Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants

Abdoulaye Sakho[1,2], Emmanuel Malherbe[1], Erwan Scornet[2]

[1]Artefact Research Center, Paris, France

[2]Sorbonne Université and Université Paris Cité, CNRS, LPSM, F-75005 Paris, France

Context and objectives

☞ Addressing class imbalance in binary classification presents a significant challenge across various machine learning applications such as fraud detection [1], medical diagnosis [2] and churn detection [3].

☞ *Synthetic Minority Oversampling Technique* (SMOTE) [4] is a common rebalancing strategy. However, few works analyse SMOTE theoretically. We adress this gap.

☞ We adapt SMOTE based on our theoretical findings to introduce two new oversampling variants.

Theoretical results on SMOTE

Let $\mathcal{D}_N = \{(X_i, Y_i)\}_{i=1}^N$ be set of N i.i.d samples distributed as (X, Y), with $\mathcal{X} \subset \mathbb{R}^d$ and Y $\in \{0, 1\}$. We suppose $\sum_{i=1}^N Y_i = n << N$.

SMOTE algorithm has a single hyperparameter, K, by default set to 5, which stands for the number of nearest neighbors considered when interpolating. We denote by $Z_{K,n}$ the distribution of samples generated by SMOTE.

Assumption: The imbalance ratio defined as n/N satisfies $\lim_{N\to\infty} \frac{n}{N} = \mathbb{P}(Y = 1)$.

Thus, we study the theoretical behavior of SMOTE, by letting $n \to \infty$ or by providing finite-sample upper bounds.

Convergence analysis of SMOTE

For all Borel sets $B \subset \mathbb{R}^d$, if $K/n \to 0$, we have

$$\lim_{n\to\infty} \mathbb{P}[Z_{K,n} \in B] = \mathbb{P}[X \in B]. \tag{1}$$

SMOTE asymptotically regenerates the distribution of the minority class when $K/n \to 0$.

Characteristic distance and copying phenomenon

Assume that there exists $R > 0$ such that $\mathcal{X} \in B(0, R)$. Besides, there exist $0 < C_2 < \infty$ such that for all $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x\in\mathcal{X}}$. For all $d \geq 2$, for all $\gamma \in (0, 1/d)$, we have

$$\mathbb{P}\left[\|Z_{K,n} - X_c\|_2 > 12R(K/n)^\gamma\right] \leq \left(\frac{K}{n}\right)^{2/d-2\gamma}. \tag{2}$$

We see that SMOTE regenerates the distribution of the minority class by almost exactly copying the original observations.

Vanishing density near the border of the support

There exists $R > 0$ such that $\mathcal{X} = B(0, R)$. Besides, there exist $0 < C_1 < C_2 < \infty$ such that for all $x \in \mathbb{R}^d$, $C_1 \mathbb{1}_{x\in\mathcal{X}} \leq f_X(x) \leq C_2 \mathbb{1}_{x\in\mathcal{X}}$. Let $\varepsilon \in (0, R)$ such that $\left(\frac{\varepsilon}{R}\right)^{1/2} \leq \frac{c_d}{\sqrt{2}dC_2}$. Then, for all $1 \leq K < n$, and all $z \in B(0, R)\backslash B(0, R-\varepsilon)$, and for all $d > 1$, we have

$$f_{Z_{K,n}}(z) \leq C_2^{3/2} \left(\frac{2^{d+2} c_d^{1/2}}{d^{1/2}}\right) \left(\frac{n-1}{K}\right) \left(\frac{\varepsilon}{R}\right)^{1/4}. \tag{3}$$

SMOTE is unable to reproduce the original density in the peripheral area: it vanishes as it approaches the boundary, whereas the original density is uniform on $B(0, R)$.

Based on our theoretical findings, we propose two slight modifications of SMOTE: tuning the hyperparameter K of SMOTE and Multivariate Gausssian SMOTE (MGS) which can be seen as a gaussian kernel estimator.

