Classification de psychotropes dans un espace de grande dimension: puissance du test et stabilité d'apprentissage

Mireille Tohmé, Régis Lengellé, Peter Boeijinga

CLASSIFICATION DE PSYCHOTROPES DANS UN ESPACE DE GRANDE DIMENSION: PUISSANCE DU TEST ET STABILITÉ D'APPRENTISSAGE

Mireille Tohmé[1,2], Régis Lengellé[2], Peter Boeijinga[1]

[1] *FORENAP Frp, 27, Rue du 4eme RSM , 68250 Rouffach, France*
[2] *Institut Charles Delaunay-LM2S (FRE CNRS 2848), Université de technologie de Troyes, BP 2060, 10010 Troyes cedex, France*
mireille.tohme@utt.fr ; regis.lengelle@utt.fr; peter.boeijinga@forenap.com

Résumé - Dans ce papier, nous proposons de classer les psychotropes à partir de mesures de potentiels évoqués en utilisant la P300. La difficulté du problème réside dans le fait que les observations sont en faible nombre et dans un espace de grande dimension, ce qui est fréquent dans les études pharmaceutiques. Il est alors très difficile de caler un modèle probabiliste et de fournir une valeur réaliste de la p-value. Notre test repose sur une approche Reconnaissance de Formes (RdF). L'objectif de ce papier est de présenter les premiers résultats relatifs aux liens entre stabilité d'apprentissage et puissance du test obtenu. Cette pré-étude nous permet de retenir le meilleur algorithme d'apprentissage du détecteur c'est à dire celui qui fournit la plus grande puissance parmi les algorithmes testés. Dans un premier temps, nous donnons les éléments justifiant ce lien. Nous étudions ensuite la stabilité de différents algorithmes d'apprentissage de détecteur, nous estimons leur puissance et confirmons expérimentalement le lien entre stabilité d'apprentissage et puissance. Notre méthode est enfin appliquée sur des données réelles (potentiels évoqués) de différentes molécules testées par rapport à un placebo.

Abstract - In this paper, we propose an approach to classify psychotropic drugs from the events related potential (ERP) signals using the P300 components. The difficulties of the problem reside essentially in the fact that traditional methods do not apply when observations are in a high dimensional space, which is a common case in biomedical engineering. Our objective is to propose new hypothesis tests that give p-values reflecting the reality of the efficacy criterion of drugs. Our test is based on a pattern recognition approach. We first study the stability of different training algorithms. We then exhibit a relationship between stability and power functions of the corresponding tests and we give a sketch of the proof of this relationship. We finally apply our method to test the efficacy of different drugs versus placebo.

Mots Clés: Apprentissage, Biostatistiques

Introduction

Il est bien connu que des médicaments, appartenant à des classes différentes et administrés à des sujets sains induisent des changements spécifiques sur l'ensemble des paramètres issus de l'EEG quantifié. Dans ce papier, nous proposons une nouvelle approche de test de comparaison de groupes en utilisant une méthode de Reconnaissance de Formes (RdF). Dans cet esprit, nous essayons de classer les observations en deux classes. Le test initial de comparaison des moyennes des populations est alors remplacé par un test sur la probabilité d'erreur d'un détecteur. La statistique utilisée pour la décision est une estimation de celle-ci. La performance du test dépend de l'estimateur d'erreur utilisé ainsi que de l'algorithme d'apprentissage du détecteur. Dans notre application, les signaux EEG, la dimension du problème est supérieure au nombre d'observations. Il est donc très facile, même pour un détecteur

linéaire, de se trouver en situation de sur-apprentissage (la dimension de Vapnik (1998) des détecteurs linéaires en dimension p est de $p+1$). Tout en restant dans le cadre linéaire, différents algorithmes d'apprentissage sont comparés. On recherche alors, pour l'estimateur d'erreur *Leave One Out* (LOO), l'algorithme d'apprentissage qui assure la meilleure puissance pour une erreur de première espèce fixée. Nous mettons en relation la puissance du test et la stabilité d'apprentissage de chaque détecteur au travers de quelques résultats expérimentaux. Nous donnons les bases qui permettent d'expliquer cette relation. Notre test est enfin appliqué sur des données réelles de potentiels évoqués.

