Data Sampling Optimization for Improved Classification of Imbalanced Phishing Datasets

José Maurício Silva[1], Carlo Marcelo R. Silva[2], Mateus L.S.D. Barros[3], João Guilherme Miranda[3], Márcio P. Basgalupp[4], and Péricles B.C. Miranda [3]

[1]CESAR School, Brazil

[2]Universidade de Pernambuco, Brazil

[3]Departamento de Computação – UFRPE, Brazil

[4]Instituto de Ciência e Tecnologia – UNIFESP, Brazil

{pericles.miranda}@ufrpe.br

Abstract. *Phishing is a social engineering attack that captures information by impersonating trusted entities. To detect it, researchers use Machine Learning as a classification task. However, phishing datasets are often imbalanced due to Concept Drift and the semantic nature of attacks. Oversampling, undersampling, and hybrid techniques address this, with hybrids combining both strategies for better results. This study examines the impact of optimization-based sequencing of sampling algorithms on phishing data and compares it to traditional methods. Results show that optimized sequences improve classifier performance and reduce the effects of imbalance.*

Resumo. *Phishing é um ataque de engenharia social que captura informações ao imitar entidades confiáveis. Para detectá-lo, usa-se Aprendizado de Máquina como tarefa de classificação. Porém, os dados costumam ser desbalanceados devido ao Concept Drift e à natureza semântica dos ataques. Técnicas de oversampling, undersampling e híbridas são aplicadas, sendo as híbridas mais eficazes ao combinar ambas. Este estudo avalia o impacto da ordenação otimizada de algoritmos de amostragem em dados de phishing, comparando com métodos tradicionais. Os resultados mostram que as sequências otimizadas melhoram o desempenho e reduzem o desbalanceamento.*

1. Introduction

Information security is a major concern in the digital era due to the growing value of data and rising cybercrime. Phishing, a prevalent social engineering attack that impersonates trusted entities to extract sensitive information, is especially effective via emails and messages [Barros et al. 2020, de Barros et al. 2019, Barros et al. 2019]. This has motivated machine learning to detect and classify such threats from patterns in malicious website datasets [Barros et al. 2020]. However, data imbalance often hinders classification performance, as legitimate instances typically outnumber phishing ones, impairing recognition of the minority class [Haixiang et al. 2017]. The evolving nature of phishing reinforces the need for adaptive detection methods. Resampling techniques like undersampling and oversampling are used to mitigate imbalance, though they may cause information loss

or noise [Haixiang et al. 2017, Barbosa et al. 2019]. To address these limitations, recent studies have explored hybrid sampling approaches that combine over- and under-sampling techniques to leverage their complementary strengths [Srivastava and Sharan 2022, Ding et al. 2021, Barbosa et al. 2020b, Miranda et al. 2022, Oliveira et al. 2023]. Results indicate that such methods are effective for handling imbalance. The sequencing of sampling algorithms reflects the problem's complexity and the need for strategies that account for both imbalance and the nuanced nature of phishing data.

This work investigates the impact of a computational intelligence approach for balancing phishing datasets using the optimized sequencing of sampling algorithms proposed in [Miranda et al. 2022]. The goal is to assess its effect on classifier performance and compare it to traditional resampling techniques regarding accuracy, precision, recall, F1-score, and Geometric mean (G-mean). The study is driven by the need to enhance phishing detection in imbalanced scenarios and evaluate sequencing as a potential solution. Offering a comparative analysis contributes to cybersecurity research, emphasizing the importance of precision and generalization in phishing classification.

2. Related Works

Identified studies on balancing phishing datasets adopt diverse approaches, from using classifiers like Optimized Random Forest, which incorporate imbalance treatments, to methods based on feature extraction and multiple classifiers that distinguish phishing from legitimate sites without necessarily applying sampling techniques [Ahsan et al. 2018]. Studies such as Ahsan et al. [Ahsan et al. 2018] applied SMOTE to balance datasets, evaluating accuracy with classifiers such as eXtreme Gradient Boost (XGBoost), Random Forest (RF), and Support Vector Machine (SVM). On the other hand, Pristyanto and Dahlan [Pristyanto and Dahlan 2019] combined preprocessing techniques, using both oversampling (SMOTE) and undersampling (One-Sided Selection – OSS) to address imbalance in phishing data. Prayogo and Karimah [Prayogo and Karimah 2020] proposed combining SMOTE with filter-based feature selection, evaluating performance using the K-Nearest Neighbor (KNN) classifier. Ding et al. [Ding et al. 2021] introduced KM-SMOTE, an enhanced SMOTE variant that employs K-means clustering to generate more coherent minority samples, aiming to reduce imbalance in phishing datasets. Srivastava and Sharan [Srivastava and Sharan 2022] explored using hybrid sampling algorithms, specifically SMOTE/ENN, for data balancing, testing effectiveness with classifiers such as XGBoost, RF, Logistic Regression (LR), and SVMs.

