Statistique et analyse des données

André Hardy
Jean-Paul Rasson

Une nouvelle approche des problèmes de classification automatique

Statistique et analyse des données, tome 7, n° 2 (1982), p. 41-56.
<http://www.numdam.org/item?id=SAD_1982__7_2_41_0>

Statistiques et Analyse de Données
1982 - Vol. 7 n° 2 pp. 41-56

UNE NOUVELLE APPROCHE DES PROBLEMES DE CLASSIFICATION AUTOMATIQUE

André HARDY - Jean-Paul RASSON

Unité de Statistique
Département de Mathématique
Facultés Universitaires Notre-Dame de la Paix à Namur
B-5000 Namur Belgique

Résumé : *Nous supposons que le problème de classification est naturel, c'est-à-dire que nous observons N points résultant d'un processus de Poisson dans K domaines de $\mathbb{R}^n$. Pour que ce problème soit bien posé statistiquement, nous devons imposer que ces domaines soient convexes. Nous prouvons que la solution de vraisemblance maximale est constituée par les K groupes de points tels que la somme des mesures de Lebesgue de leurs enveloppes convexes soit minimale. Nous proposons alors un algorithme, optimal au sens de Hartigan et montrons que notre méthode satisfait aux conditions d'admissibilité de Fisher et Van Ness. Nous analysons enfin les résultats obtenus par l'application de notre nouvelle procédure de classification à des ensembles de données "tests" parus dans la littérature scientifique.*

Abstract : *We suppose that the clustering problem is natural, i.e. that we observe N points resulting from a Poisson Process in K domains of $\mathbb{R}^n$. For that problem to be statistically well-stated, we must take the assumption that the domains are convex. We prove that the maximum likelihood solution is constituted by the k groups of points such that the sum of the Lebesgue measures of their convex hulls is minimal. We then prove that our procedure fulfils the admissibility conditions proposed by Fisher and Van Ness. We then analyze the results obtained by the application of our new clustering procedure to some data sets "tests" published in the scientific literature.*

Mots clés : *Admissibilité, Algorithme de Fisher, Classification, Enveloppe convexe, Maximum de vraisemblance, Mesure de Lebesgue, Processus de Poisson.*

1 - INTRODUCTION.

Dans différents domaines scientifiques (physique, biologie, économie, géophysique, géologie, psychiatrie, médecine, ...), les chercheurs sont confrontés à d'immenses ensembles de données dont ils désireraient tirer une information "profitable". Le problème posé est d'examiner la structure de l'ensemble de données et, plus particulièrement, voir si les objets se répartissent "naturellement" en un certain nombre restreint de groupes d'objets semblables.

Une bonne approche de ce problème peut être fournie par les techniques de classification automatique. Celles-ci consistent à découvrir une partition de l'ensemble fini des objets en sous-ensembles, que nous appellerons "classes", de telle façon que les membres de chaque classe aient certaines caractéristiques communes qui les distinguent des membres des autres classes.

Nous ne supposons donc pas que les objets peuvent raisonnablement se répartir en groupes d'objets semblables ! La conclusion d'une étude de classification peut être que cette division n'est pas possible. Mais, si des groupes distincts existent, ils doivent être découverts.

2 - LE PROBLEME DE CLASSIFICATION.

Soit $E = \{x_1, \ldots, x_n\}$ l'ensemble des individus à classer. On suppose $\#E < \infty$. Sur chaque individu, on observe la valeur de m variables. Chaque individu pourra donc être considéré comme un point dans un espace m-dimensionnel.

Le problème de classification abordé est le suivant :

- on désire trouver une partition de l'ensemble E en k classes disjointes $C_1, C_2, \ldots, C_k$;
  soit P_k l'ensemble de toutes ces partitions en k classes;
- pour que le problème soit bien défini mathématiquement, nous associons, à tout P dans P_k, la valeur d'un critère de classification W qui mesure la qualité de chaque partition P ;
- problème : trouver la partition qui minimise la valeur du critère parmi l'ensemble des partitions en k classes.

Le nouveau critère de classification considéré est le suivant : les enveloppes convexes des k classes obtenues sont disjointes et la somme des mesures de ces enveloppes est minimale.

Le problème se laisse écrire de la façon suivante :

$$W : P_k \to \mathbb{R}^+$$

$$P = \{C_1, C_2, \ldots, C_k\} \rightsquigarrow W(P) = \sum_{i=1}^{k} m(C_i)$$

où $m(C_i)$ est la mesure de Lebesgue de l'enveloppe convexe de la classe C_i .

