Comparaison de classifieurs profonds par un critère basé densité de probabilité des taux d'erreurs

Mahmoud GHORBEL[1], Molka TROUDI[2], Faouzi GHORBEL[3]

[1]Digital Research Center of Sfax
3021, Sfax, Tunisia

[2]Université de Carthage
ECSTRA Lab - IHEC Carthage
Rue Victor Hugo, 2016, Carthage, Tunisie

[3]Université de la Manouba
Cristal Lab - ENSI
Campus universitaire de la Manouba, 2010, Tunisie

mahmoud.ghorbel1991@gmail.com, molka.ghorbel@ihec.ucar.tn,
faouzi.ghorbel@gmail.com

Résumé – Généralement, les algorithmes d'apprentissage profond sont évalués en termes de taux d'erreur ou du f-score. Le biais et la variance sont souvent retenus pour des comparaisons rigoureuses entre classifieurs. Ici, nous proposons une approche plus complète basée sur les estimateurs des fonctions densités de probabilité (fdp) des taux d'erreur respectifs des classifieurs. Pour les bases de données de petite taille,les taux d'erreurs générés par les réseaux de neurones profonds comme les architectures CNN présentent des valeurs de souvent très dispersées. L'estimation de la densité de probabilité de la variable aléatoire taux d'erreur permettrait de comparer les performances des classifieurs de manière plus précise. Dans ce contexte, nous proposons de générer des échantillons de 100 taux d'erreur puis d'estimer leurs densité par des méthodes non paramétriques. Dans ce travail, l'estimateur à noyau difféomorphisme ajusté au sens de son paramètre de lissage, connu pour sa précision au sens de l'Ecart Quadratique Moyen Intégré (EQMI) tout en tenant compte des contraintes sur le support de la distribution, a été utilisé. Nous introduisons également un critère de comparaisons basé sur la convolution entre les densités associées aux classifieurs. Des expérimentations sont menées pour la comparaison dans ce sens des algorithmes Lenet et CNN sur la base d'images MPEG-7 .

Abstract – Generally, deep learning algorithms are evaluated in terms of error rate or f-score. Bias and variance are often retained for rigorous comparisons between classifiers. Here, we propose a more complete approach based on the estimators of the probability density functions (pdf) of the respective error rates of the classifiers. For small databases, the error rates generated by deep neural networks such as CNN architectures often present very scattered values. The estimation of the probability density of the random variable error rate would make it possible to compare the performances of the classifiers in a more precise way. In this context, we propose to generate samples of 100 error rates and then to estimate their densities by non-parametric methods. In this work, the diffeomorphism-adjusted kernel estimator in the sense of its smoothing parameter, known for its accuracy in the sense of the Integrated Mean Squared Deviation (IMEQ) while taking into account the constraints on the support of the distribution, has been used. We also introduce a comparison criterion based on the convolution between the densities associated with the classifiers. Experiments are carried out for the comparison in this sense of the Lenet and CNN algorithms on MPEG-7 images Databasis.

1 Introduction

Il est maintenant bien connu que les algorithmes de classification du type apprentissage profond semblent présenter des taux de reconnaissance très supérieurs aux méthodes conventionnelles telles que celles basées sur les approches statistiques. Récemment, de multiples travaux se sont orientés vers la caractérisation de leur stabilité (ref). Cette tendance peut être justifiée par le nombre très élevé de poids dans un réseau de neurones profond. L'estimation stable de ces poids lors de la phase d'apprentissage nécessite des bases de données de tailles très élevées. Des méthodes d'augmentation de données de différents types (transformation géométrique, morphing d'image, bruitage des données d'entrées, ...) sont souvent requises. Ces effets d'instabilité sont souvent observés de manière accentuée surtout lorsque les bases de données mises en jeu sont de relative faible taille à l'image de la base d'images MPEG-7 en vision. Dans ce premier travail, nous proposons une méthode de comparaison des classifieurs au sens de la précision et de la robustesse. En modélisant les taux d'erreur par des variables aléatoires absolument continues, la comparaison des performances des classificateurs de manière complète est réalisée par

