STUDENT PERSONAL WORK

Topic: Machine Learning Methods for Computer Security

ABDOUL AZIZ HAMAYADJI : 15A773FS
KANMOGNE WABO LEONEL : 13A042FS

Dr- Ing Tchakounte Francklin

University of NGAOUNDERE 2016/2017
Systems and Networks security

1. Introduction

Nowadays, computing is becoming more and more an inescapable domain, with the birth of the internet, there is a considerable increase in data every minute. Data from exchanges of information between individuals, companies, online advertisements and system recommendations have to be sorted, categorized, classified using machine learning algorithms. Violation of security in search of compromising information, privacy mitigation ... have become a race for hackers on the Internet. These last few do not cease to develop software allowing the exploitation of vulnerability of the security, According to AVTEST [1], more than 500 333 333 examples of new malware are sighted daily to detect the companies in February 2017.

Traditional security software requires a lot of human effort to identity attacks, extract characteristics from the attacks, and encode the characteristics into software to detect the attacks. This labor-intensive process can be more efficient by applying ***machine learning algorithms*** hence the theme of secure learning is emerging as a major direction of research that offers new challenges to communities of researchers. Learning-based approaches are particularly advantageous for security applications designed to counter sophisticated and evolving adversaries because learning methods can cope with large amounts of evolving and complex data. However, the assets of learning can also potentially be subverted by malicious manipulation of data. This exposes applications that use learning techniques to a new type of security vulnerability in which an adversary can attempt to evade learning-based methods. Unlike many other application domains of machine learning, security-related applications require careful consideration of their adversarial nature and novel learning methods with improved robustness against potential attacks.

The interest in learning-based methods for security and system design applications comes from the high degree of complexity of phenomena underlying the security and reliability of computer systems. As it becomes increasingly difficult to reach the desired properties solely using statically designed mechanisms, learning methods are being used more and more to obtain a better understanding of various data collected from these complex systems. However, learning approaches can be evaded by adversaries, who change their behavior in response to the learning methods. To-date, there has been limited research into learning techniques that are resilient to attacks with provable robustness guarantees. Several priority areas of research were identified. The open problems identified in the field ranged from traditional applications of machine learning in security, such as attack detection and analysis of malicious software, to methodological issues related to secure learning, especially the development of new formal approaches with provable security guarantees. Finally a number of other potential applications were pinpointed outside of the traditional scope of computer security in which security issues may also arise in connection with data-driven methods.

2. Related works

Authors	Topics	Machine learning for security (overview)	Secure machine learning (algorithms)	Different methods for security in machine learning	Security requirements	Open issues	Year

Anthony D. Joseph;
Pavel Laskov;
Fabio Roli;
J.Doug Tygar;
Blaine Nelson
	Machine learning methods for computer security	✓	✓	✓	✓	✓	2016

Sandeep V. Sabnani Supervisor: Professor Andreas Fuchsberger	Computer Security: A Machine Learning Approach	✓	✓	-	-	✓	2016
Pavel Laskov And Marius Kloft.	A Framework for Quantitative Security Analysis of Machine Learning	-	✓	-	✓	✓	2016
Anna L. Buczak And Erhan Guven	A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection	✓	✓	-	✓	✓	2016
Tadeusz Pietraszeka And Axel Tannera	Data Mining and Machine Learning Towards Reducing False Positives in Intrusion Detection	-	✓	-	✓	✓	2016
Marco Barreno; Blaine Nelson; Anthony D. Joseph And J.D. Tygar	The security of machine learning	✓	-	✓	-	✓	2016
Philip K. Chan And LIPPMANN@LL.MIT.EDU	Machine Learning for Computer Security	✓	✓	-	-	✓	2016
Ankit Kumar Jain, And B. B. Gupta	Comparative analysis of features based machine learning approaches for phishing detection	✓	✓	✓	-	✓	2016
Michael Bailey, Jon Oberheide, Jon Andersen, Z. Morley Mao, Farnam Jahanian, and Jose Nazario.	Automated classification and analysis of internet malware. In Recent Adances in	✓	✓	-	-	-	2007