To overcome oversampling and undersampling limitations, some studies propose sequencing sampling algorithms [Barbosa et al. 2020b, Barbosa et al. 2020a]. Miranda et al. [Miranda et al. 2022] introduced SASO, a multi-objective optimization method that selects optimal sequences from algorithms like ENN, NearMiss, OSS, SMOTE, SMOTE+Tomek, and Tomek Links. Though effective in other domains, SASO has not been applied to phishing. This study evaluates its impact on phishing classifiers using metrics such as accuracy, precision, recall, F1-score, and G-mean.

3. Materials and Methods

3.1. Datasets

Two public phishing classification datasets were used. The first, the Website Phishing Data Set [Abdelhamid et al. 2014] from UCI, referred to as *UCI_Multiclass*, is a widely

used multi-class dataset with 1,353 instances and 10 attributes (9 features and 1 class). It includes three imbalanced classes: Class A ("phishing detected") with 702 instances, Class B ("suspicious") with 103, and Class C ("no phishing detected") with 548. The second dataset, *Phish_Binary*, from [Barros et al. 2020, Gomes de Barros et al. 2022], was built using data from Phishtank and Openphish. It contains 284,676 instances and 12 attributes (11 features and 1 class), with two classes: "phishing detected" (Class A – 198,891 instances) and "no phishing detected" (Class B – 94,785 instances).

3.2. Adopted Algorithms

Given the problem's nature, SASO uses the NSGA-II multi-objective genetic algorithm [Miranda et al. 2022, Deb et al. 2002], effective for combinatorial optimization. SASO's parameters are N (genes per chromosome) and $D = 6$ (possible sampling algorithms). This study used $N = 3, 5, 7$, with solutions formed from: *Near Miss* (0), *ENN* (1), *SMOTE* (2), *Tomek Links* (3), *OSS* (4), and *SMOTE + Tomek Links* (5)—four *undersampling* and two *oversampling* methods. The sequencing approach was compared to traditional *under*, *over-sampling*, and a hybrid method [Srivastava and Sharan 2022]. Baseline classifiers (KNN, SVM, RF) were tested without resampling. For fair comparison, a single solution from the *Pareto front* was selected using the *Borda Count* method. All models used Scikit-Learn[1] defaults, and samplers followed Imbalanced Learn[2] settings.

3.3. Evaluation

Balancing quality was evaluated using KNN, RF, and SVM classifier performance. A consistent procedure was applied: each sampling method was used to balance the training set, and the resulting classifier was evaluated on the original unbalanced test set. This process followed a ten-fold cross-validation, producing average values for accuracy, precision, recall, F_1-score, and G-mean. Accuracy alone can be misleading in imbalanced contexts, as it favors the majority class. Precision and recall offer insights into false positives and negatives, while the F1-score balances both. G-Mean ensures a trade-off between sensitivity and specificity. Together, these metrics provide a more robust assessment of classifier performance on imbalanced datasets.

4. Experimental Results

This section presents the findings across three scenarios: (i) classification on unbalanced data, (ii) classification with traditional sampling techniques, and (iii) classification using the Self-Adaptive Sampling Optimization (SASO) method. The results related to the research questions are detailed below, followed by a comprehensive discussion.

4.1. Results

Table 1 presents the classification results for the *UCI Multiclass Dataset*. This table shows the impact of different sampling strategies, particularly the influence of the SASO algorithm on classifiers. The study evaluates three classifiers—RF, KNN, and SVM—using various sampling techniques to address the class imbalance.

[1] `https://scikit-learn.org/stable/`
[2] `https://imbalanced-learn.org/stable/`

Baseline Performance (No Sampling Applied). Without sampling, RF performs best with 89.01% accuracy, 87.16% F1-score, and 87.28% G-mean. KNN performs moderately but has low recall (81.75%), while SVM is most affected by imbalance, with a G-mean of 62.16%. This confirms RF's robustness and SVM's vulnerability to class imbalance.

