Neurocomputing 150 (2015) 458–470

Contents lists available at ScienceDirect

Neurocomputing

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

Online tree-based ensembles and option trees for regression on evolving data streams

Elena Ikonomovska [a,*], João Gama [b], Sašo Džeroski [c]

[a] Turn Inc., 835 Main St, Redwood City, CA, United States
[b] LIAAD-INESC, Rua Dr. Roberto Frias, 378 4200-378 Porto, Portugal
[c] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form
25 April 2014
Accepted 26 April 2014
Available online 4 November 2014

Keywords:
Online learning
Data streams
Online regression trees
Option trees
Online ensembles
Online random forests

ABSTRACT

The emergence of ubiquitous sources of streaming data has given rise to the popularity of algorithms for online machine learning. In that context, Hoeffding trees represent the state-of-the-art algorithms for online classification. Their popularity stems in large part from their ability to process large quantities of data with a speed that goes beyond the processing power of any other streaming or batch learning algorithm. As a consequence, Hoeffding trees have often been used as base models of many ensemble learning algorithms for online classification. However, despite the existence of many algorithms for online classification, ensemble learning algorithms for online regression do not exist. In particular, the field of online any-time regression analysis seems to have experienced a serious lack of attention. In this paper, we address this issue through a study and an empirical evaluation of a set of online algorithms for regression, which includes the baseline Hoeffding-based regression trees, online option trees, and an online least mean squares filter. We also design, implement and evaluate two novel ensemble learning methods for online regression: online bagging with Hoeffding-based model trees, and an online RandomForest method in which we have used a randomized version of the online model tree learning algorithm as a basic building block. Within the study presented in this paper, we evaluate the proposed algorithms along several dimensions: predictive accuracy and quality of models, time and memory requirements, bias–variance and bias–variance–covariance decomposition of the error, and responsiveness to concept drift.

1. Introduction

Any-time online regression analysis is a research topic which tackles problems such as predicting traffic jams, power consumption, entertainment trends, and even flu outbreaks. These types of problems require online any-time predictive modeling and responsiveness to changes in near real-time. Ensemble learning methods are considered to be one of the most accurate and robust machine learning approaches for solving classification and regression tasks. They have been successfully used in many real-world problems, including the well known Netflix Prize[1] competition, where it was shown that best results can be achieved by combining multiple models and ensembles of models, each of them specializing in different aspects of the problem.

Common knowledge on ensemble learning methods suggests that the base models should be as accurate as possible and should have as diverse as possible distributions of errors. Although this seems as a simple requirement, it is not easily achieved [1]. From the various types of models, decision trees are particularly well suited for this task because they enable a simple and effective way to create an ensemble of diverse yet accurate hypotheses. At the same time, decision and regression trees are easy to interpret and can handle both numeric and nominal attributes. The most typical approach in learning ensembles of trees is to apply basic sampling techniques such as bagging and boosting, or to randomize the learning process.

Hoeffding-based algorithms for learning decision or regression trees [2–6] are one of the most popular methods for learning trees from data streams. Due to their ability to make theoretically supported split selection and stopping decisions in a single pass over the training data, without having to store any of the data instances, they are able to process large quantities of data at a speed that goes beyond the processing abilities of batch learning algorithms. As such, they are an invaluable tool for real-time prediction and analysis of streaming data. A recent new addition to the literature of mining data streams is the paper by Rutkowski et al. [7] in which the McDiarmid's bound is proposed as a substitution and a more suitable choice than the Hoeffding bound. Their proposal is supported by a nice theoretical analysis, however the experimental results presented in the paper are not conclusive.

* Corresponding author.
[1] http://www.netflixprize.com/

http://dx.doi.org/10.1016/j.neucom.2014.04.076
0925-2312/

While online ensembles of decision trees have been extensively studied, algorithms for learning online ensembles of regression trees have not been proposed or empirically evaluated yet. Consequently, in this work, we focus on online tree-based ensembles for any-time regression. Our base models are online Hoeffding-based regression trees, produced by streaming any-time algorithms such as the FIMT-DD algorithm [6] and a randomized version of FIMT-DD, termed R-FIMT-DD, presented in this paper for the first time. We consider two different ensemble learning methods, online bagging of Hoeffding-based trees for regression (OBag) and online random forest for any-time regression (ORF). We analyze and evaluate them in terms of their ability to achieve diversity among the constituent models, improve accuracy, as well as, in terms of their computational complexity and demand for resources. We also provide an extensive empirical comparison of our ensemble methods OBag and ORF to online option trees for regression (ORTO) [8] and show that option trees are an appealing alternative to ensembles of trees in terms of both accuracy and resource allocation.