la superposition des fonctions densité (fdps) associées à chaque classificateur.
La complexité de la forme des fdps ne peut être représentée de manière précise avec les méthodes d'estimation paramétriques. Les estimations non paramétriques parmi lesquelles nous pouvons citer l'histogramme, le noyau, les fonctions orthogonales ...etc[9], semblent plus adaptées au problème posé. Pour assurer la convergence de ce type d'estimateur, il est nécessaire d'optimiser les paramètres de lissage au sens d'un critère de convergence à l'exemple de l'Ecart Quadratique Moyen Intégré (EQMI).
La méthode du noyau [8] est adaptée à l'estimation des fdps des taux d'erreur des différents classifieurs à comparer. Les formes des fdps associées à chaque classificateur permettent de repérer le meilleur classifieur au sens de la performance et de la robustesse. A partir de ces densités de probabilité, un critère de comparaison est proposé. Il correspond à la probabilité pour que le taux d'erreur d'un classifieur donné soit inférieur à un second. Nous montrons dans ce papier que cette quantité est exactement la fonction de répartition de la fonction corrélation des deux fonctions densité de probabilité calculée en 0. Son estimation est réalisée par l'intermédiaire d'une intégration numérique de la corrélation des deux fonctions de densité estimées par la méthode du noyau ajustée par l'algorithme plug-in [3]. Dans la partie expérimentations, nous montrons comment ce critère est significatif et discriminant. De plus, il rend compte à la fois de la précision et de la stabilité relative des algorithmes à comparer. Par exemple, ce critère peut permettre de quantifier l'amélioration obtenue par l'augmentation des données pour un CNN pour des bases de faibles tailles.

2 Formulation mathématique du critère

Dans ce paragraphe, nous introduisons un critère permettant de comparer deux classifieurs donnés C_1 et C_2 en nous basant sur les fdps estimées de leurs taux d'erreur. Ces derniers varient à chaque exécution du classifieur grâce à l'échantillonnage boostrap des échantillons d'apprentissage et de test. Les probabilités d'erreur peuvent être considérées comme des variables aléatoires absolument continues pouvant être estimées par le taux d'erreur connu comme étant un estimateur consistant.
Notons par T_1 et T_2 les taux d'erreurs associés respectivement aux deux classifieurs C_1 et C_2. Il est assez raisonnable de supposer que ces deux variables aléatoires sont indépendantes. Le critère proposé dans ce papier se base sur l'évaluation de la probabilité que T_1 soit inférieur à T_2 ($P[T_1 < T_2]$) ou vice-versa. Cela revient à calculer la quantité $P[T_1 - T_2 < 0]$ qui correspond à $F_{T_1-T_2}(0)$, la fonction de répartition de la variable aléatoire de $T_1 - T_2$ calculée en 0.

$$F_{T_1-T_2}(t) = \iint_{[(t_1,t_2)\in\mathbb{R}^2 \setminus t_1-t_2<t]} f_{(T_1,T_2)}(t_1,t_2)dt_1dt_2 \tag{1}$$

ou $f_{(T_1,T_2)}$ est la loi conjointe du couple aléatoire (T_1, T_2) et $\mathbb{R}^2$ le plan réel.

Notons par f_{T_1} (resp. f_{T_2}) la densité de probabilité associée à T_1 (resp. T_2).Comme le couple (T_1, T_2) est supposé indépendant, la fonction de répartition s'écrit comme suit :

$$F_{T_1-T_2}(t) = \iint_{[(t_1,t_2)\in\mathbb{R}^2 \setminus t_1-t_2<t]} f_{T_1}(t_1)f_{T_2}(t_2)dt_1dt_2 \tag{2}$$

En effectuant le changement de variable $(u, v) = (t_1 - t_2, t_2)$, l'équation 2 s'écrit :

$$F_{T_1-T_2}(t) = \iint_{[(u,v)\in\mathbb{R}^2 \setminus u<t]} f_{T_1}(u+v)f_{T_2}(v)dudv \tag{3}$$

$$F_{T_1-T_2}(t) = \int_{-\infty}^{t}\int_{-\infty}^{+\infty} f_{T_1}(u+v)f_{T_2}(v)dudv \tag{4}$$

$$F_{T_1-T_2}(0) = \int_{-\infty}^{0} f_{T_1} * \overset{\vee}{f}_{T_2}(u)du \tag{5}$$

avec $\overset{\vee}{f}_{T_2}(u) = f_{T_2}(-u)$ et $*$ l'opération de convolution entre fonctions de carré intégrables.
Nous notons que lorsque un classifieur C_1 donne de manière plus fréquente des taux d'erreur inférieurs à ceux donnés par un classifieur C_2 pour un même jeu de données, son score sera supérieur à 0,5. Ce score ne pourra être de 0.5 que dans le cas particulier ou les deux densités de probabilité sont identiques. Si les densités de probabilité des classifieurs n'ont aucun recouvrement, le score du classifieur le plus performant sera de 1. Ainsi, la configuration dans laquelle les deux classifieurs ont les mêmes biais et variance ne fournit pas nécessairement un critère égal à 0,5. Cela permet de comparer des classifieurs dont les taux d'erreur ne présentent de différences que pour leurs moments d'ordre supérieur (Skweeness, Kurtosis).

