Séparabilité des classes dans R^p

Zighed D.A. — Lallich S. — Muhlenbach F.

E.R.I.C. Lyon, Université Lumière Lyon 2
5 av. Pierre Mendès-France
69676 BRON France
zighed@univ-lyon2.fr — lallich@univ-lyon2.fr — muhlenba@eric.univ-lyon2.fr

RÉSUMÉ. Dans cet article, nous proposons une approche statistique permettant de caractériser le degré de séparabilité des classes dans R^p. Cette approche repose sur un test statistique non paramétrique, le "Test des arêtes pondérées", dont nous présentons le principe et le résultat des premières expérimentations. Dans un premier temps, nous construisons un graphe de voisinage connexe sur l'ensemble des exemples de l'échantillon d'apprentissage, ici le Graphe des Voisins Relatifs. Ensuite, nous supprimons toutes les arêtes reliant des exemples de classes différentes. Si le poids relatif des arêtes ainsi enlevées est dans l'intervalle attendu dans le cas d'étiquettes distribuées au hasard sur les sommets du graphe de voisinage, aucune méthode d'apprentissage ne pourra fournir un modèle de prévision fiable. Nous dirons alors que les formes ne sont pas séparables et donc pas apprenables.

MOTS-CLÉS : Séparabilité, Test des arêtes pondérées, Graphe de Voisins Relatifs

1. Introduction

Les méthodes d'apprentissage se proposent d'élaborer, à partir d'un échantillon d'apprentissage Ω_l, un modèle de prédiction φ. Par construction, le modèle fourni est plus ou moins fiable suivant la méthode. Cette fiabilité, généralement évaluée *a posteriori* à partir d'un échantillon test Ω_t, dépend de l'échantillon d'apprentissage, des hypothèses statistiques sous-jacentes et des outils mathématiques mis en oeuvre. Mais, il arrive qu'aucune méthode ne puisse produire un modèle fiable. Cet échec peut s'interpréter de deux manières :

– les méthodes utilisées ne sont pas adaptées au problème que l'on cherche à apprendre et l'on doit chercher une méthode plus satisfaisante.

– les classes que l'on veut prédire ne sont pas séparables dans l'espace de représentation. Dans ce cas, nous dirons que les classes ne sont pas apprenables et il est inutile de rechercher une méthode plus satisfaisante ;

Il serait ainsi intéressant de disposer d'outils mathématiques permettant de caractériser le degré de séparabilité des classes à partir d'un échantillon d'apprentissage donné, indépendamment de toute méthode d'apprentissage.

Cette problématique n'est pas nouvelle, elle a déjà été abordée sous un point de vue statistique par Rao et Bartlett (cf. [RAO 65]). Ces auteurs se placent dans le cas où les classes sont distribuées normalement et mesurent le degré d'apprenabilité des classes à travers un test d'homogénéité de populations. Dans cette même optique, Kruskall et Wallis ont défini un test non paramétrique qui repose sur l'hypothèse d'égalité des paramètres d'échelle [AIV 86]. Plus récemment, Sebban [SEB 96] puis Zighed et Sebban [ZIG 99] ont proposé un test fondé sur le nombre d'arêtes reliant des exemples de classes différentes dans un graphe de voisinage.

Ces auteurs construisent d'abord une structure de voisinage en faisant appel à des modèles particuliers comme le graphe des voisins relatifs de Toussaint [TOU 80]. Ensuite, ils calculent le nombre d'arêtes du graphe de voisinage qui doivent être supprimées pour obtenir des amas de points homogènes vis-à-vis des classes. Une fois établie la loi de la proportion d'arêtes supprimées sous l'hypothèse de distribution au hasard des étiquettes (H_0), il ne reste plus qu'à calculer la p-value du test, ou probabilité d'avoir sous H_0 une valeur aussi étonnante que celle effectivement observée.