Impact of Conventional Sampling Methods. Applying SMOTE significantly boosts KNN and SVM performance, reducing imbalance effects. KNN reaches 91.72% accuracy and 92.18% G-mean; SVM achieves 92.44% accuracy and 92.17% G-mean. RF still leads (93.15%), but the gap narrows. Combining SMOTE with ENN further improves results: KNN reaches 92.23% accuracy and SVM 92.50%, with minimal margin behind RF (93.53%). These results show that SMOTE + ENN enhances generalization by balancing the dataset and reducing noise.

Effectiveness of SASO Algorithm. SASO was applied with chromosome sizes $N = 3, 5, 7$ and $D = 6$ sampling strategies. Results show significant performance gains across classifiers. **SASO(N=3)** $(2, 4, 0$: SMOTE, OSS, Near Miss) raised RF accuracy to 98.35%, with KNN and SVM reaching 96.50% and 96.59%. **SASO(N=5)** $(4, 3, 3, 4, 3$: OSS and Tomek Links) achieved consistent results (RF: 93.48%, KNN: 92.38%), highlighting Tomek Links' balancing role. **SASO(N=7)** $(2, 3, 4, 4, 2, 0, 4)$ delivered the best performance, with KNN outperforming RF at 99.41% accuracy. These findings confirm that larger N values enable more effective sampling combinations, and that SASO's adaptive nature can achieve near-perfect classification.

Table 2 shows classification results on the *Phish_Binary Dataset*, highlighting differences in sampling effectiveness. RF performs consistently well, though its recall suggests a slight bias toward the majority class. KNN and SVM show greater sensitivity to the sampling method used.

Baseline Performance (No Sampling Applied). Without any sampling algorithm, KNN achieves 89.76% accuracy and 92.52% F1-score, while SVM attains 90.88% accuracy and 93.60% F1-score. However, their G-mean values (86.77% and 85.24%, respectively) indicate that these models struggle with correctly identifying minority class instances.

Impact of Conventional Sampling Methods. Applying *SMOTE* improves recall for KNN and SVM, increasing their G-mean scores. Specifically, KNN with SMOTE achieves 87.71% accuracy and 96.30% recall, demonstrating its ability to capture minority class instances better. The combination of *SMOTE + ENN* further refines these improvements by removing noisy samples, leading to a balanced trade-off between precision and recall.

Effectiveness of SASO Algorithm. SASO yields the highest classification improvements. With SASO($N = 3$) $(5, 4, 5)$, KNN reaches 87.21% and SVM 91.17% accuracy. SASO($N = 5$) $(0, 1, 5, 2, 4)$ boosts KNN to 97.64% and SVM to 98.05% accuracy. SASO($N = 7$) $(3, 5, 4, 4, 3, 1, 0)$ pushes KNN to 99.41%, surpassing RF (99.38%). However, SVM shows a drop in G-mean and recall, indicating SASO's effectiveness overall, though results may vary with classifier sensitivity.

These results show that traditional oversampling techniques like SMOTE and SMOTE + ENN can significantly enhance classifier performance. However, adaptive methods such as SASO provide a more dynamic and effective solution to optimizing

