Réseaux de Neurones Convolutifs avec Apprentissage Minimax pour des Proportions par classe incertaines et déséquilibrées

Marie Guyomard, Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Réseaux de Neurones Convolutifs avec Apprentissage Minimax pour des Proportions par classe incertaines et déséquilibrées

Marie Guyomard[1], Cyprien Gilet[1], Susana Barbosa[2], Lionel Fillatre[1]

[1] Université Côte d'Azur, CNRS, I3S, Sophia-Antipolis, France
[2] Université Côte d'Azur, CNRS, IPMC, Sophia-Antipolis, France

Laboratoire I3S, Euclide B, 2000 Route des Lucioles, 06900 Sophia-Antipolis
guyomard@i3s.unice.fr

Résumé

Ce papier propose une nouvelle approche ajustant les réseaux de neurones convolutifs appliqués sur des jeux de données déséquilibrés dont les proportions par classes sont incertaines. La règle de décision constitutant la sortie du réseau de neurones est remplacée par le classifieur Minimax dont la particularité est de chercher à égaliser les risques conditionnels. De ce fait, le réseau de neurones devient robuste au déséquilibre des bases de données ainsi qu'au changement de probabilités a priori. Des expériences numériques sur des images médicales mettent en évidence la pertinence de notre approche quand il est nécessaire de classifier correctement les classes les moins représentées. Les résultats obtenus sur la base de données CIFAR100 démontre l'extensibilité de notre méthode en présence d'un grand nombre de classes. [1]

Mots Clef

Réseau de Neurones Convolutifs, Bases de données déséquilibrées, Changement de probabilités *a priori*, Classifieur Minimax.

Abstract

This paper proposes a new approach for adjusting convolutional neural networks when dealing with imbalanced datasets and prior probability shifts. The output decision rule of the trained neural network is replaced by a minimax classifier that tends to equalize the class-conditional risks of misclassification. Hence, the neural network becomes robust face to imbalanced classes and prior probability shifts. Numerical experiments on medical images show the relevance of our approach when it is necessary to well classify the classes with the smallest number of training images. Results on the CIFAR100 dataset show the scalability of our method when the number of classes is large.

1. Un article présentant cette méthode a été soumis à l'*IEEE International Conference on Image Processing* 2021. Cette version contient plus de détails.

Keywords

Convolutional Neural Network, Imbalanced datasets, Prior probability shift, Minimax classifier.

1 Introduction

L'utilisation de réseaux de neurones convolutifs (CNNs) est devenue incontournable pour la classification d'images dans un grand nombre de domaines d'application. Cependant, en présence de bases de données déséquilibrées la performances de classification des CNNs est impactée [1, 2, 3, 4]. Lorsque les classes sont inégalement représentées la plupart des CNNs cherchent naturellement à prédire les classes dominantes, à savoir celles contenant le plus grand nombre d'images, et tendent à sous-estimer les moins représentées. Autrement dit, les classes minoritaires avec seulement peu d'images auront un risque conditionnel d'erreur de classification très élevé.

De plus, comme la plupart des méthodes de *Machine Learning*, les CNNs sont sensibles à d'éventuels changements de proportions par classe. Ces changements de proportions par classe se produisent lorsque la vrai distribution *a priori* évolue au cours du temps pour des raisons inconnues, et que les proportions par classe d'observations test diffèrent de celles observées dans la base d'apprentissage [5, 6]. Lorsque un changement de probabilité *a priori* survient, le risque d'erreur moyen évolue de manière linéaire et significative [7, 8]. Il est alors crucial de les prendre en considération lorsque l'on construit un CNN. Comme discuté dans [8, 9], la sensibilité d'un classifieur à un changement de probabilité *a priori* est davantage importante lorsque les risques conditionnels sont déséquilibrés.

Ainsi, les changements de probabilité *a priori* et la présence de bases de données déséquilibrées partagent un trait commun, à savoir la sensibilité aux risques conditionnels déséquilibrés. Égaliser les risques conditionnels apparaît alors essentiel à l'obtention d'un classifieur robuste face aux difficultés de proportions par classe qui plus est interviennent dans un grand nombre d'applications comme par

exemple en médecine de précision.

