A Framework for Quantitative Security Analysis of Machine Learning

Pavel Laskov*
Universität Tübingen
Sand 13
72076 Tübingen, Germany
pavel.laskov@uni-tuebingen.de

Marius Kloft
Technische Universität Berlin
Franklinstr. 28/29
10587 Berlin, Germany
mkloft@cs.tu-berlin.de

ABSTRACT

We propose a framework for quantitative security analysis of machine learning methods. Key issus of this framework are a formal specification of the deployed learning model and an attackers constraints, the computation of an optimal attack, and a derivation of an upper bound on the adversarial impact. We exemplarily apply the framework for the analysis of one specific learning scenario, online centroid anomaly detection and experimentally verify the tightness of obtained theoretical bounds.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*; I.2.6 [**Artificial Intelligence**]: Learning—*Parameter learning*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

General Terms

Algorithms, Experimentation, Security

Keywords

Machine learning, computer security, centroid anomaly detection, intrusion detection, adversarial learning

1. INTRODUCTION

Security is all about numbers. Even the strongest cryptographic algorithms that lie at the heart of computer security eventually succumb to the dumbest possible attack, the brute force. What makes a difference is a cold-blooded calculation of resources needed for a successful attack. If the resulting numbers are absurd, a method can be considered secure.

*Pavel Laskov is also affiliated with Fraunhofer Institute FIRST, Kekulestr. 7, 12489 Berlin, Germany.

AISec'09, November 9, 2009, Chicago, Illinois, USA.

This simple argument has proved to be extremely valuable in various security applications. In view of the recent surge of interest in machine learning applications in security – accompanied, however, by some skepticism whether machine learning can be ever made secure [1, 10, 12] – such reasoning may be instrumental in steering research into the right direction. The main thesis advocated in this position paper is that investigation of security properties of machine learning has to focus on *quantitative analysis* of attacker's resources needed to subvert a learning process. Furthermore, this analysis must take into account sensible constraints that exist on an attacker's part or can be externally enforced. Without such constraints, there is little hope that learning and its underlying mathematical tools can outwit adversarial impact. Most of the well-known examples of insecurity of machine learning methods or their specific applications, e.g. [4, 6, 8, 3], assume that an attacker can arbitrarily manipulate data. With rare exceptions, such examples do not clarify whether such unlimited impact is possible in practice.

To formalize the above mentioned intuitive argument, we propose a general framework that encapsulates essential issues to be addressed by quantitative security analysis of machine learning. We show how the analysis of a relatively well-studied online centroid anomaly detection naturally falls into our framework and leads to constructive results that can be experimentally verified. Finally, we show that other work concerned with security properties of machine learning has also addressed some problems formulated in our framework. Hence we believe that our framework can facilitate the convergence of existing patchwork of "security-aware" learning methods to a well-understood methodology with solid theoretical grounds.

2. THE FRAMEWORK

The main motivation behind the proposed framework is to depart from a worst-case and develop a kind of a "reasonable-case" analysis. As it is argued in the introduction, in the worst-case an attacker can always subvert a security mechanism, yet this case is irrelevant for realistic estimation of a security risk. In practice, the isssue of crucial importance is whether an attack can succeed under *reasonable assumptions* on resources available to the attacker. These assumptions may vary among applications. Hence they have to be clearly stated and the analysis must address the interdependencies between attack effectiveness and the attacker's constraints and resource consumption. To account for these issues, we propose the following four-step procedure for security anal-

ysis of machine learning methods.

1. *Axiomatic formalization of the learning and attack processes.* The first step in the analysis is to formally specify the learning and attack processes. Such formalization should include definitions of data sources and objective (risk) functions used by each party. It should specify the knowledge available to an attacker, i.e. whether he knows an algorithm, its parameters and internal state, and which data he can potentially manipulate. Finally, the attack goal should be also specified.

2. *Specification of attacker's constraints.* Potential constraints on the attacker's part may include: percentage of traffic under his control, amount of additional data to be injected, an upper bound on the norm of manipulated part, a maximal allowable false-positive rate (in case an attack must stealthy), etc. Such constraints must be incorporated into the axiomatic formalization.