On recherche la partition P^* telle que

$$W(P^*) = \min_{P \in P_k} \sum_{i=1}^{k} m(C_i) \quad .$$

L'originalité du critère vient du fait que la plupart des procédures en classification automatique généralise la notion de distance sur $\mathbb{R}$ en une distance ou une mesure de similarité sur $\mathbb{R}^m$. Nous généralisons la mesure de Lebesgue sur $\mathbb{R}$ en la mesure de Lebesgue sur $\mathbb{R}^m$ (distance sur la droite, surface dans le plan, ...).

3 - DESCRIPTION ET JUSTIFICATION DU MODELE PROBABILISTE.

3.1 - Introduction au modèle.

On peut considérer qu'une partie non négligeable des problèmes de classification concerne les répartitions, plus ou moins aléatoires, de points dans un espace euclidien $\mathbb{R}^k$. Celles-ci peuvent être modélisées par la théorie des processus ponctuels [1] .

Celle-ci n'a pas d'autre prétention que de prouver que, pour ce faire, il suffit de mettre une loi de probabilité sur l'espace probabilisé canoniquement associé à l'ensemble des répartitions ponctuelles sur l'espace euclidien considéré et que, si l'on veut trouver des densités ou exprimer des stationnarités, ceci se fait toujours par rapport à une mesure de base qui est la mesure de Lebesgue sur cet espace [2] .

3.2 - Le modèle probabiliste.

Le modèle que nous utiliserons se basera sur les hypothèses suivantes :

- les variables aléatoires qui comptent les nombres de points dans des régions disjointes sont indépendantes;
- le nombre moyen de points dans chaque région est proportionnel à la mesure de Lebesgue de cette région.

Il existe un seul processus ponctuel satisfaisant à ces deux conditions : c'est le Processus de Poisson Stationnaire [1].

Nous considérerons donc que nous avons affaire à un problème de classification lorsque, par hypothèse, les points que nous observons, engendrés par ce processus, sont distribués dans k domaines disjoints $(D_i)_{1 \leq i \leq k}$ que nous voulons retrouver. Puisque les n points que nous observons se trouvent certainement dans l'union, soit D, des k domaines disjoints D_i, ils y sont distribués indépendamment et uniformément.

3.3 - Propriétés du modèle. Justification.

Ce modèle, qui apparaît bien sûr comme restrictif si l'on considère qu'il a fallu faire des hypothèses, est néanmoins d'intérêt assez général. En effet,

- il est le plus aléatoire (entropie maximale);
- il est la limite (sous des conditions assez générales) des sommes de processus ponctuels indépendants (loi des petits nombres);
- conditionnellement au fait que n points aléatoires engendrés par ce processus se trouvent dans une même région, ces points y sont distribués indépendamment et uniformément; ceci nous permet d'envisager les répartitions les plus aléatoires.

Signalons encore que le nom de ce processus vient de ce que, pour chaque région A, la variable aléatoire de comptage possède la distribution poissonnienne dont le paramètre est la mesure de Lebesgue de cette région.

4 - SOLUTION STATISTIQUE DU MODELE.

Les points que nous observons sont donc engendrés par un processus de Poisson dans k domaines disjoints $(D_i)_{1 \leq i \leq k}$ ($D = \bigcup_{i=1}^{k} D_i$) que l'on désire estimer.

Notons le vecteur des réalisations $(x_1, \ldots, x_n)$ par $\underline{x}$ et la fonction indicatrice de A au point y par

$$1_A(y) = \begin{cases} 1 & \text{si } y \in A, \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance devient

$$f_D(x) = \frac{1}{(m(D))^n} \prod_{i=1}^{n} 1_D(x_i)$$

où $m(D)$, la mesure de Lebesgue de D , est la somme des mesures des $(D_i)_{1 \leq i \leq k}$. Le domaine D , paramètre de dimension infinie, pour lequel la vraisemblance est maximale, est, parmi ceux qui contiennent tous les points, celui dont la mesure de Lebesgue est minimale.

Si nous n'imposons pas de contrainte supplémentaire sur D , nous constatons que nous pouvons très facilement trouver k domaines D_i qui contiennent tous les points et tels que la somme de leurs mesures soit nulle (Fig. 1)

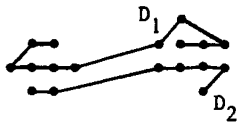


Fig. 1.

