Hybrid Models for Improved Classification Accuracy Using Negative Selection Algorithms

Ritisha Priya CS22B022[1], Sujal Acham BE22B038[2]

1 Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
2 Department of Biotechnology, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

Abstract

Imbalanced datasets pose a significant challenge in machine learning, often leading to classifiers biased towards the majority class and poor performance in identifying crucial minority instances. Traditional data augmentation techniques have limitations. This project investigates the novel application of Negative Selection Algorithms (NSAs)—inspired by the human immune system—not for direct classification, but as an intelligent data augmentation strategy. We propose converting NSA-generated 'detectors,' which define 'non-self' space based on 'self' (majority) data, into synthetic minority class instances. Several NSA variants (Real-valued NSA, V-Detector NSA, Hierarchical Detector NSA) were adapted for this purpose. Their effectiveness was evaluated by training traditional classifiers (Logistic Regression, Gaussian Naive Bayes, Random Forest, MLP) on two benchmark imbalanced datasets (Breast Cancer Wisconsin, HCV) augmented by these NSA variants. Performance was compared against raw datasets, random augmentation, and the Synthetic Minority Over-sampling Technique (SMOTE). Results demonstrate that NSA-based augmentation significantly improved classifier performance, with V-Detector NSA often matching or outperforming SMOTE. The study highlights NSA's particular strength in scenarios where the minority class is diffusely defined as 'not-self.' This research formally benchmarks NSA variants for synthetic anomaly generation, validating their potential as a principled and effective tool to enhance imbalanced classification.

Introduction

0.1 Background

Machine learning has revolutionized various fields, but its efficacy often hinges on the quality and characteristics of the training data. One common and persistent challenge is that of imbalanced datasets, where one class (the majority or 'normal' class) significantly outnumbers another (the minority or 'anomalous' class). This imbalance biases traditional classifiers towards the majority class, leading to poor performance in identifying minority instances, which are often of critical interest (e.g., fraudulent transactions, rare diseases). [1]

Standard classification algorithms aim to minimize overall error. In imbalanced scenarios, they can achieve high accuracy by simply predicting the majority class, yet fail to detect the crucial minority instances. This leads to low recall and F1-scores for the minority class, rendering the model practically useless for many real-world applications. Techniques like oversampling (e.g., SMOTE), undersampling, or cost-sensitive learning have been developed, but each has its limitations, such as overfitting, loss of information, or difficulty in assigning appropriate costs.

0.2 Negative Selection Algorithm (NSA)

Inspired by the human immune system, the Negative Selection Algorithm (NSA) offers a unique paradigm for pattern recognition. In the immune system, negative selection is a critical process occurring in the thymus. During this process, T-cells that react to 'self' (body's own) antigens are eliminated. The surviving T-cells are thus tolerant to self-antigens and are capable of identifying and responding to 'non-self' (foreign) antigens or pathogens, even without prior exposure to them. [2]

Artificially, NSA mimics this by defining a 'self' set (typically normal data) and generating a set of 'detectors'. These detectors are random patterns that do not match any 'self' samples. During the monitoring phase, if a new data sample matches any of these detectors, it is classified as 'non-self' or anomalous..

0.3 Objective

While NSA [3] has been primarily used for anomaly detection as a standalone classifier, its potential as a data generation tool is less explored, especially in a formally benchmarked hybrid setting. Traditional data augmentation techniques for minority classes sometimes generate synthetic samples that are too simplistic or do not effectively capture the boundary between classes. NSA's mechanism of generating detectors in the 'non-self' space, based on the 'self' data, offers a principled way to create synthetic anomalies that lie in regions distinct from the majority class.

This project moves beyond using NSA for direct classification. [4] Instead, we propose leveraging its detector generation capability as an intelligent data augmentation strategy. The generated detectors, representing unoccupied regions of the feature space (potential 'non-self' areas), are converted into synthetic minority class data points. This approach aims to:

- Improve the balance of the training dataset.
- Provide more diverse and representative minority samples.
- Enhance the performance of traditional classifiers on imbalanced tasks.

The novelty lies in the systematic evaluation of multiple NSA variants (R-NSA, V-NSA, HD-NSA) for this augmentation purpose and benchmarking their impact against standard models, addressing a gap identified in earlier proposals that focused on simpler NSA versions without extensive benchmarking. The primary objectives of this project are:

1. To implement and adapt several NSA variants (R-NSA, V-NSA, HD-NSA) for the purpose of synthetic anomaly generation rather than direct detection.

