Expert Systems With Applications 60 (2016) 39–50

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

Using dynamical systems tools to detect concept drift in data streams

F.G. da Costa[a,*], R.A. Rios[b], R.F. de Mello[a]

[a] *Institute of Mathematics and Computer Science, Universidade de São Paulo, São Carlos, 13566-590, Brazil*
[b] *Department of Computer Science, Universidade Federal da Bahia, Salvador, 40170-110, Brazil*

ARTICLE INFO

Article history:
Received 28 September 2015
Revised 19 April 2016
Accepted 20 April 2016
Available online 21 April 2016

MSC:
68Q32
68P20
68Wxx
65P20

Keywords:
Concept drift
Data streams
Dynamical systems
Chaos
Unsupervised learning

ABSTRACT

Real-world data streams may change their behaviors along time, what is referred to as concept drift. By detecting those changes, researchers obtain relevant information about the phenomena that produced such streams (e.g. temperatures in a region, bacteria population, disease occurrence, etc.). Many concept drift detection algorithms consider supervised or semi-supervised approaches which tend to be unfeasible when data is collected at high frequencies, due to the difficulties involved in labeling. Complementarily, current studies usually assume data as statistically independent and identically distributed, disregarding any temporal relationship among observations and, consequently, risking the quality of data modeling. In order to tackle both aspects, we employ dynamical system modeling to represent the temporal relationships among data observations and how they modify along time in attempt to detect concept drift. This approach considers Taken's immersion theorem to unfold consecutive windows of data observations into the phase space in attempt to represent and compare time dependencies. From this perspective, we proposed four new concept drift detection algorithms based on the unsupervised machine learning paradigm. The first algorithm builds dendrograms of consecutive phase spaces (every phase space represents the time relationships for the observations contained in a particular data window) and compare them out by using the Gromov–Hausdorff distance, providing enough guarantees to detect concept drifts. The second algorithm employs the Cross Recurrence Plot and the Recurrence Quantification Analysis to detect relevant changes in consecutive phase spaces and warn about relevant data modifications. We also preprocess data windows by considering the Empirical Mode Decomposition method and Mutual Information in attempt to take only the deterministic stream behavior into account. All algorithms were implemented as plugins for the Massive Online Analysis (MOA) software and then compared to well-known algorithms from literature. Results confirm the proposed algorithms were capable of detecting most of the behavior changes, creating few false alarms.

1. Introduction

Different application domains, such as climatology, network monitoring, health, and stock-market analysis, are characterized by the continuous production of large amounts of data in the form of streams, i.e., an open-ended sequence of observations whose behavior changes over time (Guha, Mishra, Motwani, & O'Callaghan, 2000). The application of traditional data mining approaches to model and extract information from such applications is prohibitive due the lack of resources to process and fully store data in main memory or even in secondary devices for later analysis (Babcock, Babu, Datar, Motwani, & Widom, 2002; Guha et al., 2000; O'callaghan, Meyerson, Motwani, Mishra, & Guha, 2002). This issue has motivated several studies to design algorithms to induce learning models by analyzing a data stream as its observations are collected. Such algorithms read observations only once and create a data summary to represent all relevant information, modeling the data stream at the current time instant (Aggarwal, Han, Wang, & Yu, 2003). Afterwards, such observations are discarded, consuming the least possible amount of main memory (Babcock et al., 2002; Guha et al., 2000). As new observations are collected, learning models are updated.

The comparison between past and current learning models makes possible to point out changes in data behavior along time. While small perturbations are seen as data instability, significant changes correspond to modifications in the phenomenon which produces the stream. The detection of these changes is the main motivation for concept drift area. Concept drift is the term used to define modifications in data streams along time, which are characterized by abrupt, usually easier to be detected, or by small changes (Tsymbal, 2004).

[*] Corresponding author.
E-mail addresses: fausto@icmc.usp.br, fausto.guzzo@gmail.com (F.G. da Costa), ricardoar@ufba.br (R.A. Rios), mello@icmc.usp.br (R.F. de Mello).

http://dx.doi.org/10.1016/j.eswa.2016.04.026
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

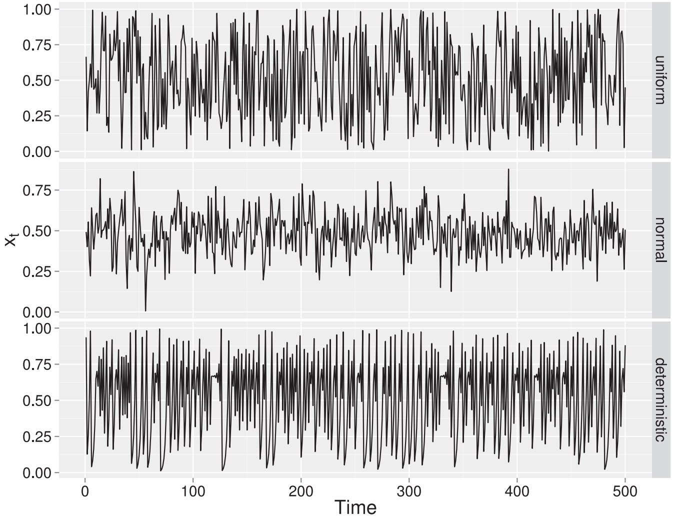

Fig. 1. Synthetic data streams produced using a Uniform distribution, a Normal distribution and a deterministic process.

