Expert Systems With Applications 82 (2017) 77–99

Contents lists available at ScienceDirect

Expert Systems With Applications

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

On the reliable detection of concept drift from streaming unlabeled data

Tegjyot Singh Sethi[a,*], Mehmed Kantardzic[a]

Data Mining Lab, University of Louisville, Louisville, USA

ARTICLE INFO

Article history:
Received 5 December 2016
Revised 15 March 2017
Accepted 3 April 2017
Available online 4 April 2017

Keywords:
Concept drift
Streaming data
Unlabeled
Margin density
Ensemble
Cybersecurity

ABSTRACT

Classifiers deployed in the real world operate in a dynamic environment, where the data distribution can change over time. These changes, referred to as concept drift, can cause the predictive performance of the classifier to drop over time, thereby making it obsolete. To be of any real use, these classifiers need to detect drifts and be able to adapt to them, over time. Detecting drifts has traditionally been approached as a supervised task, with labeled data constantly being used for validating the learned model. Although effective in detecting drifts, these techniques are impractical, as labeling is a difficult, costly and time consuming activity. On the other hand, unsupervised change detection techniques are unreliable, as they produce a large number of false alarms. The inefficacy of the unsupervised techniques stems from the exclusion of the characteristics of the learned classifier, from the detection process. In this paper, we propose the Margin Density Drift Detection (MD3) algorithm, which tracks the number of samples in the uncertainty region of a classifier, as a metric to detect drift. The MD3 algorithm is a distribution independent, application independent, model independent, unsupervised and incremental algorithm for reliably detecting drifts from data streams. Experimental evaluation on 6 drift induced datasets and 4 additional datasets from the cybersecurity domain demonstrates that the MD3 approach can reliably detect drifts, with significantly fewer false alarms compared to unsupervised feature based drift detectors. At the same time, it produces performance comparable to that of a fully labeled drift detector. The reduced false alarms enables the signaling of drifts only when they are most likely to affect classification performance. As such, the MD3 approach leads to a detection scheme which is credible, label efficient and general in its applicability.

1. Introduction

Machine Learning has ushered in an era of data deluge, with the increasing scale and reach of modern day web applications (Wu, Zhu, Wu, & Ding, 2014). Classification has been adopted as a popular technique for providing data-driven prediction/detection capabilities, at the core of several otherwise complicated or intractable tasks. The ability to generalize and extrapolate from data has made its usage attractive as a general approach to data driven problem solving. However, the generalization ability of classifiers relies on an important assumption of *Stationarity*- which states that the training and the test data should be Identically and Independently Distributed (IID), derived from the same distribution (Zliobaite, 2010). This assumption is often violated in the real world, where dynamic changes occur constantly. These changes in the data distribution, called Concept Drift, can cause the predictive performance of the classifiers to degrade over time. To ensure that classifiers operating in such dynamic environments are useful and not obsolete, an adaptive machine learning strategy is warranted, which can detect changes in data and then update the models as new data becomes available.

While several adaptive techniques have been proposed in literature (Baena-Garcıa et al., 2006; Bifet & Gavalda, 2007; Gama, Medas, Castillo, & Rodrigues, 2004; Goncalves, de Carvalho Santos, Barros, & Vieira, 2014), they rely on the unhindered and unbounded supply of human expertise, in the form of labeled data, to detect and adapt to drifting data. In a streaming environment, where data flows in constantly, such constant human intervention is impractical, as labeling is time consuming, expensive and in some cases, not a possibility at all (Krempl et al., 2014; Lughofer, Weigl, Heidl, Eitzinger, & Radauer, 2016). To highlight the problem of label dependence, consider the task of detecting hate speech from live tweets (Burnap & Williams, 2016), using a classification system facing the twitter stream (estimated at 500M daily

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

http://dx.doi.org/10.1016/j.eswa.2017.04.008
0957-4174/

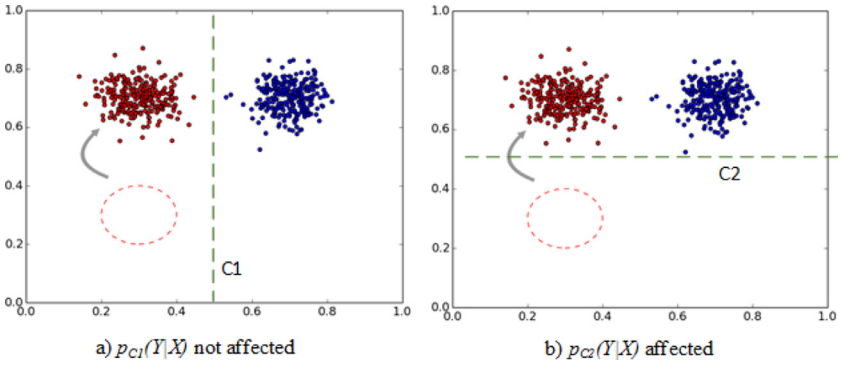

Fig. 1. Drift as a function of the learned classifier model.

tweets[1]). If 0.5% of the tweets are requested to be labeled, using crowd sourcing websites such as Amazon's Mechanical Turk,[2] this would imply a daily expenditure of $50K (each worker paid $1 for 50 tweets), and a continuous availability of 350 crowd sourced workers (assuming each can label 10 tweets per minute, and work for 12 h/day), every single day, for this particular task alone. The scale and velocity of modern day data applications makes such dependence on labeled data a practical and economic limitation. Streaming data applications need to be able to operate and detect drifts from unlabeled, or atmost sparsely labeled data, to be of any real use.

