SECURITY EVALUATION OF PATTERN CLASSIFIERS UNDER ATTACK

Tamiri Bhavani Phani Abhishek, Miss G.Keerthana, Sri.V.Bhaskara Murthy,
MCA Student, Assistant Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

ABSTRACT Pattern classification systems are commonly used in adversarial applications, like biometric authentication, network intrusion detection, and spam filtering, in which data can be purposely manipulated by humans to undermine their operation. As this adversarial scenario is not taken into account by classical design methods, pattern classification systems may exhibit vulnerabilities, whose exploitation may severely affect their performance, and consequently limit their practical utility. Extending pattern classification theory and design methods to adversarial settings is thus a novel and very relevant research direction, which has not yet been pursued in a systematic way. In this paper, we address one of the main open issues: evaluating at design phase the security of pattern classifiers, namely, the performance degradation under potential attacks they may incur during operation. We propose a framework for empirical evaluation of classifier security that formalizes and generalizes the main ideas proposed in the literature, and give examples of its use in three real applications. Reported results show that security evaluation can provide a more complete understanding of the classifier's behavior in adversarial environments, and lead to better design choices

I. INTRODUCTION

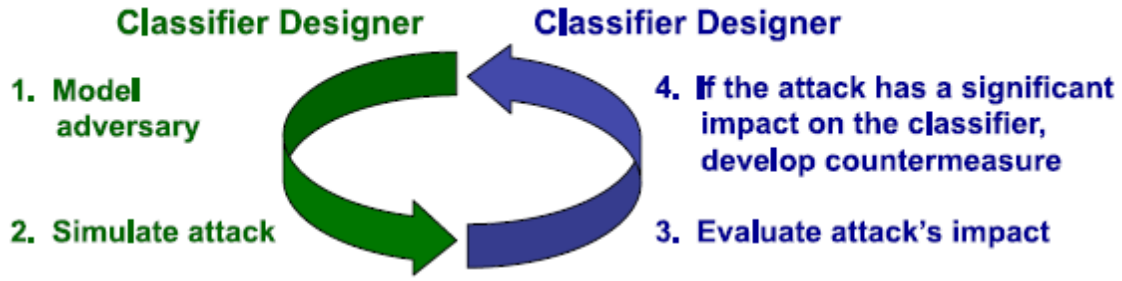
PATTERN classification systems based on machine learning algorithms are commonly used in security-related applications like biometric authentication, network intrusion detection, and spam filtering, to discriminate between a "legitimate" and a "malicious" pattern class (e.g., legitimate and spam emails). Contrary to traditional ones, these applications have an intrinsic adversarial nature since the input data can be purposely manipulated by an intelligent and adaptive adversary to undermine classifier operation. This often gives rise to an arms race between the adversary and the classifier designer. Well known examples of attacks against pattern classifiers are: submitting a fake biometric trait to a biometric authentication system (spoofing attack) [1], [2]; modifying network packets belonging to intrusive traffic to evade intrusion detection systems (IDSs) [3]; manipulating the content of spam emails to get them past spam filters (e.g., by misspelling common spam words to avoid their detection) [4], [5], [6]. Adversarial scenarios can also occur in intelligent data analysis [7] and information retrieval [8]; e.g., a malicious webmaster may manipulate search engine rankings to artificially promote her1 website.

It is now acknowledged that, since pattern classification systems based on classical theory and design methods [9] do not take into account adversarial settings, they exhibit vulnerabilities to several potential attacks, allowing adversaries to undermine their effectiveness [6], [10], [11], [12], [13], [14]. A systematic and unified treatment of this issue is thus needed to allow the trusted adoption of pattern classifiers in adversarial environments, starting from the theoretical foundations up to novel design methods, extending the classical design cycle of [9]. In particular, three main open issues can be identified: (i) analyzing the vulnerabilities of classification algorithms, and the corresponding attacks [11], [14], [15]; (ii) developing novel methods to assess classifier security against these attacks, which is not possible using classical performance evaluation methods [6], [12], [16], [17]; (iii) developing novel design methods to guarantee classifier security in adversarial environments [1], [6], [10].

