Étude théorique et expérimentale de SMOTE : limites et comparaisons des stratégies de rééquilibrage

Abdoulaye SAKHO [1,2], Erwan SCORNET [2] & Emmanuel MALHERBE [3]

[1] *Artefact Research Center, Paris, France. abdoulaye.sakho@artefact.com*
[2] *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France. erwan.scornet@sorbonne-universite.fr*
[3] *Artefact Research Center, Paris, France. emmanuel.malherbe@artefact.com*

Résumé. *Synthetic Minority Oversampling Technique* (SMOTE) est une stratégie de rééquilibrage courante pour traiter les ensembles de données déséquilibrés. Asymptotiquement, nous prouvons que SMOTE (avec la valeur par défaut de son hyperparamètre) régénère la distribution originale en copiant simplement les échantillons minoritaires initialement présents. Nous introduisons ensuite deux nouvelles stratégies liées à SMOTE et les comparons aux procédures de rééquilibrage les plus récentes. Nous montrons que les stratégies de rééquilibrage ne sont nécessaires que lorsque l'ensemble de données est fortement déséquilibré. Pour de tels ensembles de données, SMOTE, nos propositions ou les procédures de sous-échantillonnage sont les meilleures stratégies.

Mots-clés. Classification, Données déséquilibrées, SMOTE

Abstract. *Synthetic Minority Oversampling Technique* (SMOTE) is a common rebalancing strategy for handling imbalanced data sets. Asymptotically, we prove that SMOTE (with default parameter) regenerates the original distribution by simply copying the original minority samples. Then we introduce two new SMOTE-related strategies, and compare them with state-of-the-art rebalancing procedures. We show that rebalancing strategies are only required when the data set is highly imbalanced. For such data sets, SMOTE, our proposals, or undersampling procedures are the best strategies.

Keywords. Classification, Imbalanced data sets, SMOTE

1 Contexte

Les ensembles de données déséquilibrés sont un problème typique rencontré dans plusieurs applications (He and Garcia, 2009), telles que la détection de fraude (Hassan and Abraham, 2016), la prédiction de diagnostics médicaux (Khalilia et al., 2011) et même la prédiction de l'appétence au désabonnement (Nguyen and Duong, 2021). En présence d'ensembles de données déséquilibrés, la plupart des algorithmes d'apprentissage statistiques ont tendance à prédire la classe majoritaire. Plusieurs stratégies ont été développées pour traiter ce problème, comme expliqué par Krawczyk (2016) et Ramyachitra and Manikandan (2014).

SMOTE est l'algorithme central dont dérivent la plupart des algorithmes qui génèrent de nouvelles données synthétiques au sein de la classe minoritaire. En effet, à l'exception des algorithmes basés sur les réseaux de neuronnes, la plupart des autres stratégies utilisent toujours une variante de l'interpolation linéaire incluse dans la procédure SMOTE. Plusieurs variantes tentent de se concentrer sur la génération d'échantillons synthétiques proches de la limite du support de la classe minoritaire. La plus courante est *ADASYN* (He et al., 2008) dont l'idée principale est de produire davantage d'échantillons synthétiques par interpolation linéaire entre les échantillons de la classe minoritaire qui sont principalement entourés d'échantillons de la classe majoritaire. BorderLine SMOTE (Han et al., 2005) cherche à générer de nouveaux échantillons synthétiques à la frontière des deux classes. Une autre variante de SMOTE axée sur les frontières est SVM-SMOTE (Nguyen et al., 2011), dont l'idée est de commencer par appliquer un classificateur de machine à vecteur de support aux données déséquilibrées. L'interpolation linéaire est ensuite effectuée sur le vecteur de soutien de la classe minoritaire.

Il existe plusieurs travaux théoriques concernant les stratégies de rééquilibrage. La méthode de pondération par classe est étudiée théoriquement par King and Zeng (2001). King and Zeng (2001) étudient l'effet de la stratégie Random Under Sampling sur un classificateur de régression logistique. À notre connaissance, il n'existe que peu de travaux théoriques disséquant la machinerie intrinsèque de l'algorithme SMOTE. Par exemple, Elreedy and Atiya (2019) calculent l'espérance et la matrice de covariance des points générées par SMOTE. De plus, Elreedy et al. (2023) établissent une expression de la densité des échantillons générés par SMOTE à partir de la densité des échantillons originaux de la classe minoritaire.