Approches classique et RdF

Les tests d'hypothèse classique multivariés de comparaison de groupes sont, par exemple, les tests de Hotelling, Wilks, Wilcoxon, etc. Pour le test de Hotelling, les données sont supposées gaussiennes de matrices de covariance égales. L'hypothèse nulle H_0 est associée à l'égalité des moyennes alors que l'hypothèse H_1 est associée lorsqu'elles sont différentes. Pour 2 populations de même effectif m, la statistique de Hotelling T^2 est: $S(\mathbf{x}) = \frac{(2m-p-1)m}{4p(m-1)}(\hat{\mu_1} - \hat{\mu_2})'\hat{\Sigma}^{-1}(\hat{\mu_1} - \hat{\mu_2})$, où p est la dimension des données, $\hat{\mu_i}, i = 1, 2$ est la moyenne de chaque classe, $\hat{\Sigma}$ est la matrice de covariance commune estimée. Lorsque $2(m-1) < p$, $\hat{\Sigma}$ n'est pas inversible et son calcul n'est plus possible. Pour résoudre ce problème, nous proposons une approche Reconnaissance de Forme (RdF): nous essayons d'affecter chaque observation à sa classe d'appartenance. Pour chaque observation on réalise le test suivant:
$\begin{cases} \mathcal{H}_0' : \text{l'observation appartient à la classe 0} \\ \mathcal{H}_1' : \text{l'observation appartient à la classe 1} \end{cases}$

On peut démontrer que, lorsque les deux classes sont issues de la même distribution et avec des probabilités a priori égales, la probabilité d'erreur p_e de n'importe quel détecteur est égale à 1/2. En revanche, si les distributions sont différentes, il existe un détecteur optimal (par ex. le détecteur de Bayes) pour lequel $p_e < 1/2$. En conséquence, le test initial classique sur les moyennes des deux groupes devient:

$$\begin{cases} H_0 : \ p_e = 1/2 \\ H_1 : \ p_e < 1/2 \end{cases} \tag{1}$$

La règle de décision est: $\hat{p_e} \overset{D_1}{<} s$, où $\hat{p_e}$ est la valeur estimée de p_e et s est le seuil de décision déterminé par $p(D_1/H_0) = \alpha$. Il est clair que, dans le cas où $(2m \leq p)$, l'erreur empirique n'est pas applicable puisqu'elle fournira toujours, même pour un détecteur linéaire (optimal), une valeur nulle. L'estimateur retenu dans notre étude est le *Leave One Out*, en raison du fait qu'il est presque sans biais (voir ci dessous) et n'est pas coûteux en temps calcul dès lors que la base de données est de faible taille.

Leave One Out

Considérons un ensemble S regroupant m exemples de chacune des deux classes: $S = \{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}; i = 1, \ldots, 2m$ où $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^p$, $y_i \in \{0, 1\}$ est l'étiquette. Pour un algorithme d'apprentissage A produisant une sortie f, l'erreur LOO est définie par: $R_{LOO}(f, S) = \frac{1}{2m}\sum_{i=1}^{2m} l(f^{\setminus i}, \mathbf{z}_i)$, où $l(f, \mathbf{z})$ est la fonction coût, $l(f, \mathbf{z}) = 1$ si $f(\mathbf{x}_i) \neq y_i$ et 0 sinon, $f^{\setminus i}$ est la sortie du détecteur optimisé sur l'ensemble $S^{\setminus i} = S - \{(\mathbf{x}_i, y_i)\}$. Contrairement à l'erreur empirique, l'erreur LOO est presque non biaisée dans le sens où l'espérance calculée sur $2m$ observations est égale à l'erreur de généralisation calculée sur $2m-1$ observations. L'erreur de généralisation conditionnellement à l'ensemble d'apprentissage S est alors (Evgeniou et al (2004)): $R(f, S) = E_{\mathbf{z}}[l(f, \mathbf{z})]$. On définit la stabilité d'hypothèse β_S d'un algorithme d'apprentissage A (Bousquet et al (2002)) par:

$$E_{S,\mathbf{z}}\big[|l(f, \mathbf{z}) - l(f^{\setminus i}, \mathbf{z})|\big] \leq \beta_S \tag{2}$$

Un algorithme est dit stable si une petite perturbation de l'ensemble d'apprentissage ne change pas considérablement sa sortie. Un algorithme conduisant à une valeur de β_S faible fournira une estimation LOO de l'erreur de généralisation peu biaisée.