1.1 Gérer le problème des bases de données déséquilibrées

Dans [1], les auteurs apportent un panorama intéressant d'approches visant à résoudre le problème des bases de données déséquilibrées dans le *Deep Learning*. Une approche commune est de ré-échantillonner la base d'apprentissage (sur-échantillonnage ou sous-échantillonnage) en équilibrant les proportions par classe [3]. Néanmoins, cette approche introduit un biais puisque le vrai état de nature reste déséquilibré. Une autre méthode que les auteurs citent est le *cost-sensitive learning*, étudié dans [10, 11], visant à assigner différents coûts d'erreur de classification par classe afin de contrebalancer le nombre d'occurrences de chaque classe. Cependant, ces coûts sont difficiles à optimiser lorsque les bases de données contiennent un grand nombre de classes. D'autres approches comme le seuillage (*thresholding*) [12], la classification uni-classe (*one-class classification*) [13] ou encore un hybride de ces méthodes (*hybrid of methods*) [14] tentent également de résoudre le problème des proportions par classe déséquilibrées. Les auteurs dans [15] proposent de remplacer la fonction objectif standard de cross-entropy durant la procédure d'apprentissage. Dans [16] les auteurs proposent quant à eux une méthode rééquilibrant les probabilité *a priori* à l'issue de la phase d'apprentissage.

D'un point de vue théorique, une manière raisonnable de rendre un classifieur robuste aux problèmes de classes déséquilibrées et aux changements de probabilités *a priori* est de considérer un classifieur *Minimax* [8]. Le classifieur *Minimax* cherche en effet à minimiser le maximum des risques conditionnels lors de la phase d'apprentissage. Ainsi, cette méthode tend à égaliser les risques par classe ce qui induit une robustesse dur risque d'erreur global face aux changementx de proportions par classe. L'approche Minimax appliquée aux réseaux de neurones a déjà été étudiée dans [17]. Les auteurs proposent un algorithme nécessitant le rééchantillonnage de la base d'apprentissage. En revanche un tel rééchantillonnage n'est pas réalisable lorsque certaines classes ne contiennent que quelques échantillons, que le nombre total d'échantillons d'apprentissage est limité ou que le nombre de classes est important, ce qui apparaît généralement dans de nombreuses applications réelles.

1.2 Gérer le problème de changement de probabilités *a priori*

Dans le but de palier au problème de changement de probabilité *a priori*, une nouvelle méthode de classification supervisée a émergé, appelée la quantification, comme mentionnée dans [9]. A partir de l'échantillon d'apprentissage, cette approche vise à estimer les proportions pas classe de l'échantillon de test afin d'améliorer la performance générale associée à ces nouvelles observations test. Cependant les approches de quantification nécessitent que la prédiction se fasse sur un ensemble d'observation test en même temps, ce qui n'est pas toujours envisageable pour grands nombres d'application réelles comme la médecine. Dans le contexte des CNNs, la méthode proposée dans [16] convient également pour traiter le problème de changements de probabilités *a priori*.

1.3 Contributions

La contribution de ce papier est triple. Premièrement, nous couplons le classifieur *Minimax* à un CNN pré-entraîné afin de rééquilibrer les risques conditionnels de ce réseau de neurones. Deuxièmement, nous proposons un algorithme de sous-gradient projeté pour entraîner le classifieur *Minimax*. Cet algorithme peut prendre en considération toutes fonctions de perte permettant de mesurer l'erreur de classification entre les classes. Enfin, nous testons notre algorithme sur plusieurs bases de données d'images médicales pour lesquelles il est crucial d'assurer un faible risque conditionnel pour la classe minoritaire. Nous appliquons aussi notre algorithme sur la base de données CIFAR100 contenant 100 classes.

La structure du papier est la suivante. La partie 2 explique comment coupler un CNN pré-entraîné avec le classifieur *Minimax*. La section 3 illustre les bénéfices de notre approche sur différentes bases de données d'images. Enfin, la dernière partie 4 conclue le papier.

2 Couplage d'un CNN et classifieur Minimax

Cette section présente comment coupler un réseau de neurones convolutif avec le classifieur *Minimax* et décrit l'étape d'apprentissage du classifieur *Minimax*.

