2011 Fifth International Conference on Secure Software Integration and Reliability Improvement - Companion

Secure Machine Learning, A Brief Overview

Xiaofeng Liao*‡§, Liping Ding*,Yongji Wang*†
*National Engineering Research Center for Fundamental Software, Institute of Software
†State Key Laboratory of Computer Science, Institute of Software
‡Graduate University, Chinese Academy of Sciences, Beijing 100049, China
§Information Engineering School, Nanchang University, Nanchang, Jiangxi, 330031, China
{xiaofeng,liping,yongji}@nfs.iscas.ac.cn

***Abstract*—The purpose of this article is to give a brief overview on the current work towards the emerging research problem of secure machine learning. Machine learning technique has been applied widely in various applications especially in spam detection and network intrusion detection. Most existing learning schemes assume that the environment they settle in is benign. However this is not always true in the real adversarial decision-making situations where the future data sets and the training data set are no longer from the same population, due to the transformations employed by the adversaries. As more and more machine learning systems are put into use, it is imperative to consider the security of the machine learning system. As a emerging problem, it is attracting more and more researchers' attention. In this article, we present a brief overview on secure machine learning and current progress on developing secure machine learning algorithms.**

***Keywords*-Secure Machine Learning; Overview;**

I. Introduction

Statistical machine learning has been widely applied in various applications and become a valuable tool for detecting and preventing malicious activity. As a technique for building adaptive systems, machine learning enjoys several advantages over hand-crafted rules and other approaches: it can infer hidden patterns in data, it can rapidly evolve to changing and complicated situations, respond to evolving real-world inputs, adapt quickly to new signals and behaviors.

A typical learning process in the supervised classification problem is, the learner trains on a dataset of N instances, $X = \{(x, y)|x \in \mathcal{X}, y \in \mathcal{Y}\}^N$, given an instance space $\mathcal{X}$ and the label space $\mathcal{Y} = \{0, 1\}$. Given some hypothesis class Ω, the goal is to learn a classifier $f^* \in \Omega$ to minimize errors when predicting labels for new data, or if a cost function over errors exists, to minimize the total cost of errors[1].

Currently most machine learning algorithms assume that the training data and evaluation data are drawn from the same distribution. However, in many domains, including spam detection, intrusion detection, fraud detection, surveillance and counter-terrorism, this is far from the truth: there, the data is often actively manipulated by an adversary seeking to make the learner produce false negatives.

The definition of secure machine learning can be given as, when a learning algorithm succeeds in adversarial conditions, it is an algorithm for secure learning[2].

As more applications employ machine learning techniques in adversarial decision-making situations, increasing powerful attacks become possible against machine learning systems and pose a significant challenge that not yet addressed in previous research. The need to protect systems against malicious adversaries continues to increase.

The security of machine learning system is a newly emerged research problem. Different attacking cases were reported but up to now, few survey work has been conducted to reveal the current progress in this field. Besides the few existing works, this paper provides an overview over the development of secure machine learning. We group various kinds of attacking cases on machine learning systems by their application domains to give a vivid image instead by classifying them by designing a taxonomy as [1] did. The domain involved includes: spam filter, worm signature, intrusion detection, collaborative filtering, etc. We intend this paper to foster discussion about it.

II. Related Work

To our best knowledge, the only survey paper we found that summarize the current work is [2]. They develop a framework for analyzing attacks against machine learning systems and present a taxonomy that describes the space of attacks against learning systems. In the taxonomy, the attacks against learning systems are categorized along three axes:

1) Influence
    a) *Causative* attacks influence learning with control over training data.
    b) *Exploratory* attacks exploit misclassifications but do not affect training
2) Security Violation
    a) *Integrity* attacks compromise assets via false negatives
    b) *Availability* attacks cause denial of service, usually via false positives
3) Specificity

978-0-7695-4454-0/11 $26.00 © 2011 IEEE
DOI 10.1109/SSIRI-C.2011.15

a) *Targeted* attacks focus on a particular instance
b) *Indiscriminate* attacks encompass a wide class of instances

Another similar survey is [3] which focuses on a much more narrower filed of secure collaborative filtering algorithm.

