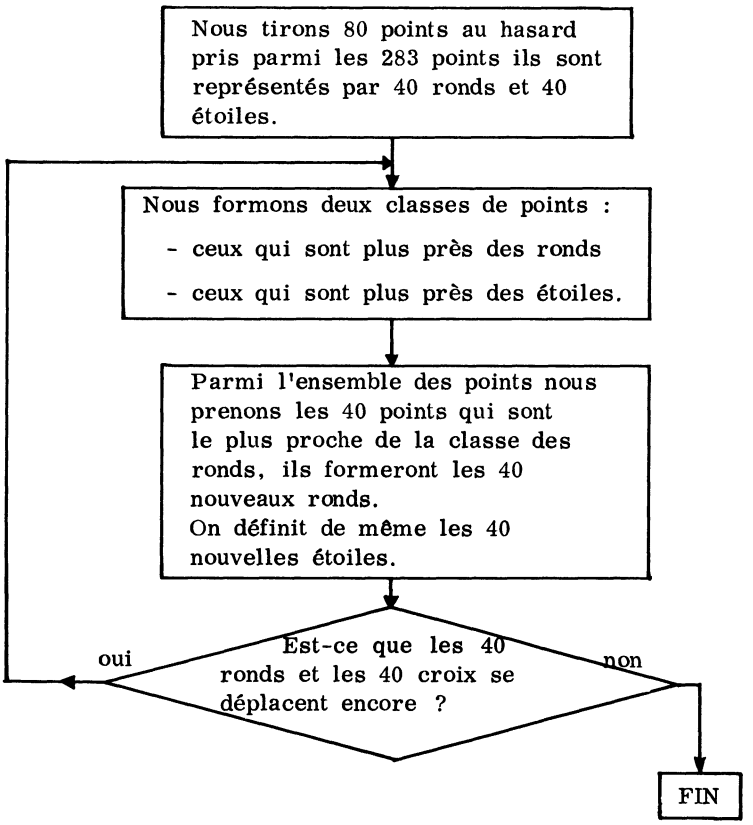
Le problème consiste à chercher les formes décrites par 283 points du plan ; nous avons utilisé la distance de Sebestien et pris 40 étalons pour chacun des deux groupes demandés.

L'organigramme schématise la méthode pour le choix de R qui a été fait : $R(x , i , L) = D(i , L)$.

La figure 1 montre les 283 points ; les 40 points de $E_1^{(0)}$ sont représentés par des ronds, $E_2^{(0)}$ est représenté par des étoiles. Ces 80 points ont été obtenus d'après un tirage au hasard.

La figure 2 montre les étalons tels qu'ils ont été obtenus à la fin de la première itération.

La figure 3 montre les étalons tels qu'ils ont été obtenus à la convergence. Les deux formes que l'œil reconnaissait ont été trouvées automatiquement. On voit l'intérêt des étalons qui jouent le rôle de "squelette" ou de sorte "d'axe factoriel discret" pour chacune des formes reconnues. En prenant trop peu d'étalons ou seulement le centre de gravité on aurait "arrondi" les formes et il n'aurait donc pas été possible de reconnaître les deux formes allongées.



IDEE INTUITIVE DE LA METHODE

VII - UNE APPLICATION EN NEUROLOGIE

Les données ont été fournies par le Docteur Marc-Vergnes*. Il s'agissait de vérifier si 25 paramètres caractéristiques du métabolisme énergétique du cerveau chez l'homme permettaient de prévoir l'évolution de malades vasculaires-cérébraux. On trouvera une étude détaillée du problème dans 11 .

Nous avons appliqué la méthode des nuées dynamiques de la façon suivante :

$$E = 70 \text{ individus}$$

$$L^{(0)} = (E_1^{(0)}, E_2^{(0)}, E_3^{(0)}) \text{ avec}$$

$$E_1^{(0)} = 15 \text{ individus tirés au hasard parmi les } 70.$$

---

* Service du Pr Gueraud - Hôpital Purpan - Toulouse.