Il y aurait donc beaucoup de solutions triviales à ce problème. On peut, cependant, se rendre compte facilement que le problème d'estimation d'un domaine est mal posé et que l'hypothèse la plus faible qui rende le domaine estimable est celle de la convexité des D_i [3] .

A une partition du nuage de points en k sous-ensembles possédant des enveloppes convexes disjointes correspond toute une famille d'estimateurs : il suffit de trouver k convexes disjoints contenant chacun un des sous-ensembles. Cependant, il est évident que, pour chaque partition, la vraisemblance possède un maximum local constitué par les enveloppes convexes des k sous-ensembles (Fig. 2).

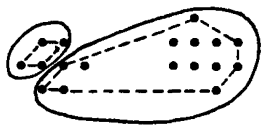


Fig. 2.

Sur cette figure, nous voyons, en trait continu, un estimateur des domaines. En trait pointillé, nous reconnaissons le maximum local correspondant. Le maximum global de la vraisemblance sera donc atteint par la partition pour laquelle la somme des mesures de Lebesgue des enveloppes convexes des k sous-ensembles est minimale (Fig. 3).

Fig. 3.

Pratiquement, si l'espace de base est $\mathbb{R}$, nous cherchons les k intervalles disjoints contenant tous les points et tels que la somme de leurs longueurs soit minimale. Dans $\mathbb{R}^2$ (ou $\mathbb{R}^3$), nous essayons de trouver les k groupes de points tels que la somme des aires (volumes) de leurs enveloppes convexes disjointes soit minimale.

5 - AUTRE APPROCHE DU MODELE.

Le modèle que nous venons de présenter peut s'interpréter comme un mélange de densités tel que ceux qu'étudie Hartigan [4] dans son chapitre V. En effet, si un point appartient à D , il sera dans D_j avec une probabilité p_j et il y sera distribué suivant la densité $f(x \mid D_j)$. On montre facilement que

$$p_j = \frac{m(D_j)}{m(D)} \quad ; \quad f(x \mid D_j) = \frac{1}{m(D_j)} \, 1_{D_j}(x) \ .$$

Suivant ce modèle, la vraisemblance devient

$$f_D(\underline{x}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{k} p_j \, f(x_i \mid D_j) \right)$$

$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{k} \frac{m(D_j)}{m(D)} \, \frac{1}{m(D_j)} \, 1_{D_j}(x_i) \right)$$

$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{k} \frac{1}{m(D)} \, 1_{D_j}(x_i) \right)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{m(D)} \, 1_D(x_i) \right)$$

$$= \frac{1}{(m(D))^n} \prod_{i=1}^{n} 1_D(x_i) \ .$$

6 - L'ALGORITHME.

6.1 - Sa description.

L'algorithme que nous proposons se base sur une généralisation multidimensionnelle de l'algorithme de Fisher [4].

Voyons comment l'algorithme se déroule dans $\mathbb{R}^2$ (pour $\mathbb{R}^k$, il suffira de remplacer "droite" par "hyperplan", "surface" par "mesure de Lebesgue", etc.).

Séparation en deux classes : ces deux classes, étant des polygones convexes, sont toujours séparées par une droite tangente aux deux classes. Il suffit donc de tracer toutes les droites comprenant deux points du nuage et de calculer la somme des surfaces des enveloppes convexes des deux classes ainsi obtenues et retenir la partition pour laquelle la somme des surfaces de leurs enveloppes convexes est minimale. Les points se trouvant sur la droite de séparation devront être répartis de toutes les façons possibles dans les deux classes (Fig. 4).

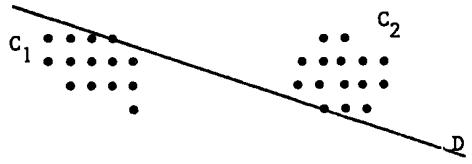


Fig. 4.

Si nous voulons une *partition optimale en trois classes*, il suffit (Fig. 5)

- de fixer un point de l'enveloppe convexe globale; ce point p appartient forcément à l'une des classes, soit C_1 ;
- d'isoler tous les convexes contenant ce point; ceci est facile : il existe toujours une tangente D_1 commune à la classe C_1 et au deuxième convexe, soit C_2 : cette droite passera par un point de C_1 et un point de C_2 ; il en va de même pour C_1 et C_3 ;
- de chercher, pour chaque convexe isolé, la meilleure partition des points restants en deux convexes disjoints entre eux et disjoints du premier; étant donnée la nature des droites utilisées, la région I (respectivement II , III) ne contient que des points de la classe C_1 (respectivement C_2 , C_3); seuls les points de la région IV devront être réaffectés, soit à C_2 , soit à C_3 ;
- de calculer, pour chacun : surface de A + somme des surfaces des enveloppes convexes pour la meilleure répartition du reste en deux; la table ainsi consti-

tuée nous donnera la meilleure partition en trois classes.

