International Journal of Multidisciplinary Engineering in Current Research - IJMEC
Volume 9, Issue 8, August-2024, http://ijmec.com/, ISSN: 2456-4265

BENCHMARK OF DATA PREPROCESSING METHODS FOR IMBALANCED CLASSIFICATION

Mohammad Samiyaan[1], Mohammed Arbaz Khan[2], Mohammed Shoaib Khan[3], Neha Hasan[4]

[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

nehahasan@lords.ac.in

ABSTRACT

Severe class imbalance poses significant challenges for machine learning in cybersecurity. Various preprocessing methods, including oversampling, undersampling, and hybrid approaches, have been developed to enhance the predictive performance of classifiers. However, a comprehensive and unbiased benchmark comparing these methods across diverse cybersecurity problems is lacking. This paper presents a benchmark of 16 preprocessing techniques evaluated on six cybersecurity datasets, alongside 17 public imbalanced datasets from other domains. We test these methods under multiple hyperparameter configurations and utilize an Auto ML system to reduce biases from specific hyperparameters or classifiers. Our evaluation focuses on performance measures that effectively reflect real-world applicability in cybersecurity Effective data preprocessing methods often improve classification performance. A baseline approach of no preprocessing outperformed many methods. 3) Oversampling techniques generally yield better results than under sampling and The standard SMOTE algorithm delivered the most significant performance gains, while more complex methods often provided only incremental improvements with reduced computational efficiency.

I. INTRODUCTION

Class imbalance is a prominent challenge in the application of machine learning to cybersecurity, where the distribution of classes is often heavily skewed. This imbalance can lead to poor predictive performance, particularly for minority classes that represent critical events such as attacks or intrusions. To address this issue, various dataset preprocessing techniques have been proposed, including oversampling, under sampling, and hybrid methods that aim to improve the training dataset's balance and, consequently, the classifiers' effectiveness. Despite the availability of these techniques, there remains a lack of comprehensive benchmarks that assess their performance across a wide range of cybersecurity problems. This gap hinders practitioners from making informed decisions about which preprocessing methods to employ. In this paper, we present an extensive benchmark of 16 preprocessing methods, evaluated on six distinct cybersecurity datasets as well as 17 public imbalanced datasets from other domains. Our approach includes rigorous testing under multiple hyperparameter configurations and utilizes an Auto ML system to mitigate biases stemming from specific hyperparameters or classifier choices. Additionally, we emphasize the importance of using appropriate performance metrics that reflect the practical effectiveness of classifiers in real-world

DOI: https://doi-ds.org/doilink/08.2024-13729438

cybersecurity scenarios. Through our analysis, we aim to provide valuable insights into the efficacy of different preprocessing methods, ultimately contributing to improved machine learning practices in cybersecurity.

II. LITERATURE REVIEW

1) He, H., & Garcia, E. A. (2009)Title: Learning from Imbalanced DataSummary: This seminal paper provides a comprehensive review of methods to handle imbalanced datasets, including sampling methods, algorithmic modifications, and performance metrics.

2) Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002)Title: SMOTE: Synthetic Minority Over-sampling Technique Summary: Introduces SMOTE, a widely-used technique that generates synthetic samples to balance class distribution, significantly impacting subsequent research in imbalanced classification.

3) Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004)Title: A Study of the Behavior of Several Methods for Balancing machine Learning Training Data Summary: Evaluates various data preprocessing methods, including over-sampling, under-sampling, and combined approaches, providing a benchmark for future research.

4) Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018)

Title: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary Summary: Reviews the progress of SMOTE and its variants, discussing challenges and future directions in imbalanced data learning.

III.SYSTEM ANALYSIS

One of the most direct and effective approaches to keep the current customers is that the company should be able to foresee potential churn in time and react to it quickly. Recognizing the indications of potential churn; satisfying customer needs, restoring and re-establishing loyalty are actions supposed to help the organization minimize the costs of gaining new customers. A big problem that encounters businesses, especially telecommunications business is 'customer churn'; this occurs when a customer decides to leave a company's landline business for another cable competitor. Therefore, our existing system beyond this study to build a model that will predict churn customer through defining the customer's precise behaviors and attributes. We will use data mining techniques such as clustering, classification and association rule. Disadvantages of existing system:

- There is no standardized approach for handling class imbalance. Techniques like oversampling and undersampling are often used in ad hoc ways, with little guidance on which is most suitable.
- Baseline techniques such as random oversampling or no preprocessing are predominantly used, despite the availability of more advanced methods.
- There is a lack of extensive benchmarking, with only limited empirical comparisons available to objectively evaluate different methods for handling imbalances.
- Current systems perform poorly on minority classes, struggling to identify rare but important cases like threats. While overall performance measures may appear decent, the performance on the minority class remains inadequate.