De façon plus générale, nous proposons dans ce papier un test non paramétrique de séparabilité des classes qui prend en compte le poids des arêtes supprimées. Pour ce faire, nous utilisons l'apport des travaux sur l'autocorrélation spatiale, plus particulièrement ceux relatifs aux statistiques "join-counts", tels qu'ils sont présentés par Cliff et Ord [CLI 86] à la suite des travaux de Moran [MOR 48], Krishna Iyer [KRI 49], Geary [GEA 54] et David [DAV 71]. Une telle démarche a déja été entreprise en classification par Lebart [LEB 00] qui utilise les travaux sur la contiguïté spatiale, notamment le coefficient de contiguïté de Geary, pour confronter structures locales et structures globales par rapport au graphe des k plus proches voisins. Nous construisons ainsi une nouvelle procédure de test qui présente un double intérêt. En premier lieu, ces travaux nous aident à dresser un cadre théorique tout à la fois rigoureux et opérationnel, qui enrichit notablement le test des arêtes proposé par Sebban [SEB 96] et Zighed, Sebban [ZIG 99]. Ensuite, ils nous aident à construire un test qui prenne en compte les poids des arêtes du graphe de voisinage.

Dans la section 2 de ce papier, nous développons notre conception de la séparabilité à partir de la notion d'amas. En section 3, nous exposons le formalisme du "test des arêtes pondérées" et nous présentons en section 4 le résultat de nos premières expérimentations, avant de conclure en section 5.

2. Séparabilité, amas et arêtes coupées

2.1. *Notations*

Les méthodes d'apprentissage à partir de données visent toutes à élaborer une fonction φ, appelée parfois règle de décision, permettant de prédire la classe d'appartenance $Y(\omega)$ inconnue d'un individu ω issu de la population Ω concernée, connaissant sa représentation $X(\omega)$.

En général, cette représentation $X(\omega)$ est fournie par un expert du domaine qui établit une liste *a priori* de variables statistiques appelées variables exogènes et notées : $X_1, X_2, ..., X_p$. Nous supposons que ces variables prennent leurs valeurs dans R, $X : \omega \in \Omega \mapsto X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_p(\omega)) \in R^p$

Dans notre contexte, toute méthode d'apprentissage Φ nécessite le recours à un échantillon d'apprentissage Ω_l et à un échantillon de test Ω_t. Le premier servira à élaborer la fonction de prédiction φ et le second servira à tester sa fiabilité.

Pour tout individu $\omega \in (\Omega_l \cup \Omega_t)$, nous supposons connues sa représentation $X(\omega)$ et sa classe $Y(\omega)$.

$$Y : \Omega \mapsto \{y_1, \cdots, y_k\}$$

La capacité d'apprentissage de toute méthode dépend fortement du degré de séparabilité des classes dans $X(\Omega)$. Nous considérerons que les classes sont d'autant plus séparables que les conditions suivantes sont réunies :

– les exemples d'une même classe apparaissent le plus en groupes dans l'espace de représentation ;
– ces groupes sont peu nombreux (au minimum, le nombre de classes) ;
– leurs frontières sont "simples".

2.2. *Graphe de voisinage et amas*

Pour rendre compte de la proximité des exemples dans l'espace de représentation, nous utilisons la notion de graphe de voisinage (cf. [ZIG 99]), ainsi le Graphe des Voisins Relatifs, le Graphe de Gabriel, la Triangulation de Delaunay ou l'Arbre de Longueur Minimale qui conduisent à des structures de graphe planaires connexes qui, de ce fait, nous paraissent naturelles. Nous utilisons ici le Graphe des Voisins Relatifs (GVR) et nous en donnons la définition ci-dessous.

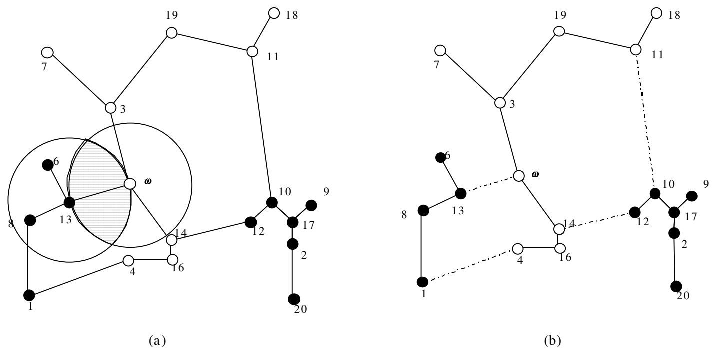

Figure 1. *Graphe des voisins relatifs et amas : deux classes sont représentées, les points noirs et les points blancs*