Notations On note par $\mathcal{U}([a,b])$ la distribution uniforme sur le segment $[a,b]$. La distribution gaussienne multivariée centré en μ et de matrice de covariance Σ est notée $\mathcal{N}(\mu,\Sigma)$. Pour tout ensemble A, on note par $Vol(A)$, la mesure de Lebesgue de A. Pour tout $z \in \mathbb{R}^d$ et$r > 0$, $B(z,r)$ est la boule centrée en z et de rayon r. On note par $c_d = Vol(B(0,1))$ le volume de la boule unitaire dans $\mathbb{R}^d$. Pour tout $p,q \in \mathbb{N}$, et tout $z \in [0,1]$, on note $\mathcal{B}(p,q;z) = \int_{t=0}^{z} t^{p-1}(1-t)^{q-1}\mathrm{d}t$ la fonction beta incomplète.

2 Étude de SMOTE

Dans cette section, nous étudions l'algorithme SMOTE, qui génère des données synthétiques par interpolation linéaire entre deux instances originales de la classe minoritaire. L'algorithme SMOTE possède un seul hyperparamètre, K, qui représente le nombre de plus proches voisins pris en compte lors de l'interpolation. Une seule itération SMOTE est détaillée dans Algorithm 1. Dans un pipeline classique d'apprentissage statistique, une itération de SMOTE est répétée autant de fois que nécessaire afin d'obtenir un ratio prédéfini entre les deux classes avant d'entraîner un classifieur.

Il a été démontré que SMOTE présentait de bonnes performances lorsqu'il était associé à des algorithmes de classification (voir par exemple Mohammed et al., 2020). Nous supposons

que $X_1, \ldots, X_n$ sont des échantillons indépendants et identiquement distribuées de la classe minoritaire (c'est-à-dire $Y_i = 1$ pour tous $i \in [n]$), avec une densité commune f_X à support borné, dénotée par $\mathcal{X}$. Enfin, on note par $f_{Z_{K,n}}$, la densité de SMOTE associé aux paramètres K, n appliquée sur les échantillons de $X_1, \ldots, X_n$ (pour tout K, n).

Algorithm 1 Une itération de SMOTE.

Input: Échantillons de la classe minoritaire $X_1, \ldots, X_n$, nombre K de plus proches voisins.
Choisir uniformément X_c (appellé **point central**) parmi $\{X_1, \ldots, X_n\}$.
On note $I = X_{(1)}(X_c,) \ldots, X_{(K)}(X_c)$, les K plus proches voisins de X_c (norme L_2).
Sélectionner $X_k \in I$ uniformément.
$w \leftarrow \mathcal{U}([0, 1])$
$Z \leftarrow X_c + w(X_k - X_c)$
Return Z

Nous nous plaçons dans le cadre de variables d'entrée continues, puisque les procédures synthétiques telles que SMOTE sont conçues à l'origine pour traiter de telles variables.

Théorème 2.1. *Supposons qu'il esixte $R > 0$, tel que $\mathcal{X} \subset B(0, R)$. De plus, supposons qu'il existe C_2 tel que, pour tout $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x \in \mathcal{X}}$. Alors, pour tout $n \geq K \geq 1$, pour tout $x_c \in \mathcal{X}$ et pour tout $\alpha > 0$, on a*

$$\mathbb{P}(|Z_{K,n} - X_c| \geq \alpha | X_c = x_c) \leq \varepsilon(n, \alpha, K, x_c), \tag{1}$$

où

$$\varepsilon(n, K, x_c, \alpha) = c_d R^d \eta(\alpha, R) \exp\left[n \left(3 \sqrt{\frac{e}{1 - \beta_{x_c,\alpha}}} \sqrt{\frac{K}{n}} + \ln\left(1 - \beta_{x_c,\alpha}\right) \right) \right], \tag{2}$$

avec $\beta_{x_c,\alpha} = \mu_X\left(B\left(x_c, \alpha\right)\right) > 0$ et

$$\eta(\alpha, R) = \begin{cases} C_2 \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1, \\ \frac{C_2}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{if } d > 1, \\ 0 & \text{if } \alpha > 2R. \end{cases}$$