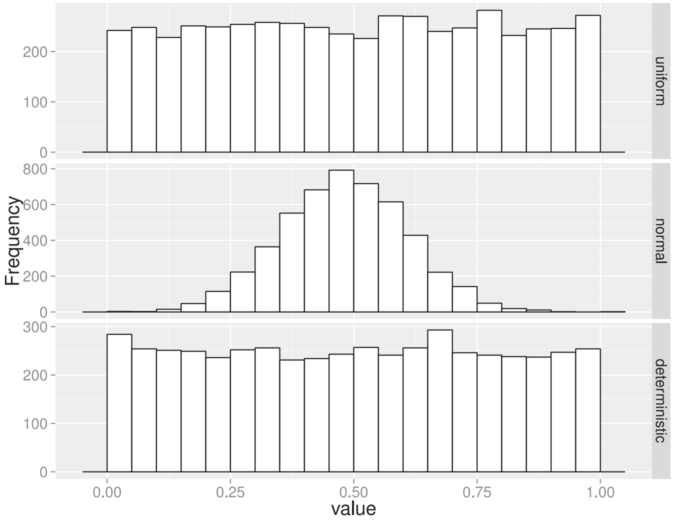

Fig. 2. Histograms produced using the data streams illustrated in Fig. 1. Observe it is possible to distinguish the Uniform from the Normal distribution, but it is not possible to separate it from the deterministic process.

Most of current studies assume that a data stream concept is composed of a set of statistically independent observations, which are typically modeled using linear and stationary processes such as the Normal and the Uniform distributions (Gama, Medas, Castillo, & Rodrigues, 2004; Klinkenberg & Joachims, 2000). Those studies rely on the estimation of probability distributions for consecutive windows of data, then they compare such estimations to point out changes (Bifet & Gavalda, 2007). We believe this characterization of a data stream concept misses the fact that data are produced along time and observations may contain dependencies which, once revealed, could better support the modeling of the stream behavior. As an example, consider three data stream concepts produced using different processes: (i) a Uniform distribution; (ii) a Normal distribution; and (iii) the deterministic process defined in Eq. 1

$$s_n = \frac{\arccos(-x_n)}{\pi}, \text{ in which } x_n = 1 - 2x_{n-1}^2. \qquad (1)$$

Fig. 1 illustrates the concepts produced by all three processes, while Fig. 2 shows the correspondent histograms used to estimate probability distributions. By analyzing the histograms, one can observe the Uniform and the Normal processes indeed characterize different concepts, however the deterministic process is quite similar to the Uniform one, making difficult to separate them out. In fact, this is among the most commonly considered approaches to tackle the problem of concept drift detection (Bifet & Gavalda, 2007).

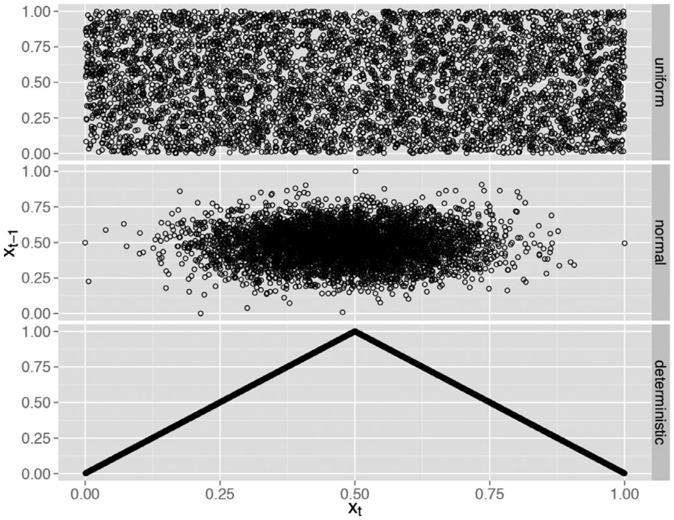

Fig. 3. Phase spaces reconstructed using the data streams illustrated in Fig. 1. Observe that all phase spaces are very different from each other, what allows us to point out three different concepts.

Aiming at overcoming this issue, we present a new approach that considers dynamical system tools. As a first step, we apply Takens' immersion theorem (Takens, 1981) to reconstruct data stream windows into a multidimensional space called phase space, making possible to observe differences as shown in Fig. 3. As one can observe, the phase space for the Uniform process spreads points all over the range, the Normal one accumulates more points around the center, while the deterministic process produces a characteristic function, commonly referred to as attractor (Alligood, Sauer, & Yorke, 1997). Thus, in this paper we consider phase spaces to model concepts and compare them out.

In addition to consider data observations as statistically independent, many of the studies employ either supervised or semi-supervised machine learning algorithms to detect concept drift (Aggarwal et al., 2003; Gama et al., 2004), i.e., they assume that the observations (or part of them) are labeled by specialists. We believe this assumption is hard to employ in many practical scenarios, once data can be produced at high frequencies or there is lack of information about the phenomenon from which the stream is generated. As a result, we decided to take only the unsupervised learning paradigm into consideration while developing this study.

We tackle this problem of concept drift detection in a different manner. First of all, we reconstruct data streams into a multidimensional space, also referred to as phase space (Alligood et al., 1997), in order to better understand the temporal relationships among observations. Then, we compare how points are distributed in such phase space using two different approaches: (i) the Permutation-Invariant Clustering Algorithm proposed by Carlsson and Mémoli (2010) and (ii) the Cross-Recurrence Plot (CRP) in conjunction with the Recurrence Quantification Analysis (RQA) measure Q_{max}, proposed by Serra, Serra, and Andrzejak (2009).

The first approach considers the Gromov–Hausdorff distance to compare ultrametric spaces, providing theoretical guarantees of representing the dissimilarities between models. The second approach allows the comparison between the dynamics of phase spaces, i.e., how the trajectories formed by two consecutive data