Apprentissage Statistique Optimal d'un Classifieur de Neyman-Pearson Paramétrique

Danny SCHMITT[1], Lionel FILLATRE[1], Michel BARLAUD[1]

[1]Université Côte d'Azur, CNRS, I3S
CS 40121 - 06903 Sophia Antipolis CEDEX, France
schmitt@i3s.unice.fr, fillatre@i3s.unice.fr, barlaud@i3s.unice.fr

Résumé – La classification de Neyman-Pearson (NP) a pour objectif l'apprentissage d'un classifieur contrôlant de manière asymétrique le risque d'erreur de chaque classe. Cet article propose une approche paramétrique qui suppose que les distributions probabilistes des échantillons de la base d'apprentissage appartiennent à une famille de distributions paramétriques. Cette approche est motivée par la difficulté d'apprendre un classifieur pour les applications disposant d'un faible nombre d'échantillons d'apprentissage. L'originalité majeure de cet article consiste à reformuler le problème de classification NP sous la forme d'un test statistique entre deux hypothèses composites. Le premier résultat significatif consiste, dans le cas de lois normales univariées, a construire un classifieur optimal uniformément le plus puissant. Ce classifieur minimise l'erreur moyenne de faux négatif par rapport aux bases d'apprentissage possibles tout en bornant l'erreur moyenne de faux positif. Un second résultat permet de régler ce test de façon à garantir que les deux types d'erreurs soient bornés en probabilité quelque soit la base d'apprentissage utilisée en pratique. Ces bornes probabilistes sont exprimées sous forme analytique.

Abstract – The Neyman-Pearson (NP) classification aims at learning a classifier that asymmetrically controls the risk of each class. This paper proposes a parametric approach which assumes that the probabilistic distributions of the learning samples belong to a family of parametric distributions. This approach is motivated by the difficulty of learning a classifier for applications with a small number of learning samples. The main novelty of this article is to state the NP classification problem under the form of a statistical test between two composite hypotheses. The first significant result consists in constructing a uniformly most powerful classifier in the case of univariate normal distrubutions. This classifier minimizes the average false negative risk with respect to all possible learning datasets while keeping the average false positive risk under a prefixed value. A second result is the possibility to tune this test in order to control precisely the two risks whatever the given learning dataset. The bounds are given in a closed-form.

1 Introduction

L'apprentissage statistique a pour objectif d'identifier la classe de nouvelles observations en se basant sur des données d'apprentissage étiquetées. Contrairement à l'apprentissage traditionnel qui minimise la probabilité globale d'erreur, la classification de Neyman-Pearson (NP) minimise la probabilité de faux négatif tout en maintenant la probabilité de faux positif inférieure à un niveau α prédéfini [8].

Pour de nombreuses applications, comme le diagnostique médical ou la surveillance des systèmes, l'approche NP s'impose puisqu'il est primordial de contraindre la probabilité de faux positif. Par exemple, ne pas détecter une maladie peut avoir de graves conséquences sur la santé du patient, alors que déclarer de façon erronée le patient malade engendre uniquement des coûts inutiles pour confirmer le diagnostique. De plus, si la base d'apprentissage n'est pas équilibrée, minimiser la probabilité d'erreur globale peut conduire à un déséquilibre inacceptable entre les deux types d'erreur.

L'approche NP en classification à été initiée dans [1]. Des résultats d'apprentissage par minimisation d'un risque empirique adapté au contexte NP ont été démontrés dans [7]. Des mesures de performances spécifiques, combinant les deux types d'erreur en un score globalement minimisé par le test NP optimal, ont été étudiées dans [6]. Étant donné que la minimisation du risque empirique est difficile d'un point de vue numérique, une solution alternative consiste à exploiter une approximation convexe de la fonction de risque [5]. Toujours en lien avec la minimisation d'un risque, la construction d'un classifieur SVM avec un nombre contrôlé de faux positifs a été proposée dans [2]. Une approche alternative [9] propose une procédure du type "plug-in" : la densité de chaque classe ainsi que le seuil du test sont estimés, puis incorporer au test NP théorique. Les méthodes citées sont non paramétriques et leur performance est en général décrite par une décroissance probabiliste de l'écart entre le classifieur NP et le test NP optimal, ce qui se formalise avec les bornes "Probablement Approximativement Correctes" (PAC). Ce type de performance est souvent très imprécise lorsque la base d'apprentissage est de taille limitée.