2. To select and preprocess suitable imbalanced benchmark datasets for experimentation.

3. To develop a robust benchmarking methodology to evaluate the performance of traditional machine learning models when trained on datasets augmented by NSA-generated anomalies.

4. To compare the effectiveness of different NSA variants as augmentation tools in terms of classification performance (accuracy, precision, recall, F1-score) improvement.

5. To provide a comprehensive report summarizing the findings, methodology, and potential of this hybrid approach.

Materials and methods

Datasets

Each of the following datasets can be found on the UCI Machine Learning Repository or the Kaggle repository:

1. Breast Cancer Wisconsin dataset: [5]The Wisconsin Diagnostic Breast Cancer dataset consists of features computed from digitized images of fine needle aspirates (FNA) of breast masses, used to classify tumors as malignant or benign. It contains 569 instances (357 benign, 212 malignant) with 30 features derived from 10 cell nucleus characteristics such as radius, texture, perimeter, and symmetry.

2. HCV dataset: [6]This dataset contains the data of 615 patients for Hepatitis C. For preprocessing, category 0 and 0s (suspected) were classified as healthy, and all stages of HepC were classified as unhealthy - 540 healthy, 74 unhealthy.

Dataset cleaning was carried out by dropping all rows with NULL values. Datasets chosen haven't been normalized. Feature reduction has been carried out to determine the two most significant and uncorrelated features so as to facilitate plotting the data points and aid visualization. For the sake of dimensional and computational simplicity, we have reduced the number of features to two in all datasets.

Test-Train split: A common split ratio of 0.25 was used. The split was stratified to ensure that the proportion of classes in both train and test sets mirrored the original dataset's distribution.

The test set was kept unseen during the NSA detector generation and synthetic sample creation process to provide an unbiased evaluation of model generalization.

0.4 Algorithms and Models

1. NSA variants used: Real-valued NSA, VD-NSA, HD-NSA [7] [8] [9]

2. Traditional models used: Logistic regression classifier, Gaussian Naive Bayes, Random forest classifier, Multilayer Perceptron

3. Data augmentation benchmarks used: Random generation, Synthetic Minority Over-sampling Technique (SMOTE)

Initial weeks were spent researching and writing the optimized codes for each of the NSA variants. Scikit-learn libraries were used for traditional models, in addition to the classical Pandas, Numpy etc. libraries. Datasets were visualised using Matplotlib, given below.

