Anomaly Detection in Data Network Traffic with Machine Learning

Master Thesis

Thomas Buck

Darmstadt University of Applied Sciences

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Introduction

We live in a time in which more and more parts of our lives are digitalized. We connect with various devices such as computers, smartphones, or tablets and are increasingly dependent on online services, as many social interactions, business matters, and financial transactions are handled over computer networks. The dependence on the digital infrastructure highlights the importance of network security in the private and public sectors as vulnerabilities or malicious activity can harm a network's integrity and compromise sensitive data. Of particular interest in the field of network security is the detection of anomalies. The detection of anomalies within network traffic has been a hot topic in the network security research community for the last decades. Significant effort has been put into the research of unsupervised anomaly detection due to the ever-changing threat landscape.
However, despite the many proposed solutions, they are rarely adopted by the industry. Despite the substantial research effort put into network anomaly detection, most adopted solutions still rely on the signature-based techniques. This discrepancy can be explained by multiple factors, such as the high error rate and the associated costs, the lack of publicly available realistic network traffic data, and the dynamic nature of network traffic. Especially the dynamic nature of network traffic is an issue that is often overlooked.

Challenges

Anomalies are abstractly defined as observations that do not conform to normal behavior. The process of identifying such deviations from the norm is referred to as anomaly detection. The definition was later extended with two, especially in the context of data mining and machine learning, important characteristics.
The two characteristics are (1) Anomalies are different from the norm with respect to their features and (2) They are rare in the dataset compared to normal instances. There are many challenges that anomaly detection faces; most of these challenges are especially prevalent in the context of network security.

1. **Encompassing every form of normal behavior**. This task quickly becomes overwhelming as networks deal with a great amount of variability. Paxon et al. formulated this problem concisely while discussing the difficulties in simulating the internet: "IP buys uniform connectivity in the face of diversity, not uniform behavior".
2. **The notion of normal.** Even a single network can show high variability in bandwidth, duration of the connections, applications, and protocol usage. Networks can show even greater variability when compared to each other.
3. **The application domain is not static.** This problem arises when the notion of normal changes over time, which is especially prevalent in the domain of computer networks.
4. **Unavailability of labeled data.** The difficulty to reliable capture or generate realistic network traffic, which is correctly labeled, still proves to be a difficult task, and although many new datasets were published in the last years, much research still relies on outdated data.
5. **High cost of error.** Missing to identify an anomaly can imply that a network attack was executed unnoticed while a normal instance which was falsely identified as an anomaly calls for further investigation. The huge amount of traffic packets can quickly render an anomaly detection approach useless.

Data

Obtaining accurate, realistic, and ideally, noise-free data has been a significant hindrance for the network security research community. One hurdle of publishing benchmark data is the sensitive nature of network traffic. Network traffic can reveal sensitive information, like private communications, confidential business information, or user access patterns. Additionally, labeling network traffic is a time intensive and expensive task. While the research community has answered these issues by providing many different publicly available datasets in the last years, a majority of today's research still utilized outdated and not realistic data. This research utilized a current dataset to evaluate the taken approach, namely the UNSW-NB15 dataset. The data was conducted by the Australian Centre for Cyber Security and includes a total of 31 hours of network traffic present in both packet capture and flow level data.

Experiment

When choosing an appropriate algorithm to detect network anomalies, a few fundamental properties were essential to us. These properties result from inherent difficulties when analyzing network traffic and stem from the challenges provided above.

- Network traffic is present as a stream of data.
- Network traffic is not static.
- Transfer-ability of detection models is not given.

Network data needs to be analyzed in an online fashion - the detection of anomalies needs to be timely. However, the stream of network data is not static; concept drifts can happen in many timescales (minutes, days, weeks). Ideally, a detection model would adapt to these concept drifts to fit the current network traffic profile. As network traffic differs from network to network, fitting a model on a network and applying it to another might not provide the desired results. Therefore, an unsupervised methodology would be ideal if a high detection rate and low false alarm rate can be achieved. To tackle these issues, we have utilized an unsupervised anomaly detection algorithm with the ability to dynamically alter the model of normality during detection, the Robust Random Cut Forest (RRCF). The algorithm detects anomalies utilizing isolation using a binary tree structure and allows for dynamic updating of the model. The algorithm was evaluated on current benchmark data utilizing a hyperparameter optimization strategy, optimizing model-specific hyperparameters, and additional data preprocessing steps to investigate the algorithm's online anomaly detection capabilities.

Results

One of the experiment's objectives was to understand the chosen hyperparameters' impact on model performance. We have gathered over 250 objective evaluations of the best optimization cycles for this analysis. The algorithm showed promising discrimination capabilities and outperformed another isolation-based, but static algorithm, the Isolation Forest.