3. *Investigation of an optimal attack policy.* Given a formal description of the problem and constraints, an optimal attack policy must be investigated. Such policy may be long-term, i.e. over multiple attack iteration, as well as short-term, for a single iteration. Investigation can be carried out either as a formal proof or numerically, by casting the search for an attack policy as an optimization problem.

4. *Bounding of attacker's gain under an optimal policy.* Finally, the progress of an attack towards its goal under an optimal policy must be analyzed. Such analysis may take different forms, for example calculation of the probability for an attack to succeed, estimation of the required number of attack iterations, calculation of the geometric impact of an attack (a shift towards an insecure state), etc.

The proposed framework enables one to quantitatively analyze and compare existing algorithms under identical conditions. It can further give a valuable intuition for hardening learning algorithms against certain attacks by identifying constraints that can be leveraged by a learner.

Another potential application of our framework is analysis of the tradeoff between security, accuracy and resource consumption on the learning side. It is well known from the statistical learning theory [11] that compact models are needed for high learning accuracy. Such models, however, may be easier for attacker to subvert. By providing a quantitative measure of security risk, our framework may be used for designing optimal risk functions for learning algorithms.

3. QUANTITATIVE ANALYSIS OF ONLINE CENTROID ANOMALY DETECTION

As an example of a practical application of the proposed framework, we outline the essential steps of a quantitative analysis of online centroid anomaly detection originally presented in [5]. Due to space limitations, we can only show the basic building blocks of this analysis; for an detailed account, the reader should consult a forthcoming publication.

Given the data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the goal of anomaly detection is to determine whether an example $\mathbf{x}$ originates from the same distribution as the set X. In the centroid anomaly detection, a Euclidean distance from an empirical mean of the data is used as a measure of anomaly:

$$f(\mathbf{x}) = ||\mathbf{x} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i||.$$

If the training data is processed online, i.e. one point at a time, the centroid model can be easily updated in presence of a new training point $\mathbf{x}$ using the following rule:

$$\mathbf{c}' = \left(1 - \frac{1}{n}\right)\mathbf{c} + \frac{1}{n}\mathbf{x}, \tag{1}$$

where c' denotes the updated centroid. Depending on whether the training set size n is allowed to grow infinitely (which eventually prohibits the model from adjusting to new data), a distinction is drawn between the infinite-horizon case, analyzed in [7], and a finite horizon case which is the subject of our model. For the finite horizon case, a rule must be given for removing some point from a working set so that the total number of data points remains n. Under the so-called "average-out" rule to be considered below, the center of mass c is removed at each iteration.

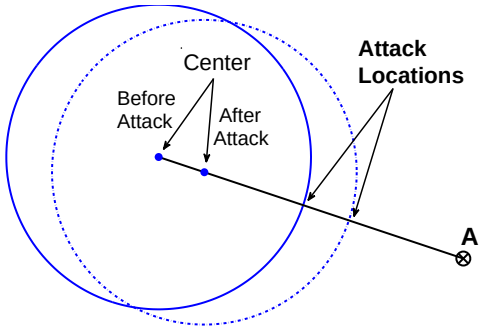

Figure 1: Illustration of a poisoning attack.

Online anomaly detection can be a subject to a poisoning attack whose goal is to force an algorithm to accept an *attack vector* $\mathbf{a}$ that lies outside of the normal sphere, i.e. $||\mathbf{a}-\mathbf{c}|| > r$. We assume that an attacker knows the anomaly detection algorithm and all the training data. However, an attacker cannot modify any existing data except for adding new data points. These assumptions model a scenario in which an attacker can sniff data on the way to a particular host and can send his own data, while not having write access to that host. As illustrated in Fig. 1, the poisoning attack attempts to inject specially crafted points that are accepted as normal and pull the center of mass in the direction of an attack vector until the latter appears normal.