Results

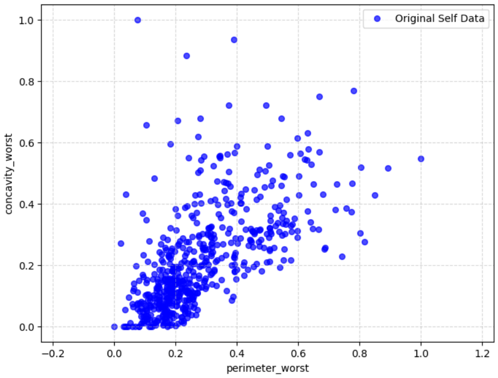

Fig 1. Original dataset, before augmentation

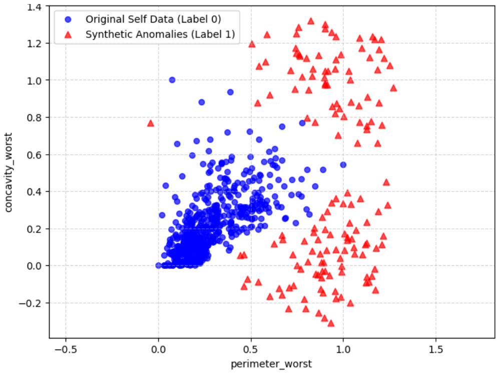

Fig 2. RNSA generated dataset

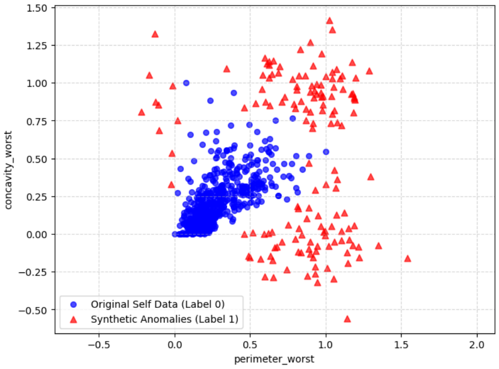

Fig 3. VD-NSA generated dataset

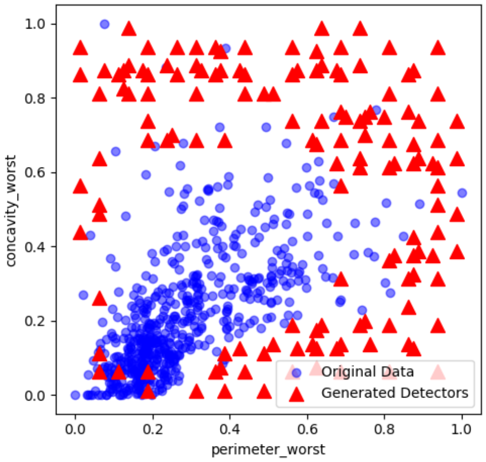

Fig 4. HD-NSA generated dataset

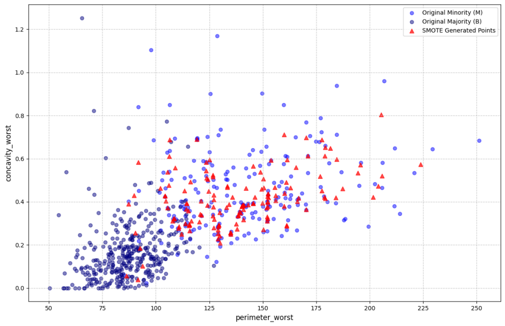

Fig 5. SMOTE generated dataset

Table 1. Comparison of accuracy and precision across techniques on Wisconsin dataset.

	Raw data	Random aug.	SMOTE	RNSA	VD-NSA	HD-NSA
LogisticRegression	A: 0.9441 P: 0.95	A: 0.9389 P: 0.94	A: 0.9777 P: 0.98	A: 0.9667 P: 0.97	A: 0.9722 P: 0.97	A: 0.8539 P: 0.86
GaussianNaiveBayes	A: 0.9510 P: 0.95	A: 0.9333 P: 0.94	A: 0.9609 P: 0.96	A: 0.9444 P: 0.95	A: 0.7944 P: 0.85	A: 0.9607 P: 0.96
RandomForest	A: 0.9021 P: 0.90	A: 0.9111 P: 0.91	A: 0.9609 P: 0.96	A: 0.9667 P: 0.97	A: 0.9667 P: 0.97	A: 0.9438 P: 0.94
MLP∗	A: 0.9510 P: 0.95	A: 0.9222 P: 0.93	A: 0.9888 P: 0.99	A: 0.9611 P: 0.96	A: 0.9667 P: 0.97	A: 0.9663 P: 0.97

*MLP: Multilayer Perceptron

Table 2. Comparison of accuracy and precision across techniques on HCV dataset.

	Raw data	Random aug.	SMOTE	RNSA	VD-NSA	HD-NSA
LogisticRegression	A: 0.9481 P: 0.93	A: 0.9686 P: 0.97	A: 0.9111 P: 0.91	A: 0.9593 P: 0.96	A: 0.9926 P: 0.99	A: 0.9177 P: 0.92
GaussianNaiveBayes	A: 0.9481 P: 0.93	A: 0.9738 P: 0.97	A: 0.8444 P: 0.86	A: 0.9815 P: 0.98	A: 0.9852 P: 0.98	A: 0.9367 P: 0.93
RandomForest	A: 0.9740 P: 0.94	A: 0.9791 P: 0.97	A: 0.9630 P: 0.96	A: 0.9778 P: 0.98	A: 0.9963 P: 0.99	A: 0.9684 P: 0.94
MLP	A: 0.9740 P: 0.94	A: 0.9843 P: 0.97	A: 0.9630 P: 0.96	A: 0.9778 P: 0.97	A: 0.9926 P: 0.99	A: 0.9747 P: 0.95

Discussion

Almost all of the NSA-augmented datasets achieved better results than raw datasets. Randomly augmented datasets performed unpredictably, mainly underperforming for the Wisconsin dataset and overperforming for the HCV dataset.

Among the three variants of NSA, VD-NSA consistently achieved the best scores. The NSA-augmented datasets dipped in performance for the Gaussian Naive Bayes method.

SMOTE achieved expected results on the first dataset, however performed poorly on the second dataset. Overall, VD-NSA was able to match or surpass the results of the SMOTE-augmented dataset.

Conclusion

This project successfully demonstrated that Negative Selection Algorithms (NSAs) can significantly enhance imbalanced classification by serving as an intelligent data augmentation technique. Our initial hypothesis—that such structured anomaly generation would improve model performance—was validated, with NSA-augmented datasets achieving results comparable, and at times superior, to established methods like SMOTE. Unlike potentially unpredictable random augmentation, NSA's immune-inspired mechanism of defining detectors in the 'non-self' space, based on 'self' data, offers a principled approach to creating synthetic anomalies that lie in regions genuinely distinct from the majority class, thereby avoiding the generation of overly simplistic or poorly positioned synthetic samples.

