A “Poisoning” Attack Against Online Anomaly Detection

Marius Kloft
Department of Computer Science
Technical University of Berlin
Franklinstr. 28/29, 10587 Berlin
mkloft@cs.tu-berlin.de

Pavel Laskov
Fraunhofer Institute FIRST IDA
Kekulestr. 7, 12489 Berlin
pavel.laskov@first.fraunhofer.de

Introduction. Online anomaly detection techniques are steadily gaining attention in the security community, as the need grows to identify novel exploits in highly non-stationary data streams. The primary goal of online anomaly detection is to dynamically adjust the concept of normality while still detecting anomalous behavior. Online anomaly detection can be seen at a high level of abstraction as a following process:

1. Evaluate the degree of anomality of an incoming data point x.
2. If x is normal, replace some previously seen data point x_i with x.
3. Update a normality model.

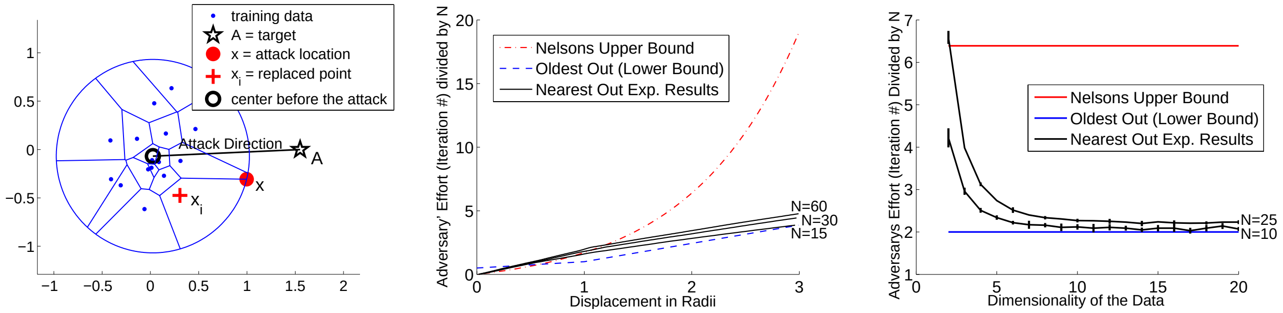
Specific realizations of this process need to make decisions on a normality model, rules for finding a data point to be removed from a working set, and a strategy for an update of a normality model.

While this online anomaly detection scheme provides an appealing possibility to adjust a normality model to possible normality drift, a question arises whether it is robust against targeted “poisoning” attacks. The latter have been first investigated by Nelson et al. [1] for a specific form of online anomaly detection, in which the concept of normality is modeled as a center of mass of all data points observed so far. The key idea of a poisoning attack is to insert specially crafted data points so that the center of mass drifts into the direction of an attack vector. After enough poison, the latter falls within a pre-defined radius of the center of mass and is treated as normal. The main result in [1] is surprisingly optimistic: in order to move the center of mass by D times normality radius R, an attacker needs to insert an exponential number of points:

$$T \geq N \cdot (\exp(D) - 1).$$

On the other hand, an infinitely growing window for computation of the center of mass prevents an algorithm from adjusting to a drift in the normal data. The main goal of this contribution is the analysis of robustness of online anomaly detection *with limited horizon* against poisoning attacks.

An optimal greedy attack. Intuitively, the success of an attack depends on a particular rule for finding a data point to be removed from a working set. A straightforward rule can, for example, remove a data



(a) (b) (c)

Figure 1: (a) Illustration of the optimization task (1). (b) Experimental results on dimensionality-reduced real data. (c) Sensitivity of the “nearest-out” learner to dimensionality.

point with an oldest timestamp ("oldest-out"). Unfortunately, this strategy is extremely sensitive against a poisoning attack. When an attacker knows which point is to be removed next, it can pre-compute the location of a new center of mass and in insert a point on a line connecting the new center of mass with an attack vector at a distance R from the center. It can be easily shown, that the amount of work needed from an attacker to evade this scheme is linear in the displacement D of an attack vector:

$$T \leq N \cdot D \quad \text{for all} \quad D \geq 1.$$

A seemingly more robust update strategy is to remove a nearest neighbor of an incoming data point ("nearest-out"). The rationale behind this strategy is to decrease susceptibility of the center of mass to targeted data points placed along the direction from the center of mass to the attack vector. Unfortunately, as the analysis below shows, even this update rule can be easily evaded by a more sophisticated attack strategy.

