GRIFIN Research Internship: Data-driven assessment of intrusion detection

Sébastien Tixeuil, Gregory Blanc

Application deadline: January 4th, 2021

Keywords — intrusion detection, artificial intelligence, assessment methods and metrics, dataset generation

Context

Intrusion detection systems (IDS) are security measures that monitor an information system and raise an alert whenever an attack or an anomaly is detected. Misuse detection designates IDS that rely on a database of signatures indicative of known attacks. On the other hand, anomaly detection is concerned with modelling the legitimate behaviour of a system, and detects any deviation. Anomaly detectors are hence capable to detect new attacks which do not have any known signature yet. However, anomaly detectors are more prone to false positives.

As a common security measure, it is important to assess the performance of an IDS: ideally, it should not miss any attack while avoiding to mis-classify legitimate events. These errors are known as *false negative* and *false positive*, respectively and can be modelled in a *confusion matrix*, which also features the number of *true positives*, that is the events that were classified as intrusions, and *true negatives*, the legitimate events that were rightfully not considered as intrusions. Leveraging the confusion matrix, one can compute common classification performance metrics such as, among others, the *accuracy*, the amount of correct detections within the set of all monitored events, or the *detection rate*, also known as *recall*, that is the amount of detected intrusions within the set of malicious events.

According to Milenkovski et al. [1], the evaluation methodologies of IDS focus on:

- the **attack detection accuracy**, measuring the accuracy of an IDS in the presence of mixed workload, that is datasets including both legitimate and malicious events;
- the **attack coverage**, measuring the amount of attacks an IDS can detect when facing a dataset of pure malicious events;
- the **resistance to evasion techniques**.

The latter is quite overlooked in comparison to the first two, as it is considered to be of limited importance from a practical perspective [2].

The emergence of artificial intelligence-based intrusion detection, in particular deep-learning based anomaly detectors (DAD), is questionable as black-box models are not adequate for applications that require domain knowledge such as cybersecurity. The state-of-the-art of DAD research works is growing fast [3], with performance results ranging from mediocre to outstanding, but no common ground truth. Indeed, while common performance metrics are computed

(at least, the accuracy), the assessment methodology usually differs in terms of input data. Best practices with respect to dataset partition (between training and testing) are not always enforced, and many issues plague the machine learning process from collection and labelling to learning, to performance evaluation, to operation [4]. In fact, anomaly detection struggles with the changing nature of anomalies itself [2], and DAD is not immune, although it usually better cope with variance in input data. Indeed, traffic changes over time, deviating from the once learned representation. This phenomenon, known as *concept drift*, has been previously discussed for malicious software classification [5] but seldom addressed in DAD.

Other issues of interest include *adversarial samples* which are specifically crafted to evade detection [6], and represent the future of cybercrime. We advocate the inclusion of adversarial assessment in the evaluation of IDS, as a way to test the robustness of IDS, especially IA-based ones, as adversarial machine learning [7] has demonstrated the ability to deceive machine learning models.

This internship aims at surveying assessment methodologies for IDS, and especially DAD, and proposing a data-driven approach able to improve the assessment coverage with respect to common data-related issues such as *time decay* [5] or lack of representativeness. To address these issues, generative approaches [8] may be used to provide more complete evaluation datasets.

Activities

- survey of IDS assessment methodologies, metrics and datasets
- survey of common data-related issues incl. *concept drift*, *distributional shift*, and *adversarial samples*
- focus on network-based IDS and their learning features
- design of a data-driven assessment approach
- generation of datasets to assess input space coverage and a set of chosen issues
- evaluation of the approach on some available IDS implementations

Practical information

The internship will take place at LIP6, a laboratory of Sorbonne Université (Paris). It will be 5 months long.

Applicants are about to complete their Master 2 level degree (or equivalent engineering school degree) and should have the following skills:

- intermediate to strong knowledge and practice of machine/deep learning
- fundamentals in networking, and basic practice of traffic analysis
- concepts in cybersecurity, in particular intrusion detection
- practice in code development

The internship topic is linked to a Ph.D offer in the context of the GRIFIN project (funded by ANR), a research collaboration between Télécom SudParis, Sorbonne Université and LORIA.

Applications (resume, motivation letter, academic transcripts, recommendation letters) must be sent to `sebastien.tixeuil[at]lip6.fr` and `gregory.blanc[at]telecom-sudparis.eu`.

References

[1] Aleksandar Milenkoski, Marco Vieira, Samuel Kounev, Alberto Avritzer, and Bryan D Payne. Evaluating computer intrusion detection systems: A survey of common practices. *ACM Computing Surveys (CSUR)*, 48(1):1–41, 2015.

[2] Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy*, pages 305–316. IEEE, 2010.

[3] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[4] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. *arXiv preprint arXiv:2010.09470*, 2020.

[5] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. {TESSERACT}: Eliminating experimental bias in malware classification across space and time. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 729–746, 2019.

[6] Maria Rigaki and Sebastian Garcia. Bringing a gan to a knife-fight: Adapting malware communication to avoid detection. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 70–75. IEEE, 2018.

[7] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.