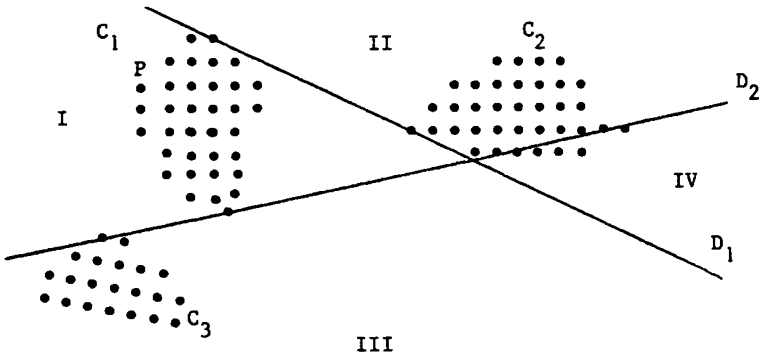


Fig. 5.

Si nous voulons trouver la *meilleure partition en k classes*, il suffit :
- de fixer un point de l'enveloppe convexe globale et d'isoler tous les convexes contenant ce point;
- d'associer, à chacun d'eux, la meilleure partition des points restants en (k-1) convexes disjoints entre eux et disjoints du premier.

Note : la partition est garantie optimale, mais pas nécessairement unique.

6.2. Sa complexité.

La difficulté d'un problème de classification automatique vient du fait que l'on est en présence d'un problème "combinatoire". Le nombre de partitions de n individus en k classes est donné par le nombre de Stirling d'ordre 2 [5]

$$\#P_k = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} k^n \quad .$$

Un algorithme passant en revue toutes les partitions de n individus en k classes sera donc de complexité exponentielle.

C'est la raison pour laquelle nous utilisons un algorithme de programmation dynamique du type "branch and bound" pour résoudre notre problème de classi-

fication. Il n'est, en général, pas possible de calculer la complexité réelle de tels algorithmes. Seule une complexité maximale peut être avancée.

Remarquons cependant les choses suivantes :

- Etant données les hypothèses du modèle, et plus particulièrement l'hypothèse de convexité des sous-domaines dans lesquels les points sont distribués, nous ne devons pas considérer toutes les partitions des n individus en k classes, mais seulement celles dont les enveloppes convexes sont disjointes. Le nombre de telles partitions dans un espace à m dimensions, est au plus égal à

$$C \cdot n^{m\binom{k}{2}}$$

où C est une constante positive indépendante de n [20]. La complexité de notre algorithme sera donc au plus polynomiale !
- De nombreux travaux furent réalisés pour le calcul d'enveloppes convexes. Graham [6] propose un algorithme permettant de calculer l'enveloppe convexe de n points dans le plan de complexité maximale $O(n \log n)$. Preparata et Hong [7] proposèrent un algorithme dans l'espace à trois dimensions de complexité maximale $O(n \log n)$. Yao [8] montra que le calcul de l'enveloppe convexe de n points dans un espace m-dimensionnel demandait une complexité minimale $\Omega(n \log n)$. Bentley et Shamos [9] montrèrent que la complexité moyenne dans le plan et l'espace à trois dimensions était linéaire pour une large classe de famille de distributions. Devroye [10] étendit cette technique à une complexité moyenne linéaire pour un espace à m dimensions, mais pour une classe plus restreinte de distributions. Tout récemment, Bentley, Faust et Preparata [11] mirent au point des algorithmes plus rapides permettant de calculer une approximation d'enveloppes convexes.
- En ce qui concerne le calcul du volume d'un polyèdre convexe dans un espace m-dimensionnel, Cohen et Hickey [12] proposent un algorithme exact de complexité maximale $O(e^m)$ et un algorithme approché de complexité $e\, m\, 2^{e(v+1)}$ où v est fonction de l'erreur que l'on commet et e représente le nombre de sommets de l'enveloppe convexe des points.

7 - ADMISSIBILITE DE NOTRE PROCEDURE DE CLASSIFICATION.

