International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Study of Evasion Attack using Feature Selection in Adversarial Environment

Swapnali Jadhav[1], Vidya Dhamdhere[2]

[1]Pune University, G.H.R.C.E.M, Wagholi, Pune, Maharashtra, India

[2]Professor, Pune University, G.H.R.C.E.M, Wagholi, Pune, Maharashtra, India

Abstract: ***Not only Pattern recognition but also machine learning techniques have been increased in adversarial settings such as intrusion, spam, and malware detection, although its security against well-crafted attacks that aims to evade detection. Spam filtering is one of the most common application examples considered in adversarial machine learning. In this task, the goal is often to design feature selection against attacks. Here we use Random Forest Classifier to find evasion attacks. The ability of rapidly evolve to changing and complex situations has helped it become a fundamental tool for computer security. Evasion attacks assumes that the attacker can arbitrarily change every feature, but they constrain the degree of manipulation, e.g., limiting the number of medications, or their total cost. Adversarial Feature Selection design phase are given in this paper.***

Keywords: Adversarial learning, classifier security, evasion attacks, feature selection, spam filtering.

1. Introduction

Machine-learning and pattern-recognition techniques are increasingly being adopted in security applications like spam filtering, network intrusion detection, and malware detection due to their ability to generalize, and to potentially detect novel attacks or variants of known ones. The main aim of feature selection (FS) is to discover a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original data [8]. Many spam detection techniques based on machine learning techniques have been proposed. As the amount of spam has been increased tremendously using bulk mailing tools, spam detection techniques should counteract with it.

If we hope to use machine learning as a general tool for computer applications, it is incumbent on us to investigate how well machine learning performs under adversarial conditions. An interesting, preliminary result is that classifier security to evasion may be even worsened by the application of feature selection.

It requires:
1) Finding potential vulnerabilities of learning before they are exploited by the adversary;
2) Investigating the impact of the corresponding attacks (i.e., evaluating classier security); and
3) Devising appropriate countermeasures if an attack is found to significantly degrade the classer's performance.

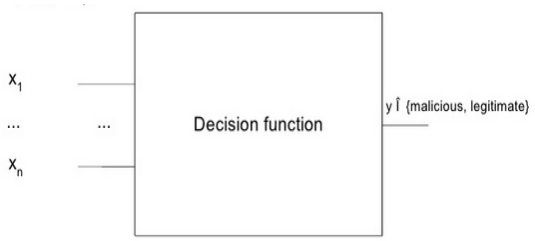

Figure 1: Structure of Adversarial Feature Selection

It shows the structure of Adversarial Feature Selection. It is now acknowledged that, since pattern classification systems based on classical theory and design methods[2] do not take into account adversarial settings, they exhibit vulnerabilities to several potential attacks, allowing adversaries to undermine their effectiveness [1], [3],[4], [5], [6], [7].

Below section 2 will give information about existing system, section 3 will give system architecture. Furthermore conclusion and reference.

2. Existing System

- An implicit assumption behind traditional machine learning and pattern recognition algorithms is that training and test data are drawn from the same, possibly unknown, distribution. This assumption is, however, likely to be violated in adversarial settings, since attackers may carefully manipulate the input data to downgrade the system's performance. It categorizes attacks according to three axes: the attack influence, the kind of security violation, and the attack specificity. The attack influence can be either causative or exploratory. Depending on the kind of security violation, an attack may compromise a system's availability, integrity, or privacy: availability attacks aim to downgrade the overall system's accuracy, causing a denial of service; integrity attacks, instead, only aim to have malicious samples misclassified as legitimate; and privacy attacks aim to retrieve some protected or sensitive information from the system.
- Bursteinas and Long 00; Thota et al. 09; Zhao and Zhu 06; Zhu 08] performed feature selection but they did not mention how they decided the number of important features, and they did not provide variable importance of each feature as a numerical value.
- 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 [6], [12], [13], and it is included in several real spam filters. In this example, we focus on model selection.

- Random Forests (RF) is a special kind of ensemble learning techniques and robust concerning the noise and the number of attributes [Breiman 01]. RF builds an ensemble of CART tree classifications using bagging mechanism [Duda *et al*. 01] [15].
- The support vector machine (SVM) is a exercise procedure for knowledge organization and reversion rubrics after statistics, for instance the SVM can be recycled to study polynomial, circular foundation purpose (RBF) then multi-layer perception (MLP) classifiers SVMs

3. System Architecture

Spam filtering discriminates between legitimate and spam emails by analyzing their textual content, exploiting the so called bag-of-words feature representation, in which each binary feature denotes the presence (1) or absence (0) of a given word in an email. Despite its simplicity, this kind of classifier has shown to be highly accurate, while also providing interpretable decisions. It has been, therefore, widely adopted in previous work.Instead of SVM Classifier here Random Forest classifier is used.

