Statistique et analyse des données

P. Baufays
J.-P. Rasson

# Propriétés théoriques et pratiques et applications d'une nouvelle règle de classement

*Statistique et analyse des données*, tome 9, n° 3 (1984), p. 1-10
<http://www.numdam.org/item?id=SAD_1984__9_3_1_0>

Statistiques et Analyse de données
1984 - vol.9 n° 3 pp.1-10

# PROPRIETES THEORIQUES ET PRATIQUES ET APPLICATIONS D'UNE NOUVELLE REGLE DE CLASSEMENT

P. BAUFAYS et J.-P. RASSON

Unité de Statistique - Département de Mathématique
Facultés Universitaires N.-D. de la Paix
Rempart de la Vierge, 8 - B-5000 NAMUR, BELGIQUE

Résumé : *Nous analysons la règle de discrimination proposée récemment par Baufays et Rasson (1984) : conditions d'admissibilité de Fisher et Van Ness (1973), estimation du taux d'erreur. Celui-ci est analysé par simulations et étudié pour des données réelles. Une variante du critère est proposée pour le cas où les coûts de mauvais classement sont différents.*

Mots clés : *Conditions d'admissibilité de Fisher et Van Ness, Estimation du taux d'erreur, Coûts de mauvais classement.*

Abstract : *We analyze the allocation rule recently proposed by Baufays and Rasson (1984) : admissibility conditions of Fisher and Van Ness (1973), and error rate estimation. The latter is analyzed by means of simulations and studied for real data sets. A modification of the rule is proposed for the case where the misclassification penalties are unequal.*

Keywords : *Admissibility conditions of Fisher and Van Ness, Error rate estimation, Misclassification penalties.*

## 1 - INTRODUCTION

Récemment, Baufays et Rasson (1984) ont proposé une règle d'affectation en analyse discriminante. Nous nous proposons d'en étudier les propriétés : conditions d'admissibilité de Fisher et Van Ness (1973), estimation du taux d'erreur, réduction de la taille de l'ensemble de référence. Nous la comparons avec des règles classiques (linéaire, quadratique et plus proche voisin) à la fois pour des données générées et pour des exemples réels, tirés de la littérature. Enfin, nous considérons une modification du critère lorsque les probabilités a priori ou les coûts de mauvais classement ne satisfont pas le modèle, dont la règle est issue.

Les individus de chacun des $k$ groupes sont uniformément distribués dans un domaine convexe borné ( $C_i$ , $i=1,\ldots,k$ ). Les probabilités a priori sont proportionnelles aux mesures des domaines et les coûts associés aux affectations erronées sont supposés égaux. Soient $X_i$ l'ensemble des individus étiquetés de la $i$-ième population et $H(X_i)$ son enveloppe convexe. Nous définissons :

$$S_i(x) = m(H(X_i \cup \{x\})) - m(H(X_i)) .$$

Si les domaines $C_i$ sont supposés disjoints, nous imposons en outre :

$$S_i(x) = +\infty$$

lorsque $H(X_i \cup \{x\})$ possède une intersection non vide avec au moins un ensemble $H(X_j)$ , $j \neq i$ . La règle affecte l'individu $x$ au groupe $i$ ssi :

$$S_i(x) < S_j(x) \quad , \quad j=1,\ldots,k \ , \ j \neq i .$$

En outre, aucune décision n'est prise à propos des individus qui appartiennent à plus d'une enveloppe convexe.

## 2 - REDUCTION DE LA TAILLE DE L'ECHANTILLON ETIQUETE

Cette règle d'affectation ne dépend que des individus étiquetés, qui sont sommets d'une enveloppe convexe. Ceci permet une réduction importante de la taille de l'échantillon étiqueté, qu'il nous est permis de quantifier grâce à un résultat de Raynaud (1971). Le nombre de sommets, de l'enveloppe convexe d'un échantillon de taille $n$ uniformément distribué

dans un espace à d dimensions, a pour ordre de grandeur $O(n^{((d-1)/(d+1))})$ lorsque n tend vers l'infini.

Le nombre d'individus étiquetés qui déterminent la règle, est donc de loin inférieur à la taille de l'ensemble de référence. L'importance de ce résultat se juge aux efforts déployés afin d'en obtenir un similaire pour la règle du plus proche voisin : règles condensée (Hart (1968)), réduite (Gates (1971)), ..., utilisation du diagramme de Voronoï (Toussaint et Poulsen (1979)).

