ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Evaluating at Design Phase the Security of Pattern Classifiers

A.Raghu
M.Tech Student,
Department of CSE,
Global Institute of Engineering and Technology,
Chilkur (V), RR District,Telangana.

Mrs. M.Jhansi Lakshmi
Associate professor,
Head of the Department CSE,
Global Institute of Engineering and Technology,
Chilkur (V), RR District,Telangana.

Abstract:

Pattern classification is a branch of machine learning that focuses on recognition of patterns and regularities in data. In adversarial applications like biometric authentication, spam filtering, network intrusion detection the pattern classification systems are used 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.

Index Terms:

Pattern classification, adversarial classification, performance evaluation, security evaluation, robustness evaluation.

INTRODUCTION:

Systems of pattern classification on the basis of machine learning algorithms are used in security associated applications to discriminate among a genuine as well as a malevolent pattern class. In opposition to traditional systems, these applications have a fundamental adversarial nature as input data is manipulated by an intelligent as well as adaptive adversary to destabilize classifier operation .

Adversarial situations take place in intelligent data analysis as well as information retrieval. Since pattern classification systems on basis of classical theory and design methods do not consider adversarial settings, they display vulnerabilities to quite a lot of potential attacks, permitting adversaries to undermine their efficiency. The majority of works were spotlighted on application-specific issues associated to spam filtering and network intrusion detection. While not many theoretical representations of adversarial classification problems were proposed in machine learning literature on the other hand, they do not recommend practical guidelines for designers of systems of pattern recognition.

Most significant open issues can be recognized such as analyzing susceptibility of classification algorithms, developing new methods to consider classifier security against these attacks, which are not likely using classical performance evaluation methods and developing new design methods to assurance classifier security in adversarial environments. We put forward a framework for empirical assessment of classifier security that generalizes most important ideas that are projected in the literature and can be functional to different classifiers, learning algorithms, as well as classification tasks. It is grounded on formal representation of adversary and on a representation of data distribution that corresponds to the entire attacks considered in earlier work; offers a systematic system for generation of training and testing sets that facilitate security evaluation; and holds application-specific methods for attack simulation.

METHODOLOGY:

Taxonomy of possible attacks against pattern classifiers was projected which is based on two most important features such as category of influence of attacks on classifier, as well as type of security violation they cause learning algorithm to cause succeeding misclassifications; if it exploits knowledge of trained classifier to cause misclassifications, devoid of affecting learning algorithm.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Causative attacks might influence training as well as testing data, or else only training data, whereas exploratory attacks have an effect on only testing data. The security violation is an integrity violation, if it permits adversary to access service protected by classifier; an accessibility violation, if it denies lawful users access to it; or else a privacy violation, if it permits adversary to get hold of secret information from the classifier. Security problems regularly guide towards a reactive arms race among the adversary and classifier designer. At every step, adversary analyzes classifier defences, and expands an attack scheme to prevail over them. The designer act in response by means of analyzing new attack samples, and, if necessary, updates classifier; by retraining it on recent collected samples, and features that can notice novel attacks. To protect a system, a general approach that is used in engineering as well as cryptography is security by means of obscurity that keeps secret some of system details towards adversary [3].

Concept of security by design advocate that systems have to be designed from ground-up to be protected, devoid of assuming that adversary might ever find out several important system details. Three most important concepts more or less openly emerged from earlier work that will be exploited in our framework in support of security evaluation are: Arms race as well as security by design: as it is not likely to expect how many and types of attacks a classifier will incur throughout operation, classifier security have to be proactively assessed by means of a what-if analysis, by means of simulating potential attack situations [4]. Adversary modelling: effectual simulation of attack situations necessitates a formal representation of the adversary. Data distribution under attack: distribution of testing information might fluctuate from that of training data, when classifier is in attack.

LITERATURE SURVEY:

1) Robustness of Multimodal Biometric Fusion Methods against Spoof Attacks

In this paper, 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

2) Multimodal Fusion Vulnerability to Non-Zero Effort (Spoof) Imposters:

In biometric systems, the threat of "spoofing", where an imposter will fake a biometric trait, has 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.

3) Polymorphic Blending Attacks:

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 defense 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 analyze 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 defense.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

4) On Attacking Statistical Spam Filters:

The efforts of anti-spammers and spammers has 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 illustrates 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.

5) Good Word Attacks on Statistical Spam Filters:

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.

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) analyzing the vulnerabilities of classification algorithms, and the corresponding attacks. (ii) developing novel methods to assess classifier security against these attacks, which is 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 her1 website.

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 (Section 3). 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 encompasses and generalizes 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 behavior, 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.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

ADVANTAGES:

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

IMPLEMENTATION:

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)

RELATED WORK:

Generation of training and test data sets from gathered data is an important task in developing a classifier with high generation ability. The investigation of Machine Learning paradigms for detecting attacks against networked computers was a response to the weaknesses of attack signatures. As a matter of fact, signatures usually capture just some characteristics of the attack, thus leaving room for the attacker to produce the same effects by applying slight variations in the way the attack is crafted.

The generalization capability of machine learning algorithms has encouraged many researchers to investigate the possibility of detecting variations of known attacks. While machine learning succeeded in achieving this goal in a number of security scenarios, it was also a source of large volumes of false alarms. We learned that to attain the trade-off between detection rate and false alarm rate was not only a matter of the selection of the learning paradigm, but it was largely dependent on the problem statement.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Pattern recognition systems are increasingly being used in adversarial environments like biometric authentication and spam filtering tasks, in which data can be manipulated by humans to understand the outcomes of the automatic analysis. Current pattern recognition design methods do not explicit the intrinsic of these problems. This may be limiting their widespread adoption as potentially useful tools in many applications. If for instance, a more secure biometric of high quality gives a low match score and a less secure biometric gives a high match score, then there is a high likelihood of a spoof attack. It is commonly understood that one of the strengths of a multimodal system is in its ability to accommodate for noisy sensor data in an individual modality. In contrast, a more secure algorithm, in order to address the issue of a spoof attack on a partial subset of the biometric modalities, must require adequate performance in all modalities. This type of algorithm would invariably negate, to some extent, the contribution of a multimodal system to performance in the presence of noisy sensor data. A multimodal system improves the performance aspect but increases the security only slightly since it is still vulnerable to partial spoof attacks. Enhanced fusion methods, which utilize approaches to improve security, will again suffer decreased performance when presented with noisy Data.

CONTRIBUTIONS:

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 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 data dependent; on the other hand, model-driven analyses 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.

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.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[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.

Author details:

Author1:

A.Raghu

M.Tech, Department of CSE, Global Institute of Engineering and Technology, Chilkur (V), RR District, Telganana

Author 2:

Mrs. M.Jhansi Lakshmi

Associate professor, Head of the Department CSE, Global Institute of Engineering and Technology Chilkur (V), RR District, Telganana