A Review on Security Evaluation for Pattern Classifier against Attack

Kunjali Pawar
M.E. Student, Department of C. E
Dr. D.Y.Patil School of
Engineering & Technology
Savitribai Phule Pune University, Pune.

Madhuri Patil
Assistant Professor
Dr. D. Y. Patil School of
Engineering & Technology
Savitribai Phule Pune University, Pune.

ABSTRACT

The systems which can be used for pattern classification are used in adversarial application, for example spam filtering, network intrusion detection system, biometric authentication. This adversarial scenario's exploitation may sometimes affect their performance and limit their practical utility. In case of pattern classification conception and contrive methods to adversarial environment is a novel and relevant research direction, which has not yet pursued in a systematic way. To address one main open issue: evaluating at contrive phase the security of pattern classifiers (for example the performance degradation under potential attacks which incurs during the operation). To propose a framework for evaluation of classifier security and also this framework can be applied to different classifiers on one of the application from the spam filtering,biometric authentication andnetwork intrusion detection.

General Terms

Security, Pattern Classification, Attack, Theory, classifier.

Keywords

Machine learning system, Security evaluation, Adversarial classification, Arms-Race, Spam Filtering.

1. INTRODUCTION

Machine learning systems provide pliability relating with unfolding the input in a number of applications. Machine learning techniques are applied to a growing number of systems and networking problems [1], particularly those problems where the intention is to discern anomalous system behavior. For instance, Network Intrusion Detection Systems (NIDS) monitor network traffic to discern abnormal movements, such as attacks against hosts or servers.Machine learning is used to prevent unlawful or unsanctioned activity which are created from the adversary [2]. Machine learning is used in security affiliated functions bring in a classification, such as intrusion detection systems, spam filters, biometric authentication, etc. Measuring the security performance of classifiers is an important part in facilitating decision making. As spam filters evolve to better classify spam, spammers canadapt their messages to avoid detection.

The input data can be manipulated by an adversary to compose classifiers to produce false negative [3]. This frequently brings about an arms race in the middle of the adversary and the classifier designer.In the case of the arms-race problem in pursuing the security it is not enough to retort to observed attacks. There is some open issues which can be identified: (i) development of methods which assess the security of classifier against the attacks (ii) Analysis of vulnerabilities and corresponding attacks of classification [4].

The security in Machine Learning Systems besides of spam filtering (spam e-mails) and network intrusion detection systems that is NIDS [5]. The Machine learning systems have been employed in different number of applications which contains Online Deputy Systems (ODS), Clump Supervisimg (cluster monitoring), toxin detection same as virus detection and some dynamic operations applications. There are some algorithms with accurate performance in the case of adversarial condition like Secure Learning Algorithms [6]. Some Classifiers are utilized to generate some contrasts which promote security intention. For example, the intention of a toxin (virus) detection system is to diminish vulnerabilities.The toxins (virus) give antecedent to contamination or by detecting the contamination.

An adversary's attempt to procure the data which are nothing but the domestic state of a Machine Learning System (MLS) to- (i) infuse the personal data which is encrypted in its domestic state otherwise (ii) originate the data which sanction the adversary to effectually onslaught the system.

The respite of the paper is unionized as follows: The section 2, scrutinize about the arms-race problem between classifier designer and the adversary. The section 3, discuss an overview of the Security Evaluation Framework. The section 4, summarize the conclusion and the future scope.

2. PREVIOUS WORK ON SECURITY EVALUATION

Previous work in the adversarial learning system can be categorized according to the two main steps, the pro-active arms race and the re-active arms race. It pivoted on identifying vulnerabilities of the adversarial learning algorithms and assessing impact of corresponding attacks [7] on the targeted classifier.

2.1 Arms-Race Problem

The Arms-Race is linking of the classifier designer and the adversary which is modeled. For example Fake biometric traits in Biometric Authentication. It analyzes the classifier defenses and develops the Attack strategy to overcome them. The role of classifier designer is to model the adversary according to the algorithms or methods. To pursue the security in the context of Arms-Race it is not enough to retort to the observed attacks, but it is also important to proactively intercept the adversary by suggesting the relevance, the potential attacks through a what-if analysis classifier.