To get a straightforward and clear image of the various attacks posed on machine learning systems, we group attack cases available according to the domains they belongs to. We classify secure machine learning systems according to the methodology they adopt.

The rest of this paper is organized as follows. Section III is a survey of attacks against learning systems categorized according to their application domain. Section IV introduces the different lines of methodology adopted in building secure machine learning system. Section V lists some perspective research direction. Finally in section VI comes the conclusion.

III. Potential Attacks

In this section we list some examples showing the threat toward machine learning systems in different domains.

First are two examples in signature generation in [4] and [5] which demonstrate the practical impact of allergy attacks, where attackers manipulate the signature generation system and turn it into an active agent for DoS attack against the protected system. While in [6] a practical attacks against learning is described, an adversary constructs labeled samples that, when used to train a learner, prevent or severely delay generation of an accurate classifier. They show that even a delusive adversary, whose samples are all correctly labeled, can obstruct learning. They simulate and implement highly effective instances of these attacks against the Polygraph automatic polymorphic worm signature generation algorithms.

Next are three examples on intrusion detection system. In literature, intrusion detection systems have been approached by various machine learning techniques. However, there is no a review paper to examine and understand the current status of using machine learning techniques to solve the intrusion detection problems. Altogether about 55 related studies in the period between 2000 and 2007 focusing on developing single, hybrid, and ensemble classifiers are reviewed in [7]. Related studies are compared by their classifier design, datasets used, and other experimental setups.

A method of evade the intrusion detection systems by a polymorphic blending attack (PBA) is demonstrated in [8]. The main idea of a PBA is to create each polymorphic instance in such a way that the statistics of attack packet(s) match the normal traffic profile. They show that in general, generating a PBA that optimally matches the normal traffic profile is a hard problem (NP-complete). However, the problem of finding a PBA can be reduced to the SAT or ILP problems so that solvers available in those domains can be used to find a near-optimal solution. We also present a heuristic (hill-climbing) to find an approximate solution.

Another way to escape intrusion detection is shown in [9], an adversary crafts an offensive mechanism that renders an anomaly-based intrusion detector blind to the presence of ongoing, common attacks. It presents a method that identifies the weaknesses of an anomaly-based intrusion detector, and shows how an attacker can manipulate common attacks to exploit those weaknesses.

Followed are two examples of spam systems. In [10], an adversary exploits statistical machine learning, as used in the SpamBayes spam filter, to render it useless. They also demonstrate a new class of focused attacks that successfully prevent victims from receiving specific email messages.

Similarly, [11] examined the general attack methods spammers used, along with challenges faced by developers and spammers. They also demonstrate an attack that, while easy to implement, attempted to more strongly work against the statistical nature behind filters.

Recommender System is a utility of machine learning. As more and more web sites using collaborative filtering methods, much users want to get a better result from the recommend system ,including the malicious user. More and more literatures pay attention to the security of this system. Here we particularly dwell on the attacks on collaborative filtering. The widespread deployment of recommender systems has lead to user feedback of varying quality. While some users faithfully express their true opinion, many provide noisy ratings which can be detrimental to the quality of the generated recommendations. The presence of noise can violate modelling assumptions and may thus lead to instabilities in estimation and prediction. Even worse, malicious users can deliberately insert attack profiles in an attempt to bias the recommender system to their benefit. While previous research has attempted to study the robustness of various existing Collaborative Filtering (CF) approaches, this remains an unsolved problem.

By experiments on a large data set of movie ratings, [12] suggested that new ways must be used to evaluate and detect shilling attacks on recommender systems.

A study from [13] presents an extension to the Valiant's model of machine learning with the presence of errors, possibly maliciously generated by an adversary, in the sample data. In Valiant's model, an algorithm succeeds if it can, with probability at least $1 - \delta$ learn a hypothesis that has at most probability ϵ of making an incorrect prediction on an example drawn from the same distribution. [13] examined the case where an attacker has arbitrary control over some fraction β of the training examples. This work provides an interesting and useful bound on the ability to succeed at PAC-learning[1].