	Intrusion Detection (RAID), pages 178–197						
Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar.	The security of Machine learning. Machine Learning, 81(2):121–148,	✓	✓	✓	✓	✓	2010
Steffen Bickel and Tobias Scheffer	Dirichlet-enhanced spam filtering based on biased samples. In Neural Information Processing Systems (NIPS), pages 161–168	-	✓	-	✓	✓	2007
Battista Biggio, Giorgio Fumera, Ignazio Pillai, and Fabio Roli	A survey and experimental evaluation of image spam filtering techniques. Pattern Recognition Letters, 32:1436–1446	-	✓	✓	-	✓	2011
Nedim Šrndić and Pavel Laskov	Evasion resistant detection of malicious PDF files based on hierarchical document structure. In Network and Distributed System Security Symposium (NDSS),	-	-	✓	✓	✓	2013

✓ : This symbol denote that the specific domain is covered.

- : This symbol denote that the specific domain is not covered.

3. Methodology employed

In this article "*Machine Learning Methods for Computer Security*", the authors focused on three major points which are:

❖ The first is *machine learning for security*; here we are going to try to solve some questions like what security problems can machine learning best help to solve? What scenarios are they ill-suited for? Certain methods are used for answer at these questions especially intrusion detection based on automatic learning, in particular anomaly detection [45, 69, 71, 126], rule inference [72, 73, 74, 78] and supervised learning. Another crucial method of systems based on machine learning for computer security is the analysis of malicious software; these systems collect masses of data that will be analyzed by appropriate algorithms.

Static analysis of JavaScript token sequences has been successfully deployed for detection of drive-by-downloads [100] and JavaScript-bearing malicious PDF documents [70]. Another form of static analysis with a focus on PDF document structure was instrumental in recently developed highly effective methods for detection of PDF malware [80, 111, 115]. Dynamic analysis of JavaScript string allocations combined with classification of string payloads has been successfully used for in-browser detection of drive-by-downloads [34].

In summary, machine learning methods are currently widely used as general reactive safety architectures. They can deal optimally with security issues with clear semantics and a well-defined scope. A key advantage of learning - is the ability to generalize the information contained in the data, although such generalization may not be easily expressible in a human-readable form.

❖ The second is "*Secure machine learning*". In this section the author present different approaches, some of which are still research domains. However the biggest challenge is to develop an algorithm with robustness to worst-case noise. Two groups of methods should be noted here. There are:

- The min-max approach (i.e., training a classifier with best performance under the worst noise) results in a relatively easy optimization problem [37, 47].
- The more complex game theoretic approach is advantageous when the learner and the attacker have different cost functions; e.g., the learner is trying to minimize the spam frequency while the attacker is boosting the attractiveness of spam content. Such problems can be solved using the Nash equilibrium approach, which, however, is tractable only under some stringent assumptions [19, 21]. Compared to the baseline learning algorithms, these advanced methods improve accuracy by 10–15%, which is far from a strong claim of security.

In summary, the most progress in the study of security of machine learning has been made in identifying the security weaknesses of learning methods. While there exist methods with improved robustness against attacks, such improvements are insufficient to claim strong security. Very little attention has been given so far to the investigation of autonomous learning methods, apart from the online anomaly detection studied by [60, 61]. Finally, the limited existing methodology for security analysis, thus far, has not provided results that have improved the design of novel learning algorithms.

❖ And the third is "*Secure learning beyond security*". In this last part of the theme, the author emphasizes that analysis of the main factors, which may make specific application domains attractive for novel attacks against learning methods and survey the situation in several domains, in which learning technologies play a crucial role however several approaches and methods based machine learning algorithms are used :

- Spam filtering is the most popular example of machine learning applications that has to deal with adversarial inputs

4. Results found and Perspectives

4.1. Results found

In spite of the difficulty and the complexity encountered in major challenges, the result is mostly achieved “at least in the immediate” to solve the problem because other people will always be there to create. Several solutions have been proposed in this area but the most of them still require improvements. According to *Machine Learning Methods for Computer Security* [10] the author present among other:

- *Detection of advanced persistent threats.* The notion of “advanced persistent threats” (APT) refers to customized malware, exemplified by the Stuxnet [120], MiniDuke [54] and Flame [119], used for espionage or cyber-warfare. The goal of such malware is to penetrate highly sensitive sites and, in some cases, to remain undetected for a substantial time period. They typically rely on novel penetration vectors (e.g., zero-day exploits) and deploy extensive obfuscation techniques. Signature-based security tools are ill-suited for advanced persistent threats since they rely on previously crafted detection patterns that are too slow for targeted attacks. Machine learning methods may be better at detecting of APTs by automatically spotting anomalous events.
- *Protection of mobile devices.* Private and sensitive data stored on mobile devices is becoming an increasingly attractive target for attackers. Mobile devices have limited resources, and common security approaches, such as matching files against signatures, cannot be applied on a regular basis. Learning methods may support mobile security by enabling lightweight and signature-independent mechanisms for detecting malicious activity, for example, by identifying unusual information leakage from a device.
- *Dynamic and continuous authentication.* User authentication is typically based on secrets that can be forgotten and stolen. Authentication normally is only done once, at the beginning of a session, and grants full access rights to a given identity. We envision a more flexible authentication scheme, which is potentially less prone to lost or broken secrets, using information stored about users on systems they log into. Machine learning can use stored data to learn the behavior of legitimate users. Then, in some cases, users could authenticate simply

by means of their usual actions, thus reducing the need for or bolstering traditional password-based authentication. Machine learning could strengthen passwords with questions derived from previous user activities (e.g., by inferring which acquaintances the user should reasonably be able to recognize based on their web-browsing). It similarly could also be used to generate secret questions for resetting forgotten passwords. Finally, it could provide for continuous and incremental authentication during the course of a session based on comparing user's activities with their profiles when users request additional privileges.

- *Assisted malware analysis.* Countering malicious software is a continual arms race. More often than not, there is a significant delay between the appearance of a new malware and the availability of a malware signature. A significant amount of time during development of such signatures is devoted to analyzing and understanding the inner workings of malicious programs. Machine learning methods could accelerate this process. Techniques for feature selection and visualization could be applied to emphasize patterns within a malware's code or behaviors, and thus, these techniques could facilitate the rapid development of detection patterns.
- Computer forensics. Forensics is one of the most data-intensive areas of computer security. Here, significant benefits can also be expected from applying machine learning to evaluate data of forensic images, system logs or logged network traffic in the aftermath of a security incident. In a related field of criminal forensic investigation, similar methods can be applied to identify evidence of human crimes. In both cases, the objective of learning methods should be to help focus investigator's attention on important information related to goals of investigation and, possibly, to other similar cases, by sifting through large volumes of forensic data and prioritizing the information.

4.2. Perspectives

Recent developments in the learning methodology, and the growing experience with its application in the security practice, have underlined the necessity for deeper understanding of the security aspects of machine learning. These developments have motivated the Perspectives which have for some purpose the improvement of the algorithms of machine learning (complexity in terms of time of detections of instructions or attacks in order to improve the time of reactions)

5. Most important publication for this work

Authors	References and topics

Anthony D. Joseph;
Pavel Laskov;
Fabio Roli;
J.Doug Tygar;
Blaine Nelson
	Machine learning methods for computer security

Sandeep V. Sabnani Supervisor: Professor Andreas Fuchsberger	Computer Security: A Machine Learning Approach
Pavel Laskov And Marius Kloft.	A Framework for Quantitative Security Analysis of Machine Learning
Anna L. Buczak And Erhan Guven	A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection
Tadeusz Pietraszeka And Axel Tannera	Data Mining and Machine Learning Towards Reducing False Positives in Intrusion Detection
Marco Barreno; Blaine Nelson; Anthony D. Joseph And J.D. Tygar	The security of machine learning
Philip K. Chan And LIPPMANN@LL.MIT.EDU	Machine Learning for Computer Security