DOI: https://doi-ds.org/doilink/08.2024-13729438

- As a result of these issues, systems are likely underperforming in real-world tasks, not reaching their full potential due to suboptimal handling of class imbalance.

Proposed system:

The proposed system aims to enhance class imbalance handling in cybersecurity by extensively benchmarking advanced preprocessing techniques. By objectively evaluating methods like SMOTE on diverse real-world datasets, the system will provide definitive guidance on best practices. Sophisticated oversampling approaches, proven to improve minority class identification, will be utilized. With unbiased empirical evidence, these techniques can become standardized, replacing basic under sampling or no preprocessing methods. The system will optimize and leverage unique advantages of methods like Borderline-SMOTE and ADASYN through rigorous testing. This is expected to significantly improve minority class performance while maintaining overall accuracy. Consequently, operational effectiveness in critical tasks like threat detection will be enhanced. Additionally, the developed benchmarking framework will support ongoing advancements as new techniques emerge. By modernizing class imbalance handling, the system aims to elevate cybersecurity systems to the state-of-the-art level seen in other machine learning domains. The outcomes will also serve as a template for imbalanced learning in other applications, such as fraud detection. In summary, through principled benchmarking and adoption of advanced techniques, the proposed system promises significant progress in addressing a long-standing deficiency. Advantages of proposed system:

1. Enhanced minority class performance: Oversampling techniques can improve the detection of rare yet crucial cases such as threats and fraud.

2. Increased real-world effectiveness: Improved handling of class imbalances directly leads to better performance in operational tasks.

3. Standardized best practices: Benchmarking offers guidance on the most effective preprocessing techniques for various scenarios.

4. Utilizes advanced algorithms: Sophisticated methods like SMOTE and Borderline-SMOTE can surpass basic approaches.

5. Ongoing development: The benchmarking framework allows for the evaluation of new techniques as they are introduced.

6. Broader applicability: The methodology can be a model for imbalanced learning in other fields, such as healthcare.

IV. RELATED WORK

Over the years, many data preprocessing methods suitable for class-imbalanced learning have been published, but in comparison, only a relatively small number of benchmarks encompassing an extensive range of both methods and datasets exist. Typically, every publication introducing a new method includes experimental evaluation, but the scope of these experiments tends to be small contains experiments datasets and compares the method and plain decision tree baseline. With that said, there already exist publications that focus mainly on comparing preprocessing methods, but usually, they tend to focus only on oversampling methods. Most of these studies are also performed on a relatively small number of datasets. which is very extensive both in

DOI: https://doi-ds.org/doilink/08.2024-13729438

terms of methods compared and datasets used. However, it focuses only on oversampling methods and also does not contain experiments in the cybersecurity domain. Additionally, none of the studies above performs as broad a search in hyperparameters and successive classifier models as we do.

In the cybersecurity domain compared several preprocessing methods on the dataset.

V METHODOLGY

1. Problem Definition and Dataset Selection Define the Problem: Clearly articulate the challenges of severe class imbalance in cybersecurity and the need for effective preprocessing methods.

Dataset Selection: Choose six cybersecurity datasets that represent diverse problems and characteristics. Include 17 public imbalanced datasets from various domains for broader comparison.

2. Preprocessing Techniques Selection

Identify Techniques: Select 16 preprocessing techniques including oversampling (e.g., SMOTE variants), undersampling, and hybrid approaches. Ensure to cover a wide range of methods to capture different strategies and complexities.

3. Experimental Setup

Hyperparameter Configuration: Define multiple configurations for each preprocessing technique to explore their robustness and sensitivity.

Auto ML Integration: Incorporate an Auto ML system to automate model selection and reduce bias introduced by specific hyperparameters or classifiers.

4. Evaluation MetricsPerformance Measures: Focus on metrics that reflect real-world applicability in cybersecurity, such as Precision, Recall, F1-score, ROC-AUC.

Statistical Rigor: Conduct statistical tests to compare the performance of preprocessing methods across datasets and hyperparameter configurations.

5. Implementation

Baseline Comparison: Establish a baseline using the dataset without any preprocessing to assess the impact of preprocessing techniques.

Implementation of Preprocessing: Apply each technique to the datasets using the defined hyperparameter configurations.

6. Model Training and Evaluation

Cross-validation: Use stratified cross-validation to ensure robust evaluation of each preprocessing technique. Performance Evaluation: Train and evaluate machine learning models (e.g., SVM, Random Forest) on preprocessed datasets using chosen metrics.

7. Analysis and Interpretation

Comparison of Results: Analyze and compare the performance of each preprocessing technique across cybersecurity datasets and other domains.

Identify Effective Techniques: Identify preprocessing methods that consistently improve classification performance, considering computational efficiency and incremental gains.

8. Reporting and Discussion

Document Findings: Present results comprehensively, including tables, figures, and statistical analyses.