Les méthodes d'analyse des données et de classification automatique existent depuis très longtemps et se sont enrichies au fil des années. Malheureuse-

ment, avant 1960, ces méthodes restaient inabordables car elles nécessitaient une masse énorme de calculs.

A partir de 1960, l'apparition et l'extraordinaire développement des ordinateurs permirent l'utilisation pratique de ces méthodes. Un chercheur, désirant classer ses données, a le choix entre une multitude de procédures de classification.

C'est la raison pour laquelle un certain nombre d'auteurs se sont attaqués au problème complexe suivant : comment choisir une meilleure méthode en classification automatique ?

Deux approches ont retenu notre attention : une approche théorique basée sur les articles de L. Fisher et J.W. Van Ness [13] et de Sidak [14] et une approche plus pratique basée sur les articles de G. Nagy [15] et E. Ruspini [16].

7.1 - Approche théorique.

Fisher et Van Ness se sont rapidement rendu compte que le problème de la détermination d'une meilleure procédure en classification automatique n'était pas suffisamment bien défini que pour en trouver une solution complète. Néanmoins, basant leur approche sur la théorie de la décision, ils suggèrent neuf conditions qu'une procédure de classification devrait vérifier pour être appelée "admissible". A notre connaissance, c'est la seule qui compare directement les procédures de classification.

Nous passerons notre procédure au crible de ces neuf conditions d'admissibilité que nous diviserons en trois groupes :

Conditions directement applicables sur $\mathbb{R}^k$.

Nous pouvons constater que ces conditions seront toutes vérifiées trivialement par notre procédure. Rappelons à cet effet quelles en sont les définitions.

Condition 1 : une procédure est dite "convexe-admissible si les enveloppes convexes des k classes produites sont deux à deux disjointes.

Condition 1' : lorsque la Condition 1 n'est pas vérifiée, on demande parfois qu'une propriété plus faible le soit. On parlera de "connexe-admissible". Rappelons qu'une procédure satisfait à 1' dès qu'elle satisfait à 1 .

Condition 2 : Une procédure est dite admissible par rapport aux proportions des points si une duplication d'un ou de plusieurs points ne modifie pas les frontières des classes.

Condition 3 : une procédure est dite admissible par rapport aux propor-

tions des classes si une duplication de chaque classe un nombre arbitraire de fois ne modifie pas les frontières des classes.

Condition 4 : si une procédure de classification donne k classes et si nous enlevons tous les points d'une classe, soit C_j , alors la procédure sera dite admissible par rapport à l'omission de classe si, lorsqu'elle est appliquée aux points restants, afin d'obtenir (k-1) classes, elle donne les mêmes classes à l'exception de C_j .

Conditions directement applicables sur $\mathbb{R}$ et généralisables à $\mathbb{R}^k$.

Comme nous l'avons déjà signalé, les procédures de classification existantes utilisent une distance ou semi-distance sur $\mathbb{R}^k$ comme généralisation de la mesure de Lebesgue sur $\mathbb{R}$. C'est la raison pour laquelle les trois conditions qui suivent ont été initialement exprimées en termes de distance. Pour chacune d'entre elles, nous allons d'abord montrer qu'elle est vérifiée sur la droite, où nous utilisons la même mesure, ensuite nous donnerons la généralisation en termes de mesure de Lebesgue sur $\mathbb{R}^k$ et montrerons que notre procédure satisfait à ces nouvelles conditions.

Condition 5 : supposons que les N données $\{x_1 , \dots , x_N\}$ à classifier soient ordonnées de telle façon que les classes produites s'écrivent sous la forme : $C_1 = \{x_1 , \dots , x_{n_1}\}$, $C_2 = \{x_{n_1+1} , \dots , x_{n_2}\}$, $\dots$, $C_k = \{x_{n_1+n_2+\dots+n_{k-1}+1} , \dots , x_N\}$.

Soit $\{y_1 , \dots , y_N\}$ une permutation quelconque de $\{x_1 , \dots , x_N\}$. Définissons les ensembles suivants : $C'_1 = \{y_1 , \dots , y_{n_1}\}$, $C'_2 = \{y_{n_1+1} , \dots , y_{n_2}\}$, $\dots$, $C'_k = \{y_{n_1+n_2+\dots+n_{k-1}+1} , \dots , y_N\}$.

Nous appellerons $C'_1 , \dots , C'_k$ une image de $C_1 , \dots , C_k$.

