A Survey On Security Evaluation Of Pattern Classifiers Under Attack

Sandhyarani Sambhaji Patil, Dnyaneshwar A. Rokade

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. Pattern classification systems may exhibit vulnerabilities if adversarial scenario is not taken into account. Multimodal biometric systems are more robust to spoofing attacks,as they combine information coming from different biometric traits.In this paper,weevaluate the security of pattern classifiersthat formalizes and generalizes the main ideas proposed in the literature and give examples of its use in three real applications.We propose a framework for evaluation of pattern security,model of adversary for defining any attack scenario.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 Term*-adversarial classification,adversarial scenario,patternclassification, security evaluation.**

I. Introduction

In Pattern classification systems machine learning algorithms are used to perform security-related applications like biometric authentication, network intrusion detection, and spam filtering, to distinguish between a "legitimate" and a "malicious" pattern class.The input data can be purposely manipulated by an adversary to make classifiers to produce false negative.This often gives rise to an arms race between the adversary and the classifier designer.Well known examples of attacks are: Spoofing attacks where one person or program purposely falsifying data and thereby gaining an illegitimate advantage[1][2],modifying network packets belonging to intrusive trafficmanipulating contents of emails[3],modifying network packets belonging to intrusive traffic.

Mainly three main open issues are identified: (i) analyzing the vulnerabilities of classification algorithms, and the corresponding attacks (ii) developing novel methods to assess classifier security against these attacks (iii) developing novel design methods to guarantee classifier security in adversarial environments.

Machine learning is used to prevent illegal or unsanctioned activity which are created from adversary. Machine learning is used in security related tasks involving classification, such as intrusion detection systems, spam filters, biometric authentication. Measuring the security performance of these classifiers is an essential part for facilitating decision making.

II. Literature Survey

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, a spammer modifies a spam message by inserting or appending words indicative of legitimate email. 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[4].

Spoof attacks consist in submitting fake biometric traits to biometric systems, and this is a major threat in security. Multi-modal biometric systems are commonly used in spoof attacks.Multimodal biometric systems for personal identity recognition is very useful from 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[1][2].

Intrusion detection systems analyze network traffic to prevent and detect malicious activities like intrusion attempts,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.

III. Methodology

Construction of Training (TR) and Testing(TS):

Generation of trainingand test data sets from gathered data is an important task in developing a classifier with high generation ability.Reassembling techniques are used instatistical analysis,are used for model selection by estimating the classification performanceof classifiers.Reassembling techniques are used for estimating statisticssuch as the mean and the median by randomly selecting data from the given data set,calculating statistics on that data and repeating above procedure many times. “Training” data refers both to the data used by the learning algorithm during classifier design, coming from D(where D is data set), and to the data collected during operation to retrain the classifier through online learning algorithms. “Testing” data refers both to the data drawn from D to evaluate classifier performance during design, and to the data classified during operation.

We propose an algorithm to sample training (TR) and testing (TS) sets of any desired size from the distributions.Training and Test sets have been obtained from distribution using a classical resampling technique like cross validation or bootstrapping.Security evaluation is carried out by averaging the performance of the trained and tested data.

IV. Architecture

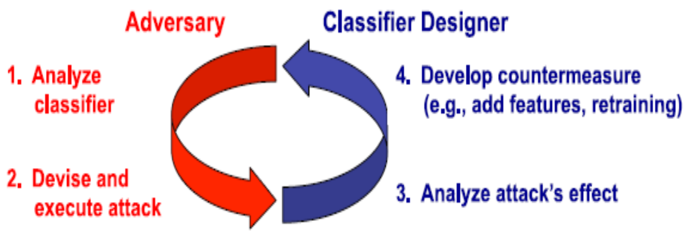
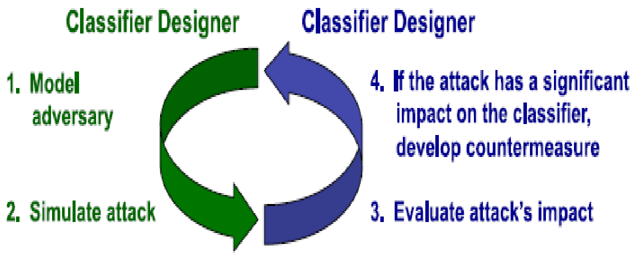

Adversary
Classifier Designer
1. Analyze classifier
2. Devise and execute attack
3. Analyze attack's effect
4. Develop countermeasure (e.g., add features, retraining)

Fig.1. A conceptual representation in arm race in adversarial classification

(a)The classical “reactive” arm race


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

(b)The “proactive” arm race

In “reactive” arms race, the designer reacts to the attack by analyzing the attack’s effects and developing countermeasures. In “proactive” arms race, the designer tries to anticipate the adversary by simulating potential attacks,evaluating their effects and developing countermeasures if necessary.

We summarize the three main concepts in our framework for security evaluation:

1) Arms race and security by design: since it is not possible to predict how many and which kinds of attacks a classifier will incur during operation, classifier security should be proactively evaluated using a what-if analysis, by simulating potential attack scenarios.

2) Adversary modeling: effective simulation of attack scenarios requires a formal model of the adversary.

3) Data distribution under attack: the distribution of testing data may differ from that of training data, when the classifier is under attack.

V. 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 performanceevaluation design step. In this paper the 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.

VI. 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] D. Lowd and C. Meek, “Good Word Attacks on Statistical Spam Filters,” Proc. Second Conf. Email and Anti-Spam, 2005.