Par conséquent, si $\lim_{n\to\infty} K/n = 0$, on a alors, pour tout $x_c \in \mathcal{X}$, $Z_{K,n}|X_c = x_c \to x_c$ en probabilité.

Nous montrons tout d'abord que la densité des points générés par la procédure SMOTE, avec la valeur par défaut pour K, converge en probalité vers la densité de la classe minoritaire, lorsque le nombre d'échantillons minoritaires augmente. Nous prouvons (Theorem 2.1) également que, sans réglage de l'hyperparamètre K (habituellement fixé à 5), SMOTE copie asymptotiquement les échantillons minoritaires originaux, manquant ainsi de la variabilité intrinsèque désirée dans toute procédure générative synthétique. Cela souligne l'importance du réglage de l'hyperparamètre dans SMOTE, lorsque le nombre d'échantillons de la classe minoritaire est suffisamment grand.

3 Nouvelles stratégies

Les limites de SMOTE mises en évidence dans la section 2 nous conduisent à deux nouvelles stratégies de rééquilibrage.

CV SMOTE Nous introduisons un nouvel algorithme, appelé CV SMOTE, qui trouve le meilleur hyperparamètre K parmi une grille prédéfinie via une procédure de validation croisée 5-fold. La grille est composée de l'ensemble $\{1, 2, \ldots, 15\}$ étendu avec les valeurs $\lfloor 0.01 n_{train} \rfloor$, $\lfloor 0.1 n_{train} \rfloor$ et $\lfloor \sqrt{n_{train}} \rfloor$, où n_{train} est le nombre d'échantillons minoritaires dans l'ensemble d'apprentissage.

Rappelons qu'à travers Theorem 2.1, nous montrons que la procédure SMOTE avec la valeur par défaut de l'hyperparamètre $K = 5$ copie asymptotiquement les échantillons originaux. L'idée de CV SMOTE est donc d'essayer plusieurs valeurs de K afin d'éviter de copier les échantillons et d'obtenir probablement une amélioration de la performance prédictive du classificateur utilisé par la suite.

Multivariate Gaussian SMOTE(K) Nous introduisons maintenant une nouvelle stratégie de suréchantillonnage que nous nommons Multivariate Gaussian SMOTE (MGS). Dans cette procédure, nous générons de nouveaux échantillons à partir de la distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$, où la moyenne empirique et la matrice de covariance ($\hat{\mu}$ et $\hat{\Sigma}$ respectivement) sont estimées à l'aide des K plus proches voisins et du point central. Nous détaillons une itération MGS dans l'algorithme 2. L'idée sous-jacente de MGS est d'exploiter au maximum le voisinage du point central. L'utilisation d'une distribution gaussienne multivariée, dont le support n'est pas borné, réduit le risque de copier simplement les échantillons originaux lorsque $K/n \to 0$.

Algorithm 2 Une itération de Multivariate Gaussian SMOTE.

Input: Échantillons de la classe minoritaire $X_1, \ldots, X_n$, nombre de plus proches voisins K.
Choisir uniformément X_c (appellé **point central**) parmi $\{X_1, \ldots, X_n\}$.
On note $I = X_{(1)}(X_c,) \ldots, X_{(K)}(X_c)$, les K plus proches voisins de X_c (norme L_2).
$\hat{\mu}(X_c) \leftarrow \frac{1}{K+1} \sum_{X_k \in I \cup \{X_c\}} X_k$
$\hat{\Sigma}(X_c) \leftarrow \frac{1}{K+1} \sum_{X_k \in I \cup \{X_c\}} (X_k - \hat{\mu})^T (X_k - \hat{\mu})$
Sample $Z \sim \mathcal{N}\left(\hat{\mu}(x_c), \hat{\Sigma}(x_c)\right)$
Return Z

4 Résultats

Protocole Nous comparons les différentes stratégies de rééquilibrage sur 11 ensembles de données réelles décrits dans la Table 1. Nous utilisons un stratified-split de 80%/20% (formation/test) des données, et appliquons chaque stratégie de rééquilibrage sur l'ensemble

Table 1: Description des ensembles de données, où n est le nombre d'échantillons et d le nombre de variables explicatives.