A key takeaway is that NSA-based augmentation shows particular strengths depending on the data. While methods like SMOTE work well when both classes form fairly clear, though uneven, groups, NSA shines in situations where the classes are defined differently. Specifically, NSA is very useful when the 'self' (majority) class is well-defined, but the 'non-self' (minority) class is more spread out or simply 'not self'—an 'X versus not-X' type of problem. In these cases, NSA's approach of intelligently filling in this 'non-self' area often provides a better way to balance the dataset.

Ultimately, this research contributes a formal benchmarking of various NSA variants (R-NSA, V-NSA, HD-NSA) for this novel augmentation role, quantifying their benefits and highlighting their applicability based on data distribution characteristics. While acknowledging the study's scope regarding datasets and parameter exploration, the findings affirm NSA's considerable value. Future work could explore hybrid strategies and further refine NSA-driven synthetic data generation, solidifying its place as a powerful and nuanced tool for addressing the pervasive challenge of class imbalance in machine learning.

Acknowledgments

This project was completed as a course project under CS6024 Algorithmic Approaches to Computational Biology, taken by Prof. Manikandan at the CSE department, IIT Madras. We would like to thank him, the teaching assistants and all the course participants for the valuable learnings gained over the duration.

We also extend our gratitude to the members of the Biotech club, IIT Madras for their support in the research conducted for the Negative Selection Algorithm.

Author Contribution

1. Ritisha Priya: Dataset collection and preprocessing, EDA, code for HD-NSA, preliminary report presentation.
2. Sujal Acham: NSA literature review and selection of variants, code for RNSA, VD-NSA, Results analysis and final report.

9. GitHub Link

You can find all the codes used for the project in this Github repository: `https://github.com/salcustium/hybrid-nsa`

10. Progress After Presentation

- Improvements to the algorithm codes
- Incorporated all suggestions from the professor and TAs, such as test-train split without augmented data, feature merging
- Fixed the visualisation graphs and normalised color coding
- Tested on more datasets
- Benchmarked against SMOTE technique, wrote the code and calculated results for all datasets

11. Similar Work in Other Course

This work is original to this course.

References

1. Philippe Rambaud RFARJTJB Adel Taleb. Binary Classification vs. Anomaly Detection on Imbalanced Tabular Medical Datasets. 2023 Congress in Computer Science, Computer Engineering, Applied Computing (CSCE). 2023;doi:https://doi.org/10.1109/CSCE60160.2023.00220.

2. et al SF. Self-nonself discrimination in a computer, in: IEEE Symposium on Research in Security and Privacy. Proceedings of 1994 IEEE Computer Society Symposium on Research in Security and Privacy. 1994;doi:https://doi.org/10.1109/RISP.1994.296580.

3. Kishor Datta Gupta DD. Negative Selection Algorithm Research and Applications in the Last Decade: A Review. IEEE Transactions on Artificial Intelligence. 2021;3(2). doi:https://doi.org/10.1109/TAI.2021.3114661.

4. F Gonzalez RK D Dasgupta. Combining negative selection and classification techniques for anomaly detection. Proceedings of the 2002 Congress on Evolutionary Computation CEC'02. 2002;doi:https://doi.org/10.1109/CEC.2002.1007012.

5. Wolberg MOSNSW W. Breast Cancer Wisconsin (Diagnostic) [Dataset]. 1993;doi:https://doi.org/10.24432/C5DW2B.

6. Lichtinghagen KFHG R. HCV data [Dataset]. 2020;doi:https://doi.org/10.24432/C5D612.

7. Fabio Gonz´alez DD, Ni˜no LF. A Randomized Real-Valued Negative Selection Algorithm. Lecture Notes in Computer Science, SPRINGER. 2003;doi:https://doi.org/10.1007/978-3-540-45192-1$_2$5.

8. Zhou Ji DD. Real-Valued Negative Selection Algorithm with Variable-Sized Detectors. ecture Notes in Computer Science, vol 3102 SPRINGER. 2004;doi:https://doi.org/10.1007/978-3-540-24854-5$_3$0.

9. Junjiang He BLXLZLYW Tao Li. An immune-based risk assessment method for digital virtual assets. Computers Security, Volume 102,. 2021;doi:https://doi.org/10.1016/j.cose.2020.102134.