Motivé par le respect de la contrainte NP pour de petits échantillons, cet article fait l'hypothèse de la connaissance d'une famille paramétrique à laquelle appartient la loi de probabilité de chaque classe [4]. En adoptant ce cadre théorique, ce papier développe plusieurs résultats originaux. Tout d'abord,

il donne la forme du classifieur qui exploite de façon optimale une base d'apprentissage d'échantillons normaux. Ce test minimise uniformément, pour toute loi de probabilité, l'erreur moyenne de faux négatif tout en bornant le niveau moyen de faux positif. Les propriétés statistiques du classifieur sont calculées de façon analytique, notamment en donnant des bornes probabilistes sur les erreurs. Enfin, les performances du classifieur sont comparées à celles des méthodes de l'état de l'art.

La section 2 présente le problème de classification NP. La section 3 décrit le classifieur NP paramétrique optimal. La section 4 présente les résultats numériques.

2 Classification NP Paramétrique

Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathcal{X} \times \{0, 1\}$ où $\mathcal{X}$ est un espace probabilisable : X correspond au signal observé, et Y correspond à la classe associée à X. Un test statistique est une fonction Borel-mesurable $h : \mathcal{X} \to \{0, 1\}$ assignant une classe à un signal observé. Soit $\mathcal{H}(\mathcal{X})$ l'ensemble des tests sur $\mathcal{X}$. Classiquement, la performance de h est mesurée par sa probabilité d'erreur

$$R(h) = P_{(X,Y)}(h(X) \neq Y) \tag{1}$$

où $P_{(X,Y)}$ est la mesure de probabilité du couple (X, Y). Notons $P_j = P_{X|Y=j}$ la loi conditionnelle de X respectivement à $Y = j$. L'approche NP s'intéresse aux probabilités de faux positif $R_0(h)$ et de faux négatif $R_1(h)$:

$$R_j(h) = P_{X|Y=j}(h(X) \neq j) = P_j(h(X) \neq j). \tag{2}$$

La probabilité d'erreur s'exprime alors sous la forme $R(h) = \pi_0 R_0(h) + \pi_1 R_1(h)$, où $\pi_j = P_Y(Y = j)$. Soit $\alpha \in]0, 1[$ un niveau de faux positif fixé. L'approche NP classique recherche le test h^*_α qui minimise $R_1(h)$ sous la contrainte $R_0(h) \leq \alpha$. Ce test est appelé le test le plus puissant de niveau α [3]. Pour éviter d'introduire la notion de test randomisé, nous supposerons que les distributions P_j sont continues par rapport à la mesure de Lebesgue. La densité de P_j est notée p_j. Soit $x \in \mathcal{X}$ l'observation à tester. Le test NP $h^*_\alpha(x)$ est alors donné par

$$h^*_\alpha(x) = \begin{cases} 1 & \text{si} \quad p_1(x) > k_\alpha p_0(x), \\ 0 & \text{si} \quad p_1(x) \leq k_\alpha p_0(x), \end{cases} \tag{3}$$

où le seuil $k_\alpha > 0$ est choisi tel que $R_0(h^*_\alpha) = \alpha$.

La base d'apprentissage permet d'apprendre le test NP en estimant les distributions de probabilité P_0 et P_1. Elle est constituée de n_0 échantillons négatifs $x^0 = (x^0_1, \ldots, x^0_{n_0})$ et de n_1 échantillons positifs $x^1 = (x^1_1, \ldots, x^1_{n_1})$. Nous considérons un échantillonage rétrospectif, ce qui signifie que n_0 et n_1 sont constants quelque soit la base d'apprentissage $z = (x^0, x^1)$ [7]. Cette base est donc considérée comme la réalisation z d'une variable aléatoire Z à valeur dans $\mathcal{Z} = \mathcal{X}^{n_0} \times \mathcal{X}^{n_1}$.

