ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

SMOTE k-out: Enhancing Class Separability through Outer Synthetic Sampling

José Luis Morillo-Salas, Laura Morán-Fernández, Verónica Bolón-Canedo and Amparo Alonso-Betanzos *

CITIC, Universidade da Coruña, A Coruña, Spain

Abstract. Oversampling techniques are commonly used to address class imbalance in supervised classification, with SMOTE being a popular approach. However, traditional SMOTE generates synthetic samples within the neighbourhood of minority instances, which can increase data complexity and hinder class separability. This work proposes *SMOTE k-out*, which creates synthetic samples outside the local neighbourhood to increase minority class sparsity. This aims to reduce overfitting and mitigate the impact of noise, thereby improving the definition of the decision boundary. Experiments on multiple imbalanced datasets demonstrate that *SMOTE k-out* consistently reduces complexity and achieves higher accuracy and F-measure, particularly with SVM and LDA classifiers.

1 Introduction

Accurate classification in machine learning depends strongly on the balance and structure of the training data. In many real-world problems, such as medical diagnosis, fraud detection, or fault prediction, the available datasets are highly imbalanced, with the minority class being underrepresented. This imbalance biases learning algorithms towards the majority class, degrading their ability to identify rare but relevant events. To mitigate this issue, resampling techniques are commonly applied during preprocessing. Among them, the Synthetic Minority Oversampling Technique (SMOTE) [1] has become a reference approach due to its simplicity and effectiveness.

SMOTE generates new minority instances by interpolating between existing samples and their nearest neighbours. Although this process balances class proportions, it does not consider the intrinsic characteristics of the dataset. As a consequence, new synthetic samples can appear in regions of high class overlap or low density, increasing dataset complexity and sometimes leading to worse generalization. Several SMOTE variants have been proposed to address these limitations, incorporating noise filtering, adaptive sampling, or local density estimation. Nevertheless, few methods explicitly exploit dataset complexity as a guiding criterion for sample generation.

*Grant ERDF/EU (PID2023-147404OB-I00) funded by MICIU/AEI/10.13039/501100011033. CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia and FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01). Grant ED431C 2022/44 funded by Xunta de Galicia.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Dataset complexity [2] quantifies the difficulty of a classification problem in terms of class separability, geometric structure, and feature overlap. Recent studies have shown that reducing complexity tends to improve classification performance [3], particularly in imbalanced domains. Trying to reduce dataset complexity, we propose *SMOTE k-out*, a variant specifically designed to generate synthetic samples that promote greater sparsity in the minority class distribution. Our intuition is that when synthetic samples are generated only between existing minority samples (as in classic SMOTE), they tend to accumulate in dense regions, producing limited diversity and little complexity reduction. By contrast, generating synthetic samples outside these neighbourhoods allows for a more dispersed and informative distribution of the minority class. This outer-sample strategy aims to achieve better class separability while maintaining the simplicity and flexibility of the SMOTE framework.

The goal of this paper is to evaluate the effectiveness of *SMOTE k-out* in reducing dataset complexity and improving classification results. We analyse its behaviour using several classifiers and standard metrics, and compare its performance against traditional SMOTE. The results confirm that the proposed method leads to lower complexity and higher accuracy, especially in cases where the minority class forms compact clusters or overlaps with the majority class.

2 Proposed method: *SMOTE k-out*

To overcome the primary limitation of synthetic sampling, which often translates to an undesirable over-generalization within the strict vicinity of the minority class [1], our proposed approach explicitly seeks to increase the diversity of synthetic samples by extending beyond the class border to improve decision boundary definition [4]. Several other works have highlighted the importance of data diversity in obtaining better and fairer classification results [5, 6]. *SMOTE k-out* is designed to reduce class overlap by generating synthetic instances that extend minority support outward. Unlike SMOTE, which increases density within existing minority clusters, *SMOTE k-out* promotes outward sampling and selects only those synthetic points that minimise a dataset complexity measure. This strategy aims to enlarge the decision regions of the minority class while improving class separability.

2.1 Inner and outer sampling

For each minority instance, SMOTE generates synthetic samples along the direction connecting it to its nearest neighbours. Traditional SMOTE creates these samples between the instance and its neighbours, increasing local density within minority clusters. In contrast, *SMOTE k-out* also generates samples beyond the neighbours, along the same direction. These "outer samples" help to disperse the minority class, reducing overlap with the majority class and potentially lowering dataset complexity. The distance of outer samples can be adjusted to control how far they extend beyond the existing minority instances. Figure 1 illustrates

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

the difference between inner and outer sampling, highlighting how outer samples can improve class separability.

