POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

ADBench: a Novel Benchmark for Streaming Anomaly Detection Algorithms

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Ledio Sheshori

Advisor: Prof. Emanuele Della Valle

Co-advisor: Alessio Bernardo, Giacomo Ziffer

Academic year: 2022-2023

1. Introduction

The digital revolution has led to an explosion of real-time streaming data from diverse sources. This poses significant challenges for traditional machine learning approaches, which struggle to handle the continuous influx of information and its evolving nature. In response to this challenge, streaming machine learning has emerged as a solution capable of adapting to evolving data streams in real time, ensuring that models remain updated. One area of particular importance in streaming data analysis is anomaly detection, which aims to identify patterns deviating from expected behavior. This research area provides valuable insights across domains such as fraud analytics, network monitoring, and industrial equipment surveillance.

This work focuses on the intersection of streaming machine learning and anomaly detection, conducting an extensive analysis of state-of-the-art algorithms in an unsupervised setting. However, a core challenge lies in the lack of standardized benchmarking environments to fairly compare streaming anomaly detection techniques.

This thesis addressed the gap by providing a comparative analysis of the main streaming anomaly detection algorithm and developing ADBench[1], a benchmark tailored for evaluating streaming anomaly detection algorithms. ADBench allows rigorous and reproducible evaluation using real-world and synthetic datasets reflecting diverse scenarios. It provides an automated framework for the systematic assessment of these models exploring various dimensions, including dataset selection, evaluation metric suitability for unbalanced datasets, data visualization techniques, and robust statistical testing. In conclusion, ADBench produces diverse results, enabling fair and clear comparisons of streaming anomaly detection models from multiple perspectives.

2. State of the Art

Streaming machine learning [1] enables real-time analysis by processing data incrementally as it arrives in a continuous stream. Algorithms must meet requirements such as limited memory, constant processing time per sample, and adaptable models. A key challenge in this scenario is concept drift, where the data distribution shifts over time. To address concept drift, model updates are essential to ensure accurate predic-

[1] https://github.com/ledio7/ADBench

tions. Techniques such as sliding windows are often employed, which involve discarding older data while continuously training the model with new data.

Exploring further, streaming anomaly detection [2] is a crucial aspect of data analysis, as it involves identifying abnormal patterns in real-time streaming data. However, the dynamic nature of streams poses challenges, especially in cases such as fraud detection and network security, arising from the frequent absence of complete labelling. Consequently, the prevailing approach often leans towards unsupervised methods to tackle these issues effectively. Given the vast scope of this field, introducing a taxonomy of streaming anomaly detection methods can provide a comprehensive understanding of the domain.

The interesting taxonomy provided by [3] identifies four big categories: statistics-based, clustering-based, nearest-neighbor-based, and isolation-based. Statistics-based approaches construct probabilistic models representing normal data behavior. New samples deviating significantly from this model are flagged as anomalies. Clustering-based approaches rely on the proximity of observations within the data set. These methods can be categorized as either distance-based or density-based, where they divide the data into clusters based on the similarity between observations. In the context of anomaly detection, the cluster that exhibits the greatest distance or the smallest density from the rest can be identified as an anomaly cluster. Nearest-neighbor approaches rely on proximity among observations and can fall into distance-based or density-based categories. These methods detect anomalies by assessing the distances between an observation and all others in the dataset. An observation is flagged as an anomaly if it is significantly distant from its k nearest neighbors. These last two categories demand substantial computational resources, leading to higher resource consumption. Isolation-based algorithms efficiently isolate anomalies using tree structures and path lengths. They have lower complexity than distance/density techniques. In addition to the four established categories, there is an emerging development in streaming anomaly detection involving neural network-based approaches. These emerging neural network methods leverage the power of deep learning to model normal data patterns and behaviors. They are able to automatically extract relevant features and identify complex relationships within data. Autoencoders, for example, learn to compress data into a lower-dimensional representation and then reconstruct it. In the context of anomaly detection, they identify anomalies by generating higher reconstruction errors for abnormal instances. In conclusion, the dynamic nature of this field is characterized by continuous innovation, which means that categorizations will consistently require updates and adaptations to remain relevant and effective.