In order to quantify the effectiveness of a poisoning attack, we define the *i-th relative displacement* of the center of mass to be $D_i = \frac{(\mathbf{c}_i - \mathbf{c}_0)\cdot\mathbf{a}}{r||\mathbf{a}||}$ (w.l.o.g. assume that $\mathbf{c}_0 = 0$). This quantity measures a relative length of the projection of $\mathbf{c}_i$ onto $\mathbf{a}$ in terms of the radius of a normal sphere. Intuitively, the relative displacement represents the portion of a total distance to be traversed by the center of mass until an attack succeeds.

If an attacker has unlimited control over the training data, i.e. he can inject every single training data point, it can be

shown that the center of mass can be shifted into an attack direction in a linear number of iterations. However, in a realistic deployment of anomaly detection, normal traffic is also present. Moreover, on some systems with heavy workload, it may be difficult for an attacker to control more than a fraction ν of the training data. This scenario is modeled by "flipping a coin" with the probability ν to decide whether a point is drawn from a benign or an attack distributions. Furthermore, we assume that the expectation of benign data coincides with the initial center of mass $\mathbf{X}_0$ and that all benign data points are accepted. The attacker's strategy is formalized by a function $f : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $\|f(x) - x\| \leq r$. Such function provides the attacker with coordinates of a new data point to be injected which falls withing the raduis of the normal sphere. Our model and its assumptions can be formalized in the following axiom.

AXIOM 1 (STEPS 1 AND 2). *$\{B_i | i \in \mathbb{N}\}$ are independent Bernoulli random variables with parameter $\nu > 0$. $\boldsymbol{\epsilon}_i$ are i.i.d. random variables in $\mathbb{R}^d$, drawn from a fixed but unknown distribution $P_{\boldsymbol{\epsilon}}$, satisfying $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$ and $\|\boldsymbol{\epsilon}_i\| \leq r = 1$ for each i. B_i and $\boldsymbol{\epsilon}_j$ are mutually independent for each i, j. $f : \mathbb{R}^d \to \mathbb{R}^d$ is an attack strategy satisfying $\|f(x) - x\| \leq r$. $X_i | i \in \mathbb{N}\}$ is a collection of random vectors such that $\mathbf{X}_0 = 0$ and*

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \frac{1}{n}\left(B_i f(\mathbf{X}_i) + (1 - B_i)\boldsymbol{\epsilon}_i - \mathbf{X}_i\right). \qquad (2)$$

According to the above axiom an adversary's strategy is formalized by an *arbitrary* function f. This gives rise to a question which attack strategies are optimal in the sense that an attacker reaches his goal of concealing a pre-defined attack vector in a minimal number of iterations. An answer to this question is provided in the following theorem.

THEOREM 1 (STEP 3). *Let $\mathbf{a}$ be an attack vector and let $\mathcal{C} = \{\mathbf{X}_i, r\}$ be a centroid learner. Then the* optimal attack strategy *f is given by*

$$f(\mathbf{X}_i) := \mathbf{X}_i + \mathbf{a}. \qquad (3)$$

To quantify the progress of an attack under the optimal strategy, we bound the mean and the standard deviations of the attack displacement, which is summarized in the following theorem.

THEOREM 2 (STEP 4). *Let $\mathcal{C}$ be a centroid learner under an optimal poisoning attack with an adversarial fraction of ν in the training data stream. Then, for the displacement D_i of $\mathcal{C}$, it holds:*

$$\text{(a)} \qquad E(D_i) \;=\; (1 - c_i)\frac{\nu}{1 - \nu}$$

$$\text{(b)} \qquad \mathrm{Var}(D_i) \;\leq\; \gamma_i \left(\frac{\nu}{1 - \nu}\right)^2 + \delta_n$$

where $\gamma_i = c_i - d_i$, $c_i := \left(1 - \frac{1-\nu}{n}\right)^i$, $d_i = \left(1 - \frac{1-\nu}{n}\left(2 - \frac{1}{n}\right)\right)^i$ and $\delta_n := \frac{\nu^2 + (1 - d_i)}{(2n-1)(1-\nu)^2}$.

