Contents lists available at ScienceDirect

Expert Systems With Applications

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

Handling adversarial concept drift in streaming data

Tegjyot Singh Sethi[*], Mehmed Kantardzic

Data Mining Lab, University of Louisville, Louisville, USA

ARTICLE INFO

Article history:
Received 16 July 2017
Revised 10 December 2017
Accepted 11 December 2017
Available online 11 December 2017

Keywords:
Adversarial machine learning
Concept drift
Streaming data
Limited labeling
Active learning
Classification

ABSTRACT

Classifiers operating in a dynamic, real world environment, are vulnerable to adversarial activity, which causes the data distribution to change over time. These changes are traditionally referred to as concept drift, and several approaches have been developed in literature to deal with the problem of drift detection and handling. However, most concept drift handling techniques approach it as a domain independent task, to make them applicable to a wide gamut of reactive systems. These techniques are developed from an adversarial agnostic perspective, where they naively assume that adversarial activity is like any other change to the data, which can be fixed by retraining the models. However, this is not the case when a malicious agent is trying to evade the deployed classification system. In such an environment, the properties of concept drift are unique, as the drift is intended to degrade the system and at the same time designed to avoid detection by traditional concept drift detection techniques. This special category of drift is termed as adversarial drift, and this paper analyzes its characteristics and impact in a streaming environment. A novel framework for dealing with adversarial concept drift is proposed, called the *Predict-Detect* streaming framework. This framework uses adversarial forethought and incorporates the context of classification into the drift detection task, to provide leverage in dynamic-adversarial domains. Experimental evaluation of the framework, on generated adversarial drifting data streams, demonstrates that this framework is able to provide early and reliable unsupervised indication of drift, and is able to recover from drifts swiftly. While traditional drift detectors can be evaded by intelligent adversaries, the proposed framework is especially designed to capture adversaries by misdirecting them into revealing themselves. In addition, the framework is designed to work on imbalanced and sparsely labeled data streams, as a limited-memory, incremental algorithm. The generic design and domain independent nature of the framework makes it applicable as a blueprint for developers wanting to implement reactive security to their classification based systems.

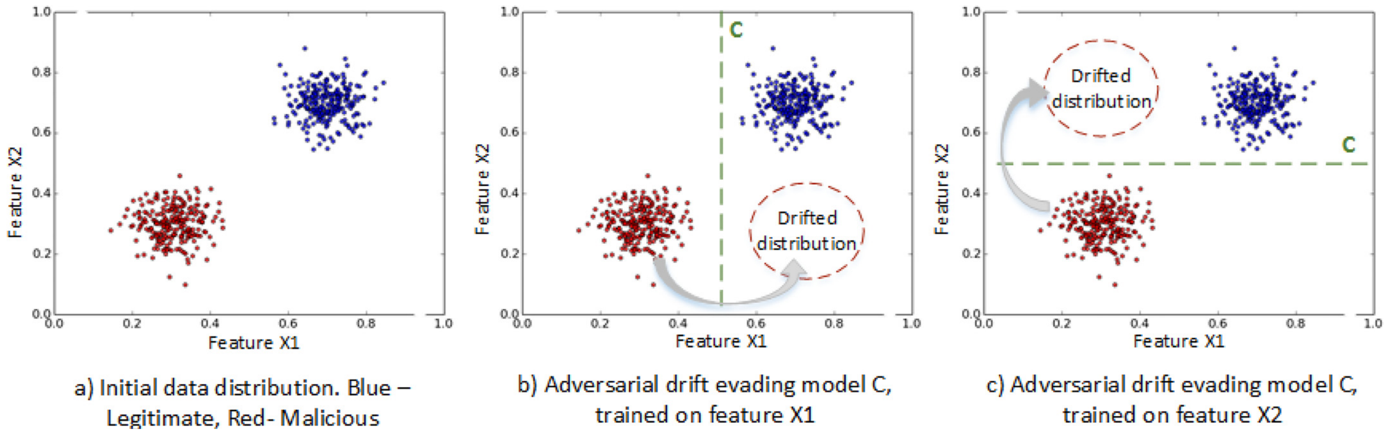
1. Introduction

Scalable data driven applications have led to scalable security problems. A potential breach now has the capability to affect billions of users and has the ability to bring mighty corporations to their knees. While more and more industries have started relying on expert systems to tackle some of the otherwise hard to approach problems, the black box nature of machine learning models leaves the systems open to unknown and unforeseen attack surfaces. The fear of losing out on the next big economic and infrastructure leverage has led to the proliferation of machine learning techniques into various application domains. However, there is a pressing need to look under the hood of these algorithms and question our assumptions and knowledge of such systems, to make sure they are not themselves introducing vulnerabilities in the systems they set out to secure.

