EVALUATION DE LA STABILITE DES CLASSIFIEURS PAR L'ALGORITHME PLUG-IN SEMI-BORNE MODIFIE

Ibtissem Ben Othman & Faouzi Ghorbel

École Nationale des Sciences de l'Informatique,
Laboratoire CRISTAL, pôle GRIFT,
Campus Universitaire de la Manouba, 2010 Manouba, Tunisie.
ben.othman.ibtissem@gmail.com & faouzi.ghorbel@ensi.rnu.tn

Résumé. En se référant au point de vue statistique de la classification, nous tenterons dans le présent travail d'évaluer le degré de stabilité des réseaux de neurones multi-couches et Bayésiens relativement au classifieur Bayésien. L'évaluation est établie par l'estimation non paramétrique de la densité de probabilité des taux d'erreur. La comparaison basée sur ce nouveau critère est valorisée par l'utilisation de l'algorithme Plug-in semi-borné modifié. L'analyse reposera aussi sur l'évaluation du biais et de la variance des taux d'erreur connu pour être un bon estimateur de la probabilité d'erreur d'un classifieur.

Mots-clés. Réseaux de neurones Bayésiens, stabilité, probabilité des taux d'erreur, algorithme Plug-in semi-borné modifié.

Abstract. Referring to the statistical point of view, we attempt in this work to evaluate the classical and Bayesian neural networks stability degree compared to the Bayesian classifier stressing their error rate probability densities. The comparison based on this new criterion is performed using the modified semi-bounded Plug-in algorithm.

Keywords. Bayesian neural networks, stability, error rate density, modified semi-bounded Plug-in algorithm.

1 Introduction

En reconnaissance de formes, on construit souvent un nombre important de descripteurs pour discriminer les objets au sens de la différence des formes. Le nombre d'exemplaires est limité alors que la dimension du vecteur descripteur observation est important (D>10). Afin de pallier à cette limitation, la solution la plus souvent préconisée consiste en une réduction de dimension préalable conservant les propriétés de dispersion ou de séparabilité des données. Les méthodes de réduction de dimensions peuvent être classées en linéaires ou non linéaires. Les plus utilisées sont les linéaires et cela pour des raisons de complexité de mise en œuvre. L'analyse discriminante linéaire (LDA) de Fisher [3] et l'analyse en composantes principales (ACP) [6] comptent parmi les techniques linéaires les plus utilisées. La LDA permet de rechercher, dans l'espace des données, les axes qui permettent de discriminer au mieux les différentes classes. Cependant la solution de l'ACP a pour objectif de chercher les axes qui décrivent au mieux les données [2]. Une fois que la réduction de dimensions est bien établie, le problème de classification se pose. Les méthodes de classification statistiques se réfèrent souvent à la règle de la décision de Bayes qui est qualifiée d'idéale puisqu'elle minimise l'erreur de classification.

Contrairement aux méthodes statistiques classiques qui essaient de chercher une sous variété dans laquelle les données sont bien représentées, les réseaux de neurones artificiels (RNAs ou RNs) permettent de réaliser des réductions de dimensions non linéaires qu'on peut qualifier de non paramétriques. Ainsi, l'approche neuronale offre une certaine flexibilité et porte un intérêt dans les applications puisqu'elle est des fois retenue comme solution technologique. Cependant, la non maîtrise de sa formulation mathématique explique l'instabilité des résultats de ses classifications.

La littérature propose diverses études comparatives entre les RNs et les méthodes statistiques (telles que l'analyse discriminante, la régression logistique et les k plus proches voisins dans [7,16,18]). Les chercheurs tentent souvent à comparer les différents types de classifieurs relativement à la performance de leurs résultats tout en oubliant que les réseaux neuronaux sont qualifiés de non stables. Notre objectif est de comparer de manière critique les différentes approches non seulement en termes de performance mais également en termes de stabilité de leurs résultats. Dans ce sens, nous avons proposé un nouveau critère d'évaluation de la performance et de la stabilité des différents classifieurs. La comparaison est basée sur l'estimation de la densité de probabilité des taux d'erreur en utilisant une nouvelle variante de l'algorithme Plug-in semi-borné. Des expérimentations basées sur des simulations stochastiques et des images de chiffres manuscrits seront à la base de ces études comparatives.