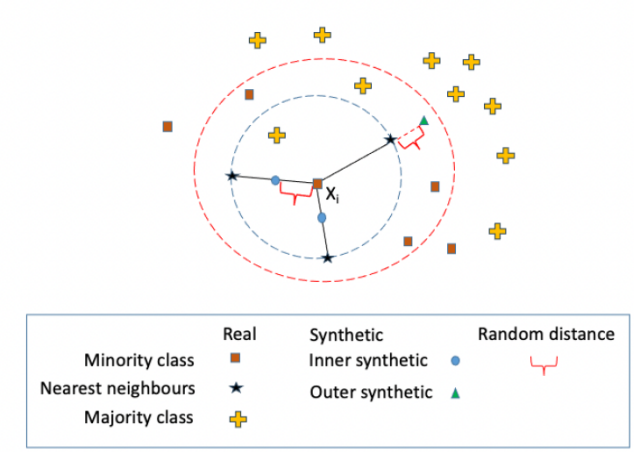

Fig. 1: Inner versus outer synthetic samples. Outer sampling disperses the minority class and tends to reduce dataset complexity.

2.2 Complexity measures

The selection of synthetic instances in *SMOTE k-out* is guided by dataset complexity. To quantify class separability, we employed two Fisher-based measures from the ECoL package [2]: the Maximum Fisher's Discriminant Ratio (F_1) and its directional variant (F_{1v}). These indicators, part of the data complexity family proposed by Lorena et al. [2], quantify how difficult a dataset is to classify in terms of feature-wise overlap and linear separability.

The Maximum Fisher's Discriminant Ratio (F_1) evaluates the overlap between class distributions for each feature. For a dataset with m features, the discriminant ratio for feature f_i is defined as $r_{f_i} = (\mu_{i,1} - \mu_{i,2})^2/\sigma_{i,1}^2 + \sigma_{i,2}^2$ where $\mu_{i,c}$ and $\sigma_{i,c}$ are the mean and variance of feature f_i in class c. The global measure is $F_1 = 1/(1 + \max_i r_{f_i})$, yielding values in $(0, 1]$, where higher values indicate higher complexity (i.e., stronger class overlap).

The directional variant F_{1v} extends this concept by searching for a projection vector $\mathbf{w}$ that maximizes the Fisher criterion in a transformed space:

$$F_{1v} = \frac{1}{1 + \frac{\mathbf{w}^\top S_W \mathbf{w}}{\mathbf{w}^\top S_B \mathbf{w}}},$$

where S_W and S_B are the within-class and between-class scatter matrices. Lower F_{1v} values correspond to datasets that are easier to separate linearly. F_1 captures local feature-level separability, while F_{1v} provides a global view of the linear discriminability of the dataset.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2.3 Complexity-guided oversampling

Using the complexity measures described above as selection criteria, *SMOTE k-out* operates as follows. For each reference minority instance, the method generates both inner and outer candidate samples along the direction connecting the instance to its nearest neighbours. The proportion of candidates generated as outer samples is controlled by the outer sampling rate $p_{\text{out}} \in [0, 100]\%$, which allows adjusting the balance between local (inner) and dispersed (outer) synthetic instances. Once candidates are generated, each one is temporarily added to the training dataset, and a dataset complexity measure $C(\cdot) \in \{F_1, F_{1v}\}$ is computed. The candidate that minimises complexity is then selected and permanently added to the dataset. Because some outer candidates may be discarded during this process, the actual fraction of synthetic samples that are outer in the final dataset may differ from p_{out}. This effective proportion is denoted as k_{out} and is reported in the experimental results. The iterative process continues until the desired oversampling percentage is reached, as can be seen in Algorithm 1.

Input: training set D, minority set S, neighbours k, outer ratio p_{out}, oversampling rate r, complexity function C
while *number of added samples* $< r \cdot |S|$ **do**
  Select $x \in S$ and its k nearest neighbours $N_k(x)$;
  Generate inner and outer candidates y along $x \leftrightarrow n$ and beyond n according to p_{out};
  Compute $C(D \cup \{y\})$ for each candidate;
  Select $y^\star = \arg\min_y C(D \cup \{y\})$;
  Add $y^\star$ to D and to S;
end

Algorithm 1: *SMOTE k-out* (complexity-guided selection)

3 Datasets and protocol

We evaluated the method on twelve binary datasets from the UCI Machine Learning Repository [7], which exhibit diverse imbalance ratios and data topologies (see Table 1). For each dataset, a holdout split is applied with two-thirds of the samples for training and one-third for testing; this procedure is repeated ten times, and results are averaged.

