Neurocomputing 000 (2018) 1–15

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Data driven exploratory attacks on black box classifiers in adversarial domains

Tegjyot Singh Sethi*, Mehmed Kantardzic

Data Mining Lab, University of Louisville, Louisville, USA

ARTICLE INFO

Article history:
Received 23 March 2017
Revised 13 November 2017
Accepted 2 February 2018
Available online xxx

Communicated by Dr Lo David

Keywords:
Adversarial machine learning
Reverse engineering
Black box attacks
Classification
Data diversity
Cybersecurity

ABSTRACT

While modern day web applications aim to create impact at the civilization level, they have become vulnerable to adversarial activity, where the next cyber-attack can take any shape and can originate from anywhere. The increasing scale and sophistication of attacks, has prompted the need for a data driven solution, with machine learning forming the core of many cybersecurity systems. Machine learning was not designed with security in mind and the essential assumption of stationarity, requiring that the training and testing data follow similar distributions, is violated in an adversarial domain. In this paper, an adversary's view point of a classification based system, is presented. Based on a formal adversarial model, the *Seed-Explore-Exploit* framework is presented, for simulating the generation of data driven and reverse engineering attacks on classifiers. Experimental evaluation, on 10 real world datasets and using the Google Cloud Prediction Platform, demonstrates the innate vulnerability of classifiers and the ease with which evasion can be carried out, without any explicit information about the classifier type, the training data or the application domain. The proposed framework, algorithms and empirical evaluation, serve as a white hat analysis of the vulnerabilities, and aim to foster the development of secure machine learning frameworks.

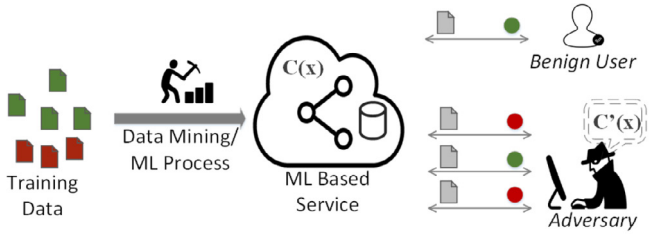
1. Introduction

The growing scale and reach of modern day web applications has increased its reliance on machine learning techniques, for providing security. Conventional security mechanisms of firewalls and rule-based black and white lists, cannot effectively thwart evolving attacks at a large scale [37]. As such, the use of data driven machine learning techniques in cybersecurity applications, has found widespread acceptance and success [15]. Whether it be for outlier detection for network intrusion analysis [51], biometric authentication using supervised classification [14], or for unsupervised clustering of fraudulent clicks [45], the use of machine learning in cybersecurity domains is ubiquitous. However, during this era of increased reliance on machine learning models, the vulnerabilities of the learning process itself have mostly been overlooked. Machine learning operates under the assumption of stationarity, i.e. the training and the testing distributions are assumed to be identically and independently distributed (IID) [53]. This assumption is often violated in an adversarial setting, as adversaries gain nothing by generating samples which are blocked by a defender's system [16]. The dynamic and contentious nature of this domain, demands a thorough analysis of the dependability and security of machine learning systems, when used in cybersecurity applications.

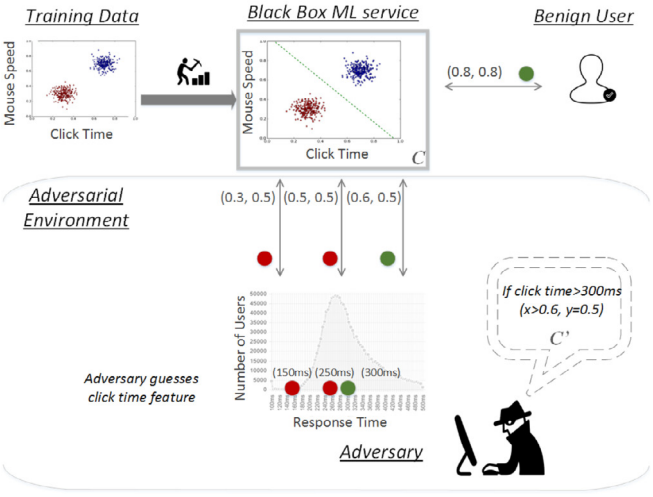
In an adversarial environment, the accuracy of classification has little significance, if an attacker can easily evade detection by intelligently perturbing the input samples [21]. Any deployed classifier is susceptible to probing based attacks, where an adversary uses the same channel as the client users, to gain information about the system, and then subsequently uses that information to evade detection [1,6]. This is seen in Fig. 1 (a), where the defender starts by learning from the training data and then deploys the classifier C, to provide services to client users. Once deployed, the model C is vulnerable to adversaries, who try to learn the behavior of the defender's classifier by submitting probes as input samples, masquerading as client users. In doing so, the defender's classifier is seen only as a black box, capable of providing tacit *Accept/Reject* feedback on the submitted samples. An adversary, backed by the knowledge and understanding of machine learning, can use this feedback to reverse engineer the model C (as C'). It can then avoid detection on future attack samples, by accordingly perturbing the input samples. It was shown recently that, deep neural networks are vulnerable to adversarial perturbations [31]. A similar phenomenon was shown to affect a wide variety of classifiers in [30], where it was demonstrated that adversarial samples are

* Corresponding author.
E-mail addresses: t0seth01@louisville.edu, tegjyotsingh.sethi@louisville.edu (T.S. Sethi), mehmedkantardzic@louisville.edu (M. Kantardzic).

https://doi.org/10.1016/j.neucom.2018.02.007
0925-2312/



(a) An adversary making probes to the black box model C, can learn it as C', using active learning.