2.1 Coupler un CNN pré-entraîné avec un classifieur en couche de sortie

Soit $\mathcal{Y} = \{1, \dots, K\}$ l'ensemble des étiquettes de chaque classe où $K \geq 2$ dénote le nombre de classes. Soit $\Phi : \mathcal{X} \to \mathcal{Y}$ un CNN qui assigne une classe à chaque image $X \in \mathcal{X}$. L'architecture d'un CNN Φ composé de s couches cachées $h_1, \dots, h_s$ peut être modélisé comme étant [18]

$$\Phi(X) = h_{s+1} \circ h_s \circ \cdots \circ h_1(X) = h_{s+1} \circ \varphi(X), \quad (1)$$

avec $h_{s+1}(\cdot)$ la couche de sortie, $\varphi(X)$ la sortie de la dernière couche cachée, et où $f \circ g(X) = f(g(X))$ dénote la composition de entre deux fonctions f et g. Dans la suite de ce papier, $Z = \varphi(X) \in \mathbb{R}^d$ correspondra aux variables explicatives *deep* et h_{s+1} dénotera le classifieur de la couche de sortie. Généralement la règle de décision h_{s+1} est un classifieur *Softmax*, dénotée ici par δ^{soft}, fondée sur une couche linéaire [18].

Ce papier a pour but de remplacer le classifieur de la couche de sortie par un classifieur *Minimax*. Pour ce faire nous étudions les réseaux de neurones profonds pouvant être modélisés comme étant

$$\Phi_\delta(X) = \delta \circ \varphi(X) = \delta(Z), \quad (2)$$

avec $\delta : \mathbb{R}^d \to \mathcal{Y}$ une règle de décision jouant le rôle de couche de sortie. En d'autres termes, $\Phi_\delta(X)$ est un CNN prenant des décisions par rapport aux variables explicatives *deep* Z. Dans cet article, nous ne souhaitons pas entraîner à nouveau les couches cachées du CNN mais seulement coupler les variables explicatives *deep* avec un classifieur spécifique (seulement ce classifieur sera entraîné). Ainsi, notre approche est une sorte de *fine tunning* puisque notre attention se concentre seulement sur le classifieur en sortie. Soit $\Delta := \{\delta : \mathbb{R}^d \to \mathcal{Y}\}$ l'ensemble de toutes les règles de décisions possibles à partir de l'ensemble des variables explicatives *deep* définies dans $\mathbb{R}^d$.

Soit $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$ l'échantillon d'apprentissage contenant m images d'entraînement étiquettées, où $\mathcal{I}$ est un ensemble d'indices fini. Soit $L : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ la fonction de perte telle que $L(k, l) := L_{kl}$ correspond au coût d'erreur de prédire la classe l quand la vraie classe est k. Ainsi, le risque empirique d'erreurs de classification du CNN Φ_δ est donné par

$$\hat{r}(\Phi_\delta) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \Phi_\delta(X_i)) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(Z_i)), \tag{3}$$

avec $Z_i = \varphi(X_i)$. Ainsi, tous les CNN de la forme $\Phi_\delta(X)$ peuvent être comparés en évaluant seulement le risque

$$\hat{r}_\varphi(\delta) = \hat{r}(\Phi_\delta) = \hat{r}(\delta \circ \varphi), \tag{4}$$

puisque $\varphi(\cdot)$ est commun à tous les CNN Φ_δ.

En d'autres termes, le risque empirique $\hat{r}(\Phi_\delta)$ d'un CNN Φ_δ est égal au risque empirique $\hat{r}(\Phi_\delta)$ de la règle de décision δ appliquée sur les variables explicatives *deep*. Il se doit de noter que la performance (*accuracy*) d'un CNN est égale à $1 - \hat{r}(\Phi_\delta)$ lorsque nous utilisons la fonction de perte classique $L_{0\text{-}1}$ définie par $L_{kl} = 0$ si $k = l$ et $L_{kl} = 1$ sinon.