	Total N	minority samples n/N	d
GA4	319 066	0.7%	7
Credit	284 315	0.2%	29
Abalone	4 177	1%	8
Phoneme	5 404	29%	5
Yeast	1 462	11%	8
Pima	768	35%	8
Wine	4 974	4%	11
Vehicule	846	23%	18
Ionosphere	351	36%	32
Haberman	306	26%	3
Breast cancer	630	36%	9

d'entrainement, afin d'obtenir un ensemble de données équilibré. Une procédure d'apprentissage (régression logistique ou Random Forest) avec des hyperparamètres par défaut est entraînée sur l'ensemble d'entraînement rééquilibré. La performance est évaluée sur l'ensemble de test via le ROC AUC. Cette procédure est répétée 100 fois et la moyenne des résultats est calculée. Nous utilisons l'implémentation de *imb-learn* (Lemaître et al., 2017) pour les stratégies état de l'art.

Méthodes de rééquilibrage Notons que tous les ensembles de données présentés dans Table 2 sont fortement déséquilibrés, avec un ratio inférieur à 1% (10% et 10% pour les ensembles de données Pima et Haberman(10%) respectivement). Alors que dans la grande majorité des scénarios, None fait partie des meilleures approches pour traiter les données déséquilibrées, elle semble être surpassée par les stratégies de rééquilibrage dédiées aux ensembles de données fortement déséquilibrés, présentées dans la Table 2. Par conséquent, en ne considérant que les variables d'entrée continue et en mesurant la performance prédictive avec l'AUC ROC, nous observons qu'appliquer une stratégie de rééquilibrage n'est nécessaire que dans un cas précis : la classe minoritaire est fortement sous-représentée. En outre, nous constatons que la stratégie RUS présente une meilleure amélioration des performances du classifier pour les très grands ensembles de données, qui devraient être moins sensibles à la perte d'informations due au sous-échantillonnage. Plusieurs articles précurseurs avaient déjà remarqué que la stratégie None était compétitive en termes de performances prédictives. He et al. (2008) comparent la stratégie None, ADASYN et SMOTE, avant l'entraînement d'un d'arbre de décision sur des ensembles de données de 5 (comprenant Vehicle, Pima, Ionosphere et Abalone). En termes de précision et de score F1, la stratégie None est à égalité avec les deux autres méthodes de rééquilibrage. Han et al. (2005) étudient l'impact de Borderline SMOTE et d'autres variantes de SMOTE sur 4 ensemble de données (y compris Pima et Haberman). La stratégie None est compétitive (en termes de score F1) sur deux de ces ensembles de données.

Table 2: ROC AUC pour Random Forest pour différentes stratégies de rééquilibrage et différents ensembles de données. Seuls les ensembles de données pour lesquels la stratégie None ne figure pas parmi les meilleures (en gras) sont affichés. Les ensembles de données artificiellement sous-échantillonnés au niveau de la classe minoritaire sont en italique.

Resampling Strategy	None	Class weight	RUS	ROS	Near Miss1	BS1	Smote	CV Smote	MGS
GA4 (1%)	0.660	0.472	**0.866**	0.500	0.848	0.652	0.506	0.720	0.650
Credit (0.2%)	0.939	0.938	**0.975**	0.941	0.906	0.945	0.954	0.954	0.950
Abalone (1%)	0.697	0.702	0.719	0.712	0.570	0.712	0.756	0.750	**0.799**
Phoneme (1%)	0.819	0.821	0.851	0.814	0.575	0.847	0.876	0.877	**0.899**
Yeast (1%)	0.906	0.928	0.931	0.929	0.806	0.946	0.967	**0.968**	0.944
Wine (4%)	0.819	0.815	**0.846**	0.810	0.748	0.827	0.828	0.822	0.822
Pima (10%)	0.797	0.804	0.802	0.800	0.680	0.812	0.807	0.806	**0.821**
Haberman (10%)	0.580	0.580	0.599	0.582	**0.634**	0.609	0.617	0.598	0.619