Une classification $C_1 , \dots , C_k$ sera admissible par rapport aux images s'il n'existe pas d'image qui lui soit uniformément meilleure dans le sens suivant :

a) $d(x_i , x_j) \geqslant d(y_i , y_j)$ lorsque le i-ième et le j-ième points sont dans la même classe;

b) $d(x_i , x_j) \leqslant d(y_i , y_j)$ lorsque le i-ième et le j-ième points sont dans des classes différentes

avec une inégalité stricte pour au moins une paire d'indices.

Notre procédure revient, sur la droite, à enlever les $k-1$ plus grandes distances entre les points et donc à laisser les k classes telles que la somme des diamètres soit minimale. Cette propriété, demandée par Šidak [14] , implique l'admissibilité par rapport aux images. En outre, elle peut être généralisée de la façon suivante : nous enlevons de l'enveloppe convexe globale une partie ne contenant aucun point dont la mesure de Lebesgue est maximale, laissant les k

convexes dont la somme des mesures de Lebesgue est minimale.

Condition 6 : les données sont dites bien structurées en k classes s'il existe une classification $C_1, \ldots, C_k$ telle que toutes les distances intra-classes sont inférieures aux distances inter-classes.

Une procédure est dite "k-groupes admissible" si, lorsque les données sont bien structurées en k classes, elle donne la classification $C_1, \ldots, C_k$ lors d'une division en k classes.

Ce critère est, bien sûr, vérifié si nous travaillons dans $\mathbb{R}$. La notion peut être généralisée, par exemple sur $\mathbb{R}^2$, de la façon suivante : tous les triangles formés par trois points appartenant à une même classe ont une surface inférieure à celles des triangles formés par trois points appartenant au moins à deux classes différentes. Pour démontrer que, dans ces conditions, notre procédure est "k-groupes admissible", il faudrait travailler cas par cas. Par exemple, une nouvelle division de nos deux classes aurait une surface supérieure à celle de $N-4$ triangles inter-classes alors que la partition initiale aurait une surface égale à celle de $N'-4$ triangles intra-classes (N', le nombre de points déterminant les enveloppes convexes, est inférieur à N).

Condition 7 : une procédure est dite admissible par rapport aux transformations monotones si une transformation monotone appliquée à chaque élément de la matrice des distances ne change pas la classification.

Sur $\mathbb{R}$, notre classification satisfait évidement à cette condition. Par contre, sur $\mathbb{R}^k$, les seules transformations monotones compatibles avec les configurations sont les homothéties pour lesquelles notre procédure est bien sûr invariante.

Condition inapplicable.

Nous devons enfin signaler une condition qui n'est applicable qu'aux méthodes hiérarchiques et qui, par conséquent, ne peut être vérifiée par notre procédure. C'est aussi le cas, comme le signalent Fisher et Van Ness, de plusieurs autres procédures.

Condition 8 : admissibilité par rapport à la structure d'arbre.

Conclusion.

Notre procédure satisfait aux conditions qui lui sont applicables; elle peut donc être appelée admissible. Aucune des procédures envisagées par Fisher et Van Ness ne vérifiait toutes les conditions qui lui étaient applicables.

7.2 - Approche pratique.

L'approche théorique de Fisher et Van Ness compare directement les procédures de classification mais uniquement qualitativement.

C'est la raison pour laquelle certains chercheurs préfèrent une approche plus pratique; celle-ci consiste à appliquer les différentes procédures de classification automatique aux mêmes ensembles de données "tests" publiées dans la littérature scientifique et à analyser les résultats obtenus. Une procédure sera appelée "admissible" si elle met en évidence la structure présente dans les ensembles de données "tests" considérés.

Les sept premiers exemples considérés sont proposés par G. Nagy [15] (Fig. 6).

