www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 3 March 2015, Page No. 10992-10997

A Machine Learning Techinque For Generative Classifier Under Attack

S.Sasikala, C.Mahesh
MCA Final Year
Veltech Technical University
Avadi,Chnnai-62
E-Mail-sasi121210@gmail.com
Head Of The Department,MCA
Veltech Technical University
Avadi,Chnnai-62
E-Mail-cmahesh@veltechuniv.edu.in

Abstract

Pattern classification systems are commonly used in adversarial applications, like biometric authentication, network intrusion detection, and spam filtering, in which data can be going on purpose manipulated by humans to undermine their operation. Extending pattern arrangement[1] theory and design methods to adversarial settings is therefore a novel and very relevant research direction, which has not yet been pursued in a systematic way. Our address one of the main open issues: evaluating at design phase the security of pattern classifiers, namely, the performance degradation below potential attacks they may incur during operation. It proposes an algorithm for the generation of training and testing sets to be used for Security evaluation . Developing a framework for the empirical evaluation of classifier security at design phase that extends the model selection and act evaluation steps of the classical design cycle. Our proposed framework for empirical evaluation of classifier security that formalizes and generalizes the main thoughts designed in the literature, and give examples of its use in three real applications. report results show that security evaluation can provide a more complete thoughtful of the classifier's behavior in adversarial environments, and lead to improved design choices .

Keywords*:* pattern classification , security evaluation ,spam filter, biometric authenticaton, robustness evaluation

I. Introduction

A logical and unified treatment of this issue is thus needed to allow the faithful adoption of pattern classifiers in adversarial environments, starting from the theoretical basics up to novel design methods, extending the classical design cycle. 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. Three main open issues can be identified. Analyzing the vulnerabilities[2] of classification algorithms, and the corresponding attacks. Developing novel method to assess classifier security next to these attacks, which is not possible using classical performance evaluation methods[3]. Developing novel design methods to guarantee classifier security in adversarial environments. The current project on security evaluation of pattern classifiers under attack is disadvantageous since it does not cater the security enhancement for classified patterns. We see that poor analyzing the vulnerabilities of classification algorithms, and the corresponding attacks. A mean webmaster may manipulate search engine rankings to naturally promote her1 website.

II. Problem Statement

- A systematic and unified dealing of this issue is thus needed to allow the trusted taking on of pattern classifiers in adversarial environments, starting from the theoretical foundations up to novel design methods, extending the classical design cycle.

- Pattern classification systems base on classical theory and design methods do not take into account adversarial settings, they exhibit vulnerabilities to some potential attacks, allowing adversaries to undermine their usefulness .
- Three main open issues can be identified:
- Analyzing the vulnerabilities of classification algorithms, and the corresponding attacks.
- Developing novel methods to assess classifier security against these attacks, which is not possible using classical performance evaluation methods.
- Developing novel design methods to promise classifier security in adversarial environments.

The Disadvantages are as following.

- reduced analyzing the vulnerabilities of classification algorithms, and the corresponding attacks.

- A mean webmaster may manipulate search engine rankings to artificially promote her1 website.

III. Proposed System

- It proposes an algorithm for the generation of training and testing sets to be used for security evaluation, which can logically accommodate application-specific and heuristic technique for simulating attacks.
- It address issues above by developing a framework for the empirical evaluation of classifier security at design phase that extend the model selection and show

evaluation steps of the classical design cycle .

- This allows one to expand suitable counter events before the attack actually occurs, according to the principle of security by design.
- The presence of carefully targeted attacks may affect the distribution of training and testing data singly[4], hence we propose a model of the data distribution that can formally characterize this behavior, and that allows us to take into account a big number of potential attacks.

The advantages are as following.

It prevent developing novel methods to assess classifier security against these attacks.

- The presence of a smart and adaptive adversary makes the classification difficulty highly non-stationary.

IV.Algorithm as Proposal

Training and Testing Set Generation

Here we propose an algorithm to sample training (TR) and testing (TS) sets of any desired size from the distribution sptrðX; Y Þ and ptsðX; Y Þ.
We assume that k _ 1 different pairs of training and testing sets ðDi
TR;DiTSÞ, i ¼ 1; . . . ; k, have been obtained from D
using a classical resampling technique, like cross-validationor bootstrapping. Accordingly, their samples follow the distributionpDðX; Y Þ. In the following, we describe how to modify each of the sets Di
TR to construct a training set TRi
that follows the distribution ptrðX; Y Þ. For the sake of simplicity,we will omit the superscript i. An identical procedurecan be followed to construct a testing set TSi from eachof the Di
TS. Security evaluation is then carried out with the
classical method, by averaging (if k > 1) the perform.

Algorithm1 construction of TR or TS.

Input:the number n of desired samples;

The distributions p(Y)and p(A/Y);

For each y ₤{L,M},a€{T,F}, the distribution p(X/Y=y,A=a), if analytically defined,or the set of samples Dy,a,otherwise.