Définition : Le Graphe des Voisins Relatifs (GVR) est un graphe connexe dans lequel, si deux points sont reliés par une arête, alors ils vérifient la propriété ci-dessous où $d(u,v)$ est la distance euclidienne entre deux points u et v dans R^p

$$d(a,b) \leq Max\left(d\left(a,c\right), d\left(b,c\right)\right) \forall c, c \neq a, b$$

Cette définition signifie que la lunule $L_{(u,v)}$ – ou zone d'influence – constituée par l'intersection des hypersphères centrées respectivement en u et en v et de rayon l'arête (u,v), ne contient aucun point (cf. figure 1 (a)).

Suivant Zighed et Sebban [ZIG 99], nous introduisons la notion d'amas pour rendre compte du fait qu'un ensemble de points proches les uns des autres sont de même étiquette. On appelle *amas* un sous-graphe connexe du graphe de voisinage dont tous les sommets sont de même étiquette. Pour construire les amas nécessaires à la caractérisation des structures de notre nuage de points, nous procédons comme suit :

1. Construction du graphe de voisinage géométrique (connexe et planaire) sur l'échantillon d'apprentissage, selon une des structures de voisinage indiquées précédemment (GG, GVR, TD ou ALM).
2. Suppression des arêtes reliant des points de classes différentes. Nous obtenons ainsi des sous-graphes connexes dont tous les éléments sont de même classe.

Le nombre d'amas obtenus donne une information partielle sur la séparabilité. En effet, si le nombre d'amas est faible, au minimum égal au nombre de classes, alors les classes sont bien séparables et nous pouvons chercher une méthode d'apprentissage susceptible d'exhiber le modèle sous-jacent. Par exemple, figure 1 (b), après avoir coupé les quatre arêtes qui relient deux points de couleurs différentes, nous avons trois amas pour deux classes. En revanche, si ce nombre tend à grandir, proche du nombre d'amas qu'on aurait en situation aléatoire, cela signifie que les classes ne peuvent être apprises.

En réalité, le nombre d'amas n'arrive pas toujours à caractériser certaines situations qui intuitivement nous paraissent différentes. Pour un même nombre d'amas, la situation peut être fort différente, suivant la facilité avec laquelle on peut dissocier ces amas dans le graphe de voisinage. Dès que $p > 1$, plutôt que de s'intéresser aux amas, on préférera s'intéresser aux *arêtes supprimées* pour construire les amas et on calculera le poids relatif de celles-ci dans l'ensemble des arêtes. Dans notre exemple, figure 1 (b), il a fallu couper quatre arêtes (représentées en pointillés) pour constituer les trois amas.

3. Le test des arêtes pondérées

De façon imagée, en analyse spatiale [CLI 86], la variable Y est assimilée à une coloration des sommets du graphe de contiguïté spatiale et l'on considère que les sommets du graphe sont de k couleurs distinctes. Il s'agit de tester l'hypothèse suivant laquelle il y a indépendance entre l'adjacence des sommets du graphe et le fait que ces sommets soient de la même couleur ou non, ce qui revient à tester l'absence d'autocorrélation spatiale entre les valeurs prises par une variable catégorielle sur les différentes zones d'un domaine plan. Pour un graphe de voisinage, cela revient à tester l'hypothèse suivant laquelle une variable catégorielle à prédire n'est pas apprenable à partir des méthodes qui mobilisent le graphe de voisinage des exemples.