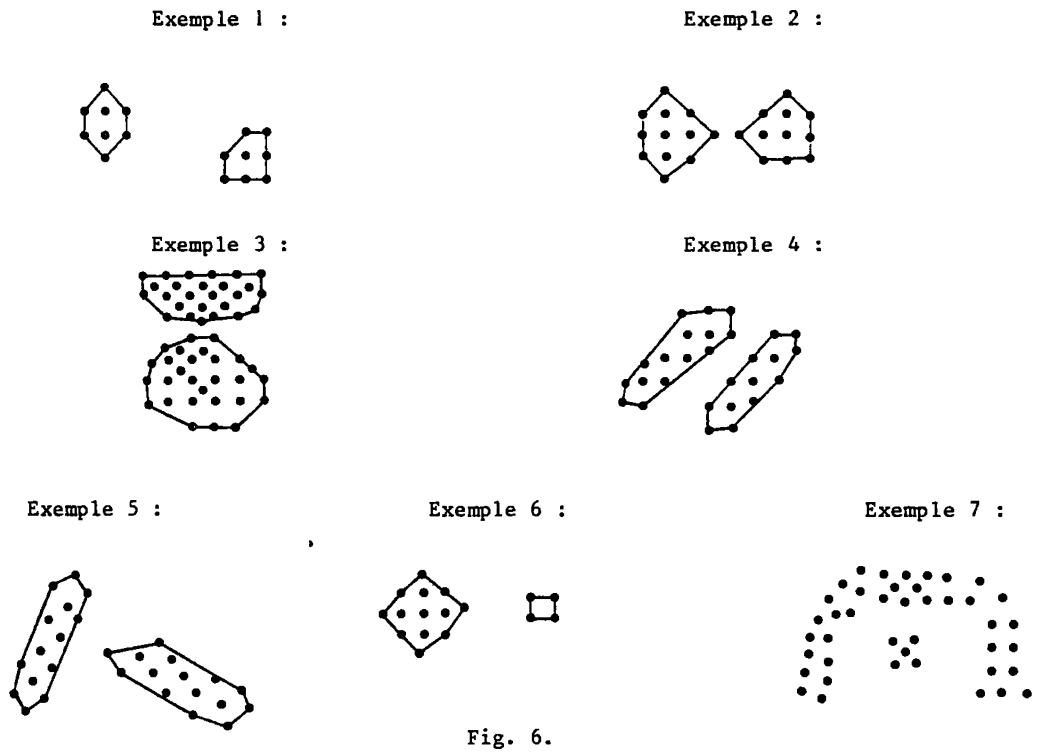


Fig. 6.

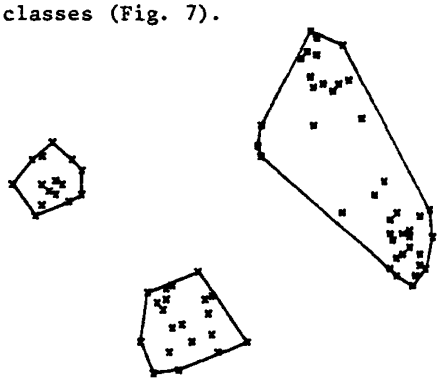
Le premier exemple représente des classes bien séparées. Le second et le troisième mettent en évidence des " ponts " entre les classes. En ce qui concerne le quatrième (respectivement le cinquième), les classes sont allongées et leurs matrices de covariance ne sont pas sphériques (respectivement proportionnelles).

Le sixième exemple présente des classes n'ayant pas le même nombre d'individus et le septième des classes naturelles non convexes.

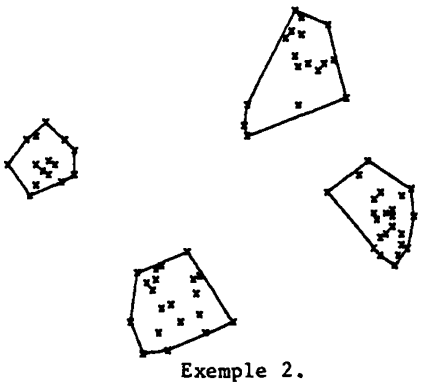
Le septième exemple n'entre pas dans les hypothèses de notre modèle puisque les domaines dans lesquels les points sont répartis ne sont pas convexes. Notre procédure retrouve trivialement les classes naturelles des six autres exemples.

L'exemple que nous allons considérer maintenant, connu dans la littérature sous le nom de "données de Ruspini" (75 points du plan), a souvent été utilisé pour tester de nouvelles procédures en classification automatique (thèses de E. Diday [17], M. Delattre [18], ...).

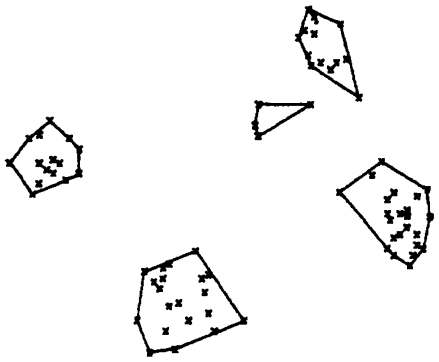
Nous allons présenter les résultats obtenus par notre procédure de classification lorsque nous recherchons successivement trois, quatre, cinq et six classes (Fig. 7).