Un algorithme d'apprentissage est une fonction $\hat{g} : \mathcal{Z} \to \mathcal{H}(\mathcal{X})$ telle que $\hat{g}_z = \hat{g}(z)$ est le classifieur sur $\mathcal{X}$ associé à la base d'apprentissage $z \in \mathcal{Z}$. Notons $\mathcal{H}(\mathcal{Z} \times \mathcal{X})$ l'ensemble des tests statistiques $h : \mathcal{Z} \times \mathcal{X} \to \{0, 1\}$. Conditionnellement à la base d'apprentissage $z \in \mathcal{Z}$, la fonction $\hat{h}_z$, définie par

$$\hat{h}_z(x) = h(z, x) \tag{4}$$

et associée au test $h \in \mathcal{H}(\mathcal{Z} \times \mathcal{X})$, est assimilée à un classifieur sur $\mathcal{X}$. Ce papier exploite cette association pour construire des classifieurs à partir de tests statistiques. Les risques $R_j(\hat{h}_z)$ de $\hat{h}_z$ sont désormais des variables aléatoires Z-mesurables :

$$R_j(\hat{h}_z) = \mathbb{E}_{X|Y=j}\left[h(Z, X)|Z\right] \tag{5}$$

où $\mathbb{E}_X$ désigne l'espérance mathématique suivant la distribution de probabilité de X. Pour se rapprocher de l'approche NP d'origine, il est raisonnable de proposer des garanties non-probabilistes de la forme

$$\mathbb{E}_Z[R_0(\hat{h}_z)] = \mathbb{E}_{Z,X|Y=0}[h(Z, X)] = R_0(h) \leq \alpha. \tag{6}$$

Le risque de faux positif de $\hat{h}_z$ est alors contrôlé en moyenne par rapport à toutes les bases d'apprentissage z possibles.

Le classifieur NP paramétrique est construit à partir d'un couple (z, x), équivalent au triplet $w = (x^0, x^1, x)$. Dans le cas paramétrique, chaque distribution P_j correspond à une distribution P_ϑ associée à la loi normale $\mathcal{N}(\vartheta, \sigma^2)$ avec σ^2 connue. Pour $\theta = (\theta_0, \theta_1, \theta_2)$, notons $P^\theta_{(Z,X)}$ la distribution de (Z, X) :

$$P^\theta_{(Z,X)} = P^{n_0}_{\theta_0} \otimes P^{n_1}_{\theta_1} \otimes P_{\theta_2}. \tag{7}$$

Les hypothèses à tester sur les distributions sont alors

$$\begin{aligned} \mathcal{H}_0 &: \left\{(z, x) \sim P^\theta_{(Z,X)} \mid \theta \in \Theta_0\right\}, \\ \mathcal{H}_1 &: \left\{(z, x) \sim P^\theta_{(Z,X)} \mid \theta \in \Theta_1\right\}, \end{aligned} \tag{8}$$

où Θ_j désigne l'ensemble

$$\Theta_j = \{\theta = (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3 \mid \theta_0 < \theta_1, \theta_2 = \theta_j\}. \tag{9}$$

L'hypothèse $\mathcal{H}_i$ signifie que l'observation à tester est distribuée suivant la loi du vecteur d'apprentissage x^i. La restriction $\theta_0 < \theta_1$ (qui sera discutée dans la section 3) est nécessaire pour obtenir un test Uniformément le Plus Puissant (UPP). Soit $\mathcal{K}_\alpha$ la classe des tests sur $\mathcal{Z} \times \mathcal{X}$ de niveau α :

$$\mathcal{K}_\alpha = \left\{h \in \mathcal{H}(\mathcal{Z} \times \mathcal{X}) \mid \sup_{\theta \in \Theta_0} R_0(h; \theta) \leq \alpha\right\}. \tag{10}$$