Classifier	Sampling Algorithm	Accuracy	F1-score	Average G-mean	Precision	Recall
RF	None	89.01% ± 0.46	87.16% ± 0.75	87.28% ± 1.01	87.31% ± 0.91	87.58% ± 0.69
KNN		87.31% ± 0.35	82.58% ± 0.77	80.60% ± 1.06	84.44% ± 0.80	81.75% ± 0.84
SVM		87.07% ± 0.27	73.83% ± 1.15	62.16% ± 0.31	82.56% ± 1.70	71.70% ± 0.92
RF	SMOTE	93.15% ± 0.30	93.11% ± 0.30	92.92% ± 0.30	93.09% ± 0.30	93.15% ± 0.30
KNN		91.72% ± 0.18	91.66% ± 0.19	92.18% ± 0.18	92.38% ± 0.18	91.72% ± 0.18
SVM		92.44% ± 0.19	92.40% ± 0.20	92.17% ± 0.19	92.36% ± 0.19	92.45% ± 0.19
RF	SMOTE + ENN[Srivastava and Sharan 2022]	93.53% ± 0.23	93.49% ± 0.23	92.33% ± 0.23	92.53% ± 0.23	93.53% ± 0.23
KNN		92.23% ± 0.30	92.17% ± 0.30	92.10% ± 0.30	92.30% ± 0.30	92.23% ± 0.30
SVM		92.50% ± 0.14	92.45% ± 0.14	92.16% ± 0.14	92.35% ± 0.14	92.50% ± 0.14
RF	SASO(N=3): 2,4,0	98.35% ± 0.26	97.55% ± 0.45	97.42% ± 0.54	97.76% ± 0.51	97.49% ± 0.44
KNN		96.50% ± 0.31	93.15% ± 0.76	91.06% ± 1.13	95.23% ± 0.85	91.85% ± 0.62
SVM		96.59% ± 0.25	92.94% ± 0.57	89.73% ± 0.78	96.63% ± 0.51	90.86% ± 0.61
RF	SASO(N=5): 4,3,3,4,3	93.48% ± 0.32	93.45% ± 0.32	93.36% ± 0.32	93.53% ± 0.60	93.48% ± 0.32
KNN		92.38% ± 0.43	92.33% ± 0.43	91.68% ± 0.43	91.89% ± 0.43	91.77% ± 0.41
SVM		92.54% ± 0.35	92.50% ± 0.35	92.47% ± 0.35	92.65% ± 0.35	92.54% ± 0.34
RF	SASO(N=7): 2,3,4,4,2,0,4	99.38% ± 0.14	99.38% ± 0.14	99.38% ± 0.14	99.40% ± 0.14	99.38% ± 0.14
KNN		**99.41% ± 0.16**	**99.41% ± 0.16**	**99.41% ± 0.16**	**99.42% ± 0.16**	**99.41% ± 0.15**
SVM		98.14% ± 0.25	98.13% ± 0.25	98.06% ± 0.25	98.13% ± 0.25	98.13% ± 0.25

Table 1. Classification results for the *UCI_Multiclass* Dataset.

Classifier	Sampling Algorithm	Accuracy	F1-score	Average G-mean	Precision	Recall
RF	None	91.42% ± 0.01	93.83% ± 0.01	87.61% ± 0.01	90.07% ± 0.01	97.92% ± 0.02
KNN		89.76% ± 0.04	92.52% ± 0.03	86.77% ± 0.03	90.17% ± 0.03	95.01% ± 0.08
SVM		90.88% ± 0.02	93.60% ± 0.01	85.24% ± 0.01	88.03% ± 0.01	98.93% ± 0.02
RF	SMOTE	88.76% ± 0.01	89.38% ± 0.01	88.56% ± 0.01	84.67% ± 0.02	94.65% ± 0.02
KNN		87.71% ± 0.02	88.68% ± 0.02	87.28% ± 0.02	82.18% ± 0.04	96.30% ± 0.07
SVM		88.49% ± 0.08	89.27% ± 0.07	88.18% ± 0.08	83.61% ± 0.10	95.77% ± 0.10
RF	SMOTE + ENN[Srivastava and Sharan 2022]	88.72% ± 0.01	89.34% ± 0.01	88.53% ± 0.01	84.70% ± 0.02	94.51% ± 0.03
KNN		87.65% ± 0.03	88.63% ± 0.03	87.21% ± 0.03	82.08% ± 0.06	96.32% ± 0.09
SVM		87.89% ± 0.08	88.80% ± 0.07	87.51% ± 0.08	82.62% ± 0.09	95.98% ± 0.09
RF	SASO(N=3): 5,4,5	91.42% ± 0.01	93.83% ± 0.01	87.61% ± 0.02	97.76% ± 0.01	97.49% ± 0.01
KNN		87.21% ± 0.04	88.11% ± 0.04	86.89% ± 0.04	82.35% ± 0.06	94.73% ± 0.11
SVM		91.17% ± 0.02	93.74% ± 0.02	85.68% ± 0.02	99.87% ± 0.01	88.23% ± 0.02
RF	SASO(N=5): 0,1,5,2,4	87.64% ± 0.04	88.62% ± 0.04	88.53% ± 0.03	84.70% ± 0.06	94.52% ± 0.09
KNN		97.64% ± 0.13	97.35% ± 0.15	97.66% ± 0.15	96.96% ± 0.04	97.76% ± 0.30
SVM		98.05% ± 0.15	98.30% ± 0.13	97.74% ± 0.17	99.66% ± 0.05	99.66% ± 0.26
RF	SASO(N=7): 3,5,4,4,3,1,0	99.38% ± 0.14	99.38% ± 0.14	**99.38% ± 0.14**	99.40% ± 0.14	**99.88% ± 0.14**
KNN		**99.41% ± 0.16**	**99.41% ± 0.16**	87.22% ± 0.04	82.13% ± 0.06	96.21% ± 0.10
SVM		87.12% ± 0.23	85.59% ± 0.18	86.17% ± 0.27	**99.95% ± 0.01**	79.56% ± 0.29

