Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Q1 Tracking concept drift using a constrained penalized regression combiner

Q2 Li-Yu Wang[a], Cheolwoo Park[a], Kyupil Yeon[b], Hosik Choi[c,*]

[a] *Department of Statistics, University of Georgia, Athens, GA 30602, USA*
[b] *Department of Applied Statistics, Hoseo University, Asan, Chungnam, 31499, Republic of Korea*
[c] *Department of Applied Information Statistics, Kyonggi University, Suwon, Kyonggi-do, 16227, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 4 November 2015
Received in revised form 4 October 2016
Accepted 1 November 2016
Available online xxxx

*Keywords:*
Ensemble methods
Clustering
Streaming data

ABSTRACT

The objective of this work is to develop a predictive model when data batches are collected in a sequential manner. With streaming data, information is constantly being updated and a major statistical challenge for these types of data is that the underlying distribution and the true input–output dependency might change over time, a phenomenon known as concept drift. The concept drift phenomenon makes the learning process complicated because a predictive model constructed on the past data is no longer consistent with new examples. In order to effectively track concept drift, we propose model-combining methods using constrained and penalized regression that possesses a grouping property. The new learning methods enable us to select data batches as a group that are relevant to the current one, reduce the effects of irrelevant batches, and adaptively reflect the degree of concept drift emerging in data streams. We demonstrate the finite sample performance of the proposed method using simulated and real examples. The analytical and empirical results indicate that the proposed methods can effectively adapt to various types of concept drift.

## 1. Introduction

We consider a sequential regression problem in data streams which constitute a series of data batches entering over time continuously. Each data batch has observations, i.e. input–output pairs, which are to be used to construct a learning model for predicting the future observations. One of the main concerns in the analysis of data streams is the so called *concept drift* phenomenon, in which the target concept to be predicted, e.g. input–output relationship, changes over time. More generally, the underlying distribution that generates data might not be the same for different batches.

The term concept drift has been informally used for referring to any changes of circumstance in a learning process. Kelly et al. (1999) used *population drift* to describe changes in the probability distributions in a classification problem, such as class priors, distributions of classes, and posterior distributions of class memberships. Widmer and Kubat (1996) stated that change of target concept can be induced by changes in the hidden contexts, not given explicitly in the form of predictive features. Lazarescu et al. (2004) defined concept drift in terms of consistency and persistence. Yang et al. (2005) used the term *concept change* and *sampling change*, each of which corresponds to the real concept drift and virtual drift, and divided concept change into concept drift and concept shift to denote gradual drift and abrupt drift, respectively. Tsymbal (2004) stated that from a practical point of view it is not essential to discriminate real and virtual concept drift since the current

[*] Corresponding author.
*E-mail address:* choi.hosik@gmail.com (H. Choi).

http://dx.doi.org/10.1016/j.csda.2016.11.002

model needs to be adjusted in both cases. Kuncheva (2004) summarized the types of change roughly into random noise, gradual changes, abrupt changes, and recurring contexts, depending on which strategy for updating the current model are required. It is noted in the paper that the noise must not be modeled but filtered out, and past knowledge should be reused for stable contexts. It is further noted that a moving window on the training data sequences may be appropriate for a gradual change, but rebuilding a classifier entirely may be preferred for an abrupt change. Therefore, regardless of the degree of drift, the current predictive model should be adjusted because past models will not produce results consistent with the current inputs.

Many researchers have suggested theoretical insights and practical learning algorithms appropriate for concept drift. Bartlett (1992) investigated the learning theory when the probability distribution of the data changes slowly but continuously throughout the learning process. Others (Valiant and Waltz, 1984; Kuh et al., 1990; Helmbold and Long, 1994) have conducted research on the ability to learn drifting concepts under restrictions on the type of admissible concept changes from a viewpoint of PAC (probably approximately correct) learning. From a practical point of view, however, these theoretical results are related only to a gradual concept drift and are driven from a simple, time window-based tracking of the change.