Proposed variant: Multivariate Gaussian SMOTE (MGS)

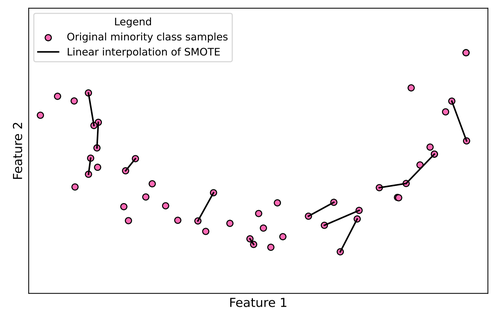


Figure: SMOTE

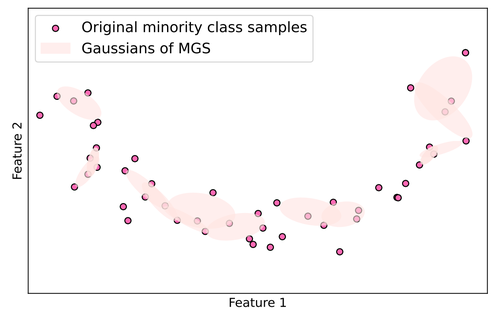


Figure: MGS

Algorithm Single MGS Sample Generation.

Require: Minority class samples $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$, number of neighbors K

Ensure: Synthetic sample Z

1: Select uniformly X_c among $X_1, \ldots, X_n$.
2: Compute the set I of the K nearest neighbors of X_c (Euclidean distance)
3: $\hat{\mu} \leftarrow \frac{1}{K+1} \sum_{x\in I} x$
4: $\hat{\Sigma} \leftarrow \frac{1}{K+1} \sum_{x\in I} (x - \hat{\mu})^T (x - \hat{\mu})$
5: Generate synthetic sample $Z \sim \mathcal{N}\left(\hat{\mu}, \hat{\Sigma}\right)$
6: **return** Z

For 11 initial data sets out of 13: applying no strategy (named None in all tables) is the best. Only data sets whose PR AUC of at least one rebalancing strategy is larger than that of None strategy are displayed in the following table.

Table: Random Forests PR AUC (tree depth tuned on PR AUC) on extremely imbalanced data sets.

Strategy	None	CW	RUS	ROS	SMOTE	SMOTE K-tuned	MGS	CT GAN
CreditCard (0.2%)	0.776	0.783	0.751	**0.787**	0.773	0.771	0.747	0.757
std	0.0	0.0	0.008	0.003	0.002	0.006	0.01	0.012
Abalone (1%)	0.048	0.055	0.052	0.052	0.046	0.051	**0.056**	0.050
std	0.006	0.009	0.017	0.011	0.011	0.012	0.014	0.016
Phoneme (1%)	0.198	0.211	0.086	0.200	0.224	0.240	**0.269**	0.130
std	0.026	0.026	0.031	0.037	0.035	0.035	0.032	0.042
Haberman (10%)	0.252	0.268	0.289	0.295	0.306	0.307	**0.311**	0.280
std	0.033	0.035	0.045	0.038	0.034	0.036	0.032	0.024
MagicTel (20%)	0.754	0.759	0.741	**0.760**	0.757	0.757	0.750	0.693
std	0.002	0.003	0.004	0.004	0.004	0.003	0.003	0.012
California (1%)	0.298	0.283	0.207	0.277	0.242	0.258	**0.322**	0.152
std	0.015	0.02	0.023	0.017	0.02	0.019	0.024	0.023
Average	0.388	0.393	0.354	0.395	0.391	0.397	**0.409**	0.344
Time (s)	6.0	5.1	1.7	5.9	6.1	120.2	7.2	32.8

Conclusion

☞ We prove that, without tuning the hyperparameter K (usually set to 5), SMOTE asymptotically copies the original minority samples, therefore lacking the intrinsic variability required in any synthetic generative procedure.

☞ Our theoretical analysis naturally leads us to introduce two SMOTE alternatives, SMOTE K-tuned and Multivariate Gaussian SMOTE (MGS). MGS is among the best strategies in our experimental setting.

☞ We extended MGS to handle both continuous and categorical features. The categorical features are handled through a generalized random forest [see 5, for details].

References

[1] A. K. I. Hassan et al. "Modeling insurance fraud detection using imbalanced data classification". In: Springer. 2016.

[2] M. Khalilia et al. "Predicting disease risks from highly imbalanced data using random forest". In: (2011).

[3] N. N. Nam et al. "Comparison of two main approaches for handling imbalanced data in churn prediction problem". In: (2021).

[4] N. V. Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: (2002).

[5] A. Sakho et al. "Harnessing Mixed Features for Imbalance Data Oversampling: Application to Bank Customers Scoring". In: Springer. 2025.