ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

A Survey on: Security Evaluation of Pattern Classifiers under Attack

Kale Tai. , Prof. Bere S. S.

P.G. Scholar, Dept. of Information Technology, DGOI, FOE, Bhigwan, Savitribai Phule University of Pune, Pune, India

Professor, Dept. of Computer Engineering, DGOI, FOE, Bhigwan, Savitribai Phule University of Pune, Pune, India

ABSTRACT: The systems which can be used for pattern classification are used in adversarial application, for example spam filtering, network intrusion detection system, biometric authentication. This adversarial scenario's exploitation may sometimes affect their performance and limit their practical utility. In case of pattern classification conception and contrive methods to adversarial environment is a novel and relevant research direction, which has not yet pursued in a systematic way. To address one main open issue: evaluating at contrive phase the security of pattern classifiers (for example the performance degradation under potential attacks which incurs during the operation). To propose a framework for evaluation of classifier security and also this framework can be applied to different classifiers on one of the application from the spam filtering,biometric authentication andnetwork intrusion detection.

KEYWORDS: Machine learning system, Security evaluation, Adversarial classification, Arms-Race, Spam Filtering.

I.INTRODUCTION

Machine learning systems provide pliability relating with unfolding the input in a number of applications. Machine learning techniques are applied to a growing number of systems and networking problems, particularly those problems where the intention is to discern anomalous system behavior. For instance, Network Intrusion Detection Systems (NIDS) monitor network traffic to discern abnormal movements, such as attacks against hosts or servers. Machine learning is used to prevent unlawful or unsanctioned activity which is created from the adversary. Machine learning is used in security affiliated functions bring in a classification, such as intrusion detection systems, spam filters, biometric authentication, etc. Measuring the security performance of classifiers is an important part in facilitating decision making. As spam filters evolve to better classify spam, spammers canadapt their messages to avoid detection [1].
The input data can be manipulated by an adversary to compose classifiers to produce false negative. This frequently brings about an arms race in the middle of the adversary and the classifier designer.In the case of the arms-race problem in pursuing the security it is not enough to retort to observed attacks. There is some open issues which can be identified: (i) development of methods which assess the security of classifier against the attacks (ii) Analysis of vulnerabilities and corresponding attacks of classification [1].
The security in Machine Learning Systems besides of spam filtering (spam e-mails) and network intrusion detection systems that is NIDS. The Machine learning systems have been employed in different number of applications which contains Online Deputy Systems (ODS), Clump Supervising (cluster monitoring), and toxin detection same as virus detection and some dynamic operations applications. There are some algorithms with accurate performance in the case of adversarial condition like Secure Learning Algorithms [2]. Some Classifiers are utilized to generate some contrasts which promote security intention. For example, the intention of a toxin (virus) detection system is to diminish vulnerabilities. The toxins (virus) give antecedent to contamination or by detecting the contamination. An adversary's attempt to procure the data which are nothing but the domestic state of a Machine Learning System (MLS) to- (i) infuse the personal data which is encrypted in its domestic state otherwise (ii) originate the data which sanction the adversary to effectually onslaught the system [2].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

II.LITERATURE SURVEY

"R.N. Rodrigues, L.L. Ling, and V. Govindaraju" Proposed [1] that, we address the security of multimodal biometric systems when one of the modes is successfully spoofed. We propose two novel fusion schemes that can increase the security of multimodal biometric systems. The first is an extension of the likelihood ratio based fusion scheme and the other uses fuzzy logic. Besides the matching score and sample quality score, our proposed fusion schemes also take into account the intrinsic security of each biometric system being fused. Experimental results have shown that the proposed methods are more robust against spoof attacks when compared with traditional fusion methods [1].

"P. Johnson, B. Tan, and S. Schuckers" Proposed [2] that biometric systems, the threat of "spoofing", where an imposter will fake a biometric trait, have lead to the increased use of multimodal biometric systems. It is assumed that an imposter must spoof all modalities in the system to be accepted. This paper looks at the cases where some but not all modalities are spoofed. The contribution of this paper is to outline a method for assessment of multimodal systems and underlying fusion algorithms. The framework for this method is described and experiments are conducted on a multimodal database of face, iris, and fingerprint match scores [2].