Algorithm	ROC AUC	PR AUC
RRCF	0.976	0.700
Isolation Forest	0.951	0.544

Table 1: Baseline Results

One of the main issues of applying the RRCF algorithm to network traffic is scaling the algorithm to accommodate large traffic volumes limiting the applicability in a real-world scenario. We have extended the algorithm with parallel processing capabilities, and while this improved the performance, the difficulty of scaling the algorithm is still an issue. Additionally, the implemented data preprocessing steps showed to reduce detection time as well as substantially improve the detection performance as seen in Fig. 1.

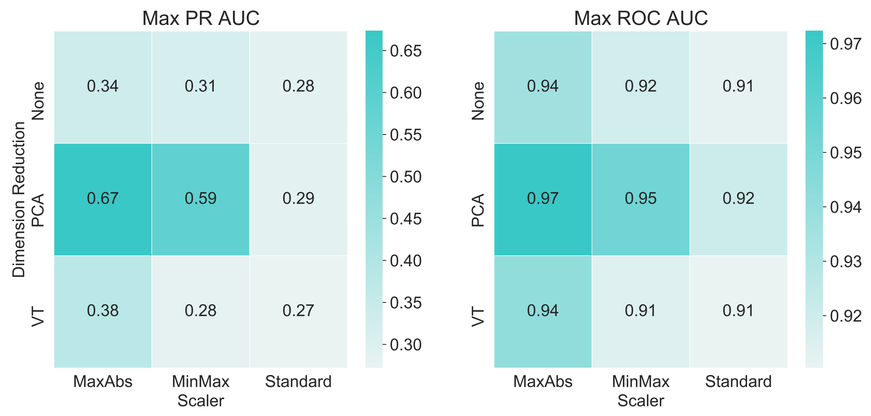

Figure 1: Max PR/ROC AUC

A combination of scaling the features according to their maximum absolute value with an additional principal component analysis application offered the best results. However, while the algorithm tends to score the instances correctly, there was no clear distinction between anomalies and normal instances, leading to a too high to tolerate false positive rate due to the large number of individual instances.

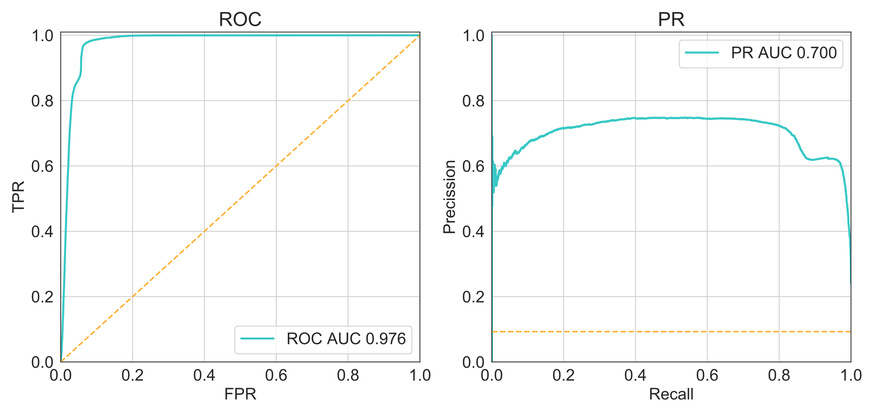

Figure 2: Best achieved ROC and PR Curve

Conclusions

We improved the algorithm's detection capabilities by the implemented subsampling extension and the implemented data preprocessing steps. The algorithm showed promising discrimination capabilities but suffered, like many other unsupervised anomaly detection techniques, from too many false alarms. One of the main issues of applying the RRCF algorithm to network traffic remains the difficulty of scaling the algorithm to accommodate large traffic volumes. We were able to substantially reduce the algorithm's runtime by extending it with parallel processing capabilities and the utilization of appropriate data preprocessing steps. Although there are issues to resolve regarding the application of the algorithm in a real-world setting, we believe that the intrinsic properties and the ability to update the model dynamically can be a valuable tool in network anomaly detection.

Forthcoming Research

With the algorithm's scalability being one of the main hindering factors, it needs to be addressed before a possible application can be considered. We propose further adaptations of the algorithm such as limiting the scope, utilization as a comprehensive network monitoring tool and decoupling the detection from the insertion process to further increase the scalability, which should be considered for further research. Although we believe that the proposed steps could enable an application of the algorithm for network anomaly detection, the general issue of too many false alarms during unsupervised anomaly detection needs to be addressed.