3.1. *Cadre statistique*

3.1.1. *Notations et abréviations.*

– Nombre de sommets du graphe : n

– Matrice d'incidence : $V = (v_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., n$; où $v_{ij} = 1$ si *i* et *j* sont reliés par une arête

– Matrice de poids : $W = (w_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., n$; où w_{ij} poids de l'arête (*i*,*j*)

On considèrera que la matrice W est symétrique. Si on doit travailler avec une matrice de poids W' non symétrique, ce qui est très intéressant pour les graphes de voisinage, on se ramène sans perte de généralité au cas symétrique en posant $w_{ij} = \frac{1}{2}\left(w'_{ij} + w'_{ji}\right)$.

– Nombre d'arêtes (joins) : a

– Proportion de sommets de la classe y$_r$: π_r, $r = 1, 2, ..., k$

On adopte, suivant Cliff et Ord, les notations simplificatrices ci-dessous :

Notations	Définition	Cas : $W = V$
$\sum_2 w_{ij}$	$\sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} w_{ij}$	$2a$
S_0	$\sum_2 w_{ij}$	$2a$
S_1	$\frac{1}{2} \sum_2 (w_{ij} + w_{ji})^2$	$4a$
S_2	$\sum_{i=1}^{n} (w_{i+} + w_{+i})^2$	$4 \sum_{i=1}^{n} v_{i+}^2$

3.1.2. *Définition de la statistique de test*

Afin de prendre en compte une éventuelle pondération des arêtes, on travaille avec la matrice de poids symétrisée W qui se réduit à la matrice d'incidence V lorsque tous les poids sont égaux à 1. Nous avons envisagé deux types de poids, ceux symétriques fondés sur les distances et ceux non symétriques fondés sur les rangs. Dans le cas des distances, nous avons choisi $w_{ij} = (1 + d_{ij})^{-1}$, alors que dans celui des rangs, nous avons écrit $w_{ij} = \frac{1}{r_j}$, où r_j est le rang du sommet j parmi les voisins du sommet i.

On distingue les arêtes reliant deux sommets de la même classe (arêtes non coupées) de celles reliant deux sommets de classes différentes (arêtes coupées pour obtenir les amas) et on calcule les sommes de poids I et J correspondant à chaque type d'arête. Désignons par I_r la somme des poids des arêtes joignant 2 sommets de classe r et par $J_{r,s}$ la somme des poids relative aux arêtes joignant deux sommets de classes r et s. Les statistiques I et J s'écrivent :

arêtes non coupées	arêtes coupées
$I = \sum_{r=1}^{k} I_r$	$J = \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} J_{r,s}$

Les statistiques I et J étant liées par la relation $I + J = \frac{1}{2} S_0$, on pourra se limiter à travailler sur la statistique J. On notera que I généralise à 2 dimensions et à k groupes le test des runs (cf. [MOO 40] et [WAL 40]).

3.1.3. *Schéma aléatoire*

On se place dans le cas d'un schéma multinomial où l'hypothèse nulle est définie par (cf. Jain et Dubes [JAI 88]) :

H_0 : les sommets du graphe sont étiquetés indépendamment, suivant une même distribution de probabilités $(\pi_r)_r$ où π_r désigne la probabilité de la classe $r, r = 1, 2, ..., k$.

On pourrait aussi envisager un schéma multi-hypergéométrique en rajoutant dans l'hypothèse nulle la contrainte d'obtenir n_r sommets de la classe $r, r = 1, 2, ...k$.

Le rejet de l'hypothèse nulle signifie soit que les classes ne sont pas attribuées indépendamment, soit que la distribution de probabilités des classes n'est pas la même pour les différents sommets. Pour tester l'hypothèse H_0 en utilisant la statistique J (ou I) , il faut d'abord étudier la loi de ces statistiques sous H_0.

3.2. *Loi de I et J sous l'hypothèse nulle*

Le test de H_0 à partir de J est un test bilatéral si l'on s'étonne tout à la fois des valeurs anormalement petites de J (grande discernabilité des classes) et de ses valeurs anormalement grandes (structuration déterministe, ou à motifs). On rejette H_0 lorsque J a une valeur exceptionnelle, compte tenu de sa loi sous H_0. Nous devons donc établir la loi de J sous H_0 afin de calculer la p-value de la valeur observée de J ou de calculer la valeur critique de J à un risque donné α_0. Ce calcul peut se faire par simulation ou par approximation normale, cette dernière option imposant de calculer la moyenne et la variance de J sous H_0.