Si $h \in \mathcal{K}_\alpha$ alors il vérifie

$$\sup_{\theta \in \Theta_0} R_0(h; \theta) = \sup_{\theta \in \Theta_0} \mathbb{E}^\theta_Z[R_0(\hat{h}_z; \theta)] \leq \alpha, \tag{11}$$

ce qui signifie que le classifieur $\hat{h}_z$ satisfait en moyenne le niveau α. Il est à noter que les notations $R_0(h; \theta)$ et $\mathbb{E}^\theta_Z[\cdot]$ soulignent le lien entre le paramètre θ et les distributions de Z et X. De même, le risque moyen $\mathbb{E}^\theta_Z[R_1(\hat{h}_z; \theta)]$ de type I satisfait

$$R_1(h; \theta) = \mathbb{E}^\theta_Z[R_1(\hat{h}_z; \theta)] = \mathbb{E}^\theta_{Z,X|Y=1}[1 - h(Z, X)], \forall \theta \in \Theta_1.$$

Pour obtenir un classifieur optimal au sens paramétrique, il suffit donc de minimiser $R_1(h; \theta)$ par rapport à $\theta \in \Theta_1$.

3 Classifieur UPP

La construction d'un test UPP est délicat puisque $\mathcal{H}_0$ et $\mathcal{H}_1$ sont composites. Le test UPP est donc construit en deux temps. Tout d'abord, le Théorème 1 établit le test UPP entre l'hypothèse composite $\mathcal{H}_0$ et une sous-hypothèse simple $\mathcal{H}_1(\theta')$ de $\mathcal{H}_1$. Ensuite, il est prouvé que ce test est UPP entre $\mathcal{H}_0$ et $\mathcal{H}_1$.

Théorème 1. *Le test le plus puissant $t_\alpha^+ \in \mathcal{K}_\alpha$ pour tester*

$$\begin{aligned} \mathcal{H}_0 &: \left\{(z,x) \sim P^{\theta}_{(Z,X)} \mid \theta \in \Theta_0\right\}, \\ \mathcal{H}_1(\theta') &: \left\{(z,x) \sim P^{\theta'}_{(Z,X)}\right\}, \end{aligned} \tag{12}$$

où $\theta' = (\theta'_0, \theta'_1, \theta'_2)$ est un vecteur fixé de Θ_1, est

$$t_\alpha^+(z,x) = \begin{cases} 1 & \textit{si} \quad \Lambda(z,x) < \sigma\sqrt{1+\frac{1}{n_0}}\,\Phi^{-1}(\alpha), \\ 0 & \textit{sinon}, \end{cases} \tag{13}$$

où $\Lambda(z,x) = \operatorname{sign}(\theta'_1 - \theta'_0)(\overline{x}_0 - x) = \overline{x}_0 - x$, $\overline{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^j$ et $\Phi(\cdot)$ est la fonction de répartition de la loi normale standard.

La restriction $\theta_1 > \theta_0$ imposée en (9) garantie que le test est Uniformément Plus Puissant (UPP) pour $\mathcal{H}_0$ contre $\mathcal{H}_1$. Si l'hypothèse $\mathcal{H}_1$ considérait la restriction inverse $\theta_0 > \theta_1$, le Théorème 1 aboutirait un test $t_\alpha^-(z,x)$ tout à fait similaire à $t_\alpha^+(z,x)$: seul le sens de l'inégalité dans (13) serait changé. Sous la restriction bilatéral $\theta_0 \neq \theta_1$, le test proposé n'est pas UPP. Pour $\theta \in \Theta_1$, la probabilité de faux négatif est

$$R_1(t_\alpha^+;\theta) = \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{\frac{n_0}{n_0+1}}\frac{\theta_1-\theta_0}{\sigma}\right). \tag{14}$$

C'est la valeur moyenne optimal qui peut être atteinte par un classifieur NP paramétrique. Le test NP optimal, dit test "oracle", est le test (3) pour tester $\mathcal{N}(\theta_0,\sigma^2)$ contre $\mathcal{N}(\theta_1,\sigma^2)$:

$$h_\alpha^*(x) = 0 \iff \operatorname{sign}(\theta_1-\theta_0)(\theta_0 - x) > \sigma\Phi^{-1}(\alpha). \tag{15}$$

Le risque $R_1(t_\alpha^+;\theta)$ peut donc être comparé à la probabilité de faux négatif du test NP "oracle" $h_\alpha^*(x)$:

$$R_1^*(\theta) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{|\theta_1-\theta_0|}{\sigma}\right). \tag{16}$$