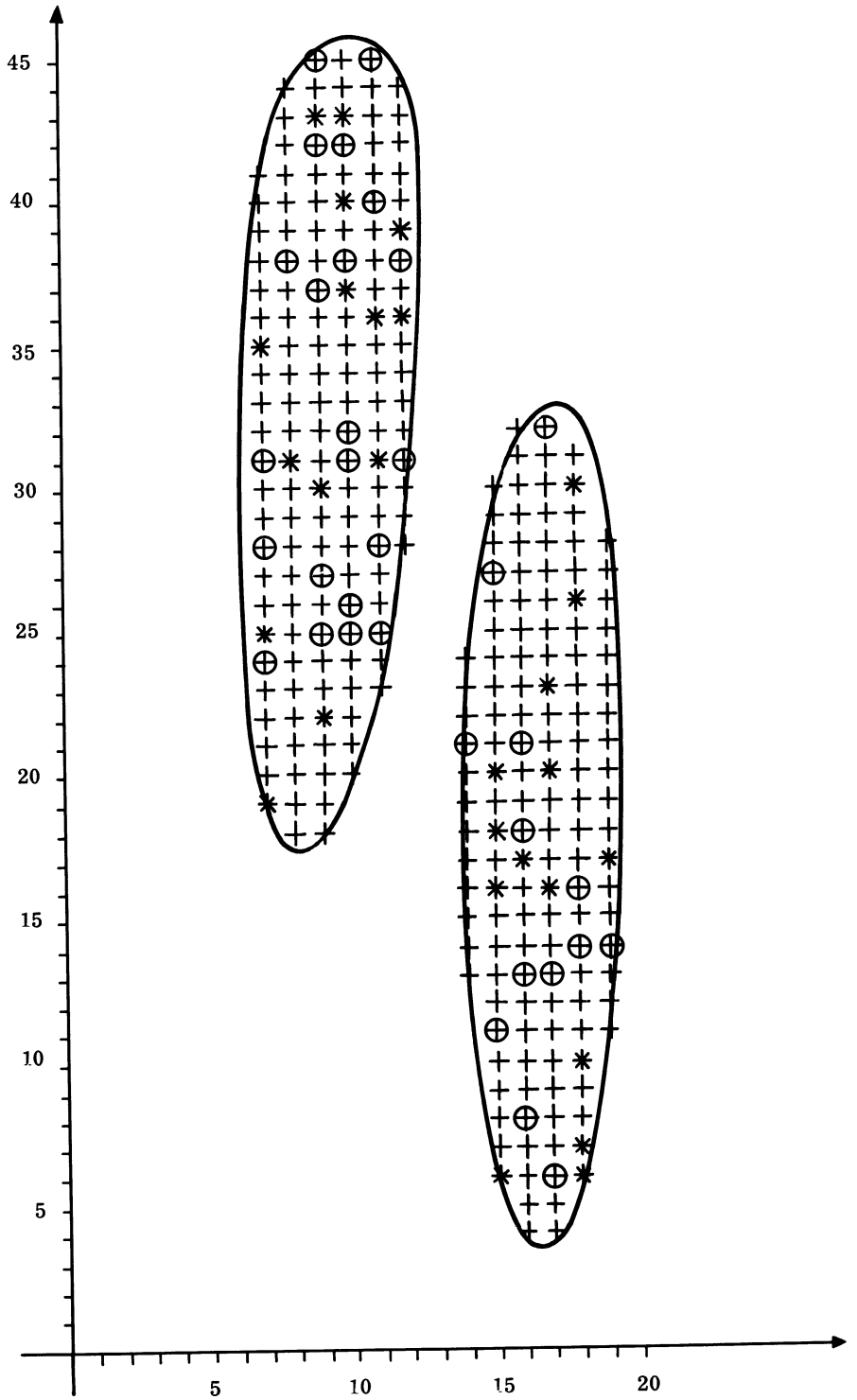


Fig. 1

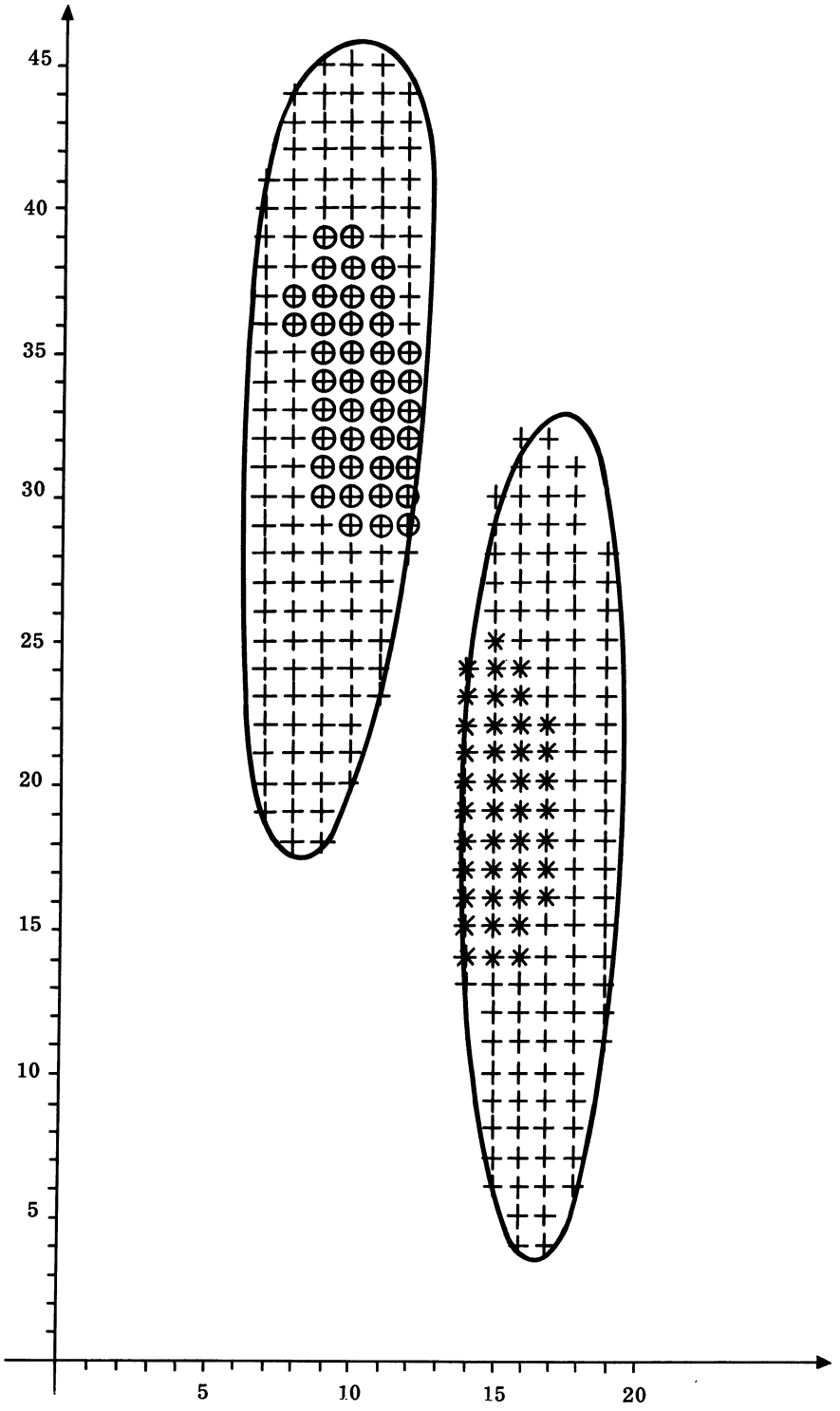


Fig. 2

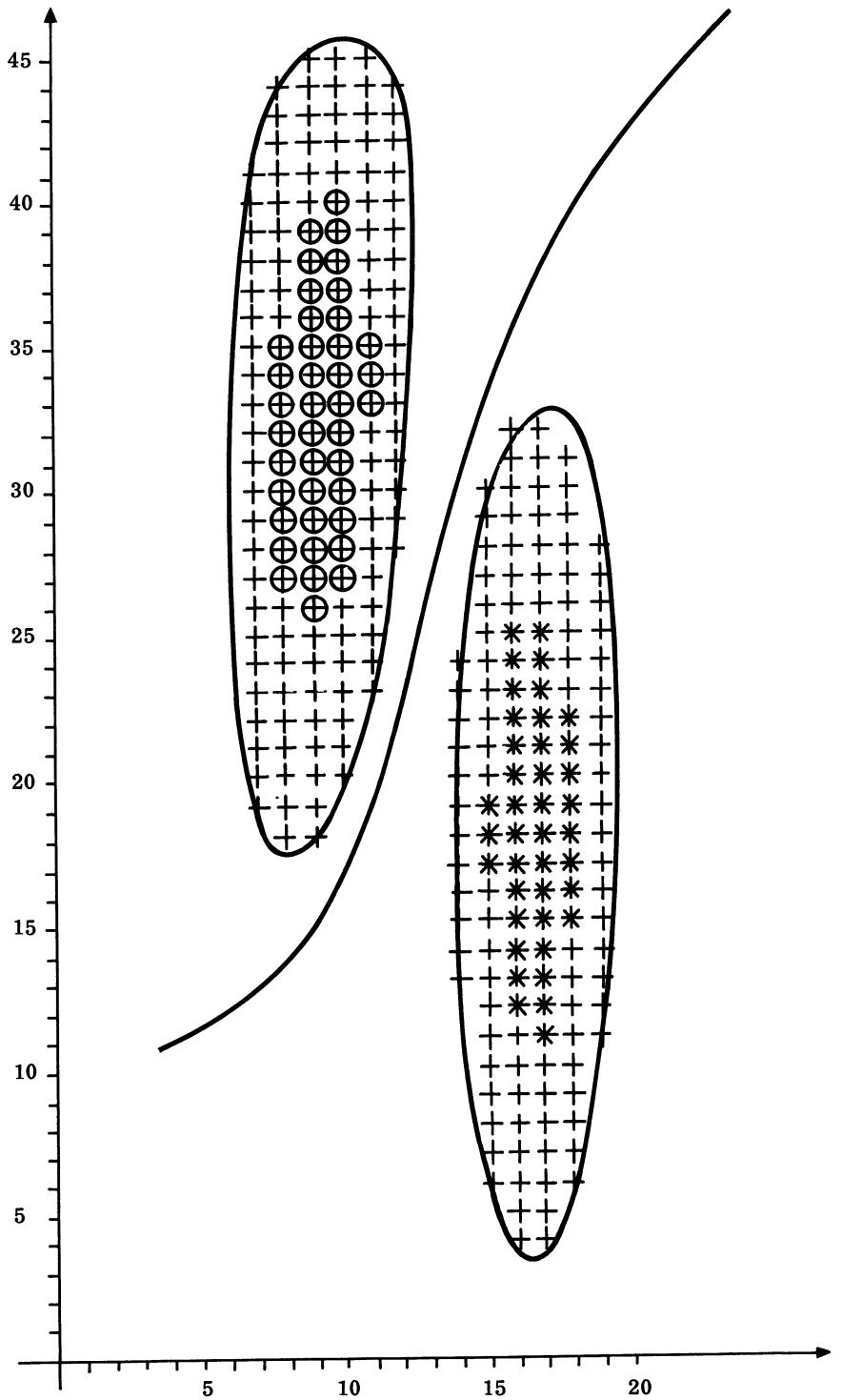


Fig. 3

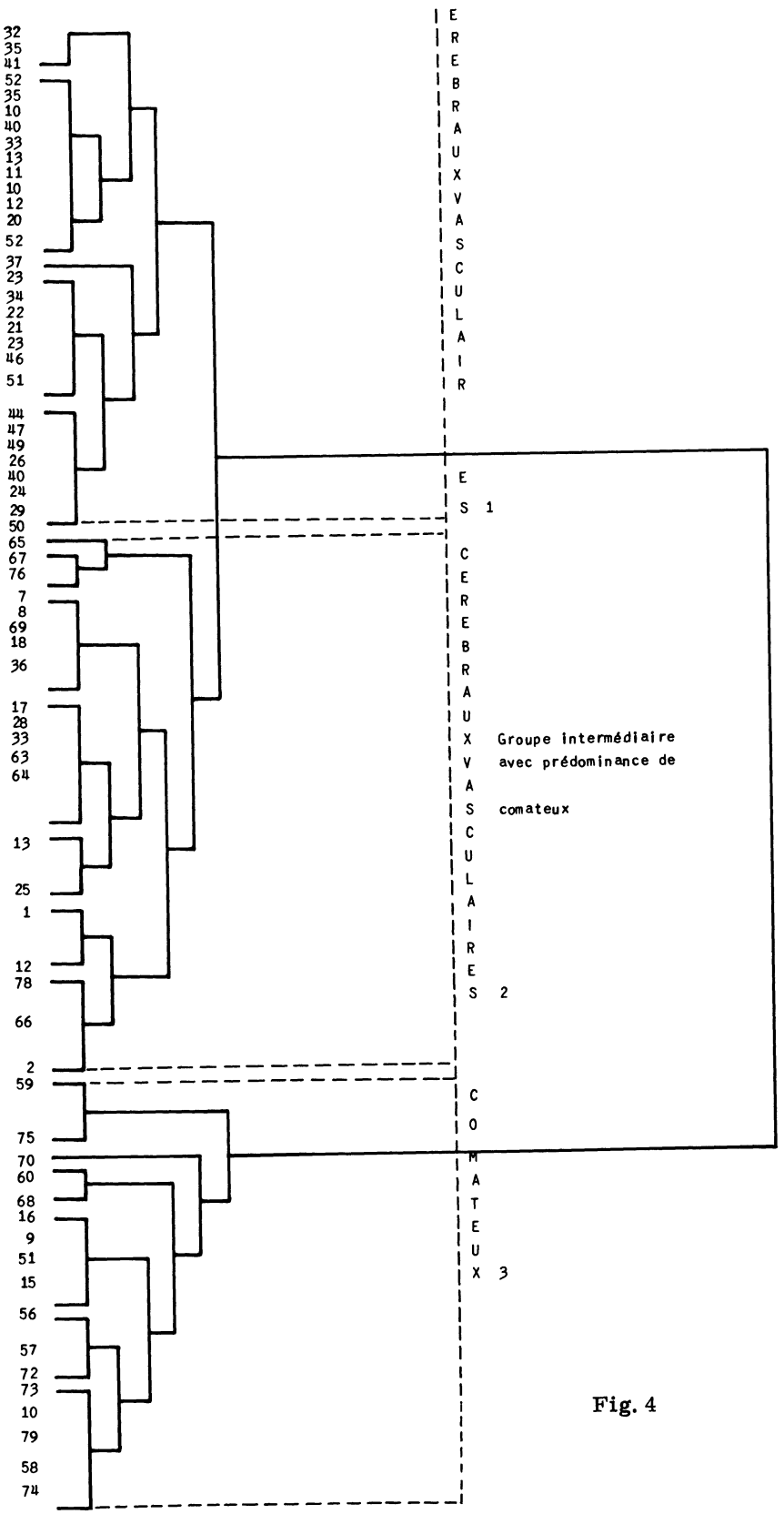

32
35
41
52
35
10
40
33
13
11
10
12
20
52
37
23
34
22
21
23
46
51
44
47
49
26
40
24
29
50
65
67
76
7
8
69
18
36
17
28
33
63
64
13
25
1
12
78
66
2
59
75
70
60
68
16
9
51
15
56
57
72
73
10
79
58
74
E R E B R A U X V A S C U L A I R
E S 1
C E R E B R A U X V A S C U L A I R E S 2
C O M A T E U X 3
Groupe intermédiaire
avec prédominance de
comateux

Fig. 4

Les résultats obtenus se recoupent exactement avec ceux donnés par l'arbre d'une analyse hiérarchique* (voir fig. 4 et Tableau 2). On trouve trois groupes à la convergence ; on constate après coup que ces groupes permettent une certaine prévision quant à l'évolution de la maladie, en effet :

Dans le premier groupe seuls quatre malades sont décédés alors que dans les deux autres groupes, on ne trouve que cinq évolutions favorables.