3 Estimation du critère

Afin d'estimer la valeur du critère proposé de la manière la plus précise possible, nous proposons de remplacer dans la formule (4) les densités de probabilité f_{T_1} et f_{T_2} par leurs estimateurs respectifs obtenus par l'estimateur à noyau dont l'expression est présentée dans l'équation (6).

$$\hat{f}_{T,N}(u) = \frac{1}{Nh_N}\sum_{i=1}^{N} K(\frac{(u-T^i)}{h_N}) \tag{6}$$

où $T^1, ..., T^N$ est un échantillon de taille N de la variable aléatoire T et K une densité de probabilité appelée noyau. Le paramètre de lissage h_N qui représente l'écart type du noyau requiert une optimisation selon un critère. Une étude asymptotique au sens de l'EQMI [10] permet d'exprimer le paramètre de lissage optimal noté h_N^* par :

$$h_N^* = N^{-\frac{1}{5}}.(J(f))^{-\frac{1}{5}}.(M(K))^{\frac{1}{5}} \tag{7}$$

Avec :

$$M(K) = \int_{-\infty}^{+\infty} K^2(x)dx \tag{8}$$

$$J(f) = \int_{-\infty}^{+\infty} (f''(u))^2 du \tag{9}$$

L'expression de $M(K)$ est simple à approcher. En revanche, l'approximation de $J(f)$ a fait l'objet de plusieurs travaux car cette entité dépend de la densité inconnue à estimer [7][3]. L'ajustement du paramètre de lissage h_N est essentiel pour aboutir à des estimations précises. L'algorithme plug-in [3] [10] permet d'optimiser le paramètre de lissage au sens de l'erreur quadratique moyenne intégrée (EQMI). Toutefois, pour les variables aléatoires sujettes à des contraintes de type support borné ou semi borné, la densité estimée peut déborder de son support naturel. Dans des travaux précédents [11][12], nous avons proposé d'ajuster la méthode du noyau-difféomorphisme en optimisant le paramètre de lissage associé par une version généralisée de l'algorithme plug-in.
Les taux d'erreur des classifieurs de type profonds peuvent donner des taux d'erreur très proches de 0. Cela nous amène à considérer le cas des contraintes semi-bornées. Dans ce qui suit, nous rappelons l'expression de l'estimateur du noyau difféomorphisme (10).

$$\hat{f}_{T,N}(u) = \frac{|\varphi'(u)|}{N h_{N_\phi}} \sum_{i=1}^{N} K\left(\frac{\varphi(u_i) - \varphi(T^i)}{h_N}\right) \tag{10}$$

où φ est un difféomorphisme du support de f_T dans R. Là aussi, h_N^* doit être ajusté pour assurer la convergence de l'estimateur. Une étude asymptotique, détaillée dans [11][12], permet d'étudier la convergence de cet estimateur et d'exprimer analytiquement $h^*_{N_\phi}$ (11)

$$h^*_{N_\phi} = [M_\Phi(K)]^{\frac{1}{5}} [J_\Phi(f)]^{-\frac{1}{5}} N^{-\frac{1}{5}} \tag{11}$$

$h^*_{N_\phi}$ dépend des entités $M_\Phi(K)$ et $J_\Phi(f)$ dont les expressions sont toutes deux dépendantes de la densité inconnue. Un algorithme plug-in adapté à l'estimateur permet d'approcher ces valeurs en minimisant l'EQMI [11][12].

4 Expérimentations

Dans ce paragraphe, nous illustrons l'approche présentée par son application sur la classification d'images de la base MPEG-7 [5]. Cette dernière contient un total de 1400 images pour 70 classes d'objets rigides ou déformables. Nous avons choisi de comparer deux classifieurs de type réseaux de neurones profonds simples. Il s'agit de Lenet qui est l'un des modèles de CNN proposés pour traiter des problèmes relativement simples tels que la reconnaissance de caractères et de formes 2D [1] [4] et d'une seconde architecture appelée S-CNN décrite dans [2].