Table 2. Classification results for *Phish_Binary* Dataset.

sampling strategies. By intelligently selecting a combination of oversampling and undersampling techniques, SASO ensures better class distribution balance and enhances classifier generalization, particularly for KNN and SVM.

5. Conclusion

This study proposes a novel method for handling dataset imbalance via sequencing sampling algorithms, with a focus on phishing detection. Leveraging computational intelligence, the approach outperforms traditional methods and achieves competitive results. SASO enhances classifier performance and accuracy, confirming its effectiveness for phishing detection and its potential applicability to other imbalance-prone domains. Future work may extend SASO to deep learning, incorporate adaptive weighting, and explore incremental learning to address Concept Drift. Enhancing feature engineering and applying SASO in online learning could improve robustness and real-time adaptation to evolving attacks.

References

Abdelhamid, N., Ayesh, A., and Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41:5948–5959.

Ahsan, M., Gomes, R., and Denton, A. (2018). Smote implementation on phishing data to enhance cybersecurity. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0531–0536.

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., and Gašević, D. (2020a). Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 605–614.

Barbosa, G., Miranda, P., Mello, R., and Silva, R. (2019). Sequenciamento de algoritmos de amostragem para aumentar o desempenho de classificadores em conjuntos de dados desequilibrados. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 413–423.

Barbosa, G., Miranda, P., Silva, R., and Mello, R. (2020b). Sequenciamento de algoritmos de amostragem para aumentar o desempenho de classificadores em conjuntos de dados desequilibrados. In *XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 413–423. SBC.

Barros, M., Silva, C., and Miranda, P. (2019). Adoção da seleção de características como mecanismo antiphishing: aplicabilidade e impactos. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 214–225.

Barros, M., Silva, C., and Miranda, P. (2020). Xphide: Um sistema especialista para a detecção de phishing. In *Anais do XX Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, pages 161–174, Porto Alegre, RS, Brasil. SBC.

de Barros, M., da Silva, C., and de Miranda, P. (2019). Aplicabilidade e impactos quanto a adoção de modelos de classificação como mecanismos anti-phishing. In *Anais Estendidos do XIX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 39–42.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Ding, X., Liu, B., Jiang, Z., Wang, Q., and Xin, L. (2021). Spear phishing emails detection based on machine learning. In *2021 IEEE 24th CSCWD*, pages 354–359.

Gomes de Barros, J. C., Revoredo da Silva, C. M., Candeia Teixeira, L., Torres Fernandes, B. J., Lorenzato de Oliveira, J. F., Luzeiro Feitosa, E., Pinheiro dos Santos, W., Ferraz Arcoverde, H., and Cardoso Garcia, V. (2022). Piracema: a phishing snapshot database for building dataset features. *Scientific Reports*, 12(1):15149.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data. *Expert Syst. Appl.*, 73(C):220–239.

Miranda, P. B., Mello, R. F., Nascimento, A. C., and Si, T. (2022). Multi-objective optimization of sampling algorithms pipeline for unbalanced problems. In *2022 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23*, pages 509–519.

Prayogo, R. D. and Karimah, S. A. (2020). Optimization of phishing website classification based on synthetic minority oversampling technique and feature selection. In *2020 International Workshop on Big Data and Information Security (IWBIS)*, pages 121–126.

Pristyanto, Y. and Dahlan, A. (2019). Hybrid resampling for imbalanced class handling on web phishing classification dataset. In *2019 4th ICITISEE*, pages 401–406.

Srivastava, J. and Sharan, A. (2022). SMOTEEN Hybrid Sampling Based Improved Phishing Website Detection. *Pre-Print*.