## 3 - CONDITIONS D'ADMISSIBILITE DE FISHER ET VAN NESS

Face au nombre sans cesse croissant de méthodes de classement proposées dans la littérature et à la difficulté de choix que cela implique, Fisher et Van Ness (1973) ont suggéré sept conditions d'admissibilité. Ce sont des propriétés que toute bonne règle d'affectation devrait vérifier. Avant de passer notre règle au crible de ces conditions, nous rappelons que parmi les règles analysées par Fisher et Van Ness (linéaire, quadratique, centroïde, "average linkage", ...), c'est la règle du plus proche voisin (PPV) qui satisfait le plus de conditions (quatre).

Les conditions de convexité, convexité-aléatoire et de répétabilité sont trivialement satisfaites. En effet, une procédure est dite convexe-admissible si lorsque les enveloppes convexes $H(X_i)$ sont disjointes, tout individu appartenant à $H(X_i)$ est classé dans le groupe correspondant ( i ). La condition de convexité-aléatoire est impliquée par la précédente. La condition de répétabilité ne s'applique qu'aux règles disposant d'une méthode de classification automatique associée. Lorsque l'échantillon étiqueté provient de cette méthode de classification, tout individu extrait de celui-ci doit être correctement classé par la règle définie à partir de l'échantillon réduit. Rappelons que l'algorithme de Hardy et Rasson (1982) partitionne l'échantillon de telle sorte que la somme des mesures de Lebesgue des enveloppes convexes soit minimale.

Les deux conditions suivantes concernent la probabilité d'un classement correct. Si les probabilités a priori sont égales, une règle de classement aléatore assigne correctement un individu dans un parmi k groupes avec probabilité 1/k . Une règle bien construite devrait au moins

faire aussi bien. Si tel est le cas, quelle que soit la taille de l'échantillon étiqueté, la règle est dite égale-admissible; si le résultat n'est vérifié que lorsque cette taille tend vers l'infini, on parlera d'asymptotique-égale- admissibilité. Désignons par $V_n$ le nombre de sommets de l'enveloppe convexe d'un échantillon de taille $n$, uniformément distribué dans le convexe $C$. Efron (1965) a montré :

$$E(m(H(X))) = (1 - E(V_{n+1}) / (n+1))\, m(C) .$$

Soit $n_i$ le nombre d'individus étiquetés de la population $i$ ; nous avons :

$$E[S_i(x)] = [E(V_{n_i+1})/(n_i+1) - E(V_{n_i+2})/(n_i+2)] \cdot m(C)$$

quand $x$ est issu de cette population. Dans ce cas, nous déduisons :

$$\lim_{n_i \to \infty} S_i(x) = 0 ,$$

ce qui prouve l'asymptotique-égale-admissibilité. Aucune preuve ou contre-exemple n'ont été trouvés à propos de la condition d'égale-admissibilité. Nous signalons toutefois le résultat suivant : deux populations sont uniformément distribuées dans deux carrés de mesure unité possédant une arête commune; le taux d'erreur calculé par simulation est 0.25 . D'autre part, il faut signaler qu'aucune règle, analysée par Fisher et Van Ness, ne vérifie cette condition.

Les deux dernières conditions ne s'appliquent directement, à notre règle, que sur la droite réelle (où elles sont vérifiées) parce que Fisher et Van Ness supposent que la règle de classement est définie en termes de distance. Nous généraliserons, dès lors, les définitions de ces conditions en termes de mesure de Lebesgue.

Une procédure est monotone-admissible si elle classe de la même façon en utilisant la distance $d$ ou la distance $f(d)$ quelle que soit la fonction $f$ continue, monotone croissante telle que $f(0) = 0$ ( $f$ est donc une homothétie de centre $0$ et de rapport positif). Nous modifions cette définition en remplaçant distance par mesure de Lebesgue.

Une partition $(E_1,\ldots,E_k)$ d'un ensemble $E$ est dite bien structurée si toutes les distances entre groupes sont plus grandes que tous les diamètres des groupes; Van Cutsem (1984) appelle une telle partition éparpillée; dans le cas contraire, il la dit enchevêtrée. Une procédure est bien structurée-admissible si chaque fois que $(X_1 \cup T_1,\ldots,X_k \cup T_k) - X_i$

étant étiqueté et $T_i$ ne l'étant pas - est bien structuré, tout individu de $T_i$ est affecté au i-ième groupe.