Les deux classifieurs comparés ont été exécutés 100 fois sur chaque jeu de données avec ou sans data augmentation. L'augmentation de la base de donnée a été réalisée en utilisant le framework KERAS de Python.

Nous observons dans la figure 1 que la densité de probabilité des taux d'erreurs associés au classifieur S-CNN sans augmentation des données devance relativement à 0 celle obtenue

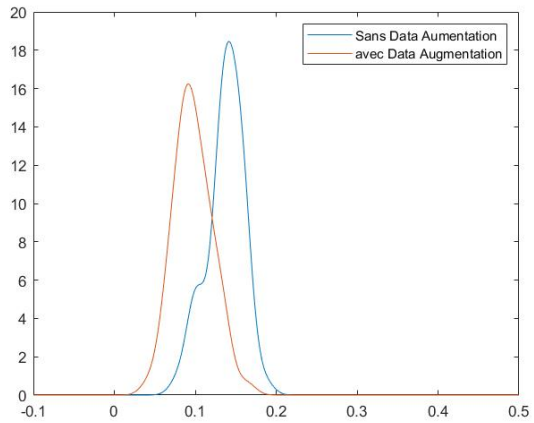

FIGURE 1 – fdps des taux d'erreurs du S-CNN avec et sans data augmentation. $P[T_1 < T_2] = 0.86$

après augmentation des données. Toutefois, un recouvrement existe entre les deux densités. Cela explique la valeur du critère $P[T_1 < T_2] = 0.86$ avec T_1 les taux d'erreur suite à l'augmentation de données et T_2 sans augmentation des données. Cela indique que l'augmentation des données améliore les performance du classifieur de 36%. Nous constatons également que l'augmentation des données n'améliore pas significativement les données. Il est à noter que la densité sans augmentation de données présente des asymétries dans sa forme ce qui signifie que certains taux d'erreurs supérieurs à sa médiane sont plus probables. Pour le classifieur Lenet, une amélioration relativement modérée des taux d'erreur suite à l'augmentation des données est observée sur la figure 2. Cette amélioration est évaluée à 22% puisque $P[T_1 < T_2] = 0.72$. La figure 3 représente les fdps des taux d'erreur des deux classifieurs après augmentation des données. Bien que les deux densités soient presque superposées, le S-CNN paraît plus performant puisque les valeurs des taux d'erreur sont moins dispersées et que la fdp de Lenet présente deux modes éloignés. Cela indique que le risque d'avoir des faibles performances du classifieur existe avec une faible mais non négligeable probabilité (de l'ordre de 4%). Le critère proposé confirme cette observation. Ainsi, $P[T_{CNN} < T_{Lenet}] = 0.56$.

5 Conclusion

Dans cet article, nous avons introduit une nouvelle démarche tenant compte des moments d'ordre supérieur à 2 de l'estimateur taux d'erreur pour la comparaison des classifieurs. Cette approche se base sur les densités de probabilité des taux d'erreur associés à chaque classifieur. L'estimateur du Noyau-difféomorphisme a été d'une grande utilité pour se doter d'une estimation à la fois ajustée et tenant compte d'information sur le support des fdps du taux d'erreur. Un critère basé sur la convolution de ces fonctions densités a été introduit permettant

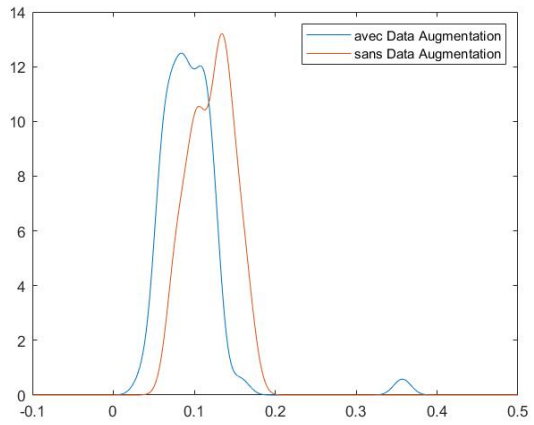

FIGURE 2 – fdps des taux d'erreurs de lenet avec et sans data augmentation. $P[T_1 < T_2] = 0.72$.