2.1.1 What-If Analysis

The process of recasting the values in the cells to see how the changes will affect the outcomes. This allows to develop the countermeasures before the attack actually occurs. The countermeasure is nothing but the action or the process or devices that mitigate or prevent the effects of the threats to computer, server or network [8]. There are two steps for determining the result, according to the generation of training and testing sets (i) Set the input values, (ii) Determine possible results.

2.1.2 *Mechanism*

2.1.2.1 'Re-active' Arms Race

In this type, the classifier designer reacts to the attack by analyzing its effects and grows the countermeasures. In this type, the classifier designer and the adversary's attempt to accomplish their aims by behaving to the changing comportant of compititor [9]. The adversary first analyzes the system as shown in Fig. 1 and manipulates data violate its security, for example, to evade detection in spam filtering, a spammer may accumulate some erudition of the words worn by the targeted antispam filter to obstruct (block) spam, and then plies spam emails consequently. Then the system designer reacts by analyzing the attack samples and updating the system consequently for example, by affixing features to discern the peculiar attacks and retraining the classifier on the recently composed samples. Security problems lead to a 'Re-active' arms-race between classifier designer and the adversary.

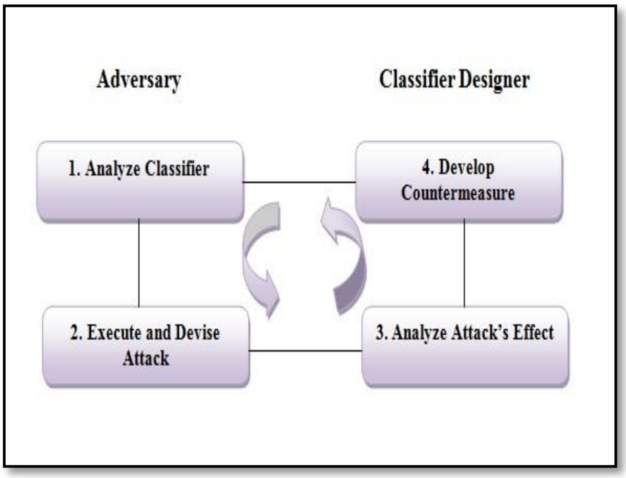
Adversary
Classifier Designer
1. Analyze Classifier
4. Develop Countermeasure
2. Execute and Devise Attack
3. Analyze Attack's Effect

Fig 1: 'Re-active' Arms Race

2.1.2.2 'Pro-active' Arms Race

In this type, the classifier designer can anticipate the adversary by simulating the potential attacks, evaluating their effects and developing the countermeasures if necessary.The 'Re-active' approaches, neither anticipates the new security vulnerabilities,nor they bid to forecast future attacks. Computer security guidelines accordingly advocate a 'Pro-active' approach in which the classifier designer also attempts to anticipate the adversary's stratagemby (i) repeating this process before system deployment, (ii) devising proper countermeasures,when required, and (iii) identifying the relevant threats.It means that, one can simulate the attacks [10] based on a model of adversaries, tocomplement the 'Re-active' arms race as shown in Fig. 2. However, the resulting systems remain effective for a longer time, with less recurrent guidene or human-intercession and with less severe vulnerabilities.

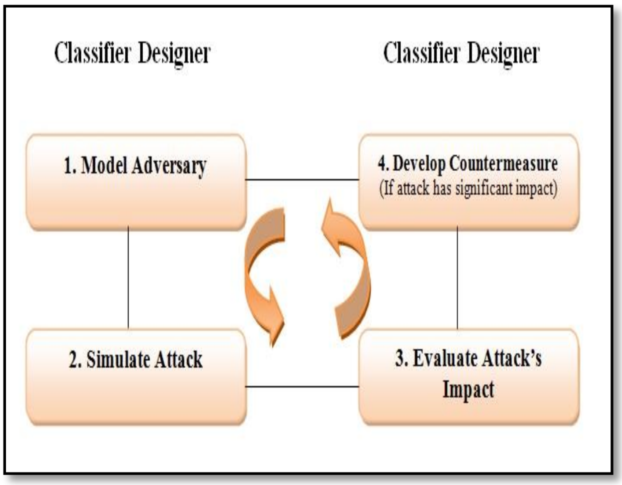
Classifier Designer
Classifier Designer
1. Model Adversary
4. Develop Countermeasure (If attack has significant impact)
2. Simulate Attack
3. Evaluate Attack's Impact

Fig 2: 'Pro-active' Arms Race

3. AN OVERVIEW OF FRAMEWORK FOR SECURITY EVALUATION

After summarizing the previous work, here take an overview of how to propose a framework for experimental C security evaluation on the basis of scenarios of potential attacks. Security evaluation is implicitly undertaken by defining an attack and assessing the impact of its given classifier.