Although this emerging field is attracting growing interest [13], [18], [19], the above issues have only been sparsely addressed under different perspectives and to a limited extent. Most of the work has focused on application-specific issues related to spam filtering and network intrusion detection, e.g., [3], [4], [5], [6], [10], while only a few theoretical models of adversarial classification problems have been proposed in the machine learning literature [10], [14], [15]; however, they do not yet provide practical guidelines and tools for designers of pattern recognition systems.

Besides introducing these issues to the pattern recognition research community, in this work we address issues (i) and (ii) above by developing a framework for the empirical evaluation of classifier security at design phase that extends the model selection and performance evaluation steps of the classical design cycle of [9].

II. SYSTEM ARCHITECTURE



III. 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. A systematic and unified treatment of this issue is thus needed to allow the 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.

DISADVANTAGES OF EXISTING SYSTEM:

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

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

IV. PROPOSED SYSTEM

In this work we address issues above by developing a framework for the empirical evaluation of classifier security at design phase that extends the model selection and performance evaluation steps of the classical design cycle .We summarize previous work, and point out three main ideas that emerge from it. We then formalize and generalize them in our framework. First, to pursue security in the context of an arms race it is not sufficient to react to observed attacks, but it is also necessary to proactively anticipate the adversary by predicting the most relevant, potential attacks through a what-if analysis; this allows one to develop suitable countermeasures before the attack actually occurs, according to the principle of security by design. Second, to provide practical guidelines for simulating realistic attack scenarios, we define a general model of the adversary, in terms of her goal, knowledge, and capability, which encompass and generalize models proposed in previous work. Third, since the presence of carefully targeted attacks may affect the distribution of training and testing data separately, we propose a model of the data distribution that can formally characterize this behaviour, and that allows us to take into account a large number of potential attacks; we also propose an algorithm for the generation of training and testing sets to be used for security evaluation, which can naturally accommodate application-specific and heuristic techniques for simulating attacks.

ADVANTAGES OF PROPOSED SYSTEM:

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.

V. IMPLEMENTATION

MODULES:

1. Attack Scenario and Model of the Adversary
2. Pattern Classification
3. Adversarial classification:
4. Security modules

MODULES DESCRIPTION:

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 to attain 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.

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.

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.

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)

VI. CONCLUSION

In this paper we focused on empirical security evaluation of pattern classifiers that have to be deployed in adversarial environments, and proposed how to revise the classical performance evaluation design step, which is not suitable for this purpose.

Our main contribution is a framework for empirical security evaluation that formalizes and generalizes ideas from previous work, and can be applied to different classifiers, learning algorithms, and classification tasks. It is grounded on a formal model of the adversary, and on a model of data distribution that can represent all the attacks considered in previous work; provides a systematic method for the generation of training and testing sets that enables security evaluation; and can accommodate application-specific techniques for attack simulation. This is a clear advancement with respect to previous work, since without a general framework most of the proposed techniques (often tailored to a given classifier model, attack, and application) could not be directly applied to other problems.

An intrinsic limitation of our work is that security evaluation is carried out empirically, and it is thus datadependent; on the other hand, model-driven analyses [12], [17], [38] require a full analytical model of the problem and of the adversary's behavior, that may be very difficult to develop for real-world applications. Another intrinsic limitation is due to fact that our method is not application-specific, and, therefore, provides only high-level guidelines for simulating attacks. Indeed, detailed guidelines require one to take into account application-specific constraints and adversary models. Our future work will be devoted to develop techniques for simulating attacks for different applications.

Although the design of secure classifiers is a distinct problem than security evaluation, our framework could be also exploited to this end. For instance, simulated attack samples can be included into the training data to improve security of discriminative classifiers (e.g., SVMs), while the proposed data model can be exploited to design more secure generative classifiers. We obtained encouraging preliminary results on this topic

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.