3. Problem Statement

A core challenge in streaming anomaly detection is the lack of standardized environments for fair algorithm comparison. Research papers often evaluate models using distinct datasets and metrics tailored to their focus, making direct comparisons difficult. This non-uniform approach provides an incomplete picture of relative strengths and weaknesses. A principled benchmark is needed to test techniques under identical conditions. Given the expansive nature of this research field, this thesis employs the following three key research questions as guiding pillars, providing a clear direction for the research to address the identified gap:

- RQ1: Is the impartial assessment of diverse anomaly detection algorithms feasible?
- RQ2: Can the evaluation process be automated for efficiency?
- RQ3: How can reproducible experiments be guaranteed?

According to [4], to be effective, a benchmark must meet several requirements:

- **Relevance**: Measure features relevant to the problem.
- **Representativeness**: Metrics broadly accepted by industry/academia.
- **Equity**: Fair comparison of all systems.
- **Repeatability**: Benchmark results should be verifiable.
- **Cost-effectiveness:** Not excessively resource intensive.
- **Scalability**: Applicable to systems of all sizes.
- **Transparency**: Easily understandable re-

sults.

While these are the core requirements, constructing a meaningful benchmark necessitates additional considerations. An effective benchmark needs to include careful data selection encompassing diverse scenarios, metrics tailored to the imbalanced nature of anomalies, visualization techniques that enhance understanding, and statistical tests to validate results. Synthetic datasets that emulate real-world conditions are crucial for controlled experimentation. The following sections will detail the approach taken to address these challenges by systematically breaking down the problem into distinct dimensions and proposing solutions for each aspect. In summary, ADBench provides a benchmarking framework enabling fair, consistent, and reproducible performance evaluation of streaming anomaly detection techniques. The insights derived will help select optimal algorithms across various streaming applications.

4. Comparative Analysis

Initially, a comprehensive comparative analysis of prominent algorithms used in streaming anomaly detection was conducted. This analysis involved a detailed examination of relevant research papers discussing the algorithms. It provided a structured assessment highlighting the strengths and limitations of each technique.

Several insights emerged regarding the algorithms' capabilities. For example, Half-Space Trees (HST) efficiently adapts to data changes through fast tree updates. Robust Random Cut Forest (RRCF) handles high-dimensional data well. Isolation Forest in Streaming (iForestASD) is slower but more precise than HST. However, the analysis also evidenced once again the challenge of comparing algorithms impartially. The techniques were often tested on different datasets and metrics based on each paper's focus. This non-uniform approach provides an incomplete picture.

Equipped with these new understandings, it is now possible to show a wider overview of streaming anomaly detection methods by incorporating promising new categories of models. The result can be seen in Figure 1: this offers a comprehensive representation of categories.

The comparative analysis established a foundation of knowledge regarding the landscape of streaming anomaly detection algorithms. This provided guidance for the subsequent development of the benchmark by identifying promising solutions and their capabilities.

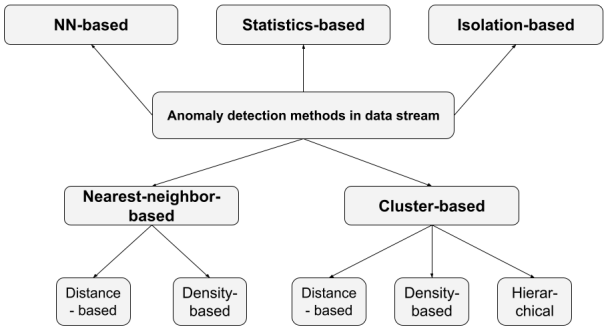


Figure 1: Updated classification of unsupervised streaming anomaly detection categories.

5. ADBench

ADBench provides researchers with a tool to impartially assess techniques on diverse datasets using standardized metrics. It automates evaluation workflows for efficiency while ensuring reproducibility. It is possible to extend it by adding new algorithms to test.

Constructing the ADBench framework required thoughtful design choices across multiple dimensions. Several key considerations were made to develop a comprehensive methodology for evaluating streaming anomaly detection algorithms. The selection of appropriate datasets is crucial, as they provide the raw materials for analysis. Both real-world and synthetic datasets have been incorporated to evaluate algorithms under realistic and controlled conditions respectively. Real datasets such as Creditcard and Covertype emulate practical scenarios and data distributions. Meanwhile, synthetic data streams induced using generators enable precise replication of desired conditions. Factors such as concept drift, noise, imbalanced classes, and high dimensionality were induced to span the spectrum of streaming challenges. In total, 20 real and 30 synthetic datasets were included to provide diverse, representative data.

