An Empirical Analysis of the Impact of Data Sharing on Anomaly Detection for Network Security

Authors

Abstract

Anomaly detection is used to discover and understand network security issues, especially pertaining to distinguishing between normal and attack traffic. Since a single entity has a limited vantage point from which to observe network traffic, the sharing of network traffic data has been proposed as a method to improve the performance of network anomaly detection. Existing work has largely focused on sharing characteristics of detected anomalies. In this paper, we take a broader view of network data sharing and aim to characterize when it can improve anomaly detection performance. In particular, we focus on the sharing of network traffic data and the models trained on them in the supervised, unsupervised and semi-supervised learning paradigms. For each mode of learning, we consider the impact of the volume of data, its heterogenity and labeling errors on the performance of anomaly detectors. Our experiments on 2 publicly available and 1 curated dataset, using 5 different types of models indicate that: TODO:add experimental conclusions here

1 Introduction

- Paragraph 1: Why is collaborative anomaly detection important?

- Paragraph 2: State of current work and need for this paper: need to understand the performance gains from different sharing modes (data, model, output) across learning paradigms.

- Paragraph 3: Data volume, heterogeneity and noise. Discuss division of data sharing into benign and malicious
- Quantifying data heterogeneity
- Benign data: If homogeneous, more data will not hinder performance? If heterogeneous, expressive enough (deep) models should be able to learn, what happens with shallow models?
- Malicious data: How is it shared?

- Paragraph 4: What are the possible learning paradigms and how do they interact with the data characteristics?
- Unsupervised learning: how is a model of ‘normal’ behavior affected by multiple data sources? (affects FP rates, will not affect TP?)

- Paragraph 5: Experiment structure: - Datasets: IDS, IoT, BAS
- Models: RF, IF
- Direct data sharing: S, US, SS; tune data het. for each
- Model sharing: S, US, SS
- Interpretation of results for different attack types

Contributions: 1. Clearly map out a taxonomy of data sharing for network security 2. Establish baselines when data sharing is helpful (with regards to data, model and output) 3. Practical method to model and understand data heterogenity

2 Background

3 Data Characteristics

3.1 CIC-IDS-2017

1. nfdump:

2. netml:

3. Flowmeter:

4 Learning Paradigms

5 Discussion

6 Related Work

6.1 Learning paradigms for network data

6.2 Data sharing for cybersecurity

6.3 Collaborative learning

Mirsky et al. [1] consider the use of an Extended Markov Model (EMM) to model anomalies across IoT devices and a blockchain for the secure aggregation of the locally trained models. Each model is just a transition matrix for the probabilities of transition from one location to another in the device, so low probability transitions are flagged as anomalous. Aggregation is done by simply adding up the local transition matrices.

7 Conclusion

Features	Unsupervised					Supervised	
	GMM	IF	KDE	OCSVM	PCA	Random Forest	SVM
nfdump	0.72	0.38	0.87	0.513	0.76	0.99	0.71
netml	0.64	0.89	0.53	0.66	0.5	0.98	0.87
flowmeter	NA	0.54	0.5		0.375	0.99	

Table 1: Malicious data classification or detection results (All 3 attacks on Friday), 1 agent

Features	Unsupervised					Supervised	
	GMM	IF	KDE	OCSVM	PCA	Random Forest	SVM
nfdump	0.36	0.14	0.94	TODO	0.36		
netml	0.87	0.35	0.87	TODO	0.86		
flowmeter	NA	0.2			0.77		

Table 2: Malicious data classification or detection results (DDoS on Friday), 1 agent

Features	Unsupervised					Supervised	
	GMM	IF	KDE	OCSVM	PCA	Random Forest	SVM
nfdump	0.51 ± 0.12	0.14 ± 0.01	0.93 ± 0.023	$0.09 \pm 1e{-5}$	0.37 ± 0.007		
netml	0.87 ± 0.003	0.35 ± 0.02	0.83 ± 0.004	$0.73 \pm 2e{-5}$	0.86 ± 0.007		
flowmeter	NA	0.21 ± 0.006	0.89 ± 0.04	$0.22 \pm 4e{-4}$	0.76 ± 0.004		

Table 3: Malicious data classification or detection results (DDoS on Friday), 10 agents

Features	Unsupervised					Supervised	
	GMM	IF	KDE	OCSVM	PCA	Random Forest	SVM
nfdump	0.82 ± 0.05	0.14 ± 0.02	0.94 ± 0.01	0.09 ± 0.003	0.29 ± 0.11		
netml	0.85 ± 0.015	0.35 ± 0.02	0.51 ± 0.01	0.73 ± 0.003	0.83 ± 0.03		
flowmeter							

Table 4: Malicious data classification or detection results (DDoS on Friday), 1000 agents

References

[1] Yisroel Mirsky, Tomer Golomb, and Yuval Elovici. “Lightweight Collaborative Anomaly Detection for the IoT using Blockchain”. In: *arXiv:2006.10587 [cs]* (June 2020). arXiv: 2006.10587. URL: `http://arxiv.org/abs/2006.10587` (visited on 03/14/2021).