"P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee" Proposed [3] that A very effective means to evade signature-based intrusion detection systems (IDS) is to employ polymorphic techniques to generate attack instances that do not share a fixed signature. Anomaly-based intrusion detection systems provide good defence because existing polymorphic techniques can make the attack instances look different from each other, but cannot make them look like normal. In this paper we introduce a new class of polymorphic attacks, called polymorphic blending attacks, that can effectively evade byte frequency-based network anomaly IDS by carefully matching the statistics of the mutated attack instances to the normal profiles. The proposed polymorphic blending attacks can be viewed as a subclass of the mimicry attacks. We take a systematic approach to the problem and formally describe the algorithms and steps required to carry out such attacks. We not only show that such attacks are feasible but also analyse the hardness of evasion under different circumstances. We present detailed techniques using PAYL, a byte frequency-based anomaly IDS, as a case study and demonstrate that these attacks are indeed feasible. We also provide some insight into possible countermeasures that can be used as defence [3].

"G.L. Wittel and S.F. Wu" Proposed [4] that the efforts of anti-spammers and spammers have often been described as an arms race. As we devise new ways to stem the flood of bulk mail, spammers respond by working their way around the new mechanisms. Their attempts to bypass spam filters illustrate this struggle. Spammers have tried many things from using HTML layout tricks, letter substitution, to adding random data. While at times their attacks are clever, they have yet to work strongly against the statistical nature that drives many filtering systems. The challenges in successfully developing such an attack are great as the variety of filtering systems makes it less likely that a single attack can work against all of them. Here, we examine the general attack methods spammers' use, along with challenges faced by developers and spammers. We also demonstrate an attack that, while easy to implement, attempts to more strongly work against the statistical nature behind filters [4].

"D. Lowd and C. Meek" Proposed [5] that Unsolicited commercial email is a significant problem for users and providers of email services. While statistical spam filters have proven useful, senders of spam are learning to bypass these filters by systematically modifying their email messages. In a good word attack, one of the most common techniques, a spammer modifies a spam message by inserting or appending words indicative of legitimate email. In this paper, we describe and evaluate the effectiveness of active and passive good word attacks against two types of statistical spam filters: naive Bayes and maximum entropy filters. We find that in passive attacks without any filter feedback, an attacker can get 50 % of currently blocked spam past either filter by adding 150 words or fewer. In active attacks allowing test queries to the target filter, 30 words will get half of blocked spam past either filter [5].

Existing System:

- Pattern classification systems based on classical theory and design methods do not take into account adversarial settings; they exhibit vulnerabilities to several potential attacks, allowing adversaries to undermine their effectiveness [1]. A systematic and unified treatment of this issue is thus needed to allow the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

trusted adoption of pattern classifiers in adversarial environments, starting from the theoretical foundations up to novel design methods, extending the classical design cycle of. In particular, three main open issues can be identified: (i) analyze the vulnerabilities of classification algorithms, and the corresponding attacks. (ii) Developing novel methods to assess classifier security against these attacks, which are not possible using classical performance evaluation methods. (iii) Developing novel design methods to guarantee classifier security in adversarial environments [1].

Disadvantages of Existing System:

1. Poor analysing the vulnerabilities of classification algorithms, and the corresponding attacks.

2. A malicious webmaster may manipulate search engine rankings to artificially promote website.

III.PROPOSED ALGORITHM

System Architecture:

The main goal is to scrutinize buttress which is difficult to represent to escape the design classifiers in the Adversarial Classification Problems with the help of the framework. An artifice for providing the security for classifier designer is to mask the data to the Adversary. A feasible fulfilment of this artifice was predicted with some soft contention which gives the identity of haphazardness in the location of classification boundaries.In the case of Arms-race, it is not possible to recommend how many and what type of attacks a classifier will incur during operation, the classifier security should proactively evaluate using a what-if analysis, by simulating potential attack scenarios.The primary goal is to formulate or model the adversary as the optimization of an actual function. The effective simulation of attack scenarios requires a formal model of the adversary. In many cases, according to the knowledge of classifier and capability of manipulation of data, the adversary acts rationally to attain a goal of security evaluation. Define the adversary in terms of attack influence (exploratory or causative feature manipulation, control on training and testing samples. Quantitative discussion on training data feature set, the learning algorithm data, classifier's decision function, feedback from classifier and some assumptions regarding on the application at hand [1].