Asymptotically, with $i, n \to \infty$, it holds that $E(D_i) \leq \frac{\nu}{1-\nu}$ and $\mathrm{Var}(D_i) \to 0$. It is interesting to observe the growth of the above bounds as a function of a number of attack iterations, shown in Fig. 2. The attack's success strongly depends on the fraction of the training data controlled by an attacker. For small ν less that a critical value ν_{crit}, the attack progress is bounded by a constant, which implies that an attack *fails even with an infinite effort*. This result provides a much stronger security guarantee than the exponential bound for the infinite horizon case [7]. The critical ratio can be computed as $\nu_{\text{crit}} = \frac{D}{D+1}$ by inverting the first bound in Theorem 2.

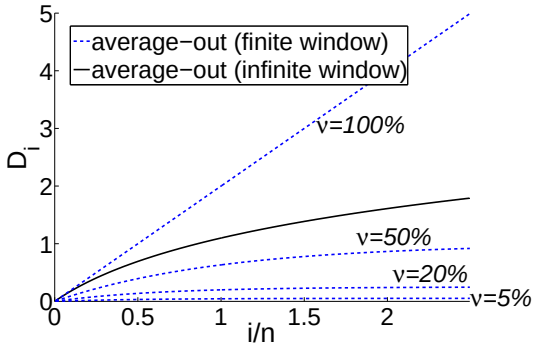

Figure 2: Theoretical progress of a poisoning attack

4. APPLICATION EXAMPLE

To illustrate how the analytical bounds can be applied in practice, we have simulated a poisoning attack against a centroid anomaly detector monitoring HTTP traffic. The payload of each inbound HTTP request is treated as a single data point. The benign dataset consists of 2950 byte strings of normal requests. We have created a malicious data set by generating various exploits using the Metasploit penetration testing framework.

Suppose the attacker wants to squeeze a particular exploits past the anomaly detector. Since he can sniff on the network, he can compute the position of the center of mass of the normal data and generate the points to be injected. Hence we can take the center of mass of our 2950 benign data points and compute the relative displacement D between the center of mass and the attack vector[1]. For example, for the IIS WebDAV 5.0 exploit, the relative displacement for is 0.179 which corresponds to $\nu_{\text{crit}} = 0.152$. Hence the attacker must control at least 15% of data in order for an attack to succeed.

We now experimentally verify that the theoretically computed critical traffic ratio indeed accurately describes the conditions under which an attack succeeds. For that, we run an attack using diffirent values of the parameter ν that controls coin-flipping in our axiomatic model. An attack succeeds if the observed relative displacement reaches the critical value of 0.179. The progression of the relative displacement over the course of attack iterations is shown in Fig. 3.

As it can be seen from the plots, the attack succeeds for $\nu = 0.16$ but fails regardless of the number of iterations for $\nu = 0.14$ and less. This experiment shows that the derived bounds are surprisingly tight in practice.

One can also see that the bounds in Theorem 2 are not constructive as such. In other words, they merely state

[1]The distance is computed using an embedding of byte strings in a space of n-grams. Technical details can be found in [9].

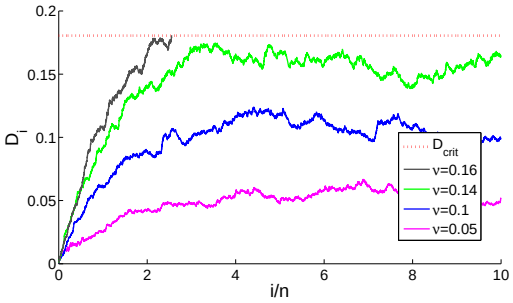

Figure 3: Simulation of a poisoning attack for various ratios of traffic controlled by an adversary.

that certain conditions have to be met for centroid anomaly detection to be immune against poisoning but provide no means for enforcing such immunity. A constructive instrument applicable regardless of the traffic ratio controlled by an attacker can be implemented by controlling the false positive rates exhibited by an algorithm. Such control, however, requires a more sophisticated model that accounts for benign data points falling outside of the sphere. Such analysis is, however, beyond the limits of this position paper and will be presented in the forthcoming publication.