Suitable metrics are crucial to effectively quantify performance in a way that captures the nuances of anomaly detection. In total, 7 key metrics were utilized, providing multifaceted insights aligned to the specifics of streaming anomaly detection. PR-AUC and ROC-AUC

are widely used metrics in anomaly detection scenarios. PR-AUC assesses the trade-off between precision and recall, crucial when anomalies are rare but critical to detect accurately. ROC-AUC evaluates the model's ability to distinguish between positive and negative instances across thresholds. However, ROC-AUC can be misleading when anomalies are scarce, while PR-AUC remains sensitive. Metrics such as F1-score and Recall provide insights into the balance between minimizing false positives and maximizing true positives. Recall specifically measures the model's sensitivity in identifying actual anomalies. The F1-score considers both precision and recall, making it significant for anomaly detection. The Geometric Mean combines recall and specificity, measuring the model's effectiveness in classifying both anomalies and normal instances correctly. Runtime and RAM-Hours capture model efficiency in terms of processing time and memory utilization.

The choice of visualization techniques aimed to facilitate intuitive interpretation of results and enable easy comparison of model performances. Heatmaps were used to instantly show model performance across multiple datasets and metrics through color-coded matrices. The color gradients allowed rapid identification of patterns. While PR and ROC curves provide a visual trade-off analysis between true positives and false positives for each model. Plotting all models on one graph enables effortless comparison. Bar plots, instead, offer granular insights into model performance for specific metric-dataset pairs. Their simplicity supported precise analysis of individual scenarios. Together, these visualizations enable multidimensional understanding, from high-level overviews using heatmaps to detailed investigations with bar plots. Effective visuals bridge metrics to actionable understanding, fulfilling a key goal of interpretable benchmarking.

The Nemenyi test was chosen as the statistical test for performance evaluation. It is specifically designed for comparing multiple machine learning models across multiple datasets, making it well-suited for this benchmarking scenario. This test enables credible conclusions by quantifying the certainty of observed differences. It prevents drawing conclusions based solely on coincidental variations.

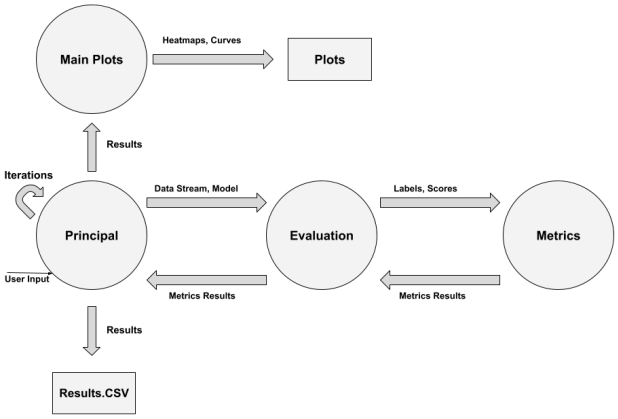


Figure 2: Block Diagram of the benchmark core architecture.

