Inférence de probabilités prédictives à l'aide des forêts aléatoires dans le contexte de classification déséquilibrée

Clément Dombry [2], Moria Mayala [*1], Charles Tillier [3] & Olivier Wintenberger [1]

[1] *Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université*
[2] *Laboratoire de Mathématiques de Besançon – Université de Franche-Comté*
[3] *Laboratoire de Mathématiques – Université de Versailles Saint-Quentin-en-Yvelines*

Résumé. Les données déséquilibrées dans les taches de classification sont aujourd'hui identifiées comme un problème majeure en apprentissage automatique. Une de ces raisons est que les algorithmes traditionnels de machine learning peuvent être mis en difficulté dans ce cadre, notamment pour détecter la classe minoritaire. Dans ce travail, nous faisons de l'inférence des probabilités prédictives reposant sur les modèles simplifiés en particulier des forêts purement aléatoires infinies (IPRF) en vue de relever les défis associés à la prédiction d'événements rares. Nous établissons notamment un théorème central limite pour cet estimateur IPRF sous certaines hypothèses de régularité sur la fonction de régression. Cependant, IPRF hérite un biais asymptotique inhérent à l'asymetrie de la distribution de classes. Nous proposons une procédure de type échantillonnage préférentielle dérivant des odd-ratio afin de réduire le biais asymptotique de IPRF. Une courte étude de simulation illustre les performances de la méthode proposée.

Mots-clés. Classification binaire, données déséquilibrées, forêts purement aléatoires infinies.

Abstract. Imbalanced data in classification tasks has been identified as an top problem in machine learning. One of the reasons for this is that traditional machine learning algorithms can get into trouble in this context, particularly when it comes to detecting the minority class. In this work, we perform predictive probability inference based on simplified models (algorithms), in particular Infinite purely random forests (IPRFs), to address the challenges associated with predicting rare events. In particular, we establish a central limit theorem for this IPRF estimator under certain regularity assumptions on the regression function. However, IPRF inherits an asymptotic bias due to the asymmetry of the class distribution. We propose a preferential sampling procedure derived from the odd-ratio in order to reduce the asymptotic bias of IPRF. A short simulation study illustrates the performance of the proposed method

Keywords. binary classification, class imbalance, infinite random forests.

1 Contexte

Il est connu que la prédiction sur les données déséquilibrées répresente une difficulté majeure en apprentissage automatique, en particulier dans les modèles de classification, qu'il s'agisse de classification binaire (ex :détecter une maladie, détection d'images, intrusion réseau, détection de fraude) ou de classification multi-classes (ex : prédire le modèle de voiture acheté), pour plus de détailles une bonne revue a été proposée par Jason Brownlee (2020). Ainsi, proposer une procedure éfficace d'estimation dans ce cadre est un enjeu d'intérêt. Nous présentons ici le cas binaire, qui est plus simple à appréhender et peut ensuite facilement se généraliser au multi-classes.

Le problème de base est le suivant. On considère $(X_i, Y_i)_{1\leq i\leq n}$ des observations i.i.d de même loi que le couple $(X, Y) \in \mathcal{X}\times\{0, 1\}$, $\mathcal{X} \subset \mathbb{R}^d$. La loi jointe de (X, Y) est complètement déterminer par la loi marginale de X et la fonction de régression μ est définie par

$$\mu(\mathbf{x}) := \mathbb{P}(Y = 1|X = \mathbf{x}), \qquad \mathbf{x} \in \mathcal{X}. \tag{1}$$

Notre objectif est d'estimer μ lorsque les classes $\{0, 1\}$ ne sont pas représentées de manière égale. Pour quantifier le degré de *déséquilibre*, nous utilisons la notion de ratio de déséquilibre (Imbalance ratio) définie comme suit

$$IR = n_0/n_1$$

où n_0 et n_1 désignent respectivement le cardinal des échantillons majoritaire et minoritaire, respectivement. Il ya un déséquilibre dans les données lorsque $IR > 1$. Dès lors que la valeur $IR \approx 1$ correspondant à des données parfaitement équilibrées (cf. Robert O'Brien et al (2019)). Sans perte de généralité, nous supposons que la classe minoritaire correspond à l'etiquette 1.