Although the use of labeled data for retraining and updating models is largely unavoidable, its use for the purpose of drift detection is superfluous. The need for constant validation of the learned model leads to wasted labels, which are discarded when the model is found to be stable (Sethi & Kantardzic, 2015). This has motivated the development of unlabeled drift detection techniques (da Costa, Rios, & de Mello, 2016; Ditzler & Polikar, 2011), which monitor changes to the feature distribution, as an early indicator of drift. However, existing methods using unlabeled data are essentially change detection techniques that detect any change to the data distribution, irrespective of its effects on the classification process (da Costa et al., 2016; Ditzler & Polikar, 2011; Kuncheva & Faithfull, 2014; Lee & Magoules, 2012; Qahtan, Alharbi, Wang, & Zhang, 2015). For the task of classification, change is relevant only when it causes model performance to degrade. This relevance is a function of the learned model, as illustrated in Fig. 1, where the same data shift resulted in diametrically opposite results. In Fig. 1a), the model performance is unaffected, while in b), there is a complete failure in the prediction capabilities of C2. The difference lies in the classier models C1 and C2, which are a result of learning on different views of the same data. The existing unlabeled techniques fail to make this distinction between the two cases, as they totally exclude the classifier from the detection process and make decisions solely on the distribution characteristics of the unlabeled data. This results in increased sensitivity to change and a large number of generated false alarms. False alarms in drift detection makes the algorithm overly paranoid and leads to wasted labeling effort, which is spent to verify relevance of the change.

From a probabilistic perspective, concept drift can be seen as a change in the joint probability distribution of the data samples X and their corresponding class labels Y, as per Eq. (1) (Gao, Fan, & Han, 2007). Unlabeled change detection techniques track changes to $P(X)$, while the labeled drift detection approaches directly track $P(Y|X)$. In this paper, an unlabeled drift detection methodology is proposed, which can vicariously track changes to $P(Y|X)$, without needing explicit labeled samples. Changes are tracked based on the distribution of sample relative to the learned classifier's boundary, to make it robust towards irrelevant changes in distribution of data.

$$P(X, Y) = P(Y|X).P(X) \tag{1}$$

The Margin Density Drift Detection (MD3) methodology, proposed in this paper, monitors the number of samples in a classifier's region of uncertainty (its margin), to detect drifts. Robust classifiers, such as Support Vector Machines (SVM) (Chang & Lin, 2011) or a feature bagged ensemble (Bryll, Gutierrez-Osuna, & Quek, 2003), after training, have regions of uncertainty called margins as depicted in Fig. 2. These regions are a result of the classifier's attempt to generalize over unseen data and they represent the model's best guess over that data space. A large margin width with a low density (given by number of samples) , is at the core of any optimization based classification process (such as SVM). While, explicit information about class distribution is learned in the training of a classifier, an additional auxiliary information also learned and often overlooked is the margin characteristics, such as the expected margin density. This information is representative of the data state and any change in it could indicate Non-Stationarity. Margin is crucial to the generalization process and any changes to the margin density is worthy of further verification. Since the margin density can be computed from unlabeled data only, it could be used as a substitute to explicit labeled drift detection techniques, for monitoring changes in $P(Y|X)$. In case of classifiers without explicit notions of margins, the generalization regions (Blindspots) are still prevalent, allowing the application of the margin density technique to a wide variety of algorithms, under an ensemble framework (Sethi, Kantardzic, & Arabmakki, 2016a). With this motivation, the MD3 methodology is proposed as a application independent, classifier independent, unlabeled and incremental approach to reliably signal concept drift from streaming data.

While false alarms in change detection are a hassle, due to the increased labeling expenditure and the need for frequent verification, this behavior is especially undesirable in cybersecurity applications because- (a) Frequent false alarms annoys experts, who provide model verification, causing the detection process to lose credibility, (b) An overly reactive system can be used by an adversary to manipulate learning, or to cause it to spend an excessive amount of money on labeling (Barreno, Nelson, Joseph, & Tygar, 2010) and (c) Increased labeling due to false alarms are expensive (even using crowd-sourcing websites at large scale, every day is expensive) and they cause delay in the detection of attacks. As such, we will evaluate the MD3 approach as a domain independent methodology first and then also evaluate its applicability as a reliable drift detection approach in adversarial streaming domains.

[1] http://www.internetlivestats.com/twitter-statistics (2016).
[2] www.mturk.com.