Algorithmes d'apprentissage

Dans cette partie, nous recensons brièvement les algorithmes d'apprentissage de détecteurs linéaires utilisés dans le cadre de cette étude. Différents algorithmes ont été étudiés : l'Analyse Factorielle Discriminante (AFD). L'AFD est connue pour être stable (Mika S. (2002)). Nous avons aussi considéré le Nearest Mean Classifier (NMC), très simple à implémenter. Une observation appartient à la classe dont la moyenne empirique est la plus proche. Cet algorithme est connu pour être stable. De plus, nous avons étudié la méthode des moindres carrés régularisés (RLS). Finalement, nous avons testé les Support Vector Machines (SVM) linéaires (Vapnik (1998)). Une expression de la stabilité de RLS et de SVM a été présentée dans Bousquet et al (2002).

Lien entre stabilité et puissance

L'expression de la stabilité β_S est indépendante de la distribution de la probabilité des données. Pour une distribution donnée des observations, nous avons $E_{S,\mathbf{z}}\big[|l(f,\mathbf{z}) - l(f^{\setminus i},\mathbf{z})|\big] = \beta_S$. Nous en déduisons : $E_S\big[(R(f,\mathbf{z}) - R_{LOO}(f,\mathbf{z}))^2\big] \geq \beta_S^2$ d'où

$$b_{LOO}^2 + V_{LOO} \geq \beta_S^2 \qquad \text{ou encore } \; V_{LOO} \geq \beta_S^2 - b_{LOO}^2 \tag{3}$$

où b_{LOO} est le biais de l'estimateur d'erreur LOO et V_{LOO} est sa variance. A supposer que l'estimateur LOO soit strictement non biaisé (il l'est faiblement), nous aurons:

$$V_{LOO} \geq \beta_S^2 \tag{4}$$

Dans ce cas, la variance de la statistique de décision est bornée inférieurement par le carré de la stabilité. La puissance d'un test dépend de la distribution de la statistique de décision sous H_1, et en particulier, à biais donné (supposé ici nul), de la variance de cette statistique. Une augmentation de cette variance se traduit en général par une diminution de la puissance du test correspondant. Nous allons essayer de faire le constat expérimental de cette relation entre puissance et stabilité.

Procédure expérimentale

Estimation de la loi de probabilité de l'estimateur LOO

Nous avons estimé la distribution de l'estimateur LOO pour divers types de loi de probabilité des données, identiques pour les deux classes (H_0). Nous présentons dans ce papier les résultats obtenus dans les cas gaussien. Les moyennes des distributions sont $\mu_0 = \mathbf{0}_p, \mu_1 = d\mathbf{1}_p$. Nous notons $\theta = d/\sqrt{(}p)$ la distance entre-classes. Pour chaque type de distribution et pour chaque valeur de θ, la distribution de l'estimateur LOO est estimée. L'hypothèse H_0 correspond à $\theta = 0$. A partir de celle-ci, nous pouvons déterminer le seuil de détection à $\alpha = p(D_1/H_0)$ fixé ou, réciproquement, calculer la p-value du test. Sous H_0, à m et p fixés, et pour toutes les distributions étudiées, nous avons observé l'invariance de la distribution de probabilité de l'estimateur LOO. Les travaux justifiant ce constat sont en cours.

Estimation de la stabilité

Pour chaque valeur de θ, nous avons estimé la stabilité d'apprentissage β_S par simulation Monte Carlo (3000 réalisations), pour différentes valeurs de m. Nous ne présentons ici que les résultats obtenus

Psychotropic drugs	m	p	RDF test	Wilcoxon	Hotelling	électrodes
Lorazepam/placebo	24	10	0.1917	0.0701	0.0412	C_Z
	24	20	0.1087	0.2831	0.0266	C_Z, F_Z
	24	30	0.0563	-	-	C_Z, F_Z, P_Z
Scopolamine/placebo	24	10	0.0200	0.0211	$2.7\ 10^{-3}$	C_Z
	24	20	0.0250	0.2540	$5.5\ 10^{-4}$	C_Z, F_Z
	24	30	0.0210	-	-	C_Z, F_Z, P_Z

Table 1: p-values obtenues pour différents électrodes

pour la dimension $p = 10$. Les simulations (illustrées très partiellement en figure 1) ont montré que la stabilité est une fonction essentiellement décroissante de la distance entre-classes θ. Nous pouvons constater sur ces figures que la stabilité est une fonction décroissante de m (l'algorithme est d'autant plus stable que m est grand), ce qui est un résultat tout à fait logique. Enfin, il apparaît que pour les valeurs de m considérées, l'algorithme NMC est le plus stable.

