Knowledge-Based Systems 104 (2016) 179–194

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Online adaptive decision trees based on concentration inequalities

Isvani Frías-Blanco[a,*], José del Campo-Ávila[b], Gonzalo Ramos-Jiménez[b], Andre C.P.L.F. Carvalho[c], Agustín Ortiz-Díaz[a], Rafael Morales-Bueno[b]

[a] Universidad de Granma, Departamento de Ciencias Informáticas, Carretera Bayamo-Manzanillo 17 $\frac{1}{2}$ km, Bayamo, Granma, Cuba
[b] Universidad de Málaga, Andalucía Tech, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, 29071 Málaga, Spain
[c] Universidade de Sao Paulo (USP), Sao Carlos, ICMC - USP Av Trabalhador Saocarlense, 400 - Centro Sao Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 12 October 2015
Revised 1 March 2016
Accepted 20 April 2016
Available online 21 April 2016

Keywords:
Adaptive learning
Concept drift
Data stream mining
Decision trees
Incremental learning
Online learning

ABSTRACT

Classification trees are a powerful tool for mining non-stationary data streams. In these situations, massive data are constantly generated at high speed and the underlying target function can change over time. The IADEM family of algorithms is based on Hoeffding's and Chernoff's bounds and induces online decision trees from data streams, but is not able to handle concept drift. This study extends this family to deal with time-changing data streams. The new online algorithm, named IADEM-3, performs two main actions in response to a concept drift. Firstly, it resets the variables affected by the change and maintains unbroken the structure of the tree, which allows for changes in which consecutive target functions are very similar. Secondly, it creates alternative models that replace parts of the main tree when they significantly improve the accuracy of the model, thereby rebuilding the main tree if needed. An online change detector and a non-parametric statistical test based on Hoeffding's bounds are used to guarantee this significance. A new pruning method is also incorporated in IADEM-3, making sure that all split tests previously installed in decision nodes are useful. The learning model is also viewed as an ensemble of classifiers, and predictions of the main and alternative models are combined to classify unlabeled examples. IADEM-3 is empirically compared with various well-known decision tree induction algorithms for concept drift detection. We empirically show that our new algorithm often reaches higher levels of accuracy with smaller decision tree models, maintaining the processing time bounded, irrespective of the number of instances processed.

1. Introduction

Online learning algorithms are suitable for mining huge amounts of data being constantly generated at high speed. Internet, phone, surveillance and sensor networks are common sources of these *data streams* [1]. Algorithms for mining data streams must perform with controlled computational resources: not all data can be stored in the available memory (a memory restriction), and learning algorithms must process data arriving at a high incoming rate (a fast processing time). Additionally, data streams can be non-stationary. For example, a user can change his/her news or clothing preferences and a learning system must track these changing preferences in order to give him/her appropriate suggestions. Thus, the target concept may change over time and a previous learning model can become outdated, a problem commonly known as *concept drift* [2].

Decision tree induction algorithms are one of the most studied machine learning techniques [3]. Despite their successful use in several classification tasks, algorithms like ID3 and C4.5 are batch learners. They are not suitable for mining data streams, as they assume that all training data are available during the learning process and that the data distribution is stationary. To overcome this limitation, several algorithms able to induce decision trees from data streams have been proposed [4–7]. The algorithms able to handle concept drift usually present a change detection method, which can trigger a forgetting mechanism to remove or to decrease the importance of outdated parts of the model, no longer significant when a concept drift occurs. This approach has been adopted for the well-known Concept-adapting Very Fast Decision Tree (CVFDT) algorithm [4,8] and many others, which were extended to be able to deal with time-changing data streams [9,10].

