Neurocomputing 166 (2015) 68–83

Contents lists available at ScienceDirect

Neurocomputing

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

Learning concept-drifting data streams with random ensemble decision trees[☆]

Peipei Li [a,*], Xindong Wu [a,b], Xuegang Hu [a], Hao Wang [a]

[a] Hefei University of Technology, Hefei, China, 230009
[b] University of Vermont, Vermont 05405, USA

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form
15 February 2015
Accepted 11 April 2015
Communicated by Hongli Dong
Available online 22 April 2015

Keywords:
Data streams
Random decision tree
Concept drift
Noisy data

ABSTRACT

Few online classification algorithms based on traditional inductive ensembling, such as online bagging or boosting, focus on handling concept drifting data streams while performing well on noisy data. Motivated by this, an incremental algorithm based on Ensemble Decision Trees for Concept-drifting data streams (EDTC) is proposed in this paper. Three variants of *random feature selection* are introduced to implement split-tests and two thresholds specified in Hoeffding Bounds inequality are utilized to distinguish concept drifts from noisy data. Extensive studies on synthetic and real streaming databases demonstrate that our algorithm of EDTC performs very well compared to several known online algorithms based on single models and ensemble models. A conclusion is hence drawn that multiple solutions are provided for learning from concept drifting data streams under noise.

1. Introduction

Industry and business applications generate a tremendous amount of data streams, such as telephone call records and sensor network data. Meanwhile, with the development of Web Services technology, streaming data from the Internet span across a wide range of domains, including shopping transactions and Internet search requests. In contrast to the traditional data sources of data mining, these streaming data present new characteristics as being continuous, high-volume, open-ended and concept drifting. It is hence a challenging and interesting issue for us to learn from concept drifting data streams.

As an important branch of data mining, classification can be solved by abundant inductive learning models [1–3]. Decision trees, due to their advantages, are one of the most popular techniques. They are widely used as a basic model in ensemble classifiers. However, it is still a challenge to learn from data streams with these traditional inductive learning models or algorithms. Thus, new models or algorithms tailored towards mining from data streams, based on decision trees, have been proposed. These recent works can be roughly classified into two main categories: *incremental decision tree based* and *random decision tree based*. Incremental decision tree based approaches for data stream classification use informed split-tests in the generation of decision trees, while random decision tree [4] based approaches select the split nodes randomly.

On one hand, main works based on incremental decision trees for data stream classification are summarized below, such as a classical continuously changing data stream algorithm of CVFDT (Concept-adapting Very Fast Decision Tree learner) [5], a Streaming Ensemble Algorithm (SEA) [6], an Accuracy Weighted Ensemble (AWE) algorithm [7], a Weighted Aging Ensemble (WAE) algorithm [8], an Additive Expert ensemble algorithm of AddExp [9] developed from the incremental ensemble method of DWM (Dynamic Weighted Majority) [10], a VFDTc [11] model extended from the VFDT (Very Fast Decision Tree) algorithm [12], a boosting-like method [13], two variants of bagging with adaptive windowing and bagging based on adaptive-size Hoeffding tree [14], the perceptron classifier based learning algorithms [15,16] and other variants of decision tree based algorithms [17,18], an incrementally Optimized Very Fast Decision Tree (iOVFDT) [19] and an Evolutionary-Adapted Ensemble (EAE) method [20]. All of the above methods enable adapting to the concept drifting data streams, but they cannot perform well in the prediction accuracy especially when handling noisy data streams. This is because these classification methods require informed split tests. It will lead to be impacted from the noisy data more significantly.

On the other hand, main works based on random decision trees for data stream classification are summarized below, such as a random decision tree ensembling method [21] with cross-validation estimation, an online algorithm of Streaming Random Forests [22] extended from

[☆] A preliminary version of this paper has been accepted for publication in the Proceeding of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2011.

[*] Corresponding author.
E-mail addresses: peipeili.hfut@gmail.com (P. Li), xwu@cs.uvm.edu (X. Wu), jsjxhuxg@hfut.edu.cn (X. Hu), jsjxwangh@hfut.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.04.024
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Random Forests [23], an algorithm using an entropy-based drift-detection technique for evolving data streams [24], an algorithm of Semi-Random Multiple decision Trees for Data Streams (SRMTDS) [25], an algorithm of Multiple Semi-Random decision Trees for concept-drifting data streams (MSRT) [26] and a Concept Drifting detection algorithm based on an ensemble model of Random Decision Trees (CDRDT) [27]. These methods mentioned above explore the efficiency and effectiveness in the classification of data streams using random split tests, which give us some insights for improving the efficiency and effectiveness in the classification of data streams with concept drifts and noisy data.