Nous proposons de définir le diamètre d'un groupe comme sa mesure de Lebesgue, et la "distance" entre deux ensembles, comme la plus petite mesure ajoutée par un point de l'un à l'enveloppe convexe de d points de l'autre (dans $R^d$ ). Sur $R$ , les deux définitions coïncident bien. D'autre part, il est évident que, quelle que soit la dimension de l'espace, la règle est monotone et bien structurée-admissible.

Nous pouvons donc affirmer que, au sens de Fisher et Van Ness, cette procédure de classement constitue une "bonne" règle.

## 4 - ESTIMATION DU TAUX D'ERREUR

Le calcul du taux d'erreur constitue un point crucial de l'analyse discriminante. S'il pouvait être estimé de manière exacte, pour chaque méthode décrite dans la littérature, on utiliserait, lorsqu'on est confronté à un problème de classement, la règle pour laquelle le taux d'erreur est minimal.

Parmi les différentes procédures d'estimation de ce taux (Toussaint (1974)), nous retenons les deux suivantes, qui fournissent des résultats intéressants. La méthode de resubstitution estime le taux d'erreur comme la proportion d'individus étiquetés mal classés. A cette méthode, souvent très optimiste, beaucoup préfèrent la méthode du "leaving-one-out" (Lachenbruch et Mickey (1968)). Chaque individu étiqueté est extrait tour à tour de l'échantillon et classé sur base de l'échantillon réduit; le taux d'erreur est estimé par la proportion d'individus mal assignés.

Pour notre règle, le taux estimé par resubstitution est toujours nul; celui estimé par "leaving-one-out" est inférieur à la proportion d'individus qui sont sommets d'une enveloppe convexe, ce qui fournit une borne supérieure, très grossière. Dans le cas où l'échantillon étiqueté est le résultat de la méthode de classification associée, ce taux est nul.

## 5 - APPLICATIONS

Nous avons décrit les propriétés théoriques de la règle de classement.

Il nous paraît opportun d'en étudier le comportement pratique. Pour ce faire, nous comparons ses performances avec celles de procédures classiques, tant pour des données simulées (où il est possible de calculer le taux d'erreur) que pour des données réelles qui ont fait l'objet d'analyses dans la littérature.

Nous considérons trois exemples simulés : dans le premier, le modèle statistique est rigoureusement respecté; dans le deuxième, nous avons modifié les probabilités a priori, et dans le troisième, les lois sont normales. Pour chaque cas, nous avons traité différentes tailles ( N ) d'ensembles de référence; pour chacune, nous avons généré cent échantillons étiquetés et classé, sur base de chacun d'eux, cent individus, pour les règles du plus proche voisin (PPV), linéaire (LD), quadratique (QD) et des mesures de Lebesgue (LB). Les taux d'erreur (exprimés en pourcents) sont consignés dans les tableaux suivants :

Exemple 1. Chacune des populations est uniformément distribuée dans un carré unitaire; les deux carrés partagent une arête.

| N | PPV | LD | QD | LB |
|---|---|---|---|---|
| 40 | 3.77 | 2.77 | 3.05 | 2.35 |
| 100 | 2.41 | 1.48 | 1.69 | 0.95 |
| 500 | 2.14 | 0.62 | 0.74 | 0.23 |

Exemple 2. Chaque population est uniformément distribuée dans un carré; la mesure du premier (resp. du second) est 81 (resp. 16 ). Nous imposons : $p_1 = p_2 = 1/2$ , alors que, suivant notre modèle :

$p_1 = 81/97$ et $p_2 = 16/97$ .

| N | PPV | LD | QD | LB |
|---|---|---|---|---|
| 40 | 2.92 | 6.81 | 1.78 | 2.58 |
| 100 | 2.31 | 6.94 | 1.62 | 1.06 |
| 500 | 2.09 | 6.88 | 1.37 | 0.10 |

Exemple 3. Les deux populations suivent une loi normale, dont la matrice de dispersion est la matrice unité. L'espérance de la première (resp. seconde) population est (0,0) (resp. (1,0) ). Dans ce cas, la meilleure règle à appliquer est la discrimination linéaire. C'est avec cette dernière, uniquement, que nous effectuons la comparaison. Les enveloppes convexes des échantillons ne sont pas nécessairement disjointes;

dans ce cas, aucune décision ne sera prise à propos de certains individus. Il nous a dès lors semblé intéressant d'introduire l'option de rejet dans la discrimination linéaire et d'étudier le taux d'erreur (première ligne) et de rejet (seconde ligne) lorsque la probabilté a porteriori doit être supérieure à 0.7 (L7), à 0.8 (L8) et à 0.9 (L9) .