Notons $\hat{\pi} := [\hat{\pi}_1, \ldots, \hat{\pi}_K]$ les proportions par classe de l'échantillon d'apprentissage telles que, pour tout $k \in \mathcal{Y}$, $\hat{\pi}_k$ est la proportions d'images observées de classe k. Comme expliqué dans [7, 8], le risque empirique $\hat{r}_\varphi(\delta)$ peut se réécrire comme étant

$$\hat{r}_\varphi(\delta) = \sum_{k \in \mathcal{Y}} \hat{\pi}_k \hat{R}_k(\delta), \tag{5}$$

$$\hat{R}_k(\delta) = \sum_{l \in \hat{\mathcal{Y}}} L_{kl} \hat{\mathbb{P}}(\delta \circ \varphi(X_i) = l \mid Y_i = k), \tag{6}$$

avec $\hat{R}_k(\delta)$ les risques conditionnels empiriques de δ associés à la classe k et $\hat{\mathbb{P}}(\cdot \mid \cdot)$ la probabilité conditionnelle estimée sur l'échantillon d'apprentissage.

Habituellement dans un CNN, la règle de décision *Softmax* a pour but d'approximer le classifieur de Bayes, dénoté par $\delta^B_{\hat{\pi}}$, minimisant $\hat{r}_\varphi(\delta)$. Ainsi, si nous remplaçons la couche *Softmax* par $\delta^B_{\hat{\pi}}$, la performance du CNN doit rester la même. De manière plus générale, soit δ^B_π le classifieur de Bayes associé aux probabilité à priori π, où π appartient au simplexe $\mathbb{S}$ de dimensions K. Nous pouvons alors définir le risque minimum de Bayes $V(\pi) := \hat{r}_\varphi(\delta^B_\pi)$ comme étant une fonction des probabilités à priori sur le simplexe $\mathbb{S}$.

2.2 Apprentissage Minimax

Étant donné une base de test $\mathcal{S}' = \{(Y'_i, X'_i), i \in \mathcal{I}'\}$, où $\mathcal{I}'$ est un ensemble d'indices fini, contenant m' images test satisfaisant des proportions par classe inconnues $\pi = [\pi_1, \ldots, \pi_K]$, le CNN Φ_δ entraîné sur l'échantillon d'apprentissage $\mathcal{S}$ est par la suite utilisé afin de prédire les classes Y'_i des images de l'échantillon test. Comme décrit dans [7, 8], le risque empirique d'erreur global du CNN Φ_δ dépend des proportions par classe π de la base de test, et est défini par

$$\hat{r}(\pi, \Phi_\delta) = \hat{r}_\varphi(\pi, \delta) = \sum_{k \in \mathcal{Y}} \pi_k \hat{R}_k(\delta). \tag{7}$$

Ainsi, ce risque d'erreur global évolue linéairement quand un changement de probabilité a priori se produit. Le risque maximum pouvant être atteint par le CNN Φ_δ est alors $M(\delta) := \max_{k \in \mathcal{Y}} \hat{R}_k(\delta)$.

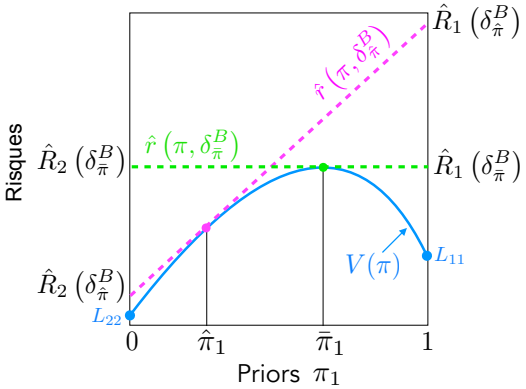

FIGURE 1 – Évolution du risque pour $K = 2$ classes. Dans ce cas, le risque (7) peut se réécrire comme étant $\hat{r}_\varphi(\pi, \delta) = \pi_1[\hat{R}_1(\delta) - \hat{R}_2(\delta)] + \hat{R}_2(\delta)$.

Afin de disposer d'un classifieur en couche de sortie du CNN qui soit robuste malgré des proportions par classe déséquilibrées ou incertaines, une approche pertinente est alors d'utiliser le classifieur *Minimax* minimisant $M(\delta)$. Comme démontré dans [7, 8], ce problème d'optimisation est équivalent à calculer les probabilités à priori $\bar{\pi} \in \mathbb{S}$ maximisant $V(\pi)$ tel que le classifieur *Minimax*, noté $\delta^B_{\bar{\pi}}$, soit le classifieur de Bayes associé aux probabilités à priori $\bar{\pi}$.