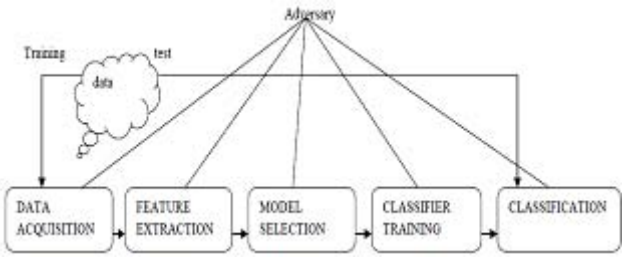

Figure 2: Design Steps

Above figure shows five elements in design phase of Adversarial Feature Selection

Input: Dataset Spam content. (eg. Email, message)

Output: feature selection and attack detection.

Process: Adversarial Feature Selection

$$\theta^{\star} = \arg\max_{\theta} \; G(\theta) + \lambda S(\theta)$$

$$\text{s.t.} \sum_{k=1}^{d} \theta_k = m$$

Where,

G = Estimate of the classifier's generalization capability

$$G(\theta) = \mathbb{E}_{\mathbf{x},y \backsim p(\mathbf{X},Y)} \; u(y, g(\mathbf{x}_{\theta}))$$

S = Security to evasion,

$$S(\theta) = \mathbb{E}_{\mathbf{x} \backsim p(\mathbf{X}|Y=+1)} \; c\left(\mathbf{x}_{\theta}^{\star}, \mathbf{x}_{\theta}\right)$$

λ =Trade off parameter to be chosen according to application-specific constraints

We compare the traditional forward feature selection wrapping algorithm with the corresponding implementation of our approach, using a linear SVM as the classification algorithm. In the latter case, instead, we consider traditional and adversarial backward feature elimination approaches, and an SVM with the RBF kernel as the wrapped classifier.

Efficiency of Random Forest Classifier:

1) Almost always have lower classification error and better f-scores than decision trees.
2) Almost always perform as well as or better than SVMs, but are far easier for humans to understand.
3) Deal really well with uneven data sets that have missing variables.
4) Gives a really good idea of which features in our data set are the most important for free.
5) Generally train faster than SVMs (though this obviously depends on implementation).

4. Advantages & Disadvantages

4.1Advantage

1) It gives Pattern recognition and machine learning techniques.
2) It gives relevant attack scenario.
3) It improves classifier security against evasion attacks.
4) It gives full analytical model of the problem and of the adversary's behavior.

4.2Disadvantage

1) To design more secure generative classifiers.
2) May be very difficult to develop for real-world applications.

5. 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 performance evaluation 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 Feature selection may be assumed a crucial step in security relating applications, like spam and malware detection. We represented about random factor classifier and system architecture. An adversarial feature selection method that optimizes the generalization capability of the classifier, but also it secure against evasion attacks. Some binary features, to the case of classification algorithms trained on either continuous or discrete feature spaces. We studied comparative all system in tabular format and we presented system model.

6. Acknowledgment

We would like to thank all the authors of different research papers referred during writing this paper. It was very knowledge gaining and helpful for the further research to be done in future.

References

[1] Kolcz and C.H. Teo, "Feature Weighting for Improved Classifier Robustness," Proc. Sixth Conf. Email and Anti-Spam, 2009.
[2] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. Wiley-Interscience Publication, 2000.
[3] 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.
[4] 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.
[5] 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.
[6] P. Laskov and R. Lippmann, "Machine Learning in Adversarial Environments," Machine Learning, vol. 81, pp. 115-119, 2010.
[7] 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.
[8] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, 1997.
[9] M. Kloft and P. Laskov, "Online anomaly detection under adversarial impact," in *Proc. 13th Int. Conf. Artif. Intell. Stat.*, Sardinia, Italy, 2010,pp. 405–412.
[10] Q. Liu, A. H. Sung, Z. Chen, and J. Xu, "Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images," *Pattern Recognit.*, vol. 41, pp. 56–66, Jan. 2008.
[11] R.N. Rodrigues, L.L. Ling, and V. Govindaraju, "Robustness ofMultimodal Biometric Fusion Methods against Spoof Attacks,"J. Visual Languages and Computing, vol. 20, no. 3, pp. 169-179, 2009.
[12] P. Johnson, B. Tan, and S. Schuckers, "Multimodal Fusion Vulnerabilityto Non-Zero Effort (Spoof) Imposters," Proc. IEEE Int'lWorkshop Information Forensics and Security, pp. 1-5, 2010.
[13] P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee,"Polymorphic Blending Attacks," Proc. 15th Conf. USENIX SecuritySymp., 2006.
[14] G.L. Wittel and S.F. Wu, "On Attacking Statistical Spam Filters,"Proc.First Conf. Email and Anti-Spam, 2004.
[15] RANDOM FORESTS Leo Breiman Statistics Department University of California Berkeley, CA 94720 January 2001