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

Identifying Security Evaluation of Pattern Classifiers under Attack

Shaik Mustafa
PG Scholar,
Department of CSE,
St.Mark Educational Institution Soceity Group of Institution, Anantapur, AP, India.

M.Venkatesh Naik
Associate Professor,
Department of CSE,
St.Mark Educational Institution Soceity Group of Institution, Anantapur, AP, India.

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

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., legitimateand 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 dataM 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 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 (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 Although this emerging field is attracting growing interest, 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 while only a few theoretical models of adversarial classification problems have been proposed in the machine learning literature 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].

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

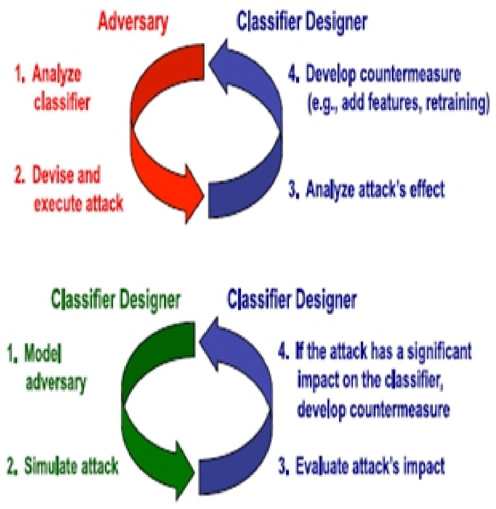
Adversary
Classifier Designer
1. Analyze classifier
2. Devise and execute attack
3. Analyze attack's effect
4. Develop countermeasure (e.g., add features, retraining)
Classifier Designer
Classifier Designer
1. Model adversary
2. Simulate attack
3. Evaluate attack's impact
4. If the attack has a significant impact on the classifier, develop countermeasure

Fig. 1. A conceptual representation of the arms race in adversarial classification. Left: the classical "reactive" arms race. The designer reacts to the attack by analyzing the attack's effects and developing countermeasures. Right: the "proactive" arms race advocated in this paper. The designer tries to anticipate the adversary by simulating potential attacks, evaluating their effects, and developing countermeasures if necessary

SPAM FILTERING:

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. This kind of classifier has been considered by several authors and it is included in several real spam filters.7 In this example, we focus on model selection. We assume that the designer wants to choose between a support vector machine (SVM) with a linear kernel, and a logistic regression (LR) linear classifier. He also wants to choose a feature subset, among all the words occurring in training emails.

A set D of legitimate and spam emails is available for this purpose. We assume that the designer wants to evaluate not only classifier accuracy in the absence of attacks, as in the classical design scenario, but also its security against the well-known bad word obfuscation (BWO) and good word insertion (GWI) attacks. They consist of modifying spam emails by inserting "good words" that are likely to appear in legitimate emails, and by obfuscating "bad words" that are typically present in spam [6]. The attack scenario can be modeled as follows.

1) Attack scenario. Goal. The adversary aims at maximizing the percentage of spam emails misclassified as legitimate, which is an indiscriminate integrity violation. Knowledge. As in [6], [10], the adversary is assumed to have perfect knowledge of the classifier, i.e.,: (k.ii) the feature set, (k.iii) the kind of decision function, and (k.iv) its parameters (the weight assigned to each feature, and the decision threshold). Assumptions on the knowledge of (k.i) the training data and (k.v) feedback from the classifier are not relevant in this case, as they do not provide any additional information. Capability. We assume that the adversary: (c.i) is only able to influence testing data (exploratory attack); (c.ii) cannot modify the class priors; (c.iii) can manipulate each malicious sample, but no legitimate ones; (c.iv) can manipulate any feature value (i.e., she can insert or obfuscate any word), but up to a maximum number nmax of features in each spam email [6], [10].

This allows us to evaluate how gracefully the classifier performance degrades as an increasing number of features is modified, by repeating the evaluation for increasing values of nmax. Attack strategy. Without loss of generality, let us further assume that x is classified as legitimate if gðxÞ ¼ Pn i¼1 wixi þ w0 < 0, where gð_Þ is the discriminant function of the classifier, n is the feature set size, xi 2 f0; 1g are the feature values (1 and 0 denote respectively the presence and the absence of the corresponding term), wi are the feature weights, and w0 is the bias. The SVM and LR classifiers perform very similarly when they are not under attack (i.e., for nmax ¼ 0), regardless of the feature set size; therefore, according to the viewpoint of classical performance evaluation, the designer could choose any of the eight models. However, security evaluation.

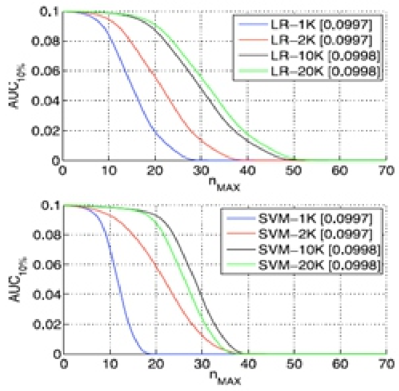

Fig. 3. AUC10 percent attained on TS as a function of nmax, for the LR (top) and SVM (bottom) classifier, with 1,000 (1K), 2,000 (2K), 10,000 (10K)

and 20,000 (20K) features. The AUC10 percent value for nmax ¼ 0, corresponding to classical performance evaluation, is also reported in the legend between square brackets highlights that they exhibit a very different robustness to the considered attack, since their AUC10 percent value decreases at very different rates as nmax increases; in particular, the LR classifier with 20,000 features clearly outperforms all the other ones, for all nmax values. This result suggests the designer a very different choice than the one coming from classical performance evaluation: the LR classifier with 20,000 features should be selected, given that it exhibit the same accuracy as the other ones in the absence of attacks, and a higher security under the considered attack

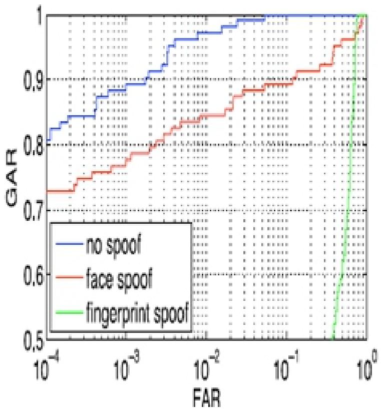

Fig. 4. ROC curves of the considered multimodal biometric system, under a simulated spoof attack against the fingerprint or the face matcher.

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.

ADVANTAGES OF PROPOSED SYSTEM:

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 .

APPLICATION EXAMPLES:

While previous work focused on a single application, we consider here three different application examples of our framework in spam filtering, biometric authentication, and network intrusion detection. Our aim is to show how the designer of a pattern classifier can use our framework, and what kind of additional information he can obtain from security evaluation. We will show that a tradeoff between classifier accuracy and security emerges sometimes, and that this information can be exploited for several purposes; e.g., to improve the model selection phase by considering both classification accuracy and security.

CONTRIBUTIONS, LIMITATIONS AND OPEN ISSUES:

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 data dependent; 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 oneto 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.

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.