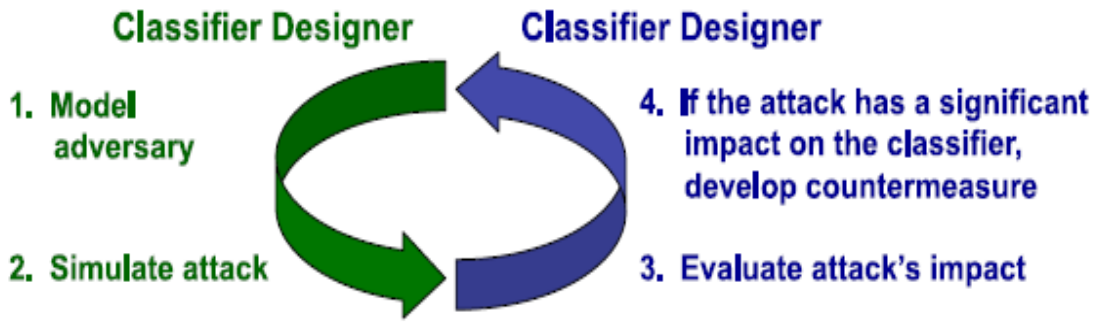

Figure 1: System Architecture of Proposed System

Modules:-

- ✓ Attack Scenario and Model of the Adversary
- ✓ Pattern Classification
- ✓ Adversarial classification:
- ✓ Security modules
- ✓

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

1. **Attack Scenario and Model of the Adversary:**
   Although the definition of attack scenarios is ultimately an application-specific issue, it is possible to give general guidelines that can help the designer of a pattern recognition system. Here we propose to specify the attack scenario in terms of a conceptual model of the adversary that encompasses, unifies, and extends different ideas from previous work. Our model is based on the assumption that the adversary acts rationally toattain a given goal, according to her knowledge of the classifier, and her capability of manipulating data. This allows one to derive the corresponding optimal attack strategy [1].

2. **Pattern Classification:**
   Multimodal biometric systems for personal identity recognition have received great interest in the past few years. It has been shown that combining information coming from different biometric traits can overcome the limits and the weaknesses inherent in every individual biometric, resulting in a higher accuracy. Moreover, it is commonly believed that multimodal systems also improve security against Spoofing attacks, which consist of claiming a false identity and submitting at least one fake biometric trait to the system (e.g., a "gummy" fingerprint or a photograph of a user's face). The reason is that, to evade multimodal system, one expects that the adversary should spoof all the corresponding biometric traits. In this application example, we show how the designer of a multimodal system can verify if this hypothesis holds, before deploying the system, by simulating spoofing attacks against each of the matchers [1].

3. **Adversarial classification:**
   Assume that a classifier has to discriminate between legitimate and spam emails on the basis of their textual content, and that the bag-of-words feature representation has been chosen, with binary features denoting the occurrence of a given set of words [1].

4. **Security modules:**
   Intrusion detection systems analyze network traffic to prevent and detect malicious activities like intrusion attempts, ROC curves of the considered multimodal biometric system under a simulated spoof attack against the fingerprint or the face matcher. Port scans, and denial-of-service attacks. When suspected malicious traffic is detected, an alarm is raised by the IDS and subsequently handled by the system administrator. Two main kinds of IDSs exist: misuse detectors and anomaly-based ones. Misuse detectors match the analyzed network traffic against a database of signatures of known malicious activities. The main drawback is that they are not able to detect never-before-seen malicious activities, or even variants of known ones. To overcome this issue, anomaly-based detectors have been proposed. They build a statistical model of the normal traffic using machine learning techniques, usually one-class classifiers, and raise an alarm when anomalous traffic is detected. Their training set is constructed, and periodically updated to follow the changes of normal traffic, by collecting unsupervised network traffic during operation, assuming that it is normal (it can be filtered by a misuse detector, and should) [1].

Advantages:

1) Proposed system prevents developing novel methods to assess classifier security against these attacks.
2) The presence of an intelligent and adaptive adversary makes the classification problem highly non-stationary.

IV. CONCLUSION AND FUTURE WORK

This paper presented an overview of work related to the security of pattern classification systems with the goal of imparting useful guidelines on how to improve their design and assess their security specific attacks. Also the paper focused on innovative security evaluation of pattern classifiers that deployed in adversarial environments. Main contribution is a framework for verifiable security evaluation that construes and establishes the notion from previous work, and can be utilized to different classifiers, learning algorithms, and classification tasks.
In the future, clustering methods can be integrated with the existing technique in order to get better results. Further, this approach can be applied to the application which makes the classification problem highly non-stationary.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