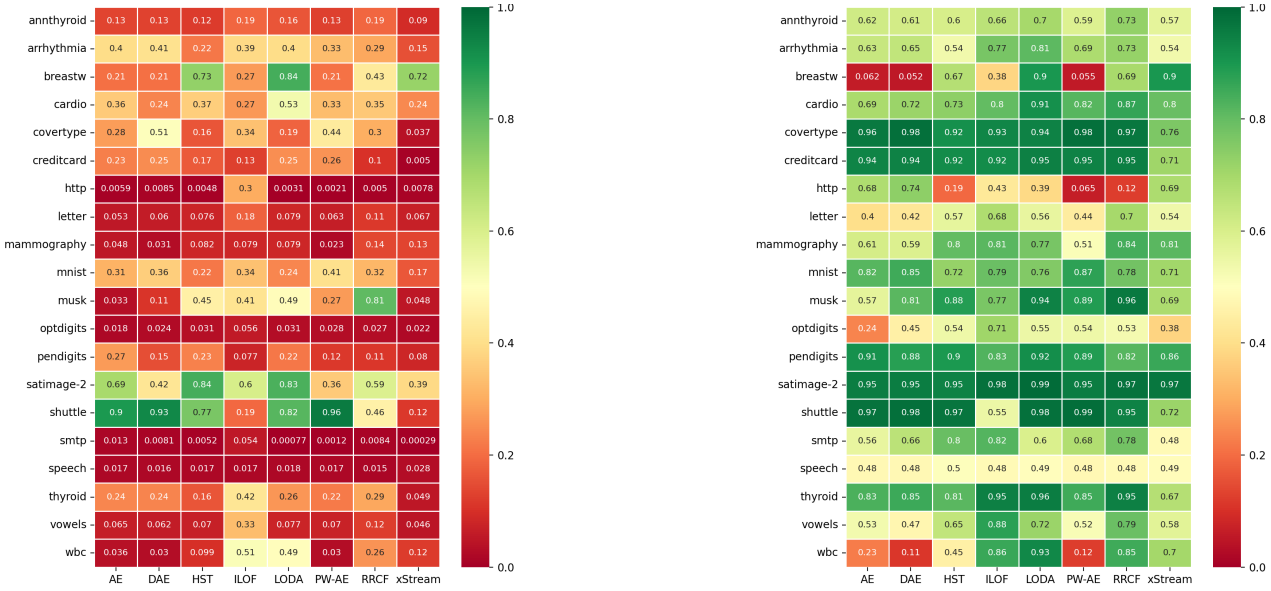
5.1. Architecture

As regards the benchmark architecture, it comprises modular components for data streaming, model execution, metric calculation, and result aggregation.

The principal workflow, shown in Figure 2, involves a main module that iterates through each dataset and model, using seeds to ensure reproducible random processes, reaching the goal outlined in RQ3. It distributes computation for efficient and scalable parallel execution, invoking the evaluation module. This second module, for each iteration, executes models on datasets via prequential evaluation, emulating real-time streaming. Computed scores and labels are passed to the metrics module. This last module calculates performance metrics aligned to streaming challenges. The computed results return to the principal module in order to be stored and used for further analysis. Subsequently, the main plots module is invoked to generate key visualizations including heatmaps, ROC curves, and PR curves for comprehensive analysis.

Additional modules enable customized bar plot visualizations for more detailed results and the Nemenyi test plots for significance evaluation. The coordinated execution started by the principal module guarantees full automation, fulfilling RQ2 and ensuring a completely hands-free workflow from start to finish without any manual intervention.

Overall, this modular architecture fulfills automation, reproducibility, scalability, and other essential elements of effective benchmarking.



(a) *PR-AUC*. (b) *ROC-AUC*.

Figure 3: Heatmaps for *PR-AUC* and *ROC-AUC* metrics.

It provides a robust framework for impartial streaming anomaly detection evaluation.

6. Results

The results of the benchmarking process revealed several interesting findings. In general, looking at the heatmaps, such as the one in Figure 3, across the same dataset, the performance of the algorithms tends to hover around similar levels, with only a few exceptions. Some PR-AUC values are lower than others: datasets with a low percentage of anomalies, such as HTTP and SMTP, posed notable challenges as reflected by their lower PR-AUC and ROC-AUC values, indicating the algorithms' struggle to distinguish anomalies from normal instances for these imbalanced datasets. It is important to note that all algorithms were run with default parameters and not extensively tuned. The Autoencoders did not consistently exhibit strong PR-AUC results across all datasets, particularly struggling on new real-world and synthetic datasets introduced in this benchmark. This suggests claims of their dominance could be context-specific.

In terms of ROC-AUC, there are generally robust values shown. For metrics such as F1-score and recall, values tended to be higher for normal classes and lower for anomalous classes, a typical pattern in anomaly detection tasks with imbalanced datasets: this highlights the inherent difficulty in detecting anomalies when they constitute a small proportion of the overall dataset. When examining Runtime and RAM-Hours, HST and RRCF easily emerged as the most resource-intensive algorithms according to the heatmaps.

PR and ROC curves showed no single algorithm consistently outperforms others across all datasets. For datasets such as Creditcard and Covertype, most models demonstrated high effectiveness, indicated by the areas under their curves, as can be seen in Figure 4; this figure serves as an illustrative example of the ROC curves results on a specific dataset. Instead, datasets with fewer anomalies, like HTTP and SMTP, exhibited lower performance with less smooth curves, implying that classification was more challenging. This underscored the absence of a single, universally top-performing algorithm across all scenarios.