Following are two examples that in general. A taxonomy of different types of attacks on machine learning techniques and systems, a variety of defenses against those attacks are provide in [14]. The other general frame is from [15], which views classification as a game between the classifier and the adversary, and produces a classifier that is optimal given the adversary's optimal strategy. Experiments in a spam detection domain shows that their approach can greatly outperform a classifier learned in the standard way, and automatically adapts the classifier to the adversary's evolving manipulations.

IV. Solutions

In this section, we list existing solutions for secure machine learning according to the methodology they adopt. Obviously, the most straightforward solution is repeated, manual, ad hoc reconstruction of the learner. however, this is not a practical solution.

Beginning from the work of [15], there is an attempt to model adversarial scenarios. They develop a formal framework for representing the interaction between the classifier and the adversary, and present a classifier that is optimal given the adversary's optimal presence. A naive Bayes classifier is extended to optimally detect and reclassify tainted instances, by taking into account the adversary's optimal feature-changing strategy. Experiments in a spam detection domain show that their approach can greatly outperform a classifier learned in a standard way. Their baseline assumption is that perfect information was available to both the classifier and the adversary.

In order to relax the assumption of perfect information, [16] assumes that the adversary has the ability to issue a polynomial number membership queries to the classifier in the form of data instances which in turn will report their labels, that is to learn sufficient information about a classifier to construct adversarial attacks. They referred to their approach as Adversarial Classifier Reverse Engineering(ACRE).

Another line of methodology is to model the interaction between the adversary and the learner as a two-person sequential Stackelberg game, the adversary modifies its strategy to avoid being detected by the current learner, while the learner then updates itself based on the new threats. Each player follows their own interest in the proposed game theoretic framework: The adversary tries to maximize its return from the false negative items, and the learner tries to minimize the error cost. Research works follow this line includes [17], [18], [19], [15].

Specifically, [19] investigate the possibility of an equilibrium in this seemingly never ending game, where neither party has an inventive to change. As it is noticeable that even solving linear Stackelberg game is NP-Hard, a simulated annealing algorithm was proposed. While [17] solve this interaction between the learner and adversary by proposing a genetic algorithm for the infinite case where the players do not need to know each others's payoff function.

As [19] assumes the two players know each other's payoff function, a relaxation of this assumption is proposed in [17] that only the adversary's payoff is required in achieving the equilibrium. Comparing to the work of [19] and [17], a following work [18] go a further step relaxing the assumption that the strategies of the adversary are stochastically sampled, instead try to optimize their payoff at each step during the play. Another relaxation is that [18] doesn't make distribution assumptions on data features as [19] and [17] both assumed that data was generated from a normal distribution.

Another line of methodology is to incorporate additional cues to help build secure machine learning system. A Recommender Systems that incorporate trust is presented in [20]. The proposed Trust-aware system takes into account the "web of trust" provided by every user. The trust-aware techniques can produce a trust score for a very high number of other users; the trust score of a user estimates the relevance of that users preferences.

A robust collaborative algorithm based on Single Value Decomposition(SVD) is described in [21]. This algorithm exploited previously established SVD based shilling detection algorithms, and combined it with SVD based-CF. It combines the detective accuracy of previously established detection models based on SVD, and is also extremely accurate on rating prediction. A variety of experiments showed that attacks of different strength were rendered much weaker by VarSelect SVD. In addition, the algorithm is very stable in the face of shilling attacks.

V. Future

Three broad research directions were presented in [22]. The first was finding bounds on adversarial influence, to better understand the limits of what an attack can and cannot do to a learning system. The second was to investigate the value of adversarial capabilities, i.e. the capabilities an attacker has and how they relate to the difficulty of performing and preventing attacks. The third was to develop some technologies for secure learning.

VI. Conclusion

Though works have been conducted on developing truly secure machine learning, it is still a open problem. We first examine different kinds of attacks on machine learning systems in recent literatures. Then, we introduce different methodology in building secure machine learning systems against these attacks. As secure machine learning itself is a newly emerging subfield, much work is still under investigation. Some perspective research direction are also discussed. This paper is one of the few works that summarizes the current works towards secure machine learning and provides a reference for future research work. We intend this paper to foster discussion about it.

Acknowledgment

This work is supported by the State Key Laboratory under Grant No.CSZZ0808, the National Science and Technology Major Project No.2010ZX01036-001-002-2 and the Grand Project Network Algorithms and Digital Information of the Institute of Software, Chinese Academy of Sciences No.YOCX285056.