Lorsque $n_0 \to +\infty$, $R_1(t_\alpha^+;\theta)$ converge vers $R_1^*(\theta)$. Le classifieur $\hat{t}_{z,\alpha}^+$ est optimal en moyenne. Cependant, pour tout classifieur $\hat{h}_z$, les écarts d'erreur

$$\begin{aligned} e_0(\hat{h}_z;\theta) &= R_0(\hat{h}_z;\theta) - \alpha, \qquad (17) \\ e_1(\hat{h}_z;\theta) &= R_1(\hat{h}_z;\theta) - R_1^*(\theta), \qquad (18) \end{aligned}$$

sont aléatoires par rapport à la base d'apprentissage z utilisée. La proposition suivante fournit des bornes probabilistes, de type PAC, sur ces écarts pour $\hat{t}_{z,\alpha}^+$.

Proposition 1. *Soit $\alpha \in]0,1[$. Pour tout $\delta_0, \delta_1 < \frac{1}{2}$, on a*

$$P_Z^\theta\left(e_0(\hat{t}_{z,\alpha}^+;\theta) \le \frac{\Phi^{-1}(1-\delta_0)}{\sqrt{2\pi n_0}}\right) \ge 1-\delta_0, \forall\theta\in\Theta_0, \tag{19}$$

$$P_Z^\theta\left(e_1(\hat{t}_{z,\alpha}^+;\theta) \le \varepsilon(\alpha,\delta_1)\right) \ge 1-\delta_1, \forall\theta\in\Theta_1, \tag{20}$$

où

$$\varepsilon(\alpha,\delta_1) = \frac{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\delta_1)}{\sqrt{2\pi n_0}}. \tag{21}$$

Pour obtenir une contrainte stricte sur le risque de faux positif avec la borne (19), il est possible d'utiliser un classifieur ajusté $\hat{t}_{z,\alpha_0}^+(x)$ qui satisfait un risque moyen α_0 donné par

$$\Phi^{-1}(\alpha_0) = \sqrt{\frac{n_0}{n_0+1}}\Phi^{-1}(\alpha) + \frac{1}{\sqrt{n_0+1}}\Phi^{-1}(\delta_0). \tag{22}$$

Ce classifieur $\hat{t}_{z,\alpha_0}^+$ satisfait alors les bornes PAC suivantes.

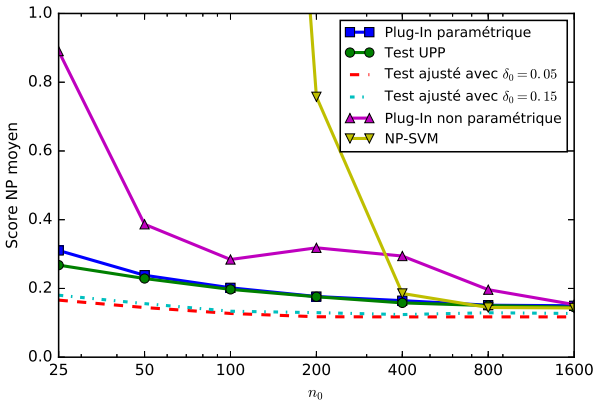

FIGURE 1 – Score NP en fonction de n_0 (échelle $\log_2$).

Proposition 2. *Soit $\alpha \in]0,1[$. Pour tout $\delta_0, \delta_1 < \frac{1}{2}$, on a*

$$P_Z^\theta\left(e_0(\hat{t}_{z,\alpha_0}^+;\theta) \le 0\right) = 1-\delta_0, \forall\theta\in\Theta_0, \tag{23}$$

$$P_Z^\theta\left(e_1(\hat{t}_{z,\alpha_0}^+;\theta) \le \epsilon_1\right) = 1-\delta_1, \forall\theta\in\Theta_1, \tag{24}$$

où

$$\Delta = \Delta(\theta) = \frac{\theta_1-\theta_0}{\sigma}, \tag{25}$$

$$\varrho = \varrho(\delta_0,\delta_1) = \frac{\Phi^{-1}(1-\delta_0)+\Phi^{-1}(1-\delta_1)}{\sqrt{n_0}}, \tag{26}$$

$$\epsilon_1 = \epsilon_1(\delta_0,\delta_1,\theta) = \Phi\left(\Phi^{-1}(1-\alpha) - \Delta + \varrho\right) - R_1^*(\theta). \tag{27}$$