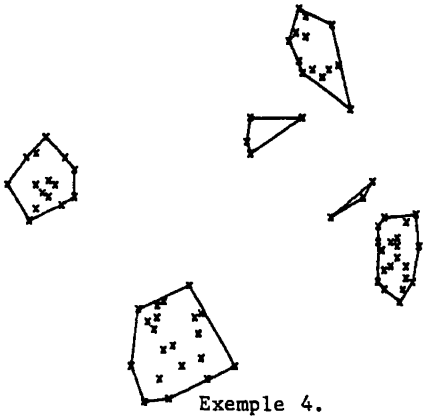
Exemple 1.



Exemple 2.



Exemple 3.



Exemple 4.

Fig. 7.

Notre procédure retrouve bien les classes naturelles dans les exemples présentés. Elle peut donc être appelée "admissible" dans ce sens.

BIBLIOGRAPHIE.

[1] P.A.W. LEWIS, "Stochastic Point Processes", Wiley, New York, 1972.

[2] K. KRICKEBERG, "Moments of point processes", Lecture Notes in Mathematics 296, pp. 70-101, 1973.

[3] J.P. RASSON, "Estimation de formes convexes du plan", Statistique et Analyse des Données 1, pp. 31-46, 1979.

[4] J.A. HARTIGAN, "Clustering Algorithms", Wiley, New York, 1975.

[5] H. SPATH, "Cluster Analysis Algorithms", Wiley, Chichester, 1980.

[6] R.L. GRAHAM, "An efficient algorithm for determining the convex hull of a finite planar set", Information Processing Letters 1, pp. 132-133, 1972.

[7] F.P. PREPARATA & S.J. HONG, "Convex hulls of finite sets in two and three dimensions", Communications of the ACM 20, 2, pp. 87-93, 1977.

[8] A.C. YAO, "A lower bound to finding convex hulls", Stanford University Computer Science Department Report STAN-CS-79-733, 1979.

[9] J.L. BENTLEY & M.I. SHAMOS, "Divide and conquer for linear expected time", Information Processing Letters 7, 2, pp. 87-91, 1978.

[10] L. DEVROYE, "A note on finding convex hulls via maximal vectors", Information Processing Letters 11, 1, pp. 53-56, 1980.

[11] J.L. BENTLEY, M.G. FAUST & F.P. PREPARATA, "Approximation Algorithms for convex hulls", Communications of the ACM 25, 1, pp. 64-68, 1982.

[12] J. COHEN & T. HICKEY, "Two algorithms for determining volumes of convex polyhedra", Journal of the ACM 26, 3, pp. 401-414, 1979.

[13] L. FISHER & J.W. VAN NESS, "Admissible clustering procedures", Biometrika 58, pp. 91-104, 1971.

[14] SIDAK, "Some ideas for the comparison of clustering procedures", 12th European Meeting of Statisticians, Varna, 1979.

[15] G. NAGY, "State of the art in pattern recognition", Proceedings of the IEEE 56, 5, pp. 836-862, 1968.

[16] E. RUSPINI, "A new approach to clustering", Information and Control 15, pp. 22-32, 1969.

[17] E. DIDAY, "Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes", Thèse d'Etat, Paris, 1972.

[18] P. HANSEN & M. DELATTRE, "Complete-link cluster analysis by graph coloring",

J. Am. Statist. Assoc. 73, pp. 397-403, 1978.

[19] D.R. CHAND & S.S. KAPUR, "An algorithm for convex polytopes", Journal of the ACM 17, pp. 78-86, 1970.

[20] T. LENGYEL, "A note on the number of clustering", 12th European Meeting of Statisticians, Varna, 1979.

[21] J.P. BENZECRI et coll., "L'Analyse des Données. Vol. I : La Taxinomie", Dunod, Paris, 1973.

[22] E. DIDAY et coll., "Optimisation en classification automatique", INRIA, 1979.

[23] S. REGNIER, "Sur quelques aspects mathématiques des problèmes de classification automatique", ICC Bulletin 4, pp. 175-191, 1965.

[24] I.C. LERMAN, "Les Bases de la Classification Automatique", Gauthier-Villars, Paris, 1970.

[25] SAID ALI HASSAN, "Recherche des sommets et arêtes de l'enveloppe convexe d'un nombre fini de points", Thèse de Docteur-Ingénieur, Toulouse, 1981.