Poster: Fixing the Foundations: Towards Generalization for Machine Learning Intrusion Detection Systems

Miel Verkerken*, Laurens D'hooge*, Tim Wauters*, Bruno Volckaert* and Filip De Turck*
*Department of Information Technology, IDLab, Ghent University - imec, Belgium
Email: {miel.verkerken, laurens.dhooge, tim.wauters, bruno.volckaert, filip.deturck}@ugent.be

Abstract

Through the ongoing digitization of the world, the number of connected devices is continuously growing without any foreseen decline in the near future. In particular, these devices increasingly include critical systems such as power grids and medical institutions, possibly causing tremendous consequences in the case of a successful cybersecurity attack. A network intrusion detection system (NIDS) is one of the main components to detect ongoing attacks by differentiating normal from malicious traffic. Anomaly-based NIDS, more specifically unsupervised methods previously proved promising for their ability to detect known as well as zero-day attacks without the need for a labeled dataset. Despite decades of development by researchers, anomaly-based NIDS are only rarely employed in real-world applications, most possibly due to the lack of generalization power of the proposed models. This article first evaluates four unsupervised machine learning methods on two recent datasets and then defines their generalization strength using a novel inter-dataset evaluation strategy estimating their adaptability. Results show that all models can present high classification scores on an individual dataset but fail to directly transfer those to a second unseen but related dataset. Specifically, the accuracy dropped on average 25.63% in an inter-dataset setting compared to the conventional evaluation approach. This generalization challenge can be observed and tackled in future research with the help of the proposed evaluation strategy in this paper.

I. Main Content

This research [1] is recently published in the special issue on "Cybersecurity Management in the Era of AI" of the Journal of Network and Systems Management (TNSM) with DOI: https://doi.org/10.1007/s10922-021-09615-7.

References

[1] M. Verkerken, L. D'hooge, T. Wauters, B. Volckaert, and F. De Turck, "Towards model generalization for intrusion detection: Unsupervised machine learning techniques," *Journal of Network and Systems Management*, vol. 30, no. 1, p. 12, Oct 2021.

FACULTY OF ENGINEERING AND ARCHITECTURE

IDLab
INTERNET & DATA LAB

DEPARTMENT OF INFORMATION TECHNOLOGY

Miel Verkerken, Laurens D'hooge, Tim Wauters, Bruno Volckaert, Filip De Turck

POSTER: FIXING THE FOUNDATIONS

Towards Generalization for Machine Learning Intrusion Detection Systems

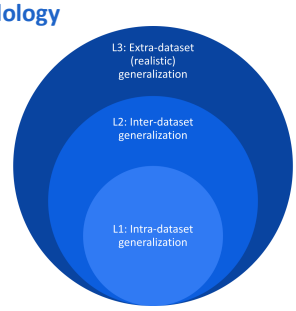
Introduction

Traditionally, network intrusion detection systems (NIDS) have relied on pattern matching against a database with signatures of known attacks. Over the last decade, security researchers have increasingly applied machine learning techniques to NIDS with outstanding results on academic benchmark datasts. However, the lack of generalizability of these results is preventing a breakthrough of machine learning based intrusion detection systems (ML-IDS). Before researchers continue to propose novel approaches, the foundations need to be fixed, starting with the academic benchmark datasets themselves.

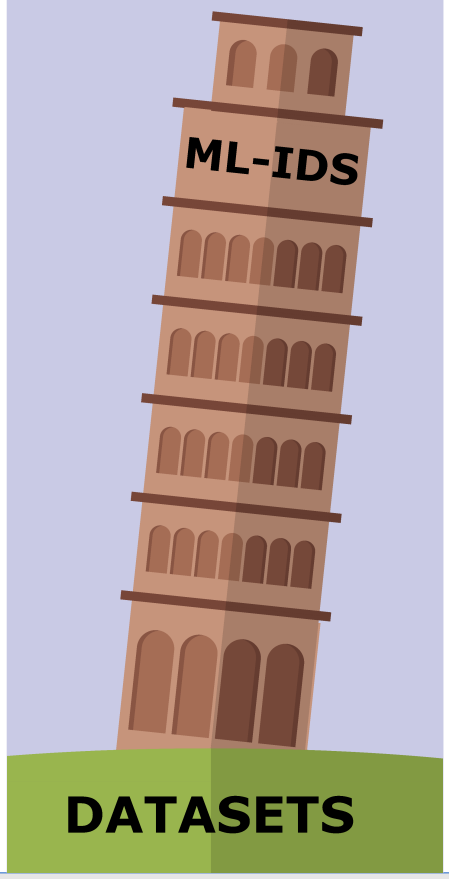
Contributions:

- Create awareness of the contrast between optimistic peer-reviewed publications and adoption of ML solutions in the real-world.
- Provide a novel strategy to evaluate the generalization strength of proposed solutions trained on academic datasets.
- Identifying common caveats for improved generalization when using academic benchmark datasets for evaluation of ML-IDS.

Methodology



The generalization strength of machine learning approaches for intrusion detection systems (ML-IDS) is estimated using our proposed inter-dataset evaluation strategy. This brings us one step closer to real-world testing. Models that generalize well are **expected** to only have a **minimal drop** in performance **between intra- and inter-dataset evaluations**.



Results

- **AUROC** decreased on average 30%
- **Accuracy** decreased on average 25%
- Best model was **OC-SVM** with only 26% decrease
- Both **supervised** and **unsupervised** models are affected
- ML models preform only slightly better than a random classifier
- Significant **disparities in exposure** between attack classes

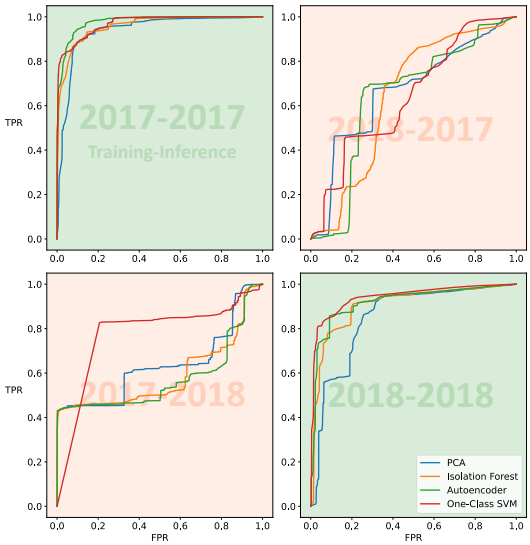


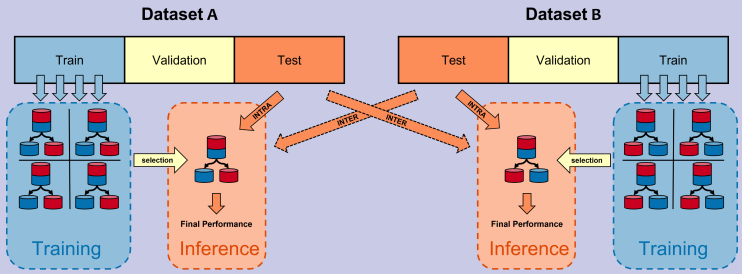
Fig. The receiver operating characteristics (ROC) curves are plotted for inter- and inter-dataset evaluation on two similar dataset with compatible features.

Evaluation strategies to estimate the generalization strength

Default: Studies assume that their results can be generalized due to the i.i.d. Property. Generally, the published results are obtained using a hold-out test set or K-fold cross-validation with samples from a single dataset.

Improved: An new evaluation strategy is proposed that trains and tests the ML models on distinct datasets with compatible features sets, such as CIC-IDS-2017 and CSE-CIC-IDS-2018.

Desired: An ideal scenario involves evaluating the trained models on real-world data that has been gathered. However, obtaining this data can be challenging due to anonymization and privacy issues.



Conclusions:

- **Do not blindly assume the generalizability** of your ML models. Evaluate them on similar datasets or real-world data.
- Proposing novel, "improved" securtiy solutions that are evaluated on existing academic ML-IDS datasets have little real-world impact due to **lack of generalizability**.

Next steps:

- Development of ContainerCap. A framework to execute, capture and label real cyber attacks in a controlled, containerized environment for **dataset generation**.
- Reduce overfitting through an **improved data preprocessing** approach, such as more sophisticated feature engineering techniques and removal of contaminated features.

Contact

Miel.Verkerken@ugent.be
www.ugent.be/ea/idlab/

Universiteit Gent
@ugent
Ghent University

Download paper here

GHENT UNIVERSITY

imec