3.2.1. *Cas booléen*

La variable Y définit deux classes, notées 1 et 2. Suivant le formalisme de Moran [MOR 48], on pose $U_i = 1$ si le i^e sommet est de classe 1 et $U_i = 0$ si le i^e sommet est de classe 2, $i = 1, 2, ..., n$. On note π_1 la proportion de sommets du type 1 et π_2 la proportion de sommets du type 2. On a alors :

$$J_{1,2} = \frac{1}{2}\sum_{2} w_{ij}\left(U_i - U_j\right)^2 = \frac{1}{2}\sum_{2} w_{ij} Z_{ij}$$

où les variables U_i suivent indépendamment la loi alternative de paramètre π_1, notée $B(1, \pi_1)$, alors que les variables $Z_{ij} = (U_i - U_j)^2$ suivent la loi $B(1, 2\pi_1\pi_2)$, mais ne sont pas indépendantes. En effet, les covariances du type $Cov(Z_{ij}, Z_{kl})$ ne sont nulles que si les 4 indices sont différents. Dans le cas où il y a un indice commun, on obtient :

$$Cov(Z_{ij}, Z_{il}) = \pi_1\pi_2(1 - 4\pi_1\pi_2)$$

Le tableau ci-dessous récapitule les différents résultats relatifs à la statistique $J_{1,2}$:

Variable	Espérance	Variance
U_i	π_1	$\pi_1\pi_2$
$Z_{ij} = (U_i - U_j)^2$	$2\pi_1\pi_2$	$2\pi_1\pi_2\left(1 - 2\pi_1\pi_2\right)$
$J_{1,2}$	$S_0\pi_1\pi_2$	$S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2\left(\frac{1}{4} - \pi_1\pi_2\right)$
$J_{1,2}$ si $w_{ij} = v_{ij}$	$2a\pi_1\pi_2$	$4a\pi_1^2\pi_2^2 + \pi_1\pi_2\left(1 - 4\pi_1\pi_2\right)\sum_{i=1}^{n} v_{i+}^2$

On peut alors centrer et réduire la valeur observée de $J_{1,2}$ pour calculer sa p-value à partir de la loi normale centrée réduite. On peut aussi calculer les valeurs critiques de $J_{1,2}$ au risque α_0 :

$$J_{1,2;\alpha_{0/2}} = S_0\pi_1\pi_2 - u_{1-\alpha_{0/2}}\sqrt{S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2\left(\frac{1}{4} - \pi_1\pi_2\right)}$$

$$J_{1,2;1-\alpha_{0/2}} = S_0\pi_1\pi_2 + u_{1-\alpha_{0/2}}\sqrt{S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2\left(\frac{1}{4} - \pi_1\pi_2\right)}$$

Par simulation, le plus commode est de calculer la p-value de la valeur observée de $J_{1,2}$. Pour simuler une réalisation de $J_{1,2}$, il suffit de simuler une réalisation de $B(1, \pi_1)$ pour chaque exemple, ce qui nécessite m nombres aléatoires entre 0 et 1, et d'appliquer ensuite la formule qui définit $J_{1,2}$. On répéte l'opération N fois et on calcule la p-value de la valeur observée de $J_{1,2}$ en calculant la proportion de valeurs simulées de $J_{1,2}$ qui sont inférieures ou égales à la valeur observée de $J_{1,2}$.

3.2.2. *Cas multiclasses*

Pour généraliser à k classes l'approche des join-counts, on raisonne sur les statistiques I et J définies précédemment. Les statistiques I et J s'écrivent :

$I = \sum_{r=1}^{k} I_r = \frac{1}{2}\sum_2 w_{ij}T_{ij}$	$J = \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} J_{r,s} = \frac{1}{2}\sum_2 w_{ij}Z_{ij}$

où les T_{ij} et les Z_{ij} sont des variables aléatoires alternatives, indicatrices de ce que les sommets i et j ont la même classe (T_{ij}) où sont de classes différentes (Z_{ij}).