2 Approche Bayésienne pour les réseaux de neurones artificiels

Certes, les réseaux neuronaux sont capables de détecter les relations complexes implicitement non linéaires entre les variables dépendantes et indépendantes. Cependant, la nature même de ces réseaux, se présentant sous forme de "boîte noire", la détermination de leurs paramètres par l'algorithme d'apprentissage et la non fixation de leurs architectures les rendent plus complexes et demandent une grande charge de calcul qui peut mener aux cas de sur-apprentissage [1,11].

Pour contourner les limites des RNAs, l'interprétation probabiliste de leur apprentissage en utilisant des techniques Bayésiennes, a été introduite par Mackay dans [11]. Une telle méthode permet d'apporter des avantages significatifs par rapport au procédé classique de l'apprentissage de ces classifieurs. L'apprentissage classique des RNAs consiste à déterminer le vecteur des poids qui minimise la fonction d'erreur. Dans l'approche bayésienne, tous les paramètres, notamment les poids du réseau, sont considérés comme des variables aléatoires issues d'une distribution de probabilité. Cette distribution est fixée initialement à une probabilité a priori et convertie en une distribution a posteriori, une fois les données d'apprentissage observées, grâce au théorème de Bayes [1]. En se basant sur l'approche Bayésienne des RNs, les auteurs présentent, dans [1,11], une valeur optimale de μ qui donne le meilleur compromis entre le sur-apprentissage et le sous-apprentissage. Dans [13, 14], nous avons démontré l'instabilité des résultats des RNAs par rapport à celles du classifieur de Bayes. Par ailleurs, l'amélioration de la stabilité et de la performance des résultats de classification des RNs classiques par l'intermédiaire de l'approche Bayésienne a été prouvée dans [15].

3 Un nouveau critère pour la comparaison des classifieurs

Dans cette étude, nous proposons un nouveau critère pour l'évaluation des stabilités et des performances des différentes approches. La comparaison est basée sur l'estimation des densités de probabilité des taux d'erreur, et valorisée par le calcul de leurs biais et variances.

Pour cela, nous considérons M ensembles de test indépendants. Pour chaque ensemble de test, nous obtenons un taux d'erreur pour les différents types de classifieurs. Soit $(\zeta_i)_{1\le i\le M}$ l'ensemble des taux d'erreur indépendants d'un classifieur quelconque. Ces taux d'erreur sont considérés comme étant des variables aléatoires indépendantes et identiquement distribuées et qui suivent la même fonction de densité de probabilité (pdf), $f(x)$. Pour l'estimation de cette pdf, les méthodes non-paramétriques telles que la méthode du noyau [3] ou les fonctions orthogonales [9] peuvent être utilisées.

3.1 Méthode du noyau conventionnelle

Dans ce papier, nous choisissons la méthode du noyau pour l'estimation de la pdf des taux d'erreur de chaque classifieur. L'estimateur à noyau de la densité de probabilité est défini comme suit:

$$\hat{f}_X(x) = \frac{1}{Mh_M}\sum_{i=1}^{M} K(\frac{x-\zeta_i}{h_M}) \quad (1)$$

où h_M est le paramètre de lissage qui dépend de la taille des échantillons observés *M*. *K (•)* est une pdf appelée noyau et supposée être une fonction paire, régulière et centrée réduite. Lors de notre étude, nous choisissons le noyau gaussien défini par : $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$.