SMOTE Nous remarquons que les performances de CV SMOTE sont comparables à celles de SMOTE avec l'hyperparamètre par défaut ($K = 5$). Cela peut s'expliquer par le choix de notre grille (qui pourrait être étendue) ou par les caractéristiques de l'ensemble des données. En effet, le seul jeu de données pour lequel nous constatons que CV SMOTE est notablement meilleur que SMOTE est GA4, qui contient le plus grand nombre d'échantillons minoritaires. Cela correspond à notre analyse théorique (Theorem 2.1) qui souligne que SMOTE, par défaut, tend à copier les échantillons minoritaires originaux, lorsque le nombre d'échantillons minoritaires est suffisamment important. Il conviendrait donc d'effectuer d'autres analyses pour étudier l'efficacité potentielle de CV SMOTE lorsque le nombre d'échantillons minoritaires est suffisamment élevé.

MGS Cette nouvelle stratégie présente de bonnes améliorations des performances prédictives. En effet, comme le montre la Table 2, MGS présente la meilleure amélioration sur 3 ensembles de données. Cela pourrait s'expliquer par l'échantillonnage gaussien des observations synthétiques qui permet aux points de données générés de se situer en dehors de l'enveloppe convexe de la classe minoritaire. MGS est potentiellement une nouvelle stratégie prometteuse, qui sera disponible sous la forme d'un logiciel libre.

5 Conclusion

Notre travail dans cet article est à la fois théorique et expérimental. Nous avons d'abord prouvé que SMOTE (avec le paramètre par défaut) régénère la distribution originale en copiant simplement les échantillons minoritaires originaux. Ces résultats théoriques nous ont permis d'introduire deux nouvelles stratégies permettant de générer des instances synthétiques au sein de la classe minoritaire.

D'autres expériences devraient être menées pour comprendre les performances surprenantes de RUS, qui surpasse systématiquement ROS, alors que les deux méthodes très similaires, car elles reposent toutes deux sur le rééchantillonnage. Enfin, afin d'analyser MGS(K) plus en détail, nous aimerions étudier l'impact d'un facteur de renormalisation λ dans l'estimation de la matrice de covariance, de sorte que la dernière étape de Algorithm 2 deviendrait $Z \sim \mathcal{N}(\hat{\mu}, \lambda\hat{\Sigma})$.

References

Elreedy, D. and A. F. Atiya (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences 505*, 32–64.

Elreedy, D., A. F. Atiya, and F. Kamalov (2023, January). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*.

Han, H., W.-Y. Wang, and B.-H. Mao (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer.

Hassan, A. K. I. and A. Abraham (2016). Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pp. 117–127. Springer.

He, H., Y. Bai, E. A. Garcia, and S. Li (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328. Ieee.

He, H. and E. A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering 21*(9), 1263–1284.

Khalilia, M., S. Chakraborty, and M. Popescu (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making 11*, 1–13.

King, G. and L. Zeng (2001). Logistic regression in rare events data. *Political Analysis 9*(2), 137–163.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence 5*(4), 221–232.

Lemaître, G., F. Nogueira, and C. K. Aridas (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*(17), 1–5.

Mohammed, A. J., M. M. Hassan, and D. H. Kadir (2020). Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering 9*(3), 3161–3172.

Nguyen, H. M., E. W. Cooper, and K. Kamei (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms 3*(1), 4–21.

Nguyen, N. N. and A. T. Duong (2021). Comparison of two main approaches for handling imbalanced data in churn prediction problem. *Journal of advances in information technology 12*(1).

Ramyachitra, D. and P. Manikandan (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR) 5*(4), 1–29.