Cette approche est illustrée dans la figure 1 pour $K = 2$ classes. La valeur maximale de V est atteinte par le classifieur égalisateur $\delta^B_{\bar{\pi}}$ tel que $\hat{R}_k\left(\delta^B_{\bar{\pi}}\right) = \max_\pi V(\pi)$ pour chaque classe $k \in \mathcal{Y}$. Ce classifieur égalisateur est donc robuste à n'importe quel changement de probabilité *a priori*.

Afin d'apprendre le classifieur *Minimax*, nous utilisons l'approche établie dans [8]. Cette méthode nécessite de discrétiser les variables explicatives *deep* Z (à l'aide de l'algorithme des *k-means*) afin d'obtenir une approximation

précise de $V(\pi)$. Comme démontré dans [8], V est une fonction concave et affine par morceaux sur le simplexe $\mathbb{S}$. Ainsi, la maximisation non-différentiable de V est réalisée par un algorithme de sous-gradient projeté [19] suivant le schéma

$$\pi^{(n+1)} = \mathrm{P}_{\mathbb{S}}\left(\pi^{(n)} + \frac{\gamma_n}{\eta_n}\, g^{(n)}\right), \tag{8}$$

avec, à chaque itération $n \geq 1$, $g^{(n)}$ le sous-gradient de V au point $\pi^{(n)}$, γ_n le pas du sous-gradient, $\eta_n = \max\{1, \|g^{(n)}\|_2\}$, et $\mathrm{P}_{\mathbb{S}}$ la projection exacte sur le simplexe probabiliste $\mathbb{S}$ [20]. Il a été démontré que cet algorithme converge fortement vers $\bar{\pi} = \mathrm{argmax}_{\pi \in \mathbb{S}} V(\pi)$.
La figure 2 résume notre approche d'ajustement d'un CNN pré-entraîné en considérant le classifieur *Minimax*. Les variables explicatives *deep* issues du CNN forment l'échantillon d'apprentissage. Cet échantillon est par la suite discrétisé par l'algorithme des *k-means*. Enfin le classifieur *Minimax* est construit avec l'algorithme de sous-gradient projeté (8).

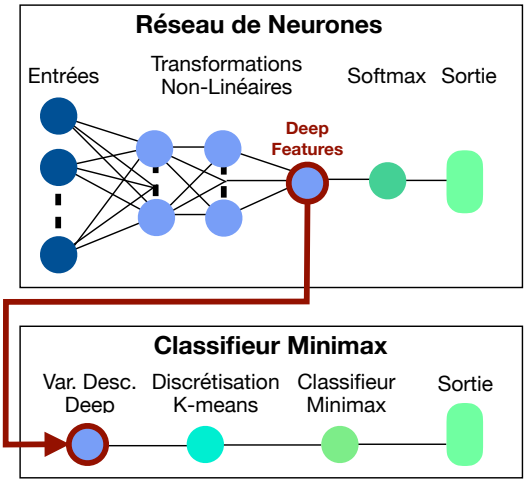

FIGURE 2 – Schéma de la méthode couplée.

3 Expériences

Cette partie illustre l'intérêt de notre approche sur trois bases de données médicales [21]. Une expérience est aussi menée sur la base de données CIFAR100 [22] contenant un très grand nombre de classes dans le but de diversifier les cas d'application et de mettre en avant les performances de notre méthode avec un grand nombres de classe.

Bases de données médicales : la base de données *DermaMNIST* [23] est une collection d'images dermatoscopiques de lésions cutanées pigmentées courantes contenant 7 catégories. *BreastMNIST* [24] quant à elle, est un recueil d'échographies du sein comprenant 2 classes tandis que *OCTMNIST* [25] regroupe en 4 catégories des images 3D du foie. Les trois bases de données médicales que nous avons considérées [21] diffèrent selon le nombre d'échantillons, de classes, mais aussi en termes de proportions par classes (voir Tableau 1). Ces bases de données font partie