| N | L | L7 | L8 | L9 | LB |
|---|---|---|---|---|---|
| 40 | 32.6 | 11.6 | 5.6 | 1.8 | 17.5 |
| | 0.0 | 50.8 | 70.9 | 87.4 | 42.3 |
| 100 | 31.1 | 10.0 | 3.9 | 0.67 | 8.8 |
| | 0.0 | 53.1 | 75.2 | 92.1 | 64.4 |
| 200 | 32.1 | 9.9 | 3.7 | 0.63 | 5.0 |
| | 0.0 | 53.3 | 76.4 | 93.6 | 76.2 |

Ces trois exemples montrent l'intérêt de la règle. Lorsque les populations sont distribuées sur des domaines convexes disjoints, son taux d'erreur tend très vite vers zéro. D'autre part, la procédure semble robuste vis-à-vis des probabilités a priori. Enfin, elle se comporte de manière très honorable lorsque les lois sont normales.

Dans le cas de données réelles, il n'est possible que d'estimer le taux d'erreur, ce que nous faisons pour les deux exemples qui ont illustré le calcul de la courbe de décision (Baufays et Rasson (1984)). Il s'agit des taux de deux substances, dans l'urine de onze hétérosexuels et de quinze homosexuels (Hand (1981)) et des longueurs et des largeurs des pétales pour cinquante représentants des variétés d'Iris Virginica et Versicolor (Fisher (1936)).

Pour chacun des exemples, sont présentés le nombre d'individus ( N ), le nombre d'individus qui sont sommets d'une enveloppe convexe ( NS ), le nombre d'individus mal classés par le méthodes de resubstitution et leaving-one-out, pour les règles linéaire ( L ), quadratique ( Q ) et pour celle de la mesure de Lebesgue ( LB ).

| DONNEES | N | NS | METHODE | PPV | L | Q | LB |
|---|---|---|---|---|---|---|---|
| HOMOSEXUELS | 26 | 10 | RESUBS. | 0 | 0 | 0 | 0 |
| | | | LEAVING | 0 | 3 | 1 | 0 |
| IRIS | 100 | 25 | RESUBS. | 0 | 6 | 4 | * 4 |
| | | | LEAVING | 7 | 6 | 5 | 5 |

* Il s'agit du nombre d'individus non classés. Le taux d'erreur est bien sûr nul.

Notre procédure fournit, dans tous les cas, des résultats au moins aussi bons que les méthodes classiques.

## 6 - UNE MODIFICATION DU CRITERE

Le critère, étudié dans cet article, suppose les coûts de mauvais classement égaux et les probabilités a priori proportionnelles aux mesures des domaines. La règle semble robuste vis-à-vis de ces dernières. Toutefois, nous proposons ici une modification de celle-ci, qui tienne compte d'autres coûts de mauvais classement et d'autres probabilités a priori.

Gordon (1984) nous a suggéré la règle suivante :

affecter $x$ au groupe 1 ssi $S_1(x) < a_{12}\, S_2(x)$

où $a_{12}$ est une fonction - à déterminer - des probabilités a priori $p_1$, $p_2$ et des coûts $c_{12}$, $c_{21}$. Rappelons que $c_{12}$ est le coût associé au classement erroné dans le groupe 1 d'un individu du groupe 2 (avec $c_{11} = c_{22} = 0$). Nous imposons que $a_{12}$ satisfasse les propriétés suivantes :

- $a_{12} \geq 0$ ;
- $a_{12} = 1/a_{21}$ ;
- si $c_{21}$ croît, il semble naturel d'imposer que $a_{12}$ croisse; et que $a_{12}$ décroisse lorsque $c_{12}$ croît;
- de même, $a_{12}$ est une fonction croissante de $p_1$ et décroissante de $p_2$ ;
- $p_1$ tend vers 0 implique que $a_{12}$ tend vers 0 ;
- $c_{21}$ tend vers l'infini implique que $a_{12}$ tend vers l'infini;
- lorsque $p_1 = m(C_1)/(m(C_1)+m(C_2))$ et $c_{12} = c_{21}$, alors $a_{12} = 1$ .