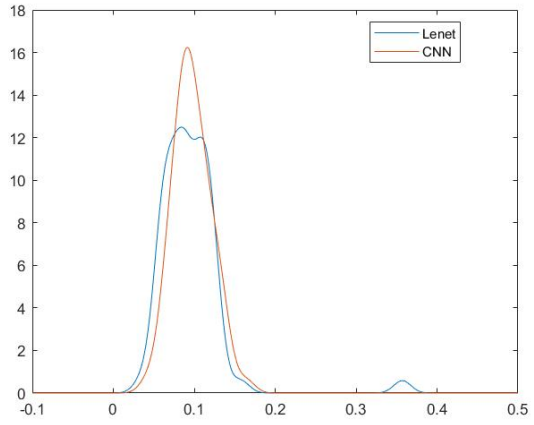

FIGURE 3 – fdps des taux d'erreurs de lenet et de S-CNN avec et data augmentation. $P[T_{CNN} < T_{Lenet}] = 0.56$

ainsi de comparer un ensemble de classifieurs de type CNN profonds avec des architectures adaptées aux données de petite taille. L'augmentation de données de type transformations géométriques a été aussi évaluée. Ce critère qui se base sur une information complète au sens statistique nous permet d'étendre les méthodes conventionnelles, se basant sur le couple Biais-variance. Dans nos futurs travaux, nous avons l'intention de mener des études de comparaisons plus larges incluant les méthodes de classification statistiques et celles basées sur la modélisation par les arbres de décision. Des comparaisons sur les bases de données larges feront aussi l'objet de nos prochaines études incluant le comportement non symétrique et d'aplatissement des fdps des taux d'erreur de certains classifieurs. Les différentes méthodes d'augmentation de données seront évaluées et comparées à l'aide du critère proposé.

Références

[1] A. El-Sawy, H. El-Bakry and M. Loey. *CNN for handwritten arabic digits recognition based on LeNet-5*. International conference on advanced intelligent systems and informatics, pages 566-575, Springer 2016.

[2] E. Ghorbel, M. Ghorbel and S. Mhiri. *Data augmentation based on invariant shape blending for deep learning classification*. Accepted in International Conference on Image Processing, IEEE 2022.

[3] M.C. Jones, S. Marron and J. Sheather. *A brief survey of bandwidth selection for density estimation*.Journal of the American Statistical Association, 91(433) :401-407, 1996.

[4] M. Kayed, A. Anter and M. Hadeer *Classification of garments from fashion MNIST dataset using CNN LeNet-5 architecture*. International Conference on Innovative Trends in Communication and Computer Engineering (ITCE). pages 238-243, IEEE 2020.

[5] L.J. Latecki, R. Lakamper and T. Eckhardt. *Shape descriptors for non-rigid shapes with a single closed contour.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(Cat. No. PR00662). Volume 1, pages 424-429, 2000.

[6] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard and L. Jackel. *Backpropagation applied to handwritten zip code recognition*. Neural computation, Volume 1, pages 541-551, 1989.

[7] A. Mugdadi and I.A. Ahmad. *A bandwidth selection for kernel density estimation of functions of random variables*. 47 :49-62,08 2004.

[8] E. Parzen. *On Estimation of a Probability Density Function and Mode*.Ann. Math. Statist. Volume 33, pages 1065-1076, https ://10.1214/aoms/1177704472.

[9] B.W. Silverman. *Density Estimation for Statistics and Data* Analysis.Chapman & Hall, London, 1986.

[10] M. Troudi,F. Ghorbel, *Ajustement des paramètres de lissage de l'estimateur ponctuel du noyau en tenant compte de la nature du support*.in F. Ghorbel, S. Derrode, O. Alata (ed.), Récentes avancées en reconnaissance de formes statistiques. ARTS-PI éditions, 2012, pp.19-71.

[11] M. Troudi, F. Ghorbel. *Extension de l'algorithme plug-in pour l'optimisation du paramètre de lissage de l'estimateur du noyau-difféomorphisme.*. Traitement du signal. n° 3-4/2014, pages 321-338. DOI :10.3166/TS.31.321-338. Lavoisier 2014.

[12] M. Troudi, F. Ghorbel. *The generalised plug-in algorithm for the diffeomorphism kernel estimate.*. International Journal of Mathematics and Computers in Simulation. Volume 15/2021, pages 128-133. DOI :10.46300/9102.2021.15.24.