REFERENCES

[1] R.N. Rodrigues, L.L. Ling, and V. Govindaraju, "Robustness of Multimodal Biometric Fusion Methods against Spoof Attacks," J. Visual Languages and Computing, vol. 20, no. 3, pp. 169-179, 2009.
[2] P. Johnson, B. Tan, and S. Schuckers, "Multimodal Fusion Vulnerability to Non-Zero Effort (Spoof) Imposters," Proc. IEEE Int'l Workshop Information Forensics and Security, pp. 1-5, 2010.
[3] P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee, "Polymorphic Blending Attacks," Proc. 15th Conf. USENIX Security Symp., 2006.
[4] G.L. Wittel and S.F. Wu, "On Attacking Statistical Spam Filters," Proc.First Conf. Email and Anti-Spam, 2004.
[5] D. Lowd and C. Meek, "Good Word Attacks on Statistical Spam Filters," Proc. Second Conf. Email and Anti-Spam, 2005.
[6] A. Kolcz and C.H. Teo, "Feature Weighting for Improved Classifier Robustness," Proc. Sixth Conf. Email and Anti-Spam, 2009.
[7] D.B. Skillicorn, "Adversarial Knowledge Discovery," IEEE Intelligent Systems, vol. 24, no. 6, Nov./Dec. 2009.
[8] D. Fetterly, "Adversarial Information Retrieval: The Manipulation of Web Content," ACM Computing Rev., 2007.
[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. Wiley-Interscience Publication, 2000.
[10] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial Classification," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 99-108, 2004.
[11] M. Barreno, B. Nelson, R. Sears, A.D. Joseph, and J.D. Tygar, "Can Machine Learning be Secure?" Proc. ACM Symp.Information, Computer and Comm. Security (ASIACCS), pp. 16-25, 2006.
[12] A.A. C_ardenas and J.S. Baras, "Evaluation of Classifiers: Practical Considerations for Security Applications," Proc. AAAI Workshop Evaluation Methods for Machine Learning, 2006.
[13] P. Laskov and R. Lippmann, "Machine Learning in Adversarial Environments," Machine Learning, vol. 81, pp. 115-119, 2010.
[14] L. Huang, A.D. Joseph, B. Nelson, B. Rubinstein, and J.D. Tygar, "Adversarial Machine Learning," Proc. Fourth ACM Workshop Artificial Intelligence and Security, pp. 43-57, 2011.
[15] M. Barreno, B. Nelson, A. Joseph, and J. Tygar, "The Security of Machine Learning," Machine Learning, vol. 81, pp. 121-148, 2010.
[16] D. Lowd and C. Meek, "Adversarial Learning," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 641-647, 2005.
[17] P. Laskov and M. Kloft, "A Framework for Quantitative Security Analysis of Machine Learning," Proc.Second ACM Workshop Security and Artificial Intelligence, pp. 1-4, 2009.
[18] NIPS Workshop Machine Learning in Adversarial Environments for Computer Security, http://mls-nips07.first.fraunhofer.de/, 2007.
[19] Dagstuhl Perspectives Workshop Mach. Learning Methods for Computer Sec., http://www.dagstuhl.de/12371/, 2012.
[20] A.M. Narasimhamurthy and L.I. Kuncheva, "A Framework for Generating Data to Simulate Changing Environments," Proc. 25th Conf. Proc. the 25th IASTED Int'l Multi-Conf.: Artificial Intelligence and Applications, pp. 415-420, 2007.
[21] S. Rizzi, "What-If Analysis," Encyclopedia of Database Systems, pp. 3525-3529, Springer, 2009.
[22] J. Newsome, B. Karp, and D. Song, "Paragraph: Thwarting Signature Learning by Training Maliciously," Proc. Ninth Int'l Conf. Recent Advances in Intrusion Detection, pp. 81-105, 2006.
[23] A. Globerson and S.T. Roweis, "Nightmare at Test Time: Robust Learning by Feature Deletion," Proc. 23rd Int'l Conf. Machine Learning, pp. 353-360, 2006.
[24] R. Perdisci, G. Gu, and W. Lee, "Using an Ensemble of One-Class SVM Classifiers to Harden Payload-Based Anomaly Detection Systems," Proc. Int'l Conf. Data Mining, pp. 488-498, 2006.
[25] S.P. Chung and A.K. Mok, "Advanced Allergy attacks: Does a Corpus Really Help," Proc. 10th Int'l Conf. Recent Advances in Intrusion Detection (RAID '07), pp. 236-255, 2007.