(b) Example task of attacking behavioral CAPTCHA. Black box model C, based on Mouse Speed and Click Time features, is used to detect benign users from bots. Adversary can reverse engineer C as C', by guessing click time feature and making probes based on the human response time chart, using the same input channels as regular users.

Fig. 1. Classifiers in adversarial environment, (a) shows the general adversarial nature of the problem and (b) shows an example considering a behavioral CAPTCHA system.

transferable across different classifier families. Cloud based machine learning services (such as Amazon AWS Machine Learning[1] and Google Cloud Platform[2]), which provide APIs for accessing predictive analytics as a service, are also vulnerable to similar black box attacks [42].

An example of the aforementioned adversarial environment is illustrated in Fig. 1(b), where a behavioral mouse dynamics based CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) system is considered. Popular examples of these systems are Google's reCAPTCHA[3] and the classifier based system developed in [14]. These systems use mouse movement data to distinguish humans from bots and provide a convenient way to do so, relying on a simple point and click feedback, instead of requiring the user to infer garbled text snippets [14]. The illustrative 2D model of Fig. 1 (b), shows a linear classifier trained on the two features of - Mouse movement speed and Click time. An adversary, aiming to evade detection by this classifier, starts by guessing the click time as a key feature (intuitive in this setting), and then proceeds to makes probes to the black box model C, to learn its behavior. Probes are made by going through the spectrum of average reaction times for humans[4], guided by the *Accept*(green)/*Reject*(red) feedback from the CAPTCHA server. The information learned by reconnaissance on the black box system, can then be used to modify the attack payload so as to subsequently evade detection. While, this example was simplistic, its purpose is to illustrate the adversarial environment in which classifiers operate. Practical deployed classifiers tend to be more complex, non linear and multidimensional. However, the same reasoning and approach can be used to evade complex systems. An example of this is the good words and bad words attacks on spam detection systems [10,27]. By launching two spam emails each differing in only one word *'Sale'*, it can be ascertained that this word is important to the classification, if the email containing that word is flagged as spam. Knowing this information, the adversary can modify the word to be *'Sa1e'*, which looks visually the same but avoids detection. These evasion attacks are non-intrusive in nature and difficult to eliminate by traditional encryption/security techniques, because they use the same access channels as regular input samples and they see the same black box view of the system. From the classification perspective, these attacks occur at test time and are aimed at increasing the false negative rate of the model, i.e. increase the number of *Malicious* (positive) samples classified as *Legitimate*(negative) by C [4,5].

In addition to increasing the false negative rate of the defender's model, an intelligent data driven adversary would be interested in creating attacks of high diversity or variability. Such attacks are difficult to stop using simple blacklisting techniques and will require a more laborious task of retraining the classifier, using newly collected and curated data. We are interested in the analysis of such adversaries, as they are sophisticated, data driven, and intend to leave the machine learning model unusable over the long run.

Data driven attacks on deployed classification systems, presents a symmetric flip side to the task of learning from data. Instead of learning from labeled data to generate a model, the task of an attacker is to learn about the model, to generate evasive data samples [1]. With this motivation, we propose the Seed-Explore-Exploit(SEE) framework in this paper, to analyze the attack generation process as a learning problem, from a purely data driven perspective and without incorporating any domain specific knowledge. Research work on detecting concept drift in data streams[38,39], motivated the need for a formal analysis of the vulnerabilities of machine learning, with an initial evaluation proposed in our work in [40]. In this paper, we extend the earlier version with: i) incorporating and evaluating effects of *Diversity* of attacks on the defender's strategy, ii) introducing adversarial metrics of attack quality and the effects of varying the parameters of attack algorithms, iii) extensive detailed experimentation of the framework using a variety of defender models and the Google Cloud Prediction Service, and iv) experimentation simulating effects of diversity on blacklisting based countermeasures. To the best of our knowledge, this is the first work which presents a comprehensive adversarial model for indiscriminate exploratory black box attacks on classifier models. This work goes beyond [40], by providing a more in depth and thorough representation of adversarial activity, which can be reused for vulnerability analysis of future developed countermeasures against adversarial machine learning. The main contributions of this paper are:

- A domain independent data driven framework is presented, to simulate attacks using an Exploration-Exploitation strategy. This generic framework and the algorithms presented, can be used to analyze simple probing attacks to more sophisticated reverse engineering attacks.

[1] https://aws.amazon.com/machine-learning/.
[2] cloud.google.com/machine-learning.
[3] www.google.com/recaptcha.
[4] www.humanbenchmark.com/tests/reactiontime.

Please cite this article as: T.S. Sethi, M. Kantardzic, Data driven exploratory attacks on black box classifiers in adversarial domains, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.007