Estimation de la puissance

Là encore, la puissance du test a été estimé par simulation de Monte Carlo, sauf pour le test de Hotelling pour lequel elle a été déterminée analytiquement. Lorsque c'est possible, c'est à dire en fonction des valeurs relatives de m et p nous présentons la courbe de puissance du test de Hotelling. Un examen des courbes de puissance de la figure 1 montre que la puissance est une fonction croissante de θ et de m, ce qui est tout à fait naturel. Nous observons aussi que le test le plus puissant, parmi ceux considérés, est celui qui repose sur l'algorithme d'apprentissage du détecteur NMC. D'une manière générale, le test est d'autant plus puissant que l'algorithme d'apprentissage est stable, ce qui est cohérent avec ce qui a été présenté précédemment. La figure 2 représente, pour l'algorithme RLS, les SVM linéaires, l'Analyse Factorielle Discriminante (AFD) et le Nearest Mean Classifier (NMC), la stabilité β_S, l'écart type σ_{LOO} de l'estimateur LOO, son biais b_{LOO} ainsi la valeur efficace de l'erreur $\sqrt{(b_{LOO}^2 + \sigma_{LOO}^2)}$.

Application aux signaux réels

Les résultats précédents ont permis de montrer que, parmi tous les détecteurs testés, NMC est le plus puissant, car le plus stable. Nous avons utilisé le test que nous avons proposé (1) afin de tester l'efficacité de deux psychotropes (le Lorazepam et la Scopolamine) par rapport au placebo pour 12 sujets de chaque groupe. Le paramètre considéré est la $S300$ (aire sous la courbe des potentiels évoqués) avec l'électrode C_Z puis C_Z, F_Z et enfin C_Z, F_Z, P_Z. La $S300$ est mesurée pour 10 latences différentes après la prise de la molécule ou du placebo. La taille de la matrice des données est donc de (24×10), (24×20) et (24×30) respectivement. Des expériences menées sur des groupes de plus grande taille et l'étude des cartographies des effets induits sur l'EEG ont démontré une influence très nette de la Scopolamine, associée à une diffusion des effets mesurés sur une zone étendue du scalp, et moindre (et plus localisée dans le temps et dans l'espace) du Lorazepam. L'expert attend donc que les tests d'hypothèse fournissent une p-value plus faible pour le test Scopolamise/placebo que Lorazepam/placebo. Par ailleurs, l'ajout d'information par l'augmentation du nombre d'électrodes (permettant de prendre en compte la diffusion spatiale) doit aider à la détection de l'efficacité du Lorazepam. Dans le cas d'une seule électrode ($p = 10$), la table 1 représente les p-values obtenues par notre test, le test de Wilcoxon multidimensionnel utilisant la loi asymptotique de la statistique