Classification performance is evaluated using Accuracy and F-measure as complementary metrics. To ensure diverse inductive biases, five standard classifiers are applied: decision tree (C4.5), instance-based learner (IB1), linear discriminant analysis (LDA), Naive Bayes, and support vector machine (SVM). All classifiers used the default hyperparameters in the Weka implementation.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Dataset	# Features	# Samples	IR
alizadeh_cath	55	303	2.48
autism adult	16	702	2.71
climate	18	540	10.73
german	20	1000	2.33
haberman	3	306	2.77
immunotherapy	7	90	3.73
planning_relax	12	182	2.50
spect	22	267	3.85
spectf	44	267	3.85
ThoracicSurgery	16	470	5.71
wholesale_channel	6	440	2.10
z_alizadeh	55	303	2.18

Table 1: Main characteristics of the datasets employed for experimentation. IR denotes the Imbalance Ratio between the majority and minority classes.

4 Experimentation

This section analyses the effect of generating synthetic outer samples in *SMOTE k-out*. The experiments aim to determine whether the introduction of outer sampling reduces dataset complexity and improves classification performance with respect to traditional SMOTE. For each dataset, several oversampling configurations were tested by varying the proportion of outer samples k_{out} in $\{0, 30, 60, 100\}\%$, while keeping the number of neighbours fixed at $k = 5$ and balancing the minority and majority classes. All results were averaged across ten runs and twelve datasets.

Table 2 summarises the average improvement obtained by *SMOTE k-out* over standard SMOTE in terms of Accuracy and F-measure for different classifiers. In all cases, generating outer samples leads to lower dataset complexity (both F_1 and F_{1v}) and to higher predictive performance. The best results are observed when $k_{out} = 60\%$, followed by $k = 100\%$, confirming that greater sparsity in the minority class enhances class separability.

Classifier	$k_{out} = 0\%$	30%	60%	100%	Best gain
C4.5	+0.8 / +0.6	+1.5 / +1.2	+2.3 / +1.9	+2.1 / +1.7	+2.3 / +1.9
IB1	+1.0 / +0.9	+2.2 / +1.8	+3.6 / +2.7	+3.5 / +2.5	+3.6 / +2.7
LDA	+1.9 / +1.4	+3.7 / +2.8	+5.1 / +3.6	+5.0 / +3.8	+5.1 / +3.6
Naive Bayes	+1.2 / +0.8	+2.9 / +2.0	+3.7 / +2.8	+3.5 / +2.6	+3.7 / +2.8
SVM	+2.5 / +1.9	+5.2 / +3.4	+7.3 / +4.4	+7.0 / +4.1	+7.3 / +4.4
Average	+1.5 / +1.1	+3.1 / +2.2	+4.4 / +3.1	+4.2 / +3.0	—

Table 2: Mean percentage gains of *SMOTE k-out* compared with standard SMOTE, expressed as Accuracy / F-measure for each classifier and outer-sample proportion (k_{out}).

All improvements are statistically significant under the Friedman test ($p < 0.05$). The benefit is consistent across classifiers, though its magnitude varies.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

SVM achieves the largest increase in both metrics, confirming that simpler and more separable data distributions facilitate the identification of optimal hyperplanes. LDA also benefits notably from the reduced complexity, while Naive Bayes and IB1 show moderate gains. C4.5 presents smaller improvements, as its rule-based structure is less sensitive to local geometric variations.

Overall, the results confirm that generating more sparse synthetic samples through the k-out mechanism yields datasets of lower complexity and, consequently, better generalisation. Even with moderate outer-sample proportions (around 60%), the proposed method consistently outperforms standard SMOTE.

5 Conclusions

This work presented *SMOTE k-out*, a variation of SMOTE designed to reduce dataset complexity by generating synthetic samples beyond the neighbourhood of minority instances. The outer sampling mechanism increases the sparsity of the minority class, producing simpler and more separable training datasets. Experimental results on twelve imbalanced datasets confirm that *SMOTE k-out* consistently improves both performance measures, Accuracy and F-measure, over standard SMOTE, with statistically significant gains for all classifiers, particularly SVM and LDA.

The method maintains the simplicity of the original SMOTE while introducing a parameter to adjust the proportion of outer samples, offering a trade-off between diversity and computational cost. Future research will explore adaptive strategies to automatically set this parameter according to dataset characteristics and will evaluate the method in multiclass and high-dimensional domains.

References

[1] N. V Chawla, K.W Bowyer, L.O Hall, and W P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[2] A.C Lorena, L. PF Garcia, J. Lehmann, Marcilio CP Souto, and T. Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34, 2019.

[3] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos. Can classification performance be predicted by complexity measures? A study using microarray data. *Knowledge and Information Systems*, 51:1067–1090, 2017.

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

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 77–91. PMLR, 2018.

[6] R. Binns. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 149–159. PMLR, 2018.

[7] D. Dua and C. Graff. UCI Machine Learning Repository. Available: `http://archive.ics.uci.edu/ml`, 2017.