International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

Machine Learning Security

Rupali Malviya[Á*] and Brajesh K. Umrao[Á]

[Á]Department of Computer Science and Engineering, UIT, Allahabad, India

Accepted 05 Nov 2014, Available online 01 Dec2014, **Vol.4, No.6 (Dec 2014)**

Abstract

The power of Machine learning to rapidly gain through experience and evolve with changing and complex situations has helped it become an essential tool for the security of computers. However, its this pliancy is also a vulnerability. It makes the machine learning systems susceptible to attacks. The attackers can exploit machine learning systems because of its nature of adaptability. In this paper we try to analyze different attacks against machine learning systems and their solutions. We examine the contemporary work in this field and present a survey of potential attacks against machine learning systems and the defenses against these attacks.

Keywords: *Machine learning, Security, Attacks, Defenses.*

1. Introduction

The techniques of Machine learning have applications in different fields. These machine learning techniques are being applied to a large number of computer networking problems. One of the most important ones is computer security. The aim in solving such problems is to detect those systems whose behavior is suspicious or which could be of an attacker. We know that NIDS (Network Intrusion Detection Systems) are used to monitor traffic in the network for detecting abnormal activities, which may probably be attacks against the hosts or servers in the network. Machine learning techniques have been merged with NIDS and these provide the benefit that they can detect anomaly in the network. They can discover the new differences in the network traffic which might be attacks. This is done by training it on normal traffic which is non-malicious or not harmful, and also on traffic that represent attack. The conventional approach used for designing an intrusion detection system is dependent on computer experts whose work is to write rules which define normal behavior and abnormal behavior that represent intrusions (Paxson,1999). According to these rules that are written the NIDS identifies the anomalous patterns in traffics and hence detect intrusions. This approach often fails to detect novel intrusions, a number of researchers have put forward the idea to apply classification techniques of machine learning in network intrusion detection systems (Androutsopoulos *et al*,2000; Lazarevic *et al*,2003; Liao *et al*,2002; Mukkamala *et al*,2002; Wehenkal,1997; Yeung *et al*,2002). But, on thing to ponder over here is that the use of machine learning opens the door to the vulnerability. This vulnerability is in terms of a possibility that an adversary can maliciously train a machine learning system in an NIDS. This training would be better called as a mis-training, as it is done by the adversary or the attacker. A million dollar question that springs up here is that how can an adversary mis-train a system. In what ways can he harm the machine learning systems? How can he confuse a learning system? How can he exploit it for his gains? This paper surveys for answers to these questions, and also the ways in which the system can be defended against such attacks.

2. Machine learning for security

The security of our computer systems, network and data is continually at risk. The massive growth of the internet, proliferation of tools, tricks and techniques for intruding and attacking systems and networks has instigated the use of machine learning incorporated NIDS over the traditional ones. Machine learning algorithms are used for misuse detection and anomaly detection. In misuse detection training data are labeled as normal or abnormal/malicious data, and then the classifier is trained to distinguish between the two. The research work in this area incorporates the application of classification algorithms, association rule mining, and cost-sensitive modeling. Where as, the anomaly detection builds patterns or models of normal behavior and detects deviations from it as attempt to intrusions. Anomaly detection research includes application of classification algorithms, outlier analysis, clustering and statistical approaches. The techniques applied are required to be effective, efficient, scalable, robust and capable of handling high volume of data with high dimensionality and heterogeneity.

A number of anomaly detection systems are being developed based on many different machine learning techniques. Some apply single learning techniques, such as support vector machines, genetic algorithms, neural

*Corresponding author: **Rupali Malviya**

networks, etc. While some other systems are based on combining different learning techniques. These are known as hybrid or ensemble techniques. These techniques are developed as classifiers, which are used to recognize whether the incoming traffic is normal or an attack.

The research work concerned with security using machine learning technique is a vast area and still needs to be researched. In order to design more sophisticated classifiers ensemble and hybrid classifiers can be examined and combined. Since the idea of coalescing multiple classifiers is to collaborate with each other instead of contention and comparison, it is worth combining the two types for intrusion detection. The performance of Machine learning algorithms depends upon certain factors. Feature selection is one of the important factors. Since there are a number of approaches to feature selection, which approach performs best for detection of intrusion with which classification techniques is also a consideration.