In this paper, we propose an incremental Ensemble algorithm based on random Decision Trees for Concept drifting data streams with noisy data called EDTC. This work is developed from our previous work in [28]. Our major contributions are as follows:

- We develop three variants of *random feature selection* to determine split-information in our algorithm, instead of informed split-tests in the aforementioned incremental decision tree based approaches. Because observations in [29,25] reveal that in comparison to the algorithms with informed split-tests in the heuristic methods, such as C4.5, bagging and boosting, the algorithms based on *random feature selection* are more effective and efficient even if in the case of noisy data.
- Contrary to most of the random decision tree based approaches mentioned above, we propose a new incremental ensemble algorithm based on random decision trees. A significant difference is that in our algorithm, each growing node in trees will not select split-features randomly until a real training instance arrives. It is conductive to avoid generating unnecessary branches. In addition, we know that the training of the ensemble classifiers can adapt to the incremental or gradual changes of data streams, but it only implicitly tells us that there are concept drifts in the data streams. To explicitly detect the types of concept drifts, in virtual of the Hoeffding Bounds inequality [30], we adopt a double-threshold-based mechanism of concept drifting detection for better tracking concept drifts and reducing the impact from noise. Instead of the global data distributions in all decision trees used in [27], we detect the concept drifts using the local data distributions in a decision tree. It is beneficial to improve the detection accuracy on the concept shifts (namely the sudden drifts) and the sampling changes [31].
- Contrary to our previous work in [28], we systematically analyze the generalization error of our EDTC algorithm impacted from the noisy data and the concept drifts. Meanwhile, we give more experiments conducted on the synthetic and real streaming data to show how our algorithm performs varying with the noisy data and the concept drifts. Extensive study shows that in comparison to the state-of-the-art online algorithms based on single and ensemble models mentioned above, our EDTC algorithm could efficiently and effectively tackle concept-drifting data streams with noisy data.

The rest of this paper is organized as follows. Section 2 reviews related work based on incremental decision trees and random decision trees. Our random ensemble random decision tree algorithm for concept drifting data streams is described in detail in Section 3. Section 4 provides the experimental study and Section 5 is the summary.

2. Related work

In this section, we first give the related works on the incremental decision tree based approaches for data stream classification, and then we give some works on random decision trees, and the random decision tree based approaches for data stream classification.

Incremental decision tree based approaches for data stream classification: We summarize main works based on incremental decision trees for data stream classification below. One classical continuously changing data stream algorithm of CVFDT (Concept-adapting Very Fast Decision Tree learner) [5] developed in 2001, using a fixed-size time window, constructs alternative sub-trees for each node to replace the old ones if concept drifts occur. Meanwhile, a Streaming Ensemble Algorithm (SEA) [6] addresses the concept drift of data streams, but it seems that the recovery time from concept drifts is demanded unnecessarily long. In 2003, an Accuracy Weighted Ensemble (AWE) algorithm was proposed for mining concept-drifting data streams using the more advanced weighted voting strategy [7]. A new algorithm called Weighted Aging Ensemble (WAE) [8] was developed from AWE in the following two modifications: (i) classifier weights depend on the individual classifier accuracies and the time they have been spending in the ensemble, (ii) individual classifiers are chosen to the ensemble on the basis of the non-pairwise diversity measure. An Additive Expert ensemble algorithm of AddExp [9] was developed from the incremental ensemble method of DWM (Dynamic Weighted Majority) [10] in 2005. It steadily updates the weights of experts and adds a new expert each time if misclassifying an instance. In 2006, a VFDTc model extended from the VFDT (Very Fast Decision Tree) algorithm [12] was designed for concept drifting data streams [11]. A boosting-like method [13] based on a classifier ensemble learned from data streams was presented in 2007 for quickly adapting to different kinds of concept drifts. Evidence shows that the advantages of boosting will be lost if there is non-trivial classification noise in the training test. In 2009, two new variants of Bagging with Adaptive Windowing and Bagging based on Adaptive-Size Hoeffding Tree [14] were developed for concept drifting data streams. In 2010 and 2011, the perceptron classifier based learning algorithms [15,16] and other variants of decision tree-based algorithms were proposed for evolving data streams [17,18,63]. In 2012, an incrementally Optimized Very Fast Decision Tree (iOVFDT) [19] was proposed for noisy data streams. In the following years, some evolutionary-adapted ensemble methods were proposed to process data streams characterized by the presence of recurring contexts [20,62]. All of the above methods are informed split-test based classification methods. They can adapt to the concept drifting data streams, but they cannot perform well in the prediction accuracy especially when handling noisy data streams. This is because the informed split-tests are more significantly impacted from the noisy data compared to the random selection in the split-tests.

Random decision trees: The concept of a *random decision forest* [32] was first proposed by Ho in 1995. It builds multiple trees in randomly selected subspaces of the feature space. In 1997, a *randomized tree* [33] was introduced by Amit. It generates feature subsets from all features randomly and uses the heuristic information entropy to select split-nodes. In 1998, Ho further designed a *random subspace* [34] technique to select random subsets of the available features in training individual classifiers of an ensemble. By using the bagging technique in tandem with random feature selection, Breiman proposed a model of *Random Forests* [23] in 2001. Contrary to the previous beliefs, a new model of *Random Decision Trees* [4] without any heuristics was designed by Fan in 2003. However, these algorithms cannot be applied in the handling of data streams due to the batch learning though they perform very well in the classification.

Random decision trees for data stream classification: New algorithms based on ensemble random decision trees subsequently have emerged for data stream handling. Main works are summarized below. A random decision tree ensembling method [21] with