Le choix du paramètre de lissage optimal est très important. Dans [5], les auteurs ont prouvé que sa valeur optimale est donnée par la minimisation de l'erreur quadratique moyenne intégrée ; $EQMI \approx \frac{L(K)}{Mh_M} + \frac{J(f)h_M^4}{4}$. Ainsi ; $h_M^* = M^{-1/5}(J(f))^{-1/5}(L(K))^{+1/5}$, avec $L(K) = \int_{-\infty}^{+\infty} K^2(x)dx$ et $J(f) = \int_{-\infty}^{+\infty}(f^{''}(x))^2 dx$. Une résolution directe de l'équation de calcul de la valeur optimale du paramètre de lissage semble très difficile. La méthode Plug-in suggérée dans [5] présente la résolution itérative de cette équation. Par ailleurs, une variante rapide de l'algorithme Plug-in a été développée en estimant directement *J(f)* à partir de l'échantillon par la double dérivation de l'expression analytique de l'estimateur à noyau [17].

3.2 Algorithme Plug-in semi-borné modifié

L'ensemble des valeurs des taux d'erreur observés $(\zeta_i)_{1\leq i\leq M}$ de chaque classifieur sont positives. Dans ce cas, la méthode du noyau conventionnelle pour l'estimation de la densité de probabilité des taux d'erreur n'est plus adéquate et peut présenter des problèmes de convergence aux bords: le phénomène de Gibbs. Plusieurs auteurs ont tenté de résoudre ce problème et ont présenté quelques méthodes pour estimer les distributions à supports bornés ou semi-bornés. Parmi lesquelles, nous pouvons citer la méthode des fonctions orthogonales et la méthode du noyau difféomorphisme [17]. Cette dernière méthode, qui est dérivée de la méthode du noyau conventionnelle, est basée sur un changement de variable appropriée par un C1-difféomorphisme. Inspiré par la méthode des noyaux, il est important de maximiser la valeur du paramètre de lissage en vue d'assurer une bonne qualité de l'estimation. L'optimisation du paramètre de lissage est effectuée par l'algorithme Plug-in difféomorphisme qui est une généralisation de l'algorithme Plug-in conventionnel [19]. Cependant, sa mise en œuvre présente des difficultés supplémentaires par rapport à l'algorithme Plug-in conventionnel puisque *L(K)* n'est plus une constante à déterminer analytiquement ou numériquement car elle dépend de la densité de probabilité à estimer. De même, *J(f)* ne dépend plus uniquement de f'', mais également de f et de f'.

Pour des raisons de complexité de mise en œuvre et de convergence, nous proposons dans cet article un algorithme Plug-in semi-borné modifié. Cette version de l'algorithme Plug-in est basée sur le changement de variable des taux d'erreur positifs : $\tau = Log(\zeta)$. Les différentes étapes de cet algorithme sont détaillées ci-dessous :

Etape 1 : En utilisant le changement de variable $\tau = Log(\zeta)$, l'expression de l'estimateur à noyau devient :

$$\hat{f}_\tau(y) = \frac{1}{Mh_M^*}\sum_{i=1}^{M} K\left(\frac{y-\tau_i}{h_M^*}\right) \quad (2)$$

Etape 2 : Itération de l'algorithme Plug-in conventionnel pour les données transformées.

Etape 3 : Calcul de $\hat{f}_X(x) = \frac{\hat{f}_\tau(Logx)}{x}$

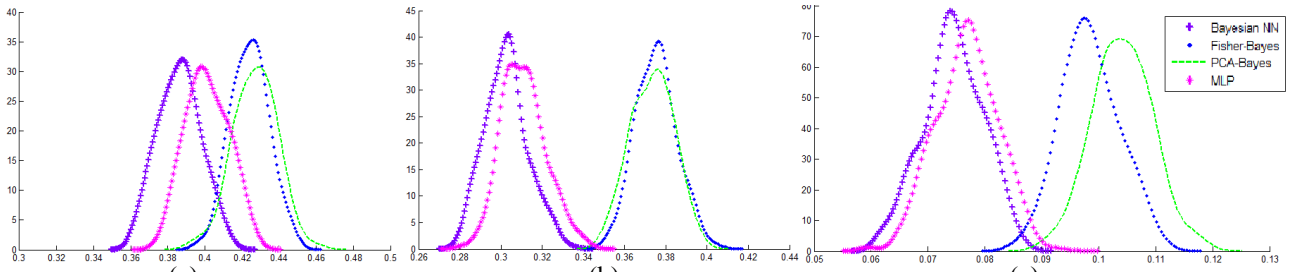
Ainsi, l'utilisation de cet algorithme tend à être un bon critère pour la comparaison de la stabilité des différents classifieurs. La raison en est que l'algorithme Plug-in modifié semi-bornée produit une précision suffisante pour l'estimation des densités et les aspects de stabilité.