3. Related Work

In (Barreno,2006) the authors have developed a theoretical account for analyzing attacks against machine learning systems and have presented a taxonomy which depicts the space of attacks against machine learning systems. In this taxonomy, the attacks against learning systems are categorized along three axes as

1. *Influence*

a. Causative - Causative attacks alter the training process through influence over the training data.
b. Exploratory - Exploratory attacks do not alter the training process but use other techniques, such as probing the learner or offline analysis, to discover information.

2. *Specificity*

a. Targeted - The specificity of an attack is a continuous spectrum. At the targeted end, the focus of the attack is on a particular point or a small set of points.
b. Indiscriminate - At the indiscriminate end, the adversary has a more flexible goal that involves a very general class of points, such as "any false negative."

3. *Security violation*

a. Integrity - An integrity attack results in intrusion points being classified as normal (false negatives).
b. Availability - An availability attack is a broader class of attack than an integrity attack. An availability attack results in so many classification errors, both false negatives and false positives, that the system becomes effectively unusable.

In causative attacks, the adversary has some measure of control over the training of the learner. An attack that causes the learner to misclassify intrusion points, for example an attack that fools an IDS into not flagging a known exploit as an intrusion, is a causative integrity attack. The distinction between targeted and indiscriminate causative integrity attacks is the difference between choosing one particular exploit or just finding any exploit. A causative availability attack causes the learner's performance to degrade. For example, an adversary might cause an IDS to reject many legitimate HTTP connections. A causative availability attack may be used to force the system administrator to disable the IDS. A targeted attack focuses on a particular service, while an indiscriminate attack has a wider scope. Exploratory attacks do not attempt to influence learning; they instead attempt to discover information about the state of the learner. Exploratory integrity attacks seek to find intrusions that are not recognized by the learner.

As far the structure organization of this paper is concerned, in the next section we present some attacks on security of machine learning. Further we discuss their defenses for building secure machine learning systems. Next we present some perspective research directions. Finally in the last section we come to the conclusion.

4. Potential Attacks on Secure learning

There are several kinds of attacks that are a threat to machine learning systems. The paper (Tan *et al*, 2002), describes an another way to circumvent intrusion detection. In this, the attacker employs a way which makes an anomaly based intrusion detection system blind towards the undergoing common attacks. It describes a technique which identifies the impuissance of intrusion detection system which is anomaly based, and demonstrates how an adversary or an attacker can exploit those impuissance.
In the paper (Fogla *et al*, 2006) the authors demonstrate a method by which an attacker can circumvent the intrusion detection system by a polymorphic blending attack. In a polymorphic blending attack i.e. PBA each polymorphic instance is created in such a way that the statistical data of packets of attacker is very similar to the profile of normal traffic. The authors in this paper have also shown that usually, the problem of generation of a PBA that matches the normal traffic profile in an optimal way is an NP-complete hard problem. In the paper (Newsome *et al*, 2006) an attack against learning is elaborated in which an attacker creates labeled samples. These labeled samples if used for training a learner, these prevent or delay for a long span of time the generation of a classifier which is accurate. The authors show that even an adversary, who has samples which are all correctly labeled, can obstruct the learning process. By simulation they have implemented these attacks against the Polygraph automatic polymorphic worm signature generation algorithms. In (Chung *et al*,2006), (Chung *et al*, 2007) two examples of signature generation manifest the practical effect of allergy attacks, in which the attackers manipulate the system that generates signature and make it an active agent for denial of service attack against the system which is protected. In the paper (Perdisci *et al*,2006), the authors have shown that if noise is introduced intentionally to mislead a worm signature generator, a much lower noise level can prevent the system from generating useful worm signatures. The authors have described a new and general class of attacks through which a worm can combine polymorphism and misleading behavior for deliberately

polluting the dataset of suspicious flows during its propagation and, thus, mislead the automatic signature generation process. The authors also suggest that unless and until an accurate and robust flow classification process is there, automatic syntactic-based signature generators are vulnerable to such kind of noise injection attacks.