À partir des résultats précédents, on obtient facilement les espérances de I et J.

Statistique	Espérance
$I = \sum_{r=1}^{k} I_r$	$\frac{1}{2}S_0\sum_{r=1}^{k}\pi_r^2$
$J = \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} J_{r,s}$	$S_0\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\pi_r\pi_s$

Du fait que $I + J = \frac{1}{2}S_0$, ces deux variables ont même variance. Le calcul de celle-ci, notée $\sigma^2 = Var(I) = Var(J)$, est compliqué par la prise en compte des covariances. En s'aidant de la démarche suivie par Cliff et Ord [CLI 86], on obtient pour le schéma binomial :

$$4\sigma^2 = S_2\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\pi_r\pi_s + (2S_1 - 5S_2)\sum_{r=1}^{k-2}\sum_{s=r+1}^{k-1}\sum_{t=s+1}^{k}\pi_r\pi_s\pi_t$$
$$+4\,(S_1 - S_2)\left[\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\pi_r^2\pi_s^2 - 2\sum_{r=1}^{k-3}\sum_{s=r+1}^{k-2}\sum_{t=s+1}^{k-1}\sum_{u=t+1}^{k}\pi_r\pi_s\pi_t\pi_u\right]$$

4. Expérimentations

4.1. *Test des arêtes pondérées sur un ensemble de bases d'apprentissage*

La statistique des arêtes pondérées a été testée sur 13 bases d'apprentissage automatique du site de l'Université de Californie à Irvine [BLA 98]. Les bases choisies ne comportent que des variables exogènes quantitatives (des données numériques continues) avec une variable endogène qualitative.

Pour chaque base, après avoir centré et réduit les données des variables quantitatives, nous construisons un graphe de voisins relatifs [TOU 80] sur l'ensemble des n individus de la base. Les résultats présentés sur le tableau 1 indiquent, pour chacune des bases testées, le nombre d'individus de la base (n), le nombre de variables exogènes (p) et le nombre de modalités de la variable endogène (k). Nous indiquons aussi les informations concernant le graphe de voisins relatifs à partir duquel nous avons pu faire nos statistiques : le nombre d'arêtes obtenues pour construire le graphe (arêtes) et le nombre d'amas obtenus après avoir coupé les arêtes du graphe reliant des points de classes différentes (amas).

Nous faisons figurer dans le tableau 1, pour chaque base, le poids relatif des arêtes coupées $(J/(I + J))$, la statistique des arêtes pondérées coupées (J Centrée et Réduite) avec son risque critique associé (p-value) dans les trois cas présentés en paragraphe 3.1.2 : quand le test se fait sans pondération, quand les arêtes sont pondérées par l'inverse de la distance entre les points ou quand les arêtes sont pondérées par l'inverse du numéro de rang d'un point aux autres points du graphe. On remarquera que pour chacune des bases les risques critiques sont extrêmement faibles, ce qui rend difficile la comparaison des différentes méthodes de pondération. Cependant l'examen des colonnes $(J/(I+J))$ montre que la pondération, notamment celle par les distances, diminue systématiquement le poids relatif des arêtes coupées pour construire les amas.

4.2. *Sensibilité du test des arêtes pondérées au bruit*

Afin d'étudier la sensibilité du test des arêtes pondérées au bruit, nous avons procédé à un retournement aléatoire d'étiquettes de la base des Iris (Iris Plants Database), une base de référence facile à apprendre.