Practical learning algorithms for drifting concepts can be categorized into single predictive model approach and model ensemble approach. In the single model approach, it is essential to select past examples or data batches consistent with the concept that is reflected in the most recent data batch; thus, an adaptive time window is frequently used to select or weight batches. See Kubat and Widmer (1995), Widmer and Kubat (1996), Klinkenberg and Joachims (2000) and Klinkenberg (2004) for examples of the single model approach. In contrast, the ensemble approach combines multiple base models and constructs or restructures an ensemble learner as a final predictive model. Relevant algorithms include Street and Kim (2001), Wang et al. (2003), Kolter and Maloof (2007), and Yeon et al. (2010). Because ensemble methods for concept drift are typically a weighted average of base learners' outputs, the method used to determine weights is critical. The algorithm proposed by Yeon et al. (2010) is to estimate the weights by training a combiner that is a constrained ridge regression applied to the level-1 data set as in stacked generalization (Wolpert, 1992). Generally, ensemble methods produce better prediction accuracy than single model approaches (Yeon et al., 2010), and they work well whether there is no or sharp concept drift.

The objective of the proposed work is to develop an ensemble method that is adaptive to both no or sudden concept drifts. We use a penalized regression combiner for the following reasons. When there are no (or little) drifts, the outputs of base models would be similar across batches, which causes high correlation among them. Thus, the weights corresponding to a similar cluster of batches will be approximately equal during this period. In contrast, if a sudden drift occurs, the outputs right after the change would be completely different from those of other batches. In this case, a combiner should be able to select the batches relevant to the current one and reduce the effects of other irrelevant batches. Yeon et al. (2010) showed that the constrained ridge regression combiner possesses some of the aforementioned desirable properties in that it can track drifting concepts well. However, their method does not select batches as groups, and the estimated weights do not seem to appropriately represent the current concept changes, as demonstrated in our numerical study. This motivates us to propose a fused penalized regression combiner with constraints that can take both group batch selection and multicollinearity into account. It groups base models with similar contributions to predicting the current target concept by forcing the weights of all the models within a group to be equal. We demonstrate its superior performance over existing methods via simulated and real examples.

The rest of the paper is organized as follows. The next section introduces the general framework of constrained penalized regression combiner for drifting concepts in a regression setting. It also discusses the ridge regression combiner proposed by Yeon et al. (2010) in detail. In Section 3, we propose a new ensemble combiner based on fused penalized regression and illustrate the implementation of its algorithm. We also analytically show the desirable properties of the proposed method. We compare the empirical performance of the proposed method with others using simulated and real examples in Section 4. Finally, we conclude and discuss a possible extension to classification problems in Section 5.

## 2. Constrained penalized regression approach

We introduce a constrained penalized regression aggregation approach in a regression setting. We borrow notations from Yeon et al. (2010) to show our generalization of their approach. Let $\{D_m,\ m = 1, 2, \ldots, M\}$ be a sequence of data batches, each consisting of input–output pairs. Suppose that observations in $D_m$ are random samples from unknown distributions $F_m(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^p$ is a predictor vector and $y \in \mathbb{R}$ is a response, i.e. they arise from a statistical model

$$y = f_m(\mathbf{x}) + \epsilon,$$

where $\epsilon$ is a random error satisfying $E(\epsilon) = 0$ and is independent of $\mathbf{x}$. The objective of learning is to construct a predictive model for future instances based on the batches $D_1, \ldots, D_M$. Note that the underlying distribution of generating data can be different for each batch, which implies the relationship between the response and predictors can change over time.

For prediction we obtain a training set from the outputs of base models and the most recent data batch $D_M = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ that best reflects the target concept to be tracked. For each $\mathbf{x}_i,\ i = 1, \ldots, n$, we denote the outputs of base models by $\hat{\mathbf{f}}(\mathbf{x}_i) = (\hat{f}_1(\mathbf{x}_i), \ldots, \hat{f}_{M-1}(\mathbf{x}_i), \hat{f}_M^{(-i)}(\mathbf{x}_i))^{\mathrm{T}}$, where $\hat{f}_M^{(-i)}(\mathbf{x}_i)$ is a leave-one-out estimate of $\hat{f}_M(\mathbf{x}_i)$ to avoid an

Please cite this article in press as: Wang, L.-Y., et al., Tracking concept drift using a constrained penalized regression combiner. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.11.002