La Proposition 2 montre que, pour $(1-\delta_0)*100\,\%$ des bases d'apprentissage z, le test $\hat{t}_{z,\alpha_0}^+$ satisfait la contrainte de type I. Elle montre également que, pour $(1-\delta_1)*100\,\%$ des bases d'apprentissage, la perte d'optimalité de $\hat{t}_{z,\alpha_0}^+$ est bornée par ϵ_1, qui tend vers zéro lorsque Δ ou n_0 deviennent arbitrairement grands. Le test ajusté permet d'obtenir un meilleur contrôle sur l'erreur de type I mais le risque $R_1(\hat{t}_{z,\alpha_0}^+;\theta)$ est en moyenne plus élevé que $R_1(\hat{t}_{z,\alpha}^+;\theta)$.

4 Résultats Numériques

Le test UPP (13) et le test ajusté sont comparés au classifieur "Plug-in" NP Non-paramétrique [9], noté NN^2, au classifieur NP-SVM [2], et à un classifieur du "Plug-in" paramétrique. L'algorithme NN^2 est utilisé suivant la configuration donnée dans [9], et l'algorithme NP-SVM a été construit numériquement suivant la démarche décrite dans [2]. Le classifieur du "Plug-in" paramétrique à été obtenu en remplaçant dans (15) les paramètres θ_i par les estimations $\overline{x}_i$. Les tests ajustés ont été testés pour satisfaire la contrainte (23) avec les valeurs $\delta_0 = 0.05$ et $\delta_0 = 0.15$. Tous les tests ont été testés au niveau $\alpha = 0.05$.

La comparaison s'appuie sur des lois normales simulées : $\mathcal{N}(0,1)$ pour la classe 0 et $\mathcal{N}(3,1)$ pour la classe 1. Les algorithmes sont entraînés sur des bases d'apprentissage où $n_0 = n_1$ et n_0 prend successivement les valeurs entre 25 et 1600 suivant une progression géométrique de raison 2. Pour chaque valeur de n_0, l'apprentissage est répété 1000 fois sur

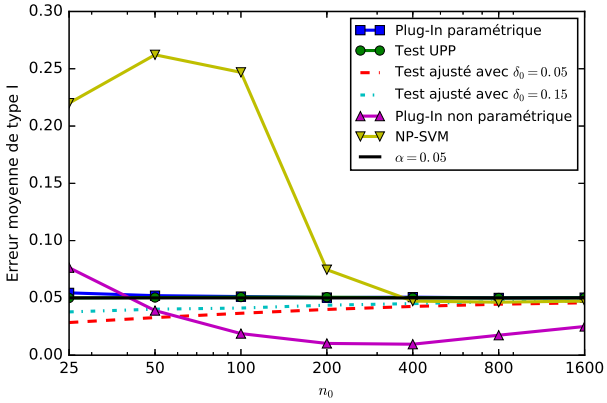

FIGURE 2 – Erreur moyenne de type I en fonction de n_0.

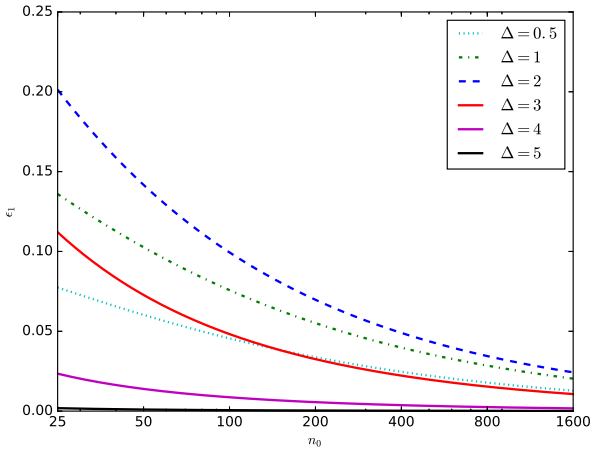

FIGURE 3 – Borne ϵ_1 de niveau exactement $1 - \delta_1$ sur l'excès d'erreur de type II $e_1(\hat{t}^+_{z,\alpha_0}; \theta)$ en fonction de n_0.