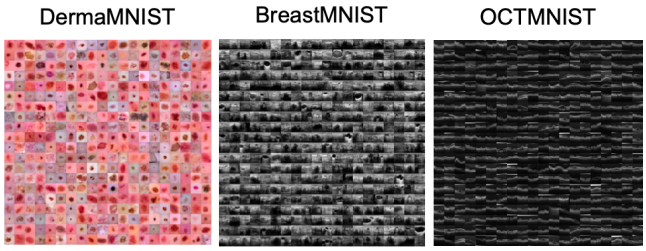

FIGURE 3 – Aperçu des bases de données médicales.

de *MedMNIST*, une collection de 10 bases de données médicales réelles en accès libre. *MedMNIST* est standardisée afin de mener différentes tâches de classification sur des images de petites tailles (28×28). Chaque base de données contient un échantillon d'apprentissage, un échantillon de validation et enfin un échantillon de test. Tout comme dans [26], nous avons entraîné le CNN sur l'échantillon d'apprentissage avec 100 itérations en utilisant la fonction de perte *cross-entropy* et l'optimiseur SGD. Nous avons utilisé les algorithmes disponibles sur [27] où les auteurs ont pris le soin de limiter le sur-apprentissage en choisissant un modèle optimal sur l'échantillon de validation. Enfin, les performances de généralisation sont évaluées sur l'échantillon de test. Une fois le modèle *ResNet* entraîné les variables explicatives *deep* du meilleur modèle sont extraites. Pour calibrer notre règle de décision minimax sur la couche de sortie du CNN, nous utilisons ensuite l'échantillon de validation afin de limiter le sur-apprentissage pouvant provenir des variables explicatives *deep* associées à l'échantillon d'apprentissage. Enfin, l'échantillon de test nous permet d'évaluer la performance du CNN ajusté. Le tableau 2 offre une comparaison des résultats sur les échantillons de validation et de test associés au CNN *ResNet* initial considérant la règle de décision *softmax*, au CNN considérant le Classifieur de Bayes Discret (DBC) $\delta^B_{\bar{\pi}}$ calculé comme dans [8], et au CNN ajusté avec notre classifieur *Minimax* discret (DMC) $\delta^B_{\bar{\pi}}$. Le DBC et le DMC ont été construit sur les mêmes variables explicatives extraites et discrétisées. Il est attendu que le DBC ait un risque global similaire à celui du CNN ayant comme couche de sortie la règle de décision *Softmax*.
Nous pouvons observer dans le tableau 2 que le DMC obtient le plus fable maximum des risques conditionnels. De plus, la différence $\psi(\delta)$ entre le maximum et le minimum des risques conditionnels étant définie par

$$\psi(\delta) := \max_{k \in \mathcal{Y}} \hat{R}_k(\delta) - \min_{k \in \mathcal{Y}} \hat{R}_k(\delta), \tag{9}$$

est la plus faible pour le DMC. Ainsi, comme nous l'avons souligné précédemment, le DMC tente de prédire toutes les classes, même les moins représentées. Cependant, pour la plupart des bases de données, le risque d'erreur global sur l'échantillon de test est le plus élevé pour le DMC que pour les deux autres méthodes. En effet, un compromis doit être fait afin d'égaliser les risques conditionnels. Le risque global est minimum lorsque les risques conditionnels sont

Bases de données	# Classes	# Apprentissage	# Validation	# Test	π^{train}	π^{val}	π^{test}
DermaMNIST	7	7,007	1,003	2,005	Min = 0.01 Max = 0. 67	Min = 0.01 Max = 0.67	Min = 0.01 Max = 0.67
BreastMNIST	2	4,709	524	624	Min = 0.27 Max = 0.73	Min = 0.27 Max = 0.73	Min = 0.27 Max = 0.73
OCTMNIST	4	97,477	10,832	1,000	Min = 0.08 Max = 0.47	Min = 0.08 Max = 0.47	Min = 0.25 Max = 0.25

TABLE 1 – Panorama des bases de données médicales (Min, resp. Max, dénote le minimum, resp. le maximum, des proportions par classe).