5. RELATED WORK

In this section we briefly review, without a claim for completeness, some previous approaches to security analysis of machine learning algorithms that contain certain building blocks of our framework. Obviously, most closely related is the work of Nelson and Joseph [7] who also addressed a poisoning attack against online anomaly detection. Their analysis based on physical analogies contains all steps of our framework except for consideration of constraints.

Perhaps the most consummate treatment of learning under an adversarial impact has be carried out by Dalvi et al. [2]. In this work, Bayesian classification is analyzed for robustness against adversarial impact. The choice of their classifier is motivated by widespread application of the naive Bayes classification in the domain of spam detection where real examples of adversarial impact have been observed for a long time. The adversarial classification is considered as a game between an attacker and a learner. Due to the complexity of analysis, only one move by each party can be analyzed. Similar to our approach, Dalvi et al. formalize the problem by defining cost functions of an attacker and a learner (Step 1) and determine an optimal adversarial strategy (Step 3). Although attacker's constraints are not explicitly treated theoretically, several scenarios using specific constraints have been tested experimentally. No analysis of attacker's gain is carried out; instead, learner's direct response to adversarial impact is considered.

6. CONCLUSIONS

Despite a significant evidence of successful attacks against certain machine learning algorithms, we believe that one should not be overly pessimistic. True, machine learning and its underlying theory has not been developed with security applications in mind. Yet precisely in security, with its highly complex phenomena such as polymorphism, obfuscation and blending, the ability of machine learning methods to uncover hidden dependencies is very promising. The key to understanding of security properties of learning algorithms lies in *quantitative analysis* of attacker's effort in relation to its goals and application-specific constraints. The framework for security analysis of learning methods advocated in this paper defines key problems to be addressed in such analysis. We have shown how this framework can be applied for the analysis of one specific learning scenario, online centroid anomaly detection, and presented experimental evidence of suprizing tightness of the resulting bounds. We have also identified several related methods containing selected building blocks of our framework. While generality and completeness of our approach needs to be further investigated, we hope that the proposed framework would facilitate the development of robust learning methods for adversarial applications.

7. REFERENCES

[1] M. Barreno, B. Nelson, R. Sears, A. Joseph, and J. Tygar. Can machine learning be secure? In *ACM Symposium on Information, Computer and Communication Security*, pages 16–25, 2006.

[2] N. N. Dalvi, P. Domingos, Mausam, S. K. Sanghai, and D. Verma. Adversarial classification. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, *KDD*, pages 99–108. ACM, 2004.

[3] P. Fogla and W. Lee. Evading network anomaly detection systems: formal reasoning and practical techniques. In *ACM Conference on Computer and Communications Security*, pages 59–68, 2006.

[4] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[5] M. Kloft and P. Laskov. A poisoning attack against online anomaly detection. In *NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security*, 2007.

[6] D. Lowd and C. Meek. Good word attacks on statistical spam filters. In *Conference on Email and Anti-Spam*, 2005.

[7] B. Nelson and A. D. Joseph. 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 (SysML)*, Saint-Malo, France, 2006.

[8] R. Perdisci, D. Dagon, W. Lee, P. Fogla, and M. Sharif. Misleading worm signature generators using deliberate noise injection. In *Proc. of IEEE Symposium on Security and Privacy*, pages 17–31, 2006.

[9] K. Rieck and P. Laskov. Linear-time computation of similarity measures for sequential data. *Journal of Machine Learning Research*, 9(Jan):23–48, 2008.

[10] Y. Song, M. Locasto, A. Stavrou, A. Keromytis, and S. Stolfo. On the infeasibility of modeling polymorphic shellcode. In *Conference on Computer and Communications Security (CCS)*, pages 541–551, 2007.

[11] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[12] S. Venkataraman, A. Blum, and D. Song. Limits of learning-based signature generation with adversaries. In *NDSS*. The Internet Society, 2008.