A optimal greedy attack can be constructed using the geometry of the "nearest-out" update rule (cf. Fig. 1(a)). Each data point x_i in a working set induces a polygon such that for any data point falling into this polygon its nearest neighbor is x_i. In order to achieve maximal displacement possible in one step, an attacker must place a point at a "corner" of some polygon (including possibly a round boundary of a hypersphere) providing the highest displacement of the center in the direction of an attack vector. This can be formulated in the language of mathematical optimization as the following optimization problem:

$$\begin{aligned} \min{}_{i \in 1,\dots,N} \quad \min{}_x \quad & \|A - \frac{1}{N}(\sum_{j=1, j \neq i}^{N} x_j + x)\| \qquad (1) \\ \text{s.t.} \quad & \|x - x_i\| \leq \|x - x_j\| \quad \forall j = 1, ..., N, \; j \neq i \\ & \|x - \frac{1}{N} \sum x_j\| \leq \|R\|. \end{aligned}$$

In the following we experimentally evaluate the effectiveness of the optimal greedy strategy on a "nearest-out" hypersphere online anomaly detection.

Experimental results. To investigate the attack in a realistic scenario, we follow the n-gram frequency approach of [2] similar to the approaches of PAYL [3] and Anagram [4]. As a training corpus, we use an 24-hour HTTP trace recorded at our institute and sanitized to remove 2% of outliers. A small working set of size N is drawn from a training corpus and mixed with the IIS 5.0 WebDAV exploit [5] used as attack vector. To reduce the dimensionality of the problem (1), we use a singular value decomposition, and project a high-dimensional space of n-gram frequencies into a subspace containing 100% of attack and 95% of the normal data variance.

The average displacement is plotted vs. the adversary's effort (attack iterations per point) in Fig. 1(b). We observe that, compared to Nelson's bound, the complexities of the attacks on the "nearest-out" and "oldest-out" learners are similar, and thus the "nearest-out" learner is not as robust against adversarial attempts of corrupting the learner as Nelson's learner.

The attack can be made even simpler by increasing the dimensionality of the subspace. To illustrate this, we have simulated attacks on 10 artificially generated data sets, each consisting of N normal points randomly drawn from the unit ball and the attacker's goal (3, 0,..., 0), for several values of N. The results are plotted in Fig. 1(c) and indicate a sensitivity of the "nearest-out" learner to dimensionality.

Conclusion. In this work we have investigated the robustness of online anomaly detection with finite horizon against poisoning attacks. We have formulated an optimal greedy attack as a mathematical optimization problem that can be solved by an attacker with reasonable effort. Our experimental evaluation showed that the "nearest-out" learner is surprisingly unsecure for attacks in high-dimensional network data.

References

[1] Nelson, B. and Joseph, A.D.: Bounding an Attack's Complexity for a Simple Learning Model. In: Proc. of the First Workshop on Tackling Computer Systems Problems with Machine Learning Techniques. (2006)

[2] Rieck, K., Laskov, P.: Language Models for Detection of Unknown Attacks in Network Traffic. In: Journal in Computer Virology, 2(4). (2007) 243–256

[3] Wang, K., Stolfo, S.: Anomalous payload-based network intrusion detection. In: Recent Advances in Intrusion Detection (RAID). (2004) 203–222

[4] Wang, K., Parekh, J., Stolfo, S.: Anagram: A content anomaly detector resistant to mimicry attack. In: Recent Advances in Intrusion Detection (RAID). (2006) 226–248

[5] Microsoft Security Bulletin MS03-007. http://www.microsoft.com/technet/security/bulletin/MS03-007.mspx.