4 Simulations

Pour l'évaluation des différentes approches, nous considérons le problème de classification binaire d'un mélange d'échantillons appartenant à des distributions gaussiennes distinctes.

Pour la phase d'apprentissage, nous avons fixé le même ensemble d'observations pour les différents classifieurs comportant 1000 échantillons générés pour chaque catégorie de classe. Ce même ensemble d'apprentissage sert à la fixation des paramètres des RNs Bayésiens et classiques. En se basant sur les résultats dans [7], un PMC ayant une seule couche cachée est généralement suffisant pour les problèmes de classification.

Il est bien connu que le meilleur critère d'évaluation des performances des classifieurs est le taux de mauvaise classification (TMC) calculé à partir de leurs matrices de confusion générées pour des ensembles de test. Toutefois, pour analyser et comparer la stabilité des différentes approches, nous avons testé leurs performances pour M ensembles de test supervisés, indépendants et comportant 1000 échantillons pour chaque classe. En pratique, M est égal à 100. L'évaluation de la performance et la stabilité des classifieurs (Bayes, PMC et RN Bayésien) est déterminée (dans le tableau 1) grâce à la comparaison des biais et des variances de leurs densités de probabilité des TMC, respectivement. Ces densités de probabilité sont estimées en utilisant l'algorithme Plug-in semi-borné modifié, et sont illustrées dans la figure 1 (Les cas de la figure 1 (a,b,c) correspondent aux cas du tableau 1).



(a) (b) (c)

Figure 1: Densités de probabilité des TMC des classifieurs; ACP-Bayes (en vert(--)), Fisher-Bayes (en bleu (..)), RN classique (en rose (*)) et RN Bayésien (en violet (+)).

Tableau 1 : Comparaison des performances et stabilités des approches neuronales et statistique.

Cas	Distributions		ACP-Bayes		Fisher-Bayes		PMC		RN Bayésien	
	Gaussienne 1	*Gaussienne 2*	*Moyenne*	*Variance*	*Moyenne*	*Variance*	*Moyenne*	*Variance*	*Moyenne*	*Variance*
a	μ_1=(1,..,1),$\sum_1$=2*Identité	μ_2=(1,..,1),$\sum_2$=3*Identité	0.4271	0.1414	0.4241	0.1037	0.4012	0.1289	0.3863	0.1261
b	μ_1=(0,..,0),$\sum_1$=Identité	μ_2=(0,..,0),$\sum_2$=2*Identité	0.3734	0.1054	0.3753	0.0983	0.3110	0.1140	0.3026	0.1094
c	μ_1=(0,0,0) $\sum_1$=[0.06 0 0 0 0.01 0 0 0 0.01]	μ_2=(0.1,0.1,0.1) $\sum_2$=[0.01 0 0 0 0.06 0 0 0 0.05]	0.1041	0.2804	0.0985	0.2612	0.0768	0.2877	0.0745	0.2709

Les résultats obtenus montrent que le PMC est moins stable que le classifieur de Bayes et le RN Bayésien. Il présente toujours une courbe de densité de probabilité des taux d'erreur la plus large, et donc la variance des taux d'erreur la plus grande. Le RN Bayésien montre une amélioration de la performance et la stabilité relativement au PMC. En effet, sa courbe de densité de probabilité d'erreur est située à gauche de celle du PMC (admettant un biais des taux d'erreur plus faible). Et ayant une courbe plus étroite, le RN Bayésien présente une variance des taux d'erreur plus petite et donc une meilleure stabilité.