Output:

A data set S (either TrorTS)drawn from p(Y)p(A/Y)p(X/y,A).

```
s←∅
for i=1,…..,n do
    sample y from p(Y)
    sample a fromp(A/Y=y)
    draw a sample x from p(X/Y=y,A=a), if analytically defined;otherwise,sample with replacement from Dy,a
    s←sU{(x,y)}
end for
return S
```

5.0System Architecture

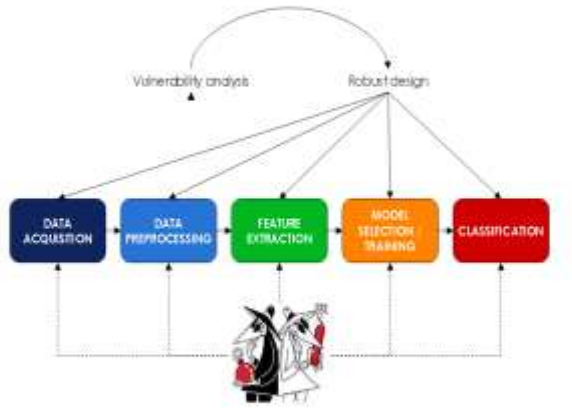

Fig 1: System Architecture

A machine learning and pattern recognition techniques have been newly adopted in security applications, like spam filtering, intrusion detection systems, and biometrics. The underlying reason is that usual security systems were not able to generalize, namely, to detect new (i.e., never-seen-before) kinds of attacks, while classification algorithms have indeed a good generalization capability. On the other hand, the introduction of pattern respect and machine learning techniques in such applications has raise itself an issue, namely, if these techniques are themselves.

An adversary[5] may find diverse ways to defeat a pattern recognition system. In particular, attacks can be devised at any stage of the design process, as well as at operation phase. As an instance, an adversary may compromise the training set used to build a classifier, by injecting carefully designed samples during the data acquisition phase. Further more he can devise some attack to mislead the data pre-processing (e.g., in spam filtering, different techniques can be used to avoid the filter to correctly parse an e-mail), as well as feature extraction (e.g., samples may be camouflaged to make the module, sensor or algorithm which performs feature extraction ineffective). Nevertheless, an adversary may exploit some characteristics of the selected classification model to design more effective attacks at operating phase. For example, a spammer may be able to get to know some among the most discriminates expressions used by a spam filter to classify legitimate e-mails, and use them to perform a more effective fine word attack. All the above mentioned issues (i.e., vulnerability identification, performance evaluation, and design of robust classifiers) raise from the fact that pattern recognition and machine learning techniques are not designed from the land up to be secure. In other words, they were not originally thought to run in adversarial environments[6]. In general, the design of a pattern recognition system should take into account explicitly that malicious adversaries can attack the system at any design stage, at least in principle. Just like a officer must think like a thief to catch a thief and a doctor must know how viruses and diseases work and behave to diagnose and counteract them, the designers of a pattern recognition system should try to identify and exploit the vulnerabilities of the system at any design stage and fix them before the system is released. In other words, the designers should put themselves in the adversary's shoes and try to anticipate the adversary's attacks.

As an instance, defence strategies may be adopted to prevent the adversary to compromise the training set, or features which are more difficult to modify for an adversary may be preferred. In general, the presence of malicious adversaries has to be considered at any level of the design of a pattern recognition system, ranging from data acquisition to classification, including feature extraction and selection, and performance evaluation. This draw near is usually referred to as security by design in security engineering[7], and it is one of the approaches exploited in this thesis to develop more secure pattern recognition systems.

VI.0 Experimental Results

Table 1.0 classification of pattern classfier potential

Attacks	pattern	classifier	Potential
0.0992	2	6	10
0.0995	5	5	20
0.0996	5	5	30
0.0997	7	8	50
1	5	10	60

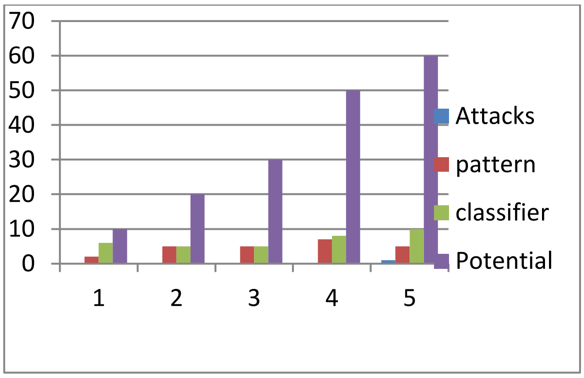

Fig 2. Function of classifier values

Each model decreases that is it drops to zero[8] for values between 3and 5 (depending on the classifier). This means that all testing spam emails gotmis classified as legitimate, after adding or obfuscating from3 to 5words.The pattern and attack classifiers perform very similarly when they are not under attack , 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

VII.0 CONCLUSION:

Our Project focused on empirical security evaluation of pattern classifiers that have to be deployed in adversarial environments, and proposed how to revise the classical arrangement 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, education 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.

VIII.0 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].A. Kolcz and C.H. Teo, “Feature Weighting for Improved Classifier Robustness,” Proc. Sixth Conf. Email and Anti-Spam, 2009.

[4]. D.B. Skillicorn, “Adversarial Knowledge Discovery,” IEEE Intelligent
Systems, vol. 24, no. 6, Nov./Dec. 2009.

[5].P. Laskov and R. Lippmann, “Machine Learning in Adversarial Environments,” Machine Learning, vol. 81, pp. 115-119, 2010.

[6]. 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.

[7] M. Barreno, B. Nelson, A. Joseph, and J. Tygar, “The Security of Machine Learning,” Machine Learning, vol. 81, pp. 121-148, 2010.

[8] Abernethy, J., O. Chapelle, and C. Castillo: 2010, ‘Graph regularization methods for Web spam detection’. *Machine Learning Journal* **81**(2). DOI: 10.1007/s10994-009-5154-2