Bases de données	Échantillons	ResNet-18 CNN			ResNet-18-DBC			ResNet-18-DMC		
		$\hat{r}$	$\max_{k\in\mathcal{Y}} \hat{R}_k$	ψ	$\hat{r}$	$\max_{k\in\mathcal{Y}} \hat{R}_k$	ψ	$\hat{r}$	$\max_{k\in\mathcal{Y}} \hat{R}_k$	ψ
DermaMNIST	**Val**	0.29	1	0.83	0.26	1	0.9	0.48	0.54	0.21
	Test	0.3	1	0.84	0.32	1	0.87	0.54	0.83	0.37
BreastMNIST	**Val**	0.17	0.43	0.36	0.14	0.43	0.39	0.17	0.19	0.03
	Test	0.16	0.5	0.46	0.18	0.57	0.54	0.19	0.19	0
OCTMNIST	**Val**	0.06	0.35	0.33	0.07	0.47	0.46	0.13	0.13	0.01
	Test	0.28	0.76	0.69	0.20	0.41	0.33	0.21	0.32	0.24

TABLE 2 – Résultats sur les échantillons de validation et de test du ResNet-18 CNN, du DBC et du DMC appliqués sur les variables explicatives extraites du ResNet-18. Pour chaque classifieur, $\hat{r}$ fait référence au risque global (3), $\max_{k\in\mathcal{Y}} \hat{R}_k$ correspond au maximum des risques conditionnels (6), et ψ est la différence entre le maximum et le minimum des risques conditionnels par classe (9).

fortement déséquilibrés et que les classes les plus représentées sont très bien prédites. Les deux méthodes comparées au DMC ne fournissent pas de prédictions précises pour les classes contenant le plus petit nombre d'images, bien que ce soient les classes d'intérêt car correspondant aux pathologies. Il est à noter ici qu'un changement de proportions par classe survient dans la bases de données OCTMNIST entre les échantillons d'apprentissage et de validation (voir Tableau 1). Par construction, le DMC apparaît significativement moins sensible à ce changement de proportions par classe que les autres méthodes.

Afin de souligner davantage ces conclusions, concentrons-nous sur l'échantillon de validation de DermaMNIST. Nous observons sur la figure 4 que malgré les proportions par classe hautement déséquilibrées (Tableau 1), le DMC permet une meilleure égalisation des risques d'erreurs par classe que les deux autres méthodes. De plus, concernant le DBC et le CNN initial, nous pouvons observer que les risques conditionnels associés aux classes les moins représentées sont situés bien au-dessus du risque moyen. Puisque la classe la plus représentée est bien prédite, le risque moyen est faible bien que les pus petites classes soient fortement mal classifiées.

Base de données CIFAR100 : Nous avons dans un dernier temps considéré la base de données CIFAR100 contenant 60 000 images avec $K = 100$ classes. Pour cette expérience nous considérons un échantillon d'apprentissage, respectivement de test, composé de 40 000 images, respectivement 20 000 images. Cette fois ci, les deux échantillons satisfont des proportions par classe parfaitement équilibrées $\hat{\pi} = [1/100, \ldots, 1/100]$. Pour cette expérience, nous considérons des variables explicatives extraites de la dernière couche cachée d'un CNN *EfficientNet-B0* [28], et nous comparons deux règles de décision en dernière couche du réseau de neurone, le DMC et la Régression Logistique Repondérée (WLR), appliquées toutes deux sur les variables explicatives *deep*. La WLR, connue pour sa capacité à faire face au problème de proportions pas classe déséquilibrées, est construite en considérant des poids par classe inversement proportionnels aux fréquences de classe.

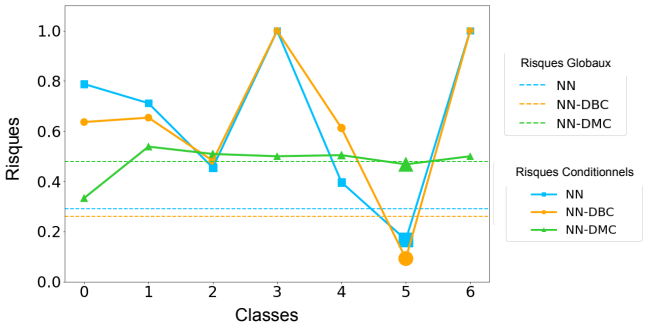

FIGURE 4 – Risques associés à l'échantillon de validation de DermaMNIST. La taille de chaque point dépend des proportions par classe.