5 Application à la reconnaissance des chiffres manuscrits

Comme application, nous étudions le problème de la reconnaissance des chiffres manuscrits qui reste actuellement l'un des sujets les plus actifs dans le tri automatique des courriers postaux et l'enregistrement des chèques bancaires. La base d'images utilisée dans le présent document est extraite de la base MNIST. Cette dernière est construite de 10 classes de chiffres manuscrits revenant à 250 écrivains. Notre ensemble d'apprentissage est composé de 10000 images isolées des chiffres de '0' à '9' choisi parmi l'ensemble d'apprentissage MNIST. Les dix classes de chiffres sont équiprobables.

L'étape la plus délicate pour la reconnaissance des chiffres manuscrits est le choix des primitives adéquates qui doit se baser sur un ensemble de critères non exhaustives tels que la stabilité, la complétude, la rapidité de calcul, la forte discrimination et l'invariance relativement aux transformations géométriques. La famille d'invariants proposée par Ghorbel dans [4] vérifie les différents critères déjà cités. Ainsi, chaque image de chiffre est décrite par ce type d'invariants et les descripteurs de Fourier (DF). Etant donnée que la taille des descripteurs retenus est élevée (D=14), une 2D-réduction de dimension est réalisée par l'ACP ou la LDA de Fisher afin de pouvoir appliquer la règle de Bayes. Pour l'approche neuronale, une architecture comportant 14 neurones dans sa couche d'entrée, 10 neurones dans la couche cachée et 10 neurones dans sa couche de sortie, a été prise en compte dans cette analyse. Comme dans les expériences précédentes, les performances des classifieurs sont évaluées 100 fois sur des ensembles de test sélectionnés par l'algorithme de la validation croisée de l'ensemble de test MNIST pour les dix classes de chiffres (1000 images pour chaque classe). Les figures 2.a et 2.b illustrent l'estimation des densités de probabilités d'erreur de chaque classifieur pour les descripteurs de Fourier et les descripteurs de Ghorbel, respectivement. Le tableau 2, ci-dessous, valorisent la comparaison des différents modèles par la présentation des biais et variances de leurs taux d'erreur.

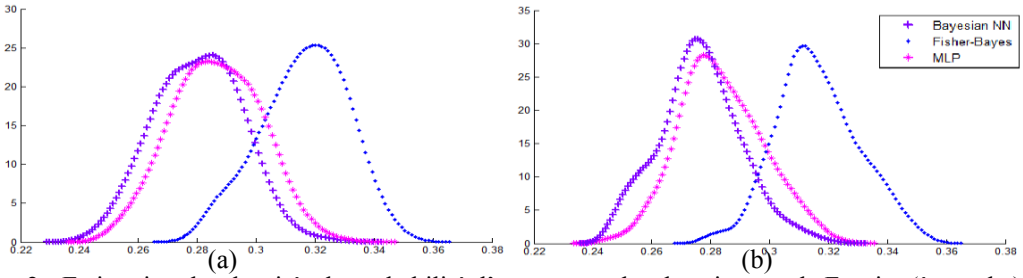


Figure 2 : Estimation des densités de probabilité d'erreur pour les descripteurs de Fourier (à gauche) et les descripteurs de Ghorbel (à droite).

Tableau 2. Résultats de comparaison des RNAs et des modèles statistiques.

	Descripteurs de Fourier		**Descripteurs de Ghorbel**	
	Moyenne	**Variance**	**Moyenne**	**Variance**
Fisher-Bayes	0.3163	0.1957	0.3157	0.1877
PMC	0.2864	0.2027	0.2836	0.1892
RN Bayésien	0.2803	0.1951	0.2765	0.1852

La figure 2 et le tableau 2 montrent que les classifieurs admettent des meilleurs résultats pour les descripteurs de Ghorbel. Etant plus performant que l'approche statistique, l'approche neuronale reste moins stable en admettant une variance des taux d'erreur la plus élevée. Cependant, le RN Bayésien présente une amélioration de la performance et la stabilité relativement au PMC.