References

[1] Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. The security of machine learning. *Mach. Learn.*, 81:121–148, November 2010.

[2] Marco Antonio Barreno. *Evaluating the Security of Machine Learning Algorithms*. PhD thesis, University of California at Berkeley, May 2008.

[3] Bhaskar Mehta and Thomas Hofmann. A survey of attack-resistant collaborative filtering algorithms. *IEEE Data Eng. Bull.*, 31(2):14–22, 2008.

[4] Simon Chung and Aloysius Mok. Allergy attack against automatic signature generation. In Diego Zamboni and Christopher Kruegel, editors, *Recent Advances in Intrusion Detection*, volume 4219 of *Lecture Notes in Computer Science*, pages 61–80. Springer Berlin / Heidelberg, 2006.

[5] Simon Chung and Aloysius Mok. Advanced allergy attacks: Does a corpus really help? In Christopher Kruegel, Richard Lippmann, and Andrew Clark, editors, *Recent Advances in Intrusion Detection*, volume 4637 of *Lecture Notes in Computer Science*, pages 236–255. Springer Berlin / Heidelberg, 2007.

[6] James Newsome, Brad Karp, and Dawn Song. Paragraph: Thwarting signature learning by training maliciously. In Diego Zamboni and Christopher Kruegel, editors, *Recent Advances in Intrusion Detection*, volume 4219 of *Lecture Notes in Computer Science*, pages 81–105. Springer Berlin / Heidelberg, 2006.

[7] Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, and Wei-Yang Lin. Review: Intrusion detection by machine learning: A review. *Expert Syst. Appl.*, 36:11994–12000, December 2009.

[8] Prahlad Fogla and Wenke Lee. Evading network anomaly detection systems: formal reasoning and practical techniques. In *Proceedings of the 13th ACM conference on Computer and communications security*, CCS '06, pages 59–68, New York, NY, USA, 2006. ACM.

[9] Kymie M. C. Tan, Kevin S. Killourhy, and Roy A. Maxion. Undermining an anomaly-based intrusion detection system using common exploits. In *Proceedings of the 5th international conference on Recent advances in intrusion detection*, RAID'02, pages 54–73, Berlin, Heidelberg, 2002. Springer-Verlag.

[10] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 7:1–7:9, Berkeley, CA, USA, 2008. USENIX Association.

[11] Gregory L. Wittel and S. Felix Wu. On attacking statistical spam filters. In *IN PROC. OF THE CONFERENCE ON EMAIL AND ANTI-SPAM (CEAS), MOUNTAIN VIEW*, 2004.

[12] Shyong K. Lam and John Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 393–402, New York, NY, USA, 2004. ACM.

[13] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, STOC '88, pages 267–280, New York, NY, USA, 1988. ACM.

[14] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, ASIACCS '06, pages 16–25, New York, NY, USA, 2006. ACM.

[15] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 99–108, New York, NY, USA, 2004. ACM.

[16] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 641–647, New York, NY, USA, 2005. ACM.

[17] Wei Liu and S. Chawla. A game theoretical model for adversarial learning. In *Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on*, pages 25 –30, 2009.

[18] Wei Liu and Sanjay Chawla. Mining adversarial patterns via regularized loss minimization. *Mach. Learn.*, 81:69–83, October 2010.

[19] Murat Kantarcioglu, Bowei Xi, and Chris Clifton. Classifier evaluation and attribute selection against active adversaries. *Data Mining and Knowledge Discovery*, 22:291–335, 2011.

[20] Paolo Massa and Bobby Bhattacharjee. Using trust in recommender systems: An experimental analysis. In *In Proceedings of iTrust2004 International Conference*, pages 221–235, 2004.

[21] Bhaskar Mehta and Wolfgang Nejdl. Attack resistant collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 75–82, New York, NY, USA, 2008. ACM.

[22] Marco Barreno, Peter L. Bartlett, Fuching Jack Chi, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, Udam Saini, and J. D. Tygar. Open problems in the security of learning. In *Proceedings of the 1st ACM workshop on Workshop on AISec*, AISec '08, pages 19–26, New York, NY, USA, 2008. ACM.