The main goal is to scrutinize buttress which is difficult to represent to escape the design classifiers in the Adversarial Classification Problems with the help of the framework. An artifice for providing the security for classifier designer is to mask the data to the Adversary. A feasible fulfillment of this artifice was predicted with some soft contention which gives the identity of haphazardness in the location of classification boundaries. [11]

3.1 Attack Scenario

In the case of Arms-race, it is not possible to recommend how many and what type of attacks a classifier will incur during operation, the classifier security should proactively evaluate using a what-if analysis, by simulating potential attack scenarios.

3.2 Adversary Modeling

3.2.1 Goal

The primary goal is to formulate or model the adversary as the optimization of an actual function [12]. The effective simulation of attack scenarios requires a formal model of the adversary. In many cases, according to the knowledge of classifier [13] and capability of manipulation of data, the adversary acts rationally to attain a goal of security evaluation.

3.2.2 Capability

Define the adversary in terms of attack influence (exploratory or causative [1], [8]), feature manipulation, control on training and testing samples.

3.2.3 Knowledge

Quantitative discussion on training data [12], feature set, the learning algorithm data, classifier's decision function, feedback from classifier [14] and some assumptions regarding on the application at hand [15].

3.3 Data Distribution under Attack

The distribution of testing data differs from training data, when the classifier is under attack. It includes discrimination between legitimate and malicious samples.

3.4 Generation of Training and Testing Data

The main task in the development of classifier is construction of Training data and Testing data. Estimation of classification performance of the classifier is used for analysis in reassembling techniques (mean and median by random selection of data). 'Testing' data refers both to the data classified during operation and to the data drawn from the data set to evaluate classifier performance during design. 'Training' data refer both to the data used by the learning algorithm throughout classifier design, coming from the data set and to the data collected throughout the operation to retrain the classifier through online learning algorithms. Training and Testing sets have been obtained from distribution using a classical reassembling technique like bootstrapping [16] or cross validation. Security evaluation can be carried out by averaging the performance of the trained and tested data.

4. APPLICATION EXAMPLE: SPAM FILTERING

In case of Spam filtering, the classifier designer is discriminate between spam emails on their basis of bag-of-words feature representation. The main aim is to unveil the use of framework obtained from Security Evaluation. There is also need to improve the model selection phase by considering classification security and accuracy. In a good word attack [17], a spammer modifies a spam message by loading or appending words indicative of legitimate email. Spam messages are obfuscated through insertion of good words or misspelling of bad words and attacks in computer security. To delineate and appraise the efficiency of active and passive good word attacks in contrast to two types of statistical spam filters maximum entropy filters and naive Bayes [9] [18].

Adversarial applications attacks in the case of spam filtering becomes a heightening defiance to the anti-spam association. The good word attack is one way to be used by the attackers (spammers) [19]. This method contains the "Good Words" which includes annexing sets. Good word is defined as the word which not specific to licit emails, but uncommon in spam (junk emails).These messages which is inoculated with some keywords are more likely to appear legitimate and bypass spam filters.

The basic model is often known as the bag of words or multivariate model. Actually, a document is configured into a set of features such as phrases, words, meta-data, etc. This set of features can be depicted as a vector whose components are multivariate (boolean) or multinomial (real values) [20]. The ordering of features is ignored. The Classification algorithm uses the feature vector on the basis upon of the document is judged. Assume that a classifier has to discriminate in the middle of legal and spam (junk) emails on the basis of literal content and that the bag-of-words feature depiction [21] has been picked, with binary features denoting the occurrence of a particular set of words. The classifier has been considered by several authors and it is included in several real spam filters.