Conclusion

Dans cet article, nous avons proposé un nouveau critère pour l'évaluation de la stabilité des RNAs relativement aux approches statistiques classiques. La comparaison est basée sur l'estimation non paramétrique de la densité de probabilité des taux d'erreur en utilisant l'algorithme Plug-in semi-borné modifié. Les résultats préliminaires trouvés montrent clairement que le classifieur de Bayes est plus stable que les RNs classiques. Cependant, l'approche Bayésienne pour la modélisation des RNAs améliore leurs stabilités et leurs performances. L'amélioration de la stabilité par la combinaison des classifieurs fera l'objectif de nos futurs travaux.

Bibliographie

[1] Bishop, CM. (1995), Neural networks for pattern recognition, *Oxford University Press*, Oxford.

[2] Drira, W. et Ghorbel, F., (2010), Réduction de dimension par un nouvel estimateur de la distance de patrick Fisher à l'aide des fonctions orthogonales, *42èmes Journées de Statistique*.

[3] Fukunaga, K. (1990), Introduction to Statistical Pattern Recognition, *Academic Press*.

[4] Ghorbel, F. (1998), Towards a unitary formulation for invariant image description: application to image coding, *Annals of telecommunication*, vol. 53, France.

[5] Hall P., Marron J., (1987), Estimation of integrated square density derivatives, *J. statistics and probability letters*, vol.74, p.567-581.

[6] Jolliffe, I.T. (1986), Principal Component Analysis, Springer-*Verlag*.

[7] Kumar, U.A. (2005), Comparison of neural networks and regression analysis: A new insight, *Expert Systems with Applications*, 29(2), 424–430.

[8] Lauret, P., Fock, E., Randrianarivony, R.N., & Manicom-Ramsamy, J.F. (2008), Bayesian neural network approach to short time load forecasting, *Energy conversion and management*, 49, 1156-1166.

[9] Lepage, R. and Solaiman, B. (2003), Les réseaux de neurones artificiels et leurs applications en imagerie et en vision par ordinateur. *Presses de l'École de technologie supérieure*, Montréal.

[10] MacKay, DJC. (1992), A practical Bayesian framework for back-propagation networks, *Neural Comput*, 4(3), 448–72.

[11] MacKay, DJC. (2003), Information theory, inference, and learning algorithms, *Cambridge University Press,* Cambridge.

[12] Nabney, IT. (2002), NETLAB: Algorithms for pattern recognition, London.

[13] Othman, I.B. and Ghorbel, F. (2013), A New criterion for Comparing Neural Networks and Bayesian Classifier, *ICCAT' 2013*, Tunisia.

[14] Othman. I.B and Ghorbel. F, (2013), A Stability Analysis of Neural and Statistical Approaches, *TAIMA'2013*, ed.8 Art-pi, Tunisia.

[15] Othman. I.B et Ghorbel. F, (2013), Stabilité des classifieurs neuronaux relativement au classifieur de Bayes, *45ème Journées de Statistique*, Toulouse.

[16] Paliwal, M. and Kumar, U.A. (2009), Neural networks and statistical techniques: A review of applications, *Expert Syst, Appl*, 36(1), 2–17.

[17] Saoudi, S., Troudi, M. and Ghorbel, F. (2009), An iterative soft bit error rate estimation of any digital communication systems using a non parametric probability density function, *Eurasip Journal on wirless Communications and Networking*.

[18] Tam, K.Y. and Kiang, M.Y. (1992), Managerial applications of neural networks: The case of bank failure predictions, *Management Science*, 38(7), 926–947.

[19] Troudi M., Ghorbel F. (2013), The generalised Plug-in algorithm for the diffeomorphism kernel estimate, *International Conference on Systems, Control, Signal Processing and Informatics*, Italy.