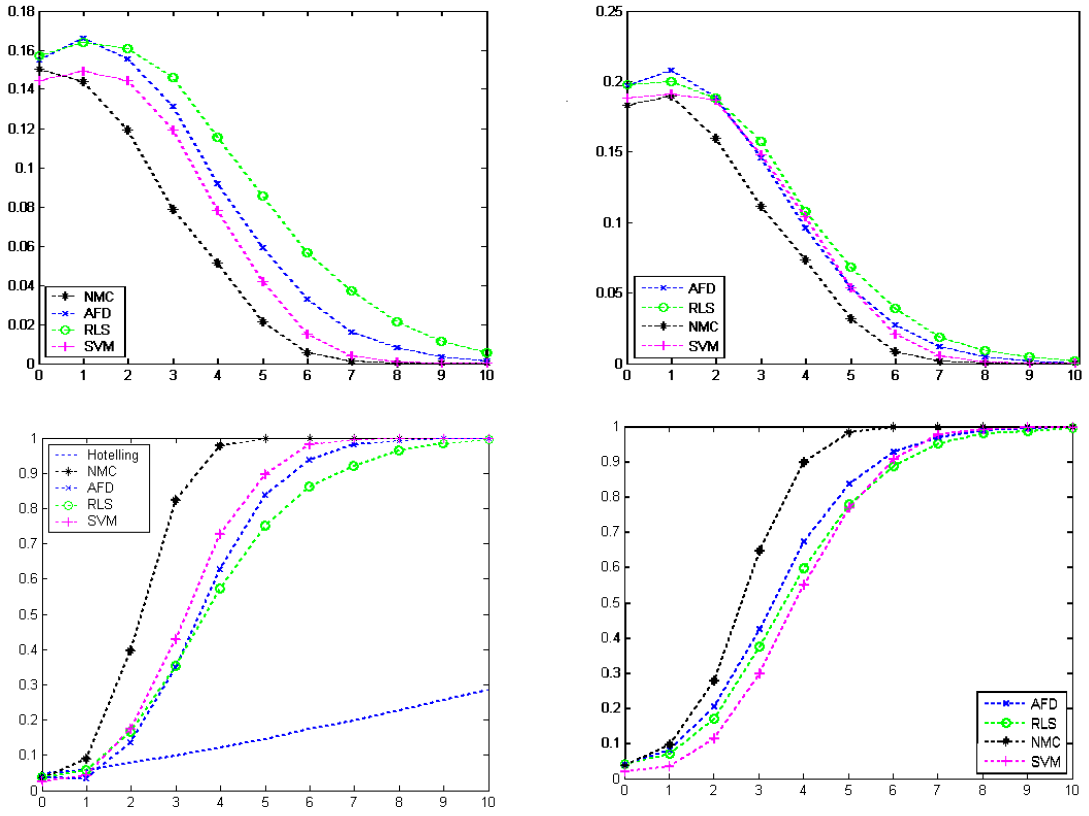

Figure 1: Comparaison de la stabilité (1ère ligne) et la puissance du test (2nde ligne) de différents algorithmes pour $m = 6$ (colonne de gauche), 4 (colonne de droite) et $p = 10$, en fonction de la distance entre-classes θ, cas gaussien

de décision (Sen, (1970)) et celui de Hotelling respectivement. L'erreur de première espèce est fixée à 5%. Seul le test de Hotelling déclare efficace le Lorazepam, ce qui n'est pas conforme aux observations des experts.
On considère maintenant les électrodes C_Z, F_Z ($p = 20$). Là encore, le test de Hotelling conduit à une prise de décision en faveur de l'efficacité du Lorazepam, ce qui parait incohérent aux experts. Toutefois, le test de Wilcoxon conduit à l'inefficacité de la Scopolamine. Ceci résulte du fait que la puissance du test de Wilcoxon devient faible lorsque $2m$ et p sont du même ordre de grandeur. Enfin, nous considérons les électrodes C_Z, F_Z, P_Z. Les tests de Hotelling et de Wilcoxon ne s'appliquent plus. Les p-values du test par RdF sont conformes aux attentes des experts, car cohérentes avec les résultats obtenus sur des groupes d'effectif plus important, en confirmant une plus grande capacité de détection des effets de la Scopolamine par rapport au Lorazepam.

Conclusion

Nous avons proposé, dans ce papier, un nouveau test de comparaison de groupes. La méthode proposée repose sur une approche de reconnaissance des formes et se substitue avantageusement aux tests de Hotelling et de Wilcoxon multidimensionnel lorsque la dimension de l'espace de représentation et le nombre d'observations disponibles sont comparables. Par ailleurs, ce test permet de traiter le cas où le nombre d'observations est inférieur à la dimension de l'espace de représentation des données.