Ici, puisque les proportions par classes sont parfaitement équilibrées, il en résulte que la WLR équivaut à considérer le classifieur initial *Softmax* composant la dernière couche du CNN. Comme illustré à la figure 5, ce classifieur est

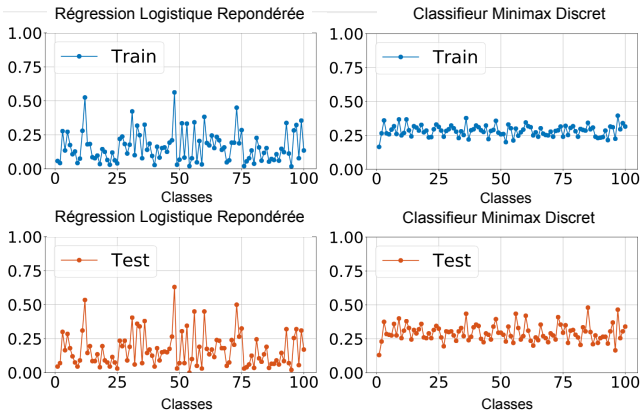

FIGURE 5 – Risques Conditionnels sur la base de données CIFAR-100.

dans l'incapacité d'égaliser les risques conditionnels. De plus, dans cet exemple les classes sont trop nombreuses pour être manuellement en mesure de calculer les poids optimaux par classe. Malgré ces difficultés, nous pouvons observer que notre approche minimax est très satisfaisant pour chercher à égaliser les risques d'erreur par classe sur cette grande base de données.

4 Conclusion

Dans cet article nous avons présenté une nouvelle approche permettant d'ajuster des réseaux de neurones convolutifs pré-entraînes pour traiter des problèmes de proportions par classe déséquilibrées ou pouvant évoluer au cours du temps. Pour ce faire, notre approche couple un réseau de neurones convolutif pré-entraîne avec une règle de décision minimax en couche de sortie. Des résultats sur plusieurs bases de données réelles ont illustré l'intérêt de notre approche. Nos prochaines recherches vont s'orienter sur l'erreur de généralisation de notre approche.

Références

[1] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, 2018.

[2] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making : The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, pp. 427–436, 2008.

[3] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1263–1284, 2009.

[4] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, 2012.

[6] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, 2008.

[7] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Springer-Verlag New York, 1994.

[8] C. Gilet, S. Barbosa, and L. Fillatre, "Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[9] P. González, A. Castaño, C. Nitesh, and J. J. Del Coz, "A review on quantification learning," *ACM Computing Surveys*, 2017.

[10] M. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," *European Conference on Artificial Intelligence*, 1998.

[11] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *EEE Transactions on Knowledge and Data Engineering*, 2006.

[12] S. Lawrence, I. Burns, A. Back, A. C. Tsoi, and C. L. Giles, *Neural Network Classification and Prior Class Probabilities*. Springer Berlin Heidelberg, 1998.

[13] H.-j. Lee and S. Cho, "The novelty detection approach for different degrees of class imbalance," in *Neural Information Processing*, I. King, J. Wang, L.-W. Chan, and D. Wang, Eds. Springer Berlin Heidelberg, 2006.

[14] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost : Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases : PKDD 2003*, 2003.

[15] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 1567–1578.

[16] J. Tian, Y.-C. Liu, N. Glaser, Y.-C. Hsu, and Z. Kira, "Posterior re-calibration for imbalanced datasets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[17] A. Guerrero-Curieses, R. Alaíz-Rodríguez, and J. Cid-Sueiro, "A fixed-point algorithm to minimax learning with neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, pp. 383–392, 2004.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http ://www.deeplearningbook.org.

[19] Y. I. Alber, A. N. Iusem, and M. V. Solodov, "On the projected subgradient method for nonsmooth convex optimization in a hilbert space," *Mathematical Programming*, vol. 81, pp. 23–35, 1998.

[20] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, vol. 158, no. 1, pp. 575–585, 2016.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "*MedMNIST* databases," https ://zenodo.org/record/4269852.X$_m dsul K i H E$.

[22] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009, https ://www.cs.toronto.edu/ kriz/cifar.html.

[23] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[24] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[25] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[26] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon : A lightweight automl benchmark for medical image analysis," *arXiv :2010.14925*, 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Github *MedMNIST*," 2020, https ://medmnist.github.io/.

[28] M. Tan and Q. V. Le, "Efficientnet : Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, 2019.