Discussion: Discuss the implications offindings for improving imbalanced classification in cybersecurity and compare with existing literature.

VI. CONCLUSION

We conducted a novel study evaluating 16 preprocessing methods across 23 datasets, including six from the cybersecurity domain. We examined

DOI: https://doi-ds.org/doilink/08.2024-13729438

both predictive and computational performance by implementing a large-scale experiment that employs AutoML to consider a wide range of classifiers and includes a hyperparameter search to eliminate potential biases present in previous benchmarks. Our main findings indicate that dataset preprocessing is often beneficial when dealing with class-imbalanced classification. However, many methods fail to consistently outperform the baseline solution of doing nothing. Generally, oversampling methods outperform undersampling methods, although there are exceptions. Among the oversampling techniques, the traditional algorithm shows the most significant performance gains, while its more advanced variants tend to offer only incremental improvements. When focusing our analysis on the cybersecurity datasets, which cover multiple cybersecurity domains, we reached the same conclusions. It is important to note that the ranking of methods is influenced by the chosen performance measure. We included multiple performance measures that are comprehensive and applicable in practical classification scenarios involving class imbalance. While the specifics of the rankings vary by measure, the main takeaways remain consistent.

VII. REFERENCES

[1] Mostofa Ahsan, Rahul Gomes, and Anne Denton. Implementing SMOTE on phishing data to enhance cybersecurity. In the 2018 IEEE International Conference on Electro/Information Technology (EIT), pages 0531–0536. IEEE, 2018.

[2] Bathini Sai Akash, Pavan Kumar Reddy Yannam, Bokkasam Venkata Sai Ruthvik, Lov Kumar, Lalita Bhanu Murthy, and Aneesh Krishna. Predicting cyber-attacks on IoT networks using deep learning and various SMOTE variants. In the International Conference on Advanced Information Networking and Applications, pages 243–255. Springer, 2022.

[3] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah, and Amir Hussain. Comparing oversampling techniques to address class imbalance: A customer churn prediction case study. IEEE Access, 4:7940–7957, 2016.

[4] Hyrum S. Anderson and Phil Roth. Ember: An open dataset for training static PE malware machine learning models. arXiv preprint arXiv:1804.04637, 2018.

[5] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In the 31st USENIX Security Symposium (USENIX Security 22), pages 3971–3988, Boston, MA, August 2022. USENIX Association.

[6] Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. 54(5), May 2021.

[7] Stefan Axelsson. The base-rate fallacy and the challenge of intrusion detection. ACM Transactions on Information and System Security (TISSEC), 3(3):186–205, 2000.

[8] Salahuddin Azad, Syeda Salma Naqvi, Fariza Sabrina, Shaleeza Sohail, and Sweta Thakur. IoT cybersecurity: Using machine learning approaches for unbalanced datasets. In the 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pages 1–6. IEEE, 2021.

DOI: https://doi-ds.org/doilink/08.2024-13729438

[9] Sikha Bagui and Kunqi Li. Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data, 8(1):1–41, 2021.

[10] Ricardo Barandela, Rosa M. Valdovinos, J. Salvador Sanchez, and Francesc J. Ferri. The imbalanced training sample problem: Under or oversampling? In Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR), pages 806–814. Springer, 2004.

[11] Jan Brabec, Tomáš Komárek, Vojtěch Franc, and Lukáš Machlica. On model evaluation under non-constant class imbalance. In International Conference on Computational Science, pages 74–87. Springer, 2020.

[12] Jan Brabec and Lukáš Machlica. Bad practices in evaluation methodology relevant to class-imbalanced problems. arXiv preprint arXiv:1812.01388, 2018.

[13] Jan Brabec and Lukáš Machlica. Decision-forest voting scheme for rare class classification in network intrusion detection. In the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 3325–3330, Oct 2018.

[14] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with applications to active user modeling and hierarchical reinforcement learning. CoRR, abs/1012.2599, 2010.

[15] Nitesh V. Chawla. Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook, pages 875–886. Springer, 2009.

[16] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357, 2002.

[17] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In the 2015 IEEE Symposium Series on Computational Intelligence, pages 159–166. IEEE, 2015.

[18] Anusha Damodaran, Fabio Di Troia, Corrado Aaron Visaggio, Thomas H. Austin, and Mark Stamp. A comparison of static, dynamic, and hybrid analysis for malware detection. Journal of Computer Virology and Hacking Techniques, 13(1):1–12, 2017.

[19] Janez Demsar. Statistical comparisons of classifiers over multiple datasets. The Journal of Machine Learning Research, 7:1–30, 2006.

[20] Stepan Dvorak, Pavel Prochazka, and Lukas Bajer. GNN-based malicious network entities identification in large-scale network data. In NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, pages 1–4, 2022.

DOI: https://doi-ds.org/doilink/08.2024-13729438