Informations générales						Sans pondération			Pondération : distance			Pondération : rang		
Nom de la base	n	p	k	amas	arêtes	J / (I + J)	J CR	p-value	J / (I + J)	J CR	p-value	J / (I + J)	J CR	p-value
Breast cancer Wisconsin	683	9	2	10	7562	0.008	-25.29	0	0.003	-24.38	0	0.014	-25.02	0
Bupa liver disorders	345	6	2	50	581	0.401	-3.89	0.0001	0.385	-4.33	1E-05	0.394	-4.08	5E-05
Glass Identification	214	9	6	52	275	0.356	-12.63	0	0.315	-12.90	0	0.342	-12.93	0
Haberman's survival	306	3	2	47	517	0.331	-1.92	0.0544	0.321	-2.20	0.028	0.331	-1.90	0.058
Image segmentation	210	19	7	27	268	0.224	-29.63	0	0.141	-29.31	0	0.201	-29.88	0
Ionosphere	351	34	2	43	402	0.137	-11.34	0	0.046	-11.07	0	0.136	-11.33	0
Iris plants	150	4	3	6	196	0.087	-17.22	0	0.074	-17.41	0	0.076	-17.14	0
Iris (Bezdek)	150	4	3	6	189	0.090	-16.82	0	0.077	-17.01	0	0.078	-16.78	0
Musk "Clean1"	476	166	2	14	810	0.167	-17.53	0	0.115	-7.69	2E-14	0.143	-18.10	0
Pima Indians diabetes	768	8	2	82	1416	0.310	-8.74	2E-18	0.282	-9.86	0	0.305	-8.93	4E-19
Waveform (Breiman)	1000	21	3	49	2443	0.255	-42.75	0	0.248	-42.55	0	0.248	-42.55	0
Wine recognition	178	13	3	9	281	0.093	-19.32	0	0.054	-19.40	0	0.074	-19.27	0
Yeast (protein localization)	1484	8	10	401	2805	0.524	-27.03	0	0.512	-27.18	0	0.509	-28.06	0

Tableau 1. *Test des arêtes pondérées sur 13 bases d'apprentissage de l'UCI.*

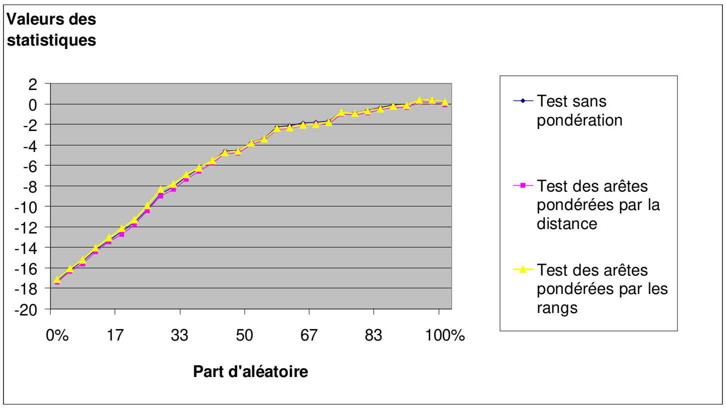

Figure 2. *Sensibilité du test au bruit sur la base Iris (statistique de test).*

Les résultats reportés figure 2 présentent les valeurs obtenues en changeant progressivement la classe initiale des exemples de la base des Iris pour les trois tests des arêtes (sans pondération, avec la pondération calculée sur les distances ou avec la pondération calculée sur les rangs) : sur le graphique, 0 % correspond à la base Iris originelle et 100 % correspond à une base entièrement aléatoire.

À partir de la statistique de test des arêtes pondérées, nous avons aussi pu calculer le risque critique de l'hypothèse nulle (répartition aléatoire des données). L'évolution du risque critique en fonction de l'aléatoire introduit dans la base d'apprentissage est représenté figure 3. Le risque critique passe la barre du seuil de significativité, suivant les tests, entre 20 et 30 étiquettes retournées sur les 150 étiquettes totales, soit 80 à 90 % d'aléatoire, le test sans pondération étant le plus sensible à la présence de bruit, le test des arêtes pondérées par la distance étant le moins sensible. Les trois tests présentent une forte robustesse au bruit : le risque critique ne "décolle" pas d'une valeur nulle ou infime en-deçà de deux tiers de retournements d'étiquettes. On retrouve là le caractère très fort de l'hypothèse de répartition aléatoire des étiquettes sur les sommets d'un graphe de voisinage connexe.

5. Conclusion

Ce travail, qui se situe dans le prolongement de nos travaux [ZIG 99], a abouti à un cadre rigoureux qui permet de prendre en compte le poids des arêtes. Nous envisageons de l'enrichir par la prise en compte de variables exogènes quelconques et de l'utiliser dans des applications réelles.

