Paper Summary
Dos and Don'ts of Machine Learning in Computer Security

Sarthak | 2020CS10379

June 2024

Problem Statement

This paper addresses the ethics of developing machine learning models by indentifying and classifying a few common pitfalls, their prevalence, impact and mitigation strategies.

Motivation

- False Positive Rate (FPR) of current network intrusion detectors often still corresponds to large number of false positives i.e. the base rate fallacy is high
- In Android Malware Detection, unrealistic sptio-temporal bias class balance inflates performance

Key Idea

- Pitfall Identification : Classification of pitfalls based on subtle issues affecting ML for security at various stages and recommending mitigations
- Prevalence Analysis : Prevalence of these pitfalls in over 30 security research papers of last decade are analyzed and feedbacks of authors are reviewed
- Impact Analysis : Case Studies demonstrating the impact of pitfalls are demonstrated for eg. mobile malware detection.

Framework & Methodology

- ML Pipeline and Pitfalls

1. Data Collection and Labeling :
   (a) Sampling Bias : The collected data does not sufficiently represent the true data distribution of the underlying security problem. In some cases, a reasonable strategy is to construct different estimates of the true distribution and analyze them individually.
   (b) Label Inaccuracy : The ground labels for classification are untrue, leading to learning errors. One must verify labels and use robust loss functions to deal with noisy labels.
2. System Design and Learning :
   (a) Data Snooping : A learning model is trained with data that is typically not available in practice. Test data should be split early and stored separately to mitigate this
   (b) Spurious Correlation : The learning model learns the shortcuts created for separating classes, instead of doing the actual task.To help with this, it is recommended to apply explanation techniues for machine learning.
   (c) Biased Parameter Selection : Test set often affect the final parameters of the learning model. Using a separate validation set is recommended.
3. Performance Evaluation :
   (a) Inappropriate baseline : This makes it difficult to measure improvement. Simple models of comparison should be employed.
   (b) Inappropriate performance measures : One must keep in mind the application-specific context while deciding metrics.
   (c) Base Rate Fallacy : This leads to an overestimation of performance. In detecting rare events, precision and recall as indicative measures are recommended.
4. Deployment and Operation:
   (a) Lab-Only Evaluation : This often leads to the oversight of practical limitations. It is necessary to simulate a near real-world simulation for testing.
   (b) Inappropriate Threat Model : Threat models should be defined precisely and systems evaluated with respect to them. In most cases, it is necessary to assume an adaptive adversary

- Prevalence Analysis : All of the pitfalls are pervasive in security research, affecting between 17 % and 90 % of the selected papers. Each paper suffers from at least three of the pitfalls with discussions accompanied only in 22 % .

Contributions & Strengths

- 10 subtle pitfalls are identified and their mitigation strategies are discussed for building of safer, more accurate and robust model.

- Vulnerabilities even in top-researches are discussed, hence cautioning beforehand for upcoming research.

Weaknesses & Limitations

- Few pitfalls like Sampling Bias cannot be mitigated completely
- Few of the pitfalls, opposite to intuition, are not highly prevalent.
- A pitfall is only counted if its presence is clear from the text or the associated artifacts, such as code or data, making the performance analysis conservative.