Sur base de ces propriétés, nous proposons :

$$a_{12} = m(C_2)\, p_1\, c_{21} \,/\, m(C_1)\, p_2\, c_{12} .$$

Nous remarquons que toute puissance positive de $a_{12}$ satisfait les propriétés précédentes.

Nous ne manquerons pas de comparer cette modification du critère de classement avec celle de la règle du PPV, proposée par Brown et Koplowitz

(1979). En effet, la règle du PPV suppose implicitement que les propriétés a priori sont égales aux proportions d'individus des différents groupes dans l'échantillon étiqueté. Quand tel n'est pas le cas, ces auteurs suggèrent d'appliquer la règle avec la distance modifiée $d_m$ :

$$d_m(x,y_k) = (n_i/p_i n)^{1/d}\, d_e(x,y_k)$$

où $d_e$ est la distance euclidienne, $d$ la dimension de l'espace, $n_i$ le nombre d'individus étiquetés du groupe $i$ et $y_k$ un individu de ce groupe. La distance entre $x$ et un individu du groupe $j$ est aussi égale à :

$$d_m(x,y_k) = (p_i n_j/p_j n_i)^{1/d}\, d_e(x,y_k) .$$

Si $q_i$ (resp. $r_i$ ) désigne la probabilité a priori supposée par notre modèle (resp. par la règle du PPV), le facteur de pondération du critère est égal à :

$(q_2/q_1)\,(p_1/p_2)$ (resp. $((r_2/r_1)(p_1/p_2))^{1/d}$ ).

## 7 - REFERENCES


[1] BAUFAYS, P., et RASSON, J.P. Une nouvelle règle de classement, utilisant l'enveloppe convexe et la mesure de Lebesgue. **Statistique et Analyse des Données**, 1984.

[2] BROWN, T.A., et KOPLOWITZ, J. The weighted nearest neighbor rule for class dependent sizes. **IEEE-IT** 25 : 617-619, 1979.

[3] EFRON, B. The convex hull of a random set of points. **Biometrika** 52 : 331-343, 1965.

[4] FISHER, L., et VAN NESS, J.W. Admissible discriminant analysis. **JASA** 68 : 603-607, 1973.

[5] FISHER, R.A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics** 7 : 179-188, 1936.

[6] GATES, G.W. The reduced nearest neighbor rule. **IEEE-IT** 18 : 431, 1972.

[7] GORDON, A.D. Communication personnelle, 1984.

[8] HAND, D.J. **Discrimination and Classification**, Wiley, 1981.

[9] HARDY, A., et RASSON, J.P. Une nouvelle approche des problèmes de classification automatique. **Statistique et Analyse des Données** 7 : 41-56, 1982.

[10] HART, P.E. The condensed nearest neighbor rule. **IEEE-IT** 14 : 515-516, 1968.

[11] LACHENBRUCH, P.A., et MICKEY, M.R. Estimation of error rate in discriminant analysis. **Technometrics** 10 : 1-11, 1966.

[12] RASSON, J.P. Estimation de formes convexes du plan. **Statistique et Analyse des Données** 1 : 31-46, 1979.

[13] RAYNAUD, H. Sur l'enveloppe convexe des muages de points aléatoires dans $R^n$ , 1. **Journal of Applied Statistics** 7 : 35-48, 1970.

[14] TOUSSAINT, G.T. Bibliography on estimation of misclassification. **IEEE-IT** 20 : 472-479, 1974.

[15] TOUSSAINT, G.T., et POULSEN, R.S. Some new algorithms and software implementation methods for pattern recognition. Dans : **Proc. Third Int. Computing Software and Applications Conference**, pp. 55-63, Chicago, 1979.

[16] VAN CUTSEM, B. Ultramétriques supérieures minimales et algorithme du lien complet. Dans : Société Française de Classification (ed.), **Actes des Journées Statistiques de la Grande Motte**, ASU, 1985.

Nous remercions Allan D. GORDON (St. Andrews, Scotland) pour les commentaires qu'il nous a formulés. Nous lui devons, entre autres, la suggestion de la modification du critère. Nous sommes redevables à André HARDY (Namur, Belgique) d'échanges fructueux durant la préparation de ce travail. B. VAN CUTSEM (Grenoble, France) nous a transmis plusieurs travaux concernant les partitions éparpillées, discutées en 3 .