Furthermore, statistical tests of the results also do not identify an algorithm or group of algorithms with statistically significant better performance based on the Nemenyi plots. This aligns with the "No Free Lunch" theorem [5], which asserts no algorithm universally outperforms in all cases. Overall, results emphasized the importance of considering dataset properties and goals to inform algorithm selection for a given scenario or domain. As a result, RQ1 was

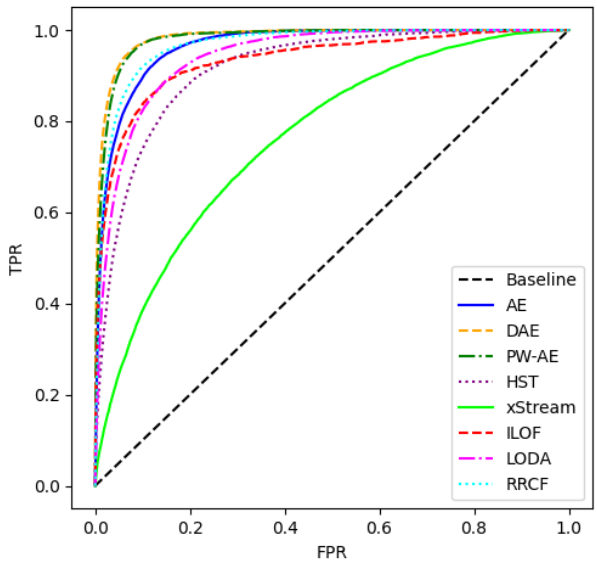


Figure 4: *ROC* curves for dataset Covertype.

fulfilled by conducting a rigorous and fair evaluation of diverse anomaly detection algorithms under this benchmarking methodology.

7. Conclusions and Future Work

This thesis presented a comprehensive study of streaming anomaly detection algorithms and their evaluation. It began by outlining the problem domain and the lack of fair comparison methods, introducing three research questions to guide the investigation. To address this, a set of algorithms was introduced, followed by a comprehensive comparative analysis. A systematic approach was then taken to break down the problem and understand related facets like evaluation metrics, data visualization, and statistical testing.

Based on these explorations, a benchmarking framework was developed meeting key requirements of relevance, automation (RQ2), reproducibility (RQ3), and flexibility. Extensive experiments evaluated algorithms across diverse datasets using varied metrics. Results were communicated through insightful visualizations and statistical analyses.

No single algorithm emerged as a definitive leader across all cases, aligning with the "No Free Lunch" theorem. However, certain algorithms maintained relatively robust rankings, indicating potential for domain-specific optimization. Overall, the research effectively demonstrated that impartial, reproducible evaluation of streaming anomaly detection algorithms is achievable through a meticulous benchmarking approach, fulfilling the most relevant research question, RQ1. The insights derived help select suitable techniques for given scenarios. By fulfilling the outlined research questions, outcomes offered substantial value for researchers selecting suitable techniques and advancing the field overall.

Future work involves extending benchmark capabilities with more datasets, models, and functionalities to propel progress in this important domain. Additionally, the benchmark could be extended to perform parameter tuning or facilitate model selection.

References

[1] Albert Bifet, Ricard Gavalda, Geoffrey Holmes, and Bernhard Pfahringer. *Machine learning for data streams: with practical examples in MOA*. MIT press, 2023.

[2] Lucas Cazzonelli and Cedric Kulbach. Detecting anomalies with autoencoders on data streams. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 258–274. Springer, 2022.

[3] Maurras Ulbricht Togbe, Mariam Barry, Aliou Boly, Yousra Chabchoub, Raja Chiky, Jacob Montiel, and Vinh-Thuy Tran. Anomaly detection for data streams based on isolation forest using scikit-multiflow. In *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20*, pages 15–30. Springer, 2020.

[4] Wei Dai and Daniel Berleant. Benchmarking contemporary deep learning hardware and frameworks: A survey of qualitative metrics. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, pages 148–155. IEEE, 2019.

[5] David Gómez and Alfonso Rojas. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. *Neural computation*, 28(1): 216–228, 2016.