1.1 Cadre statistique

On considère deux (2) modèles: le premier modèle est définit sous les mêmes hypothèses que celle de eq. (1). Il s'agit du modèle original d'intérêt, pour lequel nous aimerions prédire avec précision la classe d'appartenance des observations qui sont supposés être déséquilibrées. Ainsi, si on note

$$p := \mathbb{P}(Y = 1) = \int_{\mathcal{X}} \mu(\mathbf{x}) d\mathbb{P}_X(\mathbf{x}).$$

la probabilité d'observer la classe 1, on suppose donc que $p < q := \mathbb{P}(Y = 0)$.

Afin de gérer ce déséquilibre, on considère que l'on a également accès à des données provenant d'un deuxième modèle. Dans ce dernier, on suppose que les observations ne sont pas déséquibrées c'est à dire

$$p^* := \mathbb{P}(Y^* = 1) = \int_{\mathcal{X}} \mu^*(\mathbf{x}) d\mathbb{P}_{X^*}(\mathbf{x}),$$

avec $\mu^*(\mathbf{x}) = \mathbb{P}(Y^* = 1|X^* = \mathbf{x})$, pour $\mathbf{x} \in \mathcal{X}$, la fonction de régression et on a $p^* = q^*$. Nous supposons en outre la condition suivante :

- La probabilité de classe $p^* \in (0,1)$ satisfait $p^* > p$. Un choix pratique courant est $p^* \approx 0,5$.

A partir d'observations provenant de ces deux modèle, notre stratégie consiste à sous-échantillonner la classe majoritaire dans le premier modèle dans le but d'équlibrer les données.

.

1.2 Méthode: IPRF avec échantillonnage préférentiel

Introduit par Leo Breiman (2001) les forêts aléatoires, constituent un algorithme d'apprentissage très populaire qui offre d' excellente performances, et une grande flexibilité dans sa capacité à gérer tous les types de données. Ainsi dans le contexte des données déséquilibrées (imbalanced data), Leo Breiman et al (2004) proposent une nouvelle variante de forêts aléatoires, simplifiées appelée forêts aléatoires équilibrées (balanced random forests, PRF) qui traite à tous points de vue le probleme d'imbalanced data en sous-échantillonnant la classe majoritaire afin d'améliorer les performances de classification par rapport à la classe minoritaire. Sans perte de généralité, les forêts purement aléatoires infinies (IPRF) interviennent dans le même contexte en vue de relever les défis associés à la prédiction d'événements rares. Notre approche s'appuit sur les travaux antérieurs de Wager et Athey (2018) et de Peng et al. (2022) qui ont établi la normalité asymptotique de variantes infinies similaires de différents estimateurs à l'aide des outils de U-statistique.

Notre objectif est de construire un classifier $\widehat{\mu}_{IS}(\mathbf{x})$ de $\mu(\mathbf{x})$. Etant donné, p et p^* les proportions de la classe 1 dans les modèles original et biaisé(second modèle) . A partir des ratio de chances on établit une connexion entre $\mu^*(\mathbf{x})$ et $\mu(\mathbf{x})$ donnée par

$$\frac{\mu(\mathbf{x})/\left(1-\mu(\mathbf{x})\right)}{p/(1-p)} = \frac{\mu^*(\mathbf{x})/\left(1-\mu^*(\mathbf{x})\right)}{p^*/(1-p^*)}\,, \qquad \mathbf{x} \in \mathcal{X}\,. \tag{2}$$

Par ailleurs, on estime les proportions de la classe 1 dans les deux modèles.