5. CONCLUSION AND FUTURE SCOPE

This paper presented an overview of work related to the security of pattern classification systems with the goal of imparting useful guidelines on how to improve their design and assess their security specific attacks. Also the paper focused on innovative security evaluation of pattern classifiers that deployed in adversarial environments. Main contribution is a framework for verifiable security evaluation that construes and establishes the notion from previous work, and can be utilized to different classifiers, learning algorithms, and classification tasks.

In the future, clustering methods can be integrated with the existing technique in order to get better results. Further, this approach can be applied to the application which makes the classification problem highly non-stationary.

6. ACKNOWLEDGMENTS

The authors would like to thank the Department of Computer Engineering of D. Y. Patil School of Engineering and Technology, Pune for its generous support. They would also like to thank the Principal Dr. Uttam B. Kalwane, HOD Mrs. Arti Mohanpurkar and all Staff Members of Computer Engineering department for their valuable guidance.

7. REFERENCES

[1] A.D. Joseph, L. Huang, B. Nelson, J.D. Tygar and B. Rubinstein, "Adversarial Machine Learning," Proc. Fourth ACM Workshop Artificial Intelligence and Security, pp. 43-57, 2011.

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

[3] D. Lowd and C. Meek, "Adversarial Learning," Proc. 11thACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 641-647, 2005.

[4] Battista Biggio, Giorgio Fumera and Fabio Roli, " Security Evaluation of Pattern Classifiers under Attack," IEEE Transactions On Knowledge And Data Engineering, VOL.26, NO.4, APRIL 2014.

[5] K. Seamon and J.S. Baras, "A Framework for the Evaluation of Intrusion Detection Systems," Proc. IEEE Symp. Security and Privacy, pp. 63-77, 2006.

[6] B. Nelson, M. Barreno and A. Joseph, "The Security ofMachine Learning," Machine Learning, vol. 81, pp. 121-148, 2010.

[7] B. Nelson , M. Barreno, R. Sears, J.D. Tygar and A.D. Joseph, "Can Machine Learning be Secure?," Proc. ACM Symp. Information, Computer and Comm. Security (ASIACCS), pp. 16-25, 2006.

[8] S. Rizzi, "What-If Analysis," Encyclopedia of Database Systems, pp. 3525-3529, Springer, 2009.

[9] A. Wahi and C. Prabakaran, "A Literature Survey on Security Evaluation of the Pattern Classifiers under Attack," International Journal of Advance Research in the Computer Science and Management Studies, VOL.2, Issue 10, October 2014.

[10] A.M. Narasimhamurthy and L.I. Kuncheva, "A Framewo-rk 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.

[11] B. Biggio, F. Roli and G. Fumera, “Adversarial Pattern Classification Using Multiple Classifiers and Randomisation,” Proc. Joint IAPR Int’l Workshop Structural, Syntactic, and Statistical Pattern Recognition, pp. 500-509, 2008.

[12] 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.

[13] Mausam, S. Sanghai, N. Dalvi, and D. Vermaand P. Domingos,“Adversarial Classification,” Proc. 10th ACM SIGKDD Int’l Conference Knowledge Discovery and Data Mining, pp. 99-108, 2004.

[14] A. Adler, “Vulnerabilities in Biometric Encryption Syste-ms,” Proc. Fifth Int’l Conf. Audio- and Video-Based Biometric Person Authentication, pp. 1100-1109, 2005.

[15] D.B. Skillicorn, “Adversarial Knowledge Discovery,” IE-

[16] EE Intelligent Systems, vol. 24, no. 6, Nov. /Dec. 2009. R.J. Tibshirani and B. Efron, An Introduction to the Bootstrap, Chapman & Hall, 1993.

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

[18] I. Mani and J. Zhang, “A Multiple Instance Learning Strategy for Combating Good Word Attacks on Spam Filters,” J. Machine Learning Research, vol. 9, pp. 1115-1146, 2008.

[19] J. Zhang and I. Mani, “KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction,” Proc. Int’l Conf. Machine Learning (ICML’2003), Workshop Learning from Imbalanced Data Sets, 2003.

[20] G.L. Wittel and S.F. Wu, “On Attacking Statistical Spam Filters,” Proc. First Conf. Email and Anti-Spam, 2004.

[21] C.H. Teo and A. Kolcz, “Feature Weighting forImproved

[22] ClassifierRobustness,” Proc. Sixth Conf. Email And Anti-Spam, 2009.