The authors in (Wittel,2004) have analysed the attack methods which the spammers used generally. They have also demonstrated an attack which despite of being easy in implementation, tries to work against the statistical nature of filters more powerfully. In (Nelson *et al*, 2008), the authors present a way in which an adversary takes the advantage of statistical machine learning, as is used in the SpamBayes spam filter, to make it futile. They have also presented a new class of focused attacks which successfully hinder the receiving of specific email messages to the victims. The authors in (Dalvi *et al*,2004), view classification as a game between the attacker and the classifier. It produces an optimal classifier given the optimal strategy of the adversary. The experiments in the domain of spam detection show that their approach can surpass a classifier that has learnt in a standard way. It also automatically adjusts the classifier to the evolving manipulations of the adversary. In the paper (Huang *et al*, 2011) the authors have given a taxonomy for classifying attacks against online machine learning algorithms. The authors have discussed some application specific factors which limit the capabilities of an adversary. They have also given two models for capabilities of an adversary. They have explored the limits of an adversary's knowledge about the algorithm, the feature space, the training and input data. In addition to this they have also explored some vulnerabilities in machine learning algorithms. They have also discussed some countermeasures against attacks, introduced the evasion challenge and have also discussed learning techniques that preserve privacy.

Recommender System is an application of machine learning. The exponential growth in usage of recommender systems has lead to user feedback of varying quality. The genuine users express their true opinion, where as some naughty provide noisy ratings which degrade the quality of the recommendations that are generated based upon them. The inclusion of noise can contradict assumptions made for modeling and finally may lead to deviation in estimated and predicted results. Similarly the attackers as users can intentionally inject wrong opinion or data to bias the results of the recommender system for their own benefit and harm the victims.

5. Defenses

The problem of machine learning security is severe. But there are certain solutions that several researchers have worked upon. These solutions provide the defenses against attacks for machine learning security according to the methodology that they have adopted. In the paper (Dalvi *et al*,2004), the authors have attempted to model the adversarial scenario. They have given a framework which shows the interaction between the adversary and the classifier. They have shown an optimal classifier given the optimal presence of the adversary. They have extended the NaiveBayes classifier to detect and reclassify sullied instances in an optimal way. This takes into account the optimal feature-changing strategy of the adversary. The experiments that have been performed for detection of spam have shown that their approach is capable of surpassing a classifier which has learned in standard way. Here the supposition is perfect availability of information to both the attacker and the classifier. In the paper (Lazarevic,2003) the authors have loosened up the supposition of perfect information and in turn have assumed that the attacker is capable of issueing a polynomial number membership queries to the classifier. This is in the form of data instances that will report their labels, which is to learn ample amount of information about a classifier for the construction of adversarial attacks. The authors have called their approach as ACRE i.e. Adversarial Classifier Reverse Engineering. A different approach is to model the adversary and the learner interaction as a two entity sequential Stackelberg game. As per this approach the attacker modifies its strategy for avoiding to be detected by the learner, meanwhile the learner updates itself depending upon the novel threats. In this model, every player plays by his own interest. The attacker tries to get the most out of its return from the items that are false negative, on the other hand the learner endeavours to minimize the cost of error. In the paper (Kantarcioglu *et al*,2011) the authors examine the presence of an equilibrium, if possible, in this apparantly never ending game in which neither the adversary nor the learner seems to change. The linear Stackelberg game is NP-Hard problem. A simulated annealing algorithm is proposed. The authors in (Liu *et al*,2009) have worked out on this adversary and learner interaction and have put forward a genetic algorithm. This algorithmic approach is for the infinite case in which the players are not required to have information regarding one another's payoff function. While the authors in (Kantarcioglu *et al*,2011) have supposed that the two players i.e. the adversary/attacker and the learner have information about one another's payoff function, the authors in (Liu *et al*, 2009) have relaxed this supposition. The authors in (Liu *et al*,2009) have supposed that only payoff function of the adversary is needed in order to accomplish equilibrium. There is also an another approach proposed by authors in (Liao *et al*,2002). In this paper the authors have put forward the idea of relaxing the supposition according to which the adversary's strategies are sampled stochastically, which is contrary to the idea of optimizing payoff in each step. One more thing to point out here is that in (Liao *et al*,2002) the authors do not make distribution assumptions on features of data. Whereas, the authors in (Kantarcioglu *et al*,2011) and (Liu *et al*,2009) have assumed that the data has been is from a normal distribution.

Taking into consideration some more defense techniques incorporates some other mechanisms for security of machine learning systems. For security of recommender systems, authors in (Massa *et al*,2004) have proposed recommender systems which integrate trust. This is a trust aware system. The trust aware mechanism produces a trust score for high number of other users. A

user's trust score estimates or figures out the relevancy of the preferences of that user. In (Mehta *et al*, 2008) the authors have presented a collaborative algorithm that has been established on SVD i.e. Single Value Decomposition. The authors have exploited the already established SVD based shilling detection algorithm and have integrated it with SVD based CF. This algorithm has merged the detective accuracy of already established detection models that are based upon SVD. It is also accurate on rating the predictions. Many experiments have shown that different attacks of varying strength were rendered weaker by VarSelect SVD.