des échantillons indépendants. Les erreurs de type I, II et le score NP sont ensuite estimés sur des bases de test de 1000 exemples indépendants, puis ces estimations sont moyennées sur les 1000 répétitions de la base d'apprentissage. Le score NP est une mesure de performance [6] adaptée à la classification NP :

$$N(\hat{h}_z; \theta) = \frac{1}{\alpha} \max\{0, R_0(\hat{h}_z; \theta) - \alpha\} + R_1(\hat{h}_z; \theta) \quad (28)$$

pour tout paramètre θ. Le score NP est minimum pour h^*_α.

Le score NP est présenté sur la figure 1. La performance des classifieurs paramétriques dépend de leur capacité à maîtriser l'erreur de type I, comme cela est visible sur la figure 2. Le test UPP est légèrement plus efficace que celui du "Plug-in" paramétrique pour les faibles valeurs de n_0. Les tests ajustés, malgré leur légère perte de puissance, gagnent en score NP par leur contrôle stricte de l'erreur de type I. Le classifieur NN^2 échoue à contrôler la fausse alarme pour $n = 25$. A contrario, lorsque l'erreur de type I est sévèrement contrôlée, l'erreur de type II augmente significativement, ce qui affecte le score NP. Le classifieur NP-SVM contrôle mal l'erreur de type I. Il faut noter que la comparaison entre les classifieurs paramétriques et non-paramétriques n'est pas équitable étant donné qu'ils ne font pas les mêmes hypothèses sur les données.

La figure 3 montre la borne ϵ_1 de la Proposition 2 en fonction de n_0 pour $\delta_1 = 0.05$. Pour n_0 fixé, il est à noter que la borne ne décroit pas uniformément vers 0 en fonction de Δ mais elle converge bien vers 0 lorsque Δ devient arbitrairement grand.

5 Conclusion

Pour garantir l'efficacité d'un classifieur NP appris sur une base d'apprentissage de petite taille, cet article propose de tirer profit d'un a priori d'appartenance des distributions à une famille paramétrique. La théorie des tests d'hypothèses est utilisée pour construire un classifieur NP paramétrique UPP. Dans de futurs travaux, cette nouvelle approche sera étendue au cas d'un problème de classification bilatérale.

Références

[1] Adam Cannon, James Howse, Don Hush, and Clint Scovel. Learning with the Neyman-Pearson and min-max criteria. Technical Report LA-UR-02–2951, Los Alamos National Laboratory, 2002.

[2] Mark A. Davenport, Richard G. Baraniuk, and Clayton D. Scott. Tuning support vector machines for minimax and neyman-pearson classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(10) :1888–1898, 2010.

[3] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer Science & Business Media, New York, third edition, 2006.

[4] Evgeny Levitan and Neri Merhav. A competitive neyman-pearson approach to universal hypothesis testing with applications. *IEEE Trans. Inf. Theory*, 48(8) :2215–2229, 2002.

[5] Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. *J. Mach. Learn. Res.*, 12 :2831–2855, November 2011.

[6] Clayton Scott. Performance measures for neyman-pearson classification. *IEEE Trans. Inf. Theory*, 53(8) :2852–2863, 2007.

[7] Clayton Scott and Robert Nowak. A neyman-pearson approach to statistical learning. *IEEE Trans. Inf. Theory*, 51(11) :3806–3819, 2005.

[8] Xin Tong, Yang Feng, and Anqi Zhao. A survey on neyman-pearson classification and suggestions for future research. *Wiley Interdisciplinary Reviews : Computational Statistics*, 8(2) :64–81, 2016.

[9] Anqi Zhao, Yang Feng, Lie Wang, and Xin Tong. Neyman-pearson classification under high-dimensional settings. *J. Mach. Learn. Res.*, 17(1) :7469–7507, 2016.