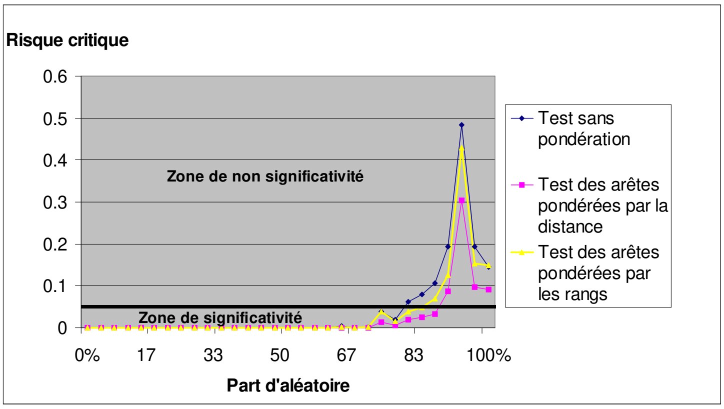

Figure 3. *Sensibilité du test au bruit sur la base Iris (risque critique).*

Une des questions qui reste en suspens est l'influence de la structure de voisinage sur la construction des amas. Nous avons déjà abordé cette question dans des travaux antérieurs et nous allons la reprendre à la lumière de cette nouvelle formulation. Cette question est cruciale, notamment dans le cas où les points d'apprentissage sont plongés dans un espace sans structure mathématique particulière.

6. Bibliographie

[AIV 86] AIVAZIAN S., ENUKOV I., MECHALKINE L., *Eléments de modélisation et traitement primaire des données*, MIR, Moscou, 1986.

[BLA 98] BLAKE C. L., MERZ C. J., « UCI Repository of machine learning databases », Irvine, CA: University of California, Department of Information and Computer Science [http://www.ics.uci.edu/∼mlearn/MLRepository.html], 1998.

[CLI 86] CLIFF A. D., ORD J. K., *Spatial processes, models and applications*, Pion Limited, London, 1986.

[DAV 71] DAVID F. N., « Measurement of diversity », *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, USA, 1971, p. 109–136.

[GEA 54] GEARY R. C., « The contiguity ratio and statistical mapping », *The Incorporated Statistician*, vol. 5, 1954, p. 115–145.

[JAI 88] JAIN A. K., DUBES R. C., *Algorithms for clustering data*, Prentice Hall, 1988.

[KRI 49] KRISHNA IYER P. V. A., « The first and second moments of some probability distribution arising from points on a lattice, and their applications », *Biometrika*, n° 36, 1949, p. 135–141.

[LEB 00] LEBART L., « Data Anlysis », GAUL W., OPITZ O., SCHADER M., Eds., *Contiguïty analysis and classification*, Berlin, 2000, Springer, p. 233–244.

[MOO 40] MOOD A., « The distribution theory of runs », *Ann. of Math. Statist.*, vol. 11, 1940, p. 367–392.

[MOR 48] MORAN P. A. P., « The interpretation of statistical maps », *Journal of the Royal Statistical Society*, série B, 1948, p. 246–251.

[RAO 65] RAO C., *Linear statistical inference and its applications*, Wiley, New-York, 1965.

[SEB 96] SEBBAN M., « Modèles théoriques en reconnaissance des formes et architecture hybride pour machine perceptive », PhD thesis, Université Lyon 2, 1996.

[TOU 80] TOUSSAINT G., « The relative neighborhood graph of a finite planar set », *Pattern recognition*, vol. 12, 1980, p. 261–268.

[WAL 40] WALD A., WOLFOWITZ J., « On a test wether two samples are from the same population », *Ann. of Math. Statist.*, vol. 11, 1940, p. 147–162.

[ZIG 99] ZIGHED D. A., SEBBAN M., « Sélection et validation statistique de variables et de prototypes », SEBBAN M., VENTURINI G., Eds., *Apprentissage automatique*, Hermès Science, 1999.