In sum, we make the following contributions to the area of online learning for any-time regression:

1. We give a systematic overview of existing methods for learning tree-based ensembles for online classification.
2. We implement and empirically evaluate two new methods for learning tree-based ensembles for online regression: online bagging of FIMT-DD trees (OBag) and an online RandomForest method for regression, based on the randomized algorithm R-FIMT-DD (ORF). To the best of our knowledge, there is no other work that studies methods for learning tree-based ensembles for online regression on data streams.
3. We further extend the empirical evaluation by performing a comparison with online option trees for regression and show that option trees achieve better accuracy using fewer resources and less computational power.
4. By using a theoretical analysis of the sources of error we study the ways these different techniques improve the accuracy over the base models and try to correlate the diversity of the ensemble to its generalization power.

The remainder of the paper is organized as follows. In Section 2 we begin with an overview of ensemble methods for online classification. To the best of our knowledge ensembles methods for online regression have not been published yet. Section 3 introduces two novel ensemble methods for online regression on data streams. Section 4 presents an extended version of the work on online option trees for regression [8]. These algorithms are empirically evaluated using the evaluation methodology described in Section 5. In Section 6, we give the results from the empirical evaluation of OBag, ORF, ORTO and Hoeffding-based regression trees in different learning scenarios. Finally, Section 6.4 presents the differences between the two ensemble learning methods OBag and ORF in terms of the source of error and the amount of diversity they were able to introduce in their base models.

2. Online ensembles of decision trees

Although tree-based ensemble methods for online regression have not been studied nor analyzed yet, there is a plethora of tree-based ensemble learning methods for online classification [9–17]. In this section, we present a survey of the existing ensemble learning methods for online classification. In the following subsections, we will discuss *online bagging* and *online boosting* as main representatives of the category of methods which introduce diversity by diversification of the training data. These methods represent the basis for implementing online versions of bagging and boosting for learning various types of ensembles. Separate subsections discuss online RandomForest and stacked generalization with restricted Hoeffding trees.

2.1. Online bagging

The simplest method to introduce diversity among the base models that constitute an ensemble is to generate different samples from the training dataset, i.e., modify the training dataset each time a base model is learned. The batch *bagging* method generates a number of bootstrapped training sets from the original dataset. A bootstrap sample can be obtained by random sampling with replacement according to a Binomial probability distribution. In the next step, a separate model is induced on each training dataset. In the testing phase, the predictions from the base models are aggregated using majority voting. Online versions of sampling-based methods for learning ensembles have been studied and proposed for the first time in the work of [18].

2.1.1. Bootstrap sampling and online bagging

Before looking into the details of online bagging, let us first discuss the details of bootstrap sampling. The main effect of the bootstrap sampling procedure is to simulate repeated observations from an unknown population using the available sample as a basis. A training sample generated with batch bootstrap sampling contains K copies of each of the original training examples, where K is a random variable distributed according to the Binomial distribution. In order to transform the batch procedure into an online one, the number of occurrences of each training example for each training dataset has to be estimated at its first occurrence, without the possibility to examine the rest of the training dataset. Oza and Russel in [18] have observed that when the size of the training dataset tends to infinity; i.e., $N \to \infty$, the Binomial distribution tends to be more and more similar to the Poisson distribution Poisson(1) which is defined as

$$Poisson(\lambda) \sim \frac{e^{-1}}{k!} \quad \text{for } \lambda = 1,$$

where k is the number of occurrences of the discrete stochastic variable K. The formula gives us the means to compute the value of K for a given randomly chosen probability for each training example. At the same time, it applies perfectly to the online learning setup in which the size of the training set is unbounded and tends to infinity. [18] have proven that if the batch and the online bagging algorithms use the same training set that grows to infinity, the distributions over the bootstrapped training sets will converge to the same distribution.

The algorithm receives as input a set of initialized base models $H_M = \{h_1, \ldots, h_M\}$, i.e., one-leaf decision trees. For every training example e, received from the data stream, and for each base model h_i, the number of occurrences K of the example e in the online training set used to update h_i is set to $k = Poisson(1)$. In the next step, the example is given K times to the procedure which refines the corresponding model h_i. This procedure is repeated as long as there are training examples available to update the set of models.

2.2. Online bagging for concept drift management

The *online bagging* meta-algorithm has been used by several authors and mainly for the task of online classification. [19] have proposed two interesting algorithms for learning ensembles of Hoeffding trees for online classification: ADWIN Bagging and Adaptive-Size Hoeffding Tree (ASHT) Bagging. Both ADWIN Bagging and Adaptive-Size Hoeffding Tree (ASHT) Bagging have been developed to address the problem of learning under non-