One such assumption which needs further scrutiny, is that of *Stationarity* of the data distribution. Data in the real world is non stationary, which can cause the trained models to represent a different view of the world than the one it is trying to generalize over at test time. The changes in the data distribution are referred to as concept drift, and they can cause the performance of the predictive model to drop over time (Gama, Medas, Castillo, & Rodrigues, 2004; Žliobaitė, 2010). While existing works have approached this problem as a domain agnostic one, regarding any change in the data distribution equally, the intricacies and causes of such changes are often ignored. One such ubiquitous cause of data changes is adversarial activity, where an attacker changes its tactics to evade detection by the deployed classifier system. A typical attacker starts by learning the behavior of the defender's classifier model,

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

https://doi.org/10.1016/j.eswa.2017.12.022



a) Initial data distribution. Blue – Legitimate, Red- Malicious

b) Adversarial drift evading model C, trained on feature X1

c) Adversarial drift evading model C, trained on feature X2

Fig. 1. Illustration of adversarial drift, as a function of the defender's classifier model *C*.

using crafted probes, and then exploits this information to generate attack samples, to evade classification (Barreno, Nelson, Joseph, & Tygar, 2010; Biggio et al., 2013; Sethi & Kantardzic, 2017a; Sethi, Kantardzic, & Ryu, 2017). These attacks leads to a change in the distribution of the data, at test time, and also leads to a drop in the prediction capabilities of the defender's model. From the perspective of streaming data mining, we refer to such changes in the data distribution at test time as *Adversarial Drifts* (Kantchelian et al., 2013). For a classifier operating in such an adversarial and dynamic environment it is essential to understand the specific nature of such drifts, so as to develop robust reactive intelligent system.

Adversarial drift is a special type of concept drift. The main characteristics of adversarial drift which distinguishes it from traditional concept drift are: a) The drift is a result of changes to the malicious class samples only, b) The drift is a function of the deployed classifier model, as the adversary reverse engineers and actively tries to evade it Barreno, Nelson, Sears, Joseph, and Tygar (2006), and c) The drift is always targeted towards subverting the deployed classifier (i.e., it is relevant only if it leads to a drop in the performance of the deployed model) (Sethi et al., 2017). The model dependent nature of adversarial drift is shown in Fig. 1, where the deployed classifier *C* dictates the possibility of adversarial drifts in the data space. The figures b) and dummyTXdummy- c) demonstrate adversarial drifts, which are caused by an attacker trying to subvert *C*. The two scenarios are a result of the different defender models, which the adversary is trying to learn and circumvent. The nature of the drift is dependent on the choice for *C* and as such the model designer has a certain degree of control over the possible space of drifts, at test time.

Handling concept drift is challenging due to the scarce availability of labeled data, when operating in a fast moving streaming data milieu. Traditional concept drift handling mechanism use labeled data to continuously validate the performance of a trained model (Gonalves Jr., Santos, Barros, & Vieira, 2014). A drift is detected when performance over a current chunk of labeled data is seen to drop. However, this is not a practical solution in a streaming environment, as human expertise in the form of labeled data is often expensive and time taking to obtain (Masud et al., 2012; Sethi & Kantardzic, 2017b). To overcome this limitation, unsupervised drift detection methodologies have been proposed (Lee & Magoules, 2012; Lindstrom, Mac Namee, & Delany, 2013; Masud, Gao, Khan, Han, & Thuraisingham, 2011; Ryu, Kantardzic, Kim, & Khil, 2012; Sethi, Kantardzic, & Hu, 2016b; Spinosa, de Leon F de Carvalho, & Gama, 2007). These methodologies monitor the feature distribution of data to indicate possible drifts. The problem with these approaches is that they are fundamentally prone to false alarms, originating from their excessive pessimism. These methods cannot distinguish between changes that affect a classifier's performance and those which do not.

The problem of reliability of the drift detection was addressed in the recent works of Sethi and Kantardzic (2017b), where the Margin Density Drift Detection (MD3) approach was developed. By including the context of classification, to the drift detection task, the MD3 approach was shown to provide domain independent and reliable indication of drift from unlabeled data. Like most approaches in the area of concept drift detection (Gonalves Jr. et al., 2014), the domain agnostic nature of the MD3 approach was shown to be an advantage in Sethi and Kantardzic (2017b). However, in an adversarial environment, disregarding of domain characteristics leads to missed opportunity and can make the detection process vulnerable to adversarial evasion. The drift is characterized by an attacker continuously trying to hide its trail, by learning about the behavior of the detection system first. As such, the drift detection can itself be vulnerable to adversarial manipulation at test time. Most drift detection methodologies are designed as adversarial agnostic approaches, where they consider drift to be independent of the deployed classifier. In an adversarial domain, the relation between the type of drift and the choice of the classifier *C* is strongly coupled (Barreno et al., 2006). It is necessary to understand and acknowledge this intricacy to make unsupervised drift detection applicable in such a domain.

An adversarial-aware, unsupervised drift detection approach will take preemptive steps at the design of the classifier model, to ensure that future detection and retraining is made easier. In this paper, the *Predict-Detect* framework is proposed, as an adversarial-aware unsupervised drift detection methodology, capable of signaling malicious activity with high reliability. The framework incorporates the ideas of reliability, which makes the MD3 (Sethi & Kantardzic, 2017b) methodology effective, and presents a novel streaming data system capable of providing life-long learning in adversarial environments. The developed methodology incorporates attack foresight into a preemptive design to provide long term benefits for reactive security. The framework is developed as an ensemble-based, application and classifier-type independent approach capable of working on streaming data with limited labeling. To the best of our knowledge, this is the first work which directly addresses the problem of adversarial concept drift in streaming data. The main contributions of the proposed work are as follows:

- Analyzing the characteristics of adversarial concept drift and the impact of the deployed classifier on the drifts generated at test time.
- Developing the *Predict-Detect* classifier framework, as a novel approach to dealing with adversarial concept drift in streaming data with limited labeling.