Hypothès 1 *Soient les tailles des sous-échantillons dans la classe 0 et la classe 1 s_0 et s_1 dans le modèle biaisé et n_0 et n_1 les tailles d'échantillons de la classe 0 et la classe 1 dans le modèle original*

$$\widehat{p^*} := \frac{s_1}{s_1+s_0} \longrightarrow p^* > 0\,, \qquad n \to \infty\,.$$

Remarquons que une valeur de $p^* = 0.5$ peut correspondre au cas des données équilibrées. De manière analogue la loi des grands nombre fournie l'estimation suivante de p

$$\widehat{p} := \frac{n_1}{n_0+n_1} \to p\,, \qquad n \to \infty\,.$$

Par échantillonnage préferentiel, nous obtenons cet estimateur

$$\widehat{\mu}_{IS}(\mathbf{x}) := \frac{n_1 s_0 \widehat{\mu^*}(\mathbf{x})}{n_0 s_1\left(1-\widehat{\mu^*}(\mathbf{x})\right) + n_1 s_0 \widehat{\mu^*}(\mathbf{x})}\,, \qquad \mathbf{x} \in \mathcal{X}\,, \tag{3}$$

Théorème 1 *Soit* $\widehat{\mu}_{IS}(\mathbf{x})$ *un estimateur par échantillonnage d'importance comme défini dans* (3). *Sous l' Hypothèse 1 , nous obtenons le théorème limite central suivant*

$$\sqrt{np_n}\left(\frac{\widehat{\mu}_{IS}(\mathbf{x})}{\mu(\mathbf{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{1-\mu(\mathbf{x})}{\mu(\mathbf{x})}\right),\ \textit{quand } n \to \infty. \tag{4}$$

Où $p_n = \mathbb{P}(X \in L_b(\mathbf{x})) \leq \mathbb{E}\Big[\mathrm{Diam}(L_b(\mathbf{x}))^d\Big]$, *si* $X \sim \mathcal{U}([0,1]^d)/\mathcal{X} = [0,1]^d$.

1.3 Etude numérique

Dans cette section, nous donnons à l'aide de données synthétiques, une illustration numérique de nos résultats théoriques. L'objectif est de montrer que notre procédure d'inférence peut être mise en œuvre et de confirmer que les propriétés asymptotiques dérivées sont empiriquement pertinentes. Considérons les observations $(X_i, Y_i)_{1\leq i\leq n}$ i.i.d, où les covariables et la variable cible sont définies comme suit

- $X_i \sim \mathcal{U}([0,1]^2)$, et $\ Y_i \in \{0,1\}$
- $n = \{100 \ldots 1000\}$ et 30 répétitions Monte-Carlo
- Utilisation du package `RandomUniformForest` (Saip Ciss,(2014)).

Nous évaluons la performance de $\widehat{\mu}_{IS} := g(\text{PRF-balanced})$ en terme de l'erreur quadratique moyenne (MSE). Cette métrique emblématique peut s'écrire comme une somme de la variance du modèle, du biais du modèle et de l'incertitude irréductible (compromis biais-variance). Par soucis de clarté, notons PRF $:= \widehat{\mu}^*$.

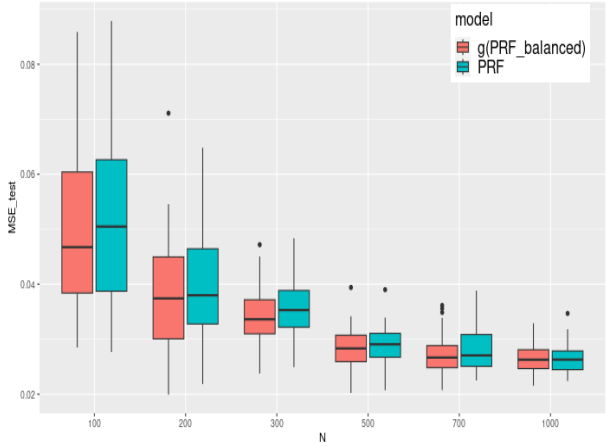

Figure 1: MSE de l'estimateur par échantillonnage d'importance $\widehat{\mu}_{IS}$ et de l'estimateur de forêts aléatoires équilibrées $\widehat{\mu}^*$ pour , $d = 2$, toutes les forêts ont $B = 500$ arbres.