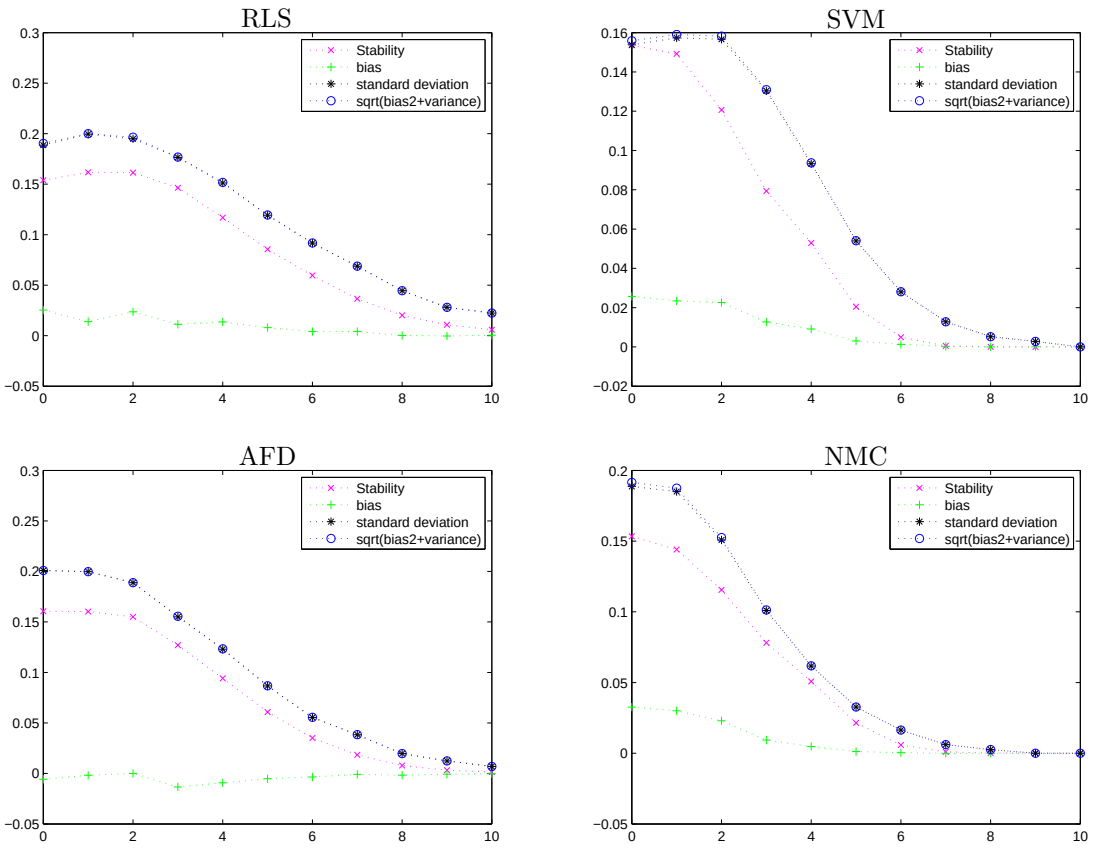

Figure 2: Comparaison de la stabilité, l'écart type et le biais pour $m = 6, p = 10$ en fonction de la distance entre-classes θ - cas gaussien

Nous avons posé les premières bases illustrant le fait que la stabilité d'apprentissage et la puissance d'un tel test sont liées et confirmé ce fait expérimentalement. Nous avons observé que notre test est très peu sensible au type de loi de probabilité des données sous H_0, ce qui permet de fournir des valeurs de seuils de détection ou de p-values qui soient valides pour une large classe de distributions. Nous avons appliqué notre approche à la mesure d'efficacité de deux psychotropes, les résultats obtenus sont tout à fait compatibles avec les observations des experts. La justification de l'invariance de la loi de probabilité de l'estimateur d'erreur LOO à la distribution des données sous H_0 est en cours.

Bibliographie

[1] Evgeniou T., Pontil M., and Elisseeff A., (2004), *Leave one out error, stability, and generalization of voting combinations of classifiers*, Machine Learning, vol. 55, no. 1, pp. 7197, 2004.
[2] Mika S., (2002) *Kernel Fisher Discriminants*, PhD thesis, University of Technology, Berlin.
[3] Bousquet O. and Elisseeff A., (2002), *Stability and generalization*, Journal of Machine Learning Research, vol. 2, pp. 499-526.
[4] Vapnik V. N. (1998), Statistical Learning Theory, Wiley, New York.
[5] Sen P.K (1970), Asymptotic distribution of a class of multivariate rank order statistics. CSAB, 19: 22-32.