TABLEAU 2

Neurologie

Les trois classes obtenues par la méthode des nuées dynamiques

GROUPES

Groupe I		Groupe II	Groupe III
51	49	65	59
32	26	67	75
35	40	76	70
77	24	7	60
80	29	8	68
38	50	69	16
30		N 18	9
46		36	81
39		17	15
43		28	56
11		33	57
14		63	72
N 19		64	73
31		13	10
42		25	79
N 20		1	58
52		12	74
37		78	
23		66	
34		2	
N 22			
N 21			
27			
48			
51			
44			
47			

L'arbre donné par la hiérarchie intriguait le Docteur Marc-Vergnes car cet arbre laissait supposer que le deuxième groupe était plus proche du premier que du troisième.

---

* Méthode de Johnson.

Nous avons fait 10 tirages différents au départ en demandant deux classes : on retrouvait à chaque fois (à 3 ou 4 individus près) le groupe 1 comme première classe, la deuxième classe étant constituée des groupes 2 et 3. Ainsi on pouvait dire que le groupe 1 constituait une forme forte vis-à-vis des groupes 2 et 3 réunis.

Faisant à nouveau 10 tirages mais cette fois-ci en demandant que le nombre maximum de classes soit trois, on obtenait (à 3 ou 4 individus près) le groupe 1 séparé des deux autres, ces deux derniers apparaissant souvent très mélangés : ainsi on pouvait conclure que le groupe 1 constituait une forme forte vis-à-vis des groupes 2 et 3, ces deux derniers formant des formes faibles l'un vis-à-vis de l'autre.

CONCLUSION

La méthode a été utilisée avec succès sur des exemples concrets (médecine, sociologie, architecture, documentation automatique) permettant de traiter de grands tableaux de données (les techniques classiques ne peuvent dépasser en général 350 objets) ; pour les petits tableaux, elle est venue confirmer et parfois enrichir les résultats obtenus par des méthodes plus classiques pour lesquelles le praticien reste souvent perplexe devant la rigidité et le peu d'informations qui lui sont apportées sur la solution proposée, aussi nous avons l'espoir d'apporter par notre méthode un outil supplémentaire pour approcher un peu plus la réalité vivante.

Cette étude a été réalisée à l'Institut de Recherche d'Informatique et d'Automatique dans le cadre du Département Informatique Appliquée dirigé par Monsieur Donio que je tiens à remercier ainsi que les stagiaires du DEA de statistique mathématique qui ont utilisé cette méthode sur de nombreux exemples concrets permettant ainsi de confirmer son intérêt.

BIBLIOGRAPHIE

[1] S. REGNIER - Sur quelques aspects mathématiques des problèmes de classification automatique. I.C.C. Bulletin 1965. Vol. 4, pp. 175-191.

[2] E.H. RUSPINI - Numerical Method For Fuzzy Clustering Information Sciences 2 (1970) 319-350.

[3] Robert E. JENSEN - A dynamic programming algorithm for cluster analysis 1969 - University of Maine - Orono Maine.

[4] J.J. ROCCHIO - Harvard University Doctoral Thesis, Report N° ISR-10. To the National Science Foundation, Avril 66.

[5] R.E. BONNER - On some clustering techniques, IBM Journal of Research and Development. Vol. 8, Jan 1964, N° 1, pp. 22-32.

[6] D.R. HILL - Mechanized Information storage, retrieval and dissemination Proccedings of the F.I.D/I.F.I.P. - Joint Conférence Rome June 14 - 17, 1967.

[7] H. BALL et J. HALL - A clustering technique for summerizing Multivariate Data, Behavioural Science. Vol. 12, N° 2, Mars 1967.

[8] FREEMAN - Experiment in discrimination and classification. Vol. 1. Pattern Recognition Journal 1969.

[9] E. DIDAY - La Méthode des nuées dynamiques et la reconnaissance des formes Fascicule D.I.A., IRIA Rocquencourt (78).

[10] ROMEDER - Thèse de 3e Cycle - I.S.U.P. Faculté des Sciences de Paris, quai St-Bernard.

[11] E. DIDAY, J.C. JACOB, J. PICARD, A. SCHROEDER - Etude statistique sur une enquête concernant le métabolisme énergétique du cerveau. Fascicule DIA, IRIA.