6. Future Scope

In (Barreno,2008) there could be several possible research directions. One possible task of research is to find bounds on the influence of the adversary. This is in order to understand the limits of what an attack can and cannot do to a learning system. Another possible idea of research is to investigate the value of adversarial capabilities. These are the capabilities which an attacker has and how they associate or logically relate to the problem of attack prevention. Further the focus of research work is development of novel techniques for machine learning security. In addition to this alleviating the problems in existing mechanisms of security is also a research direction.

Conclusion

Although a lot of work has been done for dealing with the problem of machine learning security, it is still a state of difficulty that needs to be resolved. We first analysed different kinds of attacks possible on machine learning systems described in current literatures. Next we presented different approaches against these attacks for security of machine learning systems. While security of machine learning systems is an emerging field of study, a lot of work is still to be done. We have also suggested some research ideas. This paper surveys and summarizes the work done for machine learning security in present times.

References

Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropolous(2000). An evaluation of naive Bayesian anti-spam filtering. *Proceedings of the Workshop on Machine Learning in the New Information Age*, pages 9–17.

V. Paxson(1999), Bro: a system for detecting network intruders in real-time, *Computer Networks: the International Journal of Distributed Informatique*, 31(23):2435–2463, Dec.

A. Lazarevic, L. Ert¨oz, V. Kumar, A. Ozgur, and J. Srivastava(2003), A comparative study of anomaly detection schemes in network intrusion detection. *Proceedings of the Third SIAM International Conference on Data Mining, May*.

Y. Liao and V. R. Vemuri(2002), Using text categorization techniques for intrusion detection, *Proceedings of the 11th USENIX Security Symposium, pages 51–59, Aug*.

S. Mukkamala, G. Janoski, and A. Sung(2002), Intrusion detection using neural networks and support vector machines, *Proceedings of the International Joint Conference on Neural Networks (IJCNN'02)*, pages 1702–1707.

L. Wehenkel(1997), Machine learning approaches to power system security assessmenet. *IEEE Intelligent Systems and Their Applications*, 12(5):60–72, Sept.–Oct.

D.-Y. Yeung and C. Chow(2002), Parzen-window network intrusion detectors, *Proceedings of the Sixteenth International Conference on Pattern Recognition*, pages 385–388, Aug.

Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar(2010), The security of machine learning. *Mach. Learn.*, 81:121–148, Nov.

Simon Chung and Aloysius Mok(2006), Allergy attack against automatic signature generation, *Recent Advances in Intrusion Detection*, volume 4219 of *Lecture Notes in Computer Science*, pages 61–80, Springer Berlin/ Heidelberg.

Simon Chung and Aloysius Mok(2007), Advanced allergy attacks: Does a corpus really help? *Recent Advances in Intrusion Detection*, volume 4637 of *Lecture Notes in Computer Science*, pages 236–255. Springer Berlin/ Heidelberg.

James Newsome, Brad Karp, and Dawn Song(2006),Paragraph: Thwarting signature learning by training maliciously, *Recent Advances in Intrusion Detection, of Lecture Notes in Computer Science*, volume 4219 pages 81–105. Springer Berlin/ Heidelberg.

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

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

Gregory L. Wittel and S. Felix Wu(2004), On attacking statistical spam filters, *proceedings of the conference on email and anti-spam (ceas)*, mountain view.

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

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

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

Wei Liu and S. Chawla(2009), A game theoretical model for adversarial learning, *Data Mining Workshop ICDMW '09, IEEE International Conference on*, pages 25–30.

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

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

Xiaofeng Liao, Liping Ding Yongji Wang(2011),Secure Machine Learning, A Brief Overview, *Fifth International Conference on Secure Software Integration and Reliability Improvement – Companion, IEEE*.

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

Roberto Perdisci, David Dagon, WenkeLee, Prahlad Fogla, Monirul Sharif(2006), Misleading worm signature generators using deliberate noise injection, *Security and Privacy, IEEE Symposium*, 15 pp. – 31.

Ling Huang, Antony D. joseph, Blaine Nelson, Benjamin Rubinstein, J. D. Tygar(2011), Adversarial machine learning, *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43-58, ACM, New York.