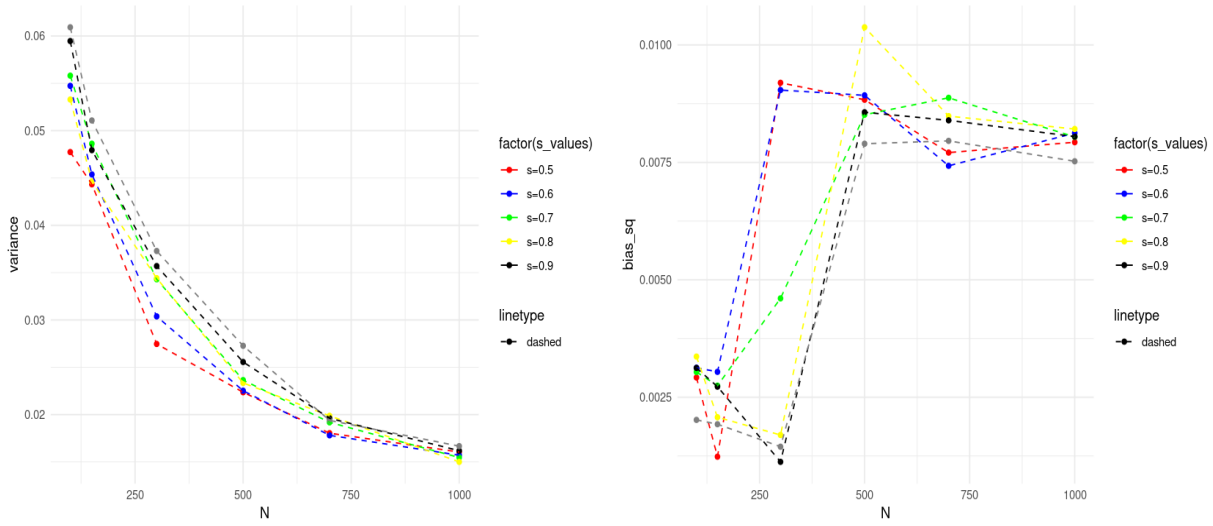

Figure 2: Variance (à gauche), Biais (à droite) de l'estimateur par échantillonnage d'importance $\widehat{\mu}_{IS}$ pour les différentes valeurs de $s \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\} * n$, $d = 2$, toutes les forêts ont $B = 500$ arbres.

Remerciements

Nous remercions Saip Ciss pour l'amélioration substantielle des hyperparamètres dans le package `randomUniformForest`. Nous remercions également Francesco Bonacina pour toute l'aide qu'il nous a apportée dans la réalisation de ces simulations. En outre, nous aimerions remercier plus particulièrement les membres du projet ANR T-REX pour le financement de nos multiples ateliers à Vienne et ailleurs au cours desquels des discussions fructueuses ont permis d'améliorer grandement ce document. Nous remercions également le "International Emerging Actions 2022" et le "Centre National de la Recherche Scientifique (CNRS)" pour le support de voyage de Moria Mayala et Charles Tillier à Vienne.

Bibliographie

Leo Breiman.(2001), Random forests.,*Machine learning*, 45(1):5–32.

Chao Chen, Andy Liaw, Leo Breiman, et al.(2004), Using random forest to learn imbalanced data.,*University of California, Berkeley*, 10(1-12):24.

Jason Brownlee. (2020), *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning.*, Machine Learning Mastery.

Robert O'Brien and Hemant Ishwaran.(2019), A random forests quantile classifier for class imbalanced data.,*Pattern recognition*, 90:232–249.

Saïp Ciss. (2014), *Forêts uniformément aléatoires et détection des irrégularités aux cotisations sociales.*, PhD thesis, Paris 10.

Stefan Wager and Susan Athey. (2018), Estimation and inference of heterogeneous treatment effects using random forests., *Journal of the American Statistical Association* , 113 (523):1228–1242.

Sylvain Arlot and Robin Genuer.(2014), Analysis of purely random forests bias.,*arXiv preprint* , arXiv:1407.3939.

Wei Peng, Tim Coleman, and Lucas Mentch. (2022), ates of convergence for random forests via generalized u-statistics, *Electronic Journal of Statistics*, 16(1):232–292.