Information Sciences 265 (2014) 50–67

Contents lists available at ScienceDirect

Information Sciences

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

Combining block-based and online methods in learning ensembles from concept drifting data streams

Dariusz Brzezinski *, Jerzy Stefanowski

Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60–965 Poznan, Poland

ARTICLE INFO

Article history:
Received 21 March 2013
Received in revised form 30 September 2013
Accepted 14 December 2013
Available online 25 December 2013

Keywords:
Concept drift
Data stream
Online classifier
Ensemble

ABSTRACT

Most stream classifiers are designed to process data incrementally, run in resource-aware environments, and react to concept drifts, i.e., unforeseen changes of the stream's underlying data distribution. Ensemble classifiers have become an established research line in this field, mainly due to their modularity which offers a natural way of adapting to changes. However, in environments where class labels are available after each example, ensembles which process instances in blocks do not react to sudden changes sufficiently quickly. On the other hand, ensembles which process streams incrementally, do not take advantage of periodical adaptation mechanisms known from block-based ensembles, which offer accurate reactions to gradual and incremental changes. In this paper, we analyze if and how the characteristics of block and incremental processing can be combined to produce new types of ensemble classifiers. We consider and experimentally evaluate three general strategies for transforming a block ensemble into an incremental learner: online component evaluation, the introduction of an incremental learner, and the use of a drift detector. Based on the results of this analysis, we put forward a new incremental ensemble classifier, called Online Accuracy Updated Ensemble, which weights component classifiers based on their error in constant time and memory. The proposed algorithm was experimentally compared with four state-of-the-art online ensembles and provided best average classification accuracy on real and synthetic datasets simulating different drift scenarios.

1. Introduction

For the last decades, several machine learning and data mining algorithms have been proposed to discover knowledge from data [26,18,17]. However, such algorithms are usually applied to static, complete datasets, while in many new applications one faces the problem of processing massive data volumes in the form of transient data streams. Example applications involving processing data generated at very high rates include sensor networks, telecommunication, GPS systems, network traffic management, and customer click logs. The processing of streaming data implies new requirements concerning limited amount of memory, small processing time, and one scan of incoming examples [10,33,22], none of which are sufficiently handled by traditional learning algorithms and, therefore, require the development of new solutions.

However, the greatest challenge in learning classifiers from data streams is reacting to concept drifts, i.e., changes in distributions and definitions of target classes over time. Such changes are reflected in the incoming learning instances and deteriorate the accuracy of classifiers trained from past examples. Therefore, classifiers that deal with concept drifts are forced to implement forgetting, adaptation, or drift detection mechanisms in order to adjust to changing environments. Moreover,

* Corresponding author. Tel.: +48 61 665 29 43.
E-mail addresses: dariusz.brzezinski@cs.put.poznan.pl (D. Brzezinski), jerzy.stefanowski@cs.put.poznan.pl (J. Stefanowski).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.12.011

depending on the rate of these changes, concept drifts are usually divided into sudden or gradual ones, both of which require different reactions [34].

As standard data mining algorithms are not capable of dealing with concept drifts and rigorous processing requirements posed by data streams, several new techniques have been proposed [14,24]. Out of many algorithms proposed to tackle evolving data streams, ensemble methods play an important role. Due to their modularity, they provide a natural way of adapting to change by modifying their structure, either by retraining ensemble members, replacing old component classifiers with new ones, or updating rules for aggregating component predictions [23]. Current adaptive ensembles can be further divided into block-based and online approaches [14].

Block-based approaches are designed to work in environments were examples arrive in portions, called blocks or chunks. Most block ensembles periodically evaluate their components and substitute the weakest ensemble member with a new (candidate) classifier after each block of examples [33,35]. Such approaches are designed to cope mainly with gradual concept drifts. Furthermore, when training their components block-based methods often take advantage of batch algorithms known from static classification. The main drawback of block-based ensembles is the difficulty of tuning the block size to offer a compromise between fast reactions to drifts and high accuracy in periods of concept stability.

In contrast to block-based approaches, online ensembles are designed to learn in environments where labels are available after each incoming example. With class labels arriving online, algorithms have the possibility of reacting to concept drift much faster than in environments where processing is performed in larger blocks of data. Many researchers tackle this problem by designing new online ensemble methods, which are incrementally trained after each instance and try to actively react to concept changes [2,31]. Some of these newly proposed ensembles are usually characterized by higher computational costs than block-based methods and the used drift detection mechanisms often require problem-specific parameter tuning. Furthermore, online ensembles ignore weighting mechanisms known from block-based algorithms and do not introduce new components periodically, thus, they require specific strategies for frequent updates of incrementally trained components.

However, we argue that block-based weighting mechanisms as well as periodical component evaluations could be still of much value in online environments. We claim that the periodical introduction of new candidate classifiers and incremental updates of component classifiers should improve the ensemble's reactions to both sudden and gradual drifts in reasonable balance with computational costs. Our previous work concerning data stream ensembles suggests that by modifying block-based ensembles towards incremental classifiers one can improve classification performance [5,6]. These motivations led us to research questions, which should be examined prior to the construction of a new type of online ensemble: Would a modification of block-based ensembles towards incremental learners also be beneficial in an online processing environment? Is it profitable to retain periodic evaluations and weighting mechanisms known from block-based algorithms while constructing on-line ensembles for concept drifting data streams? Are periodical component evaluations and new classifier insertions better than incorporating an online drift detector? Additionally, can error-based weighting proposed for block-based methods be performed after each example, without the need of dividing data into blocks?

The first aim of our paper is to answer the presented research questions by reviewing existing block-based ensemble methods, considering different ways of adapting them to online learning, and experimentally evaluating the impact of proposed adaptation strategies. To the best of our knowledge, no such analysis has been previously done. Based on the results of this experimental study, we propose and experimentally evaluate a new online algorithm, called Online Accuracy Updated Ensemble, which tries to combine the best elements of block-based weighting and online processing. The contributions of our paper are as follows:

- In Section 3, we put forward three general strategies for transforming block-based ensembles into online learners. More precisely, we investigate: (1) the use of a windowing technique which updates component weights after each example, (2) the extension of the ensemble by an incremental classifier which is trained between component reweighting, and (3) the use of an online drift detector which allows to shorten drift reaction times. We identify which of these approaches are the most beneficial for creating a new online ensemble.
- In Section 4, we introduce a new incremental error-based weighting function which evaluates component classifiers as they classify incoming examples. Next, we put forward the Online Accuracy Updated Ensemble (OAUE), an algorithm which uses the proposed function to incrementally train and weight component classifiers.
- In Section 5, we experimentally compare the three proposed general transformation strategies and verify whether block-based algorithms can be successfully transformed into incremental learners. Furthermore, we perform a sensitivity analysis of the OAUE algorithm, analyze its weighting function, and experimentally compare OAUE with popular online ensembles on several real and synthetic datasets simulating environments containing sudden, gradual, incremental, and mixed drifts.
- In Section 6, we discuss the most important issues in transforming block-based ensembles and draw lines of further research.

2. Background and related works

We assume that learning examples from a stream $\mathcal{S}$ appear incrementally as a sequence of labeled examples $\{\mathbf{x}^t, y^t\}$ for $t = 1, 2, \ldots, T$, where $\mathbf{x}$ is a vector of attribute values and y is a class label ($y \in \{K_1, \ldots, K_l\}$). In this paper, we consider a completely supervised framework, where a new incoming example $\mathbf{x}^t$ is classified by a classifier C which predicts its class label.