A key point in the induction of decision trees is the selection of split tests for the decision nodes. This selection, often based on a heuristic measure, such as information gain or gini index, tries

* Corresponding author. Tel.: +5516982077235.
E-mail addresses: ifriasb@udg.co.cu (I. Frías-Blanco), jcampo@lcc.uma.es (J.d. Campo-Ávila), ramos@lcc.uma.es (G. Ramos-Jiménez), andre@icmc.usp.br (A.C.P.L.F. Carvalho), aortizd@udg.co.cu (A. Ortiz-Díaz), morales@lcc.uma.es (R. Morales-Bueno).

http://dx.doi.org/10.1016/j.knosys.2016.04.019
0950-7051/

to optimize the predictive accuracy of the induced decision tree. In open-ended data, some algorithms use interval estimates to select the best splitting attribute. The so-called Hoeffding trees (e.g., VFDT) addressed this problem by using Hoeffding's bound, taking the information gain as a sum of independent random variables. However, Rutkowski et al. [7] recently showed that Hoeffding's bound is not an adequate tool to solve the underlying problem, since the information gain is not a sum of independent random variables. They proposed a family of algorithms based on McDiarmid's inequality. However, the probabilistic bounds obtained are too large and thus, a large amount of data is required to make a decision for a given splitting attribute.

Most of the studies on learning decision trees from time-changing data streams have been based on VFDT. One exception is OnlineTree2 [11], which stores some previous data for the tree induction (the algorithm uses a partial instance memory) and considers various heuristics in the detection and forgetting mechanisms.

The IADEM family of algorithms [5,6,12,13] can also induce decision trees from data streams. To do so, these algorithms store in the leaf nodes only relative frequencies ($\overline{X}$) and estimate confidence intervals ($[E(\overline{X}) - \varepsilon, E(\overline{X}) + \varepsilon]$) by using Chernoff's and Hoeffding's bounds. Different from VFDT, the IADEM family assumes that a relative frequency is a sum of independent random variables, not the heuristic measure. The induction of the tree by these algorithms takes these estimates into account when executing different actions (expansion of a node, selection of most appropriate attribute to expand, etc.). Various algorithms based on IADEM have been proposed to improve the quality of the induced model and to extend its areas of application [6,13]. However, none of these algorithms is able to handle concept drift.

This paper aims to overcome these drawbacks by presenting IADEM-3, a learner based on IADEM-2 [6], for data stream classification in the presence of concept drift. The new algorithm satisfies the common requirements for online learning: it has a constant time and space computational complexity per example processed, and learns with a simple scan over the training data.IADEM-3 follows a common strategy to handle concept drift by constantly monitoring the consistency of the induced model. When it detects an inconsistency regarding the most recent data, a concept drift is suspected and some actions are performed to reconstruct parts of the model affected by the change (e.g., inducing alternative models). In the literature, this strategy is viewed as a temporal window [14] moving over the most recent data and according to the current concept. Thus, the learning algorithm is trained with instances contained in this window. The main idea behind this technique is to reduce the window size (e.g., deleting outdated models) when a new concept arrives, and to allow the window size to grow as much as possible in stable concepts. Previous learning algorithms have fixed the window size [8] or adjusted this size dynamically to effectively deal with both stable and drifting concepts [8,11,15].

IADEM-3 uses well-founded and online statistical approaches for change detection, error estimates and predictive performance comparison between learning models. The algorithm handles the most common types of concept drift explicitly, allowing the learner to recover quickly when a concept drift occurs. IADEM-3 also uses alternative models to provide more accurate predictions, similar to option trees [16,17]. Option trees can be viewed as classifier ensembles, but the first ones provide a more efficient representation of various learning models. In option trees, an unlabeled example can take multiple paths and thus make multiple predictions. Therefore, an additional method to provide a single prediction is needed.

Additionally, IADEM-3 incorporates a mechanism to speed up the convergence of the model with respect to IADEM-2, allowing splits to occur faster. The mechanism aims to obtain more accurate decision tree models with fewer training instances. This method relaxes the statistical test used by IADEM-2 to make a decision on the best split and thus, a new pruning method is also proposed. Various pruning methods have been explored in batch learning [3], such as two-stage search methods, thresholds on impurity and trees to rules conversion. However, these methods cannot be applied to online learning because of their computational cost. Techniques for obtaining the right-sized trees have received little attention in this context. The pruning method proposed in this paper ensures that all split tests previously installed in decision nodes significantly contribute to predictive accuracy.

The paper is organized as follows. The next section defines concept drift and shows common types of change. Section 3 reviews adaptive learning algorithms based on decision trees and discusses the main problems to be considered in their design, how existing algorithms attempt to solve these problems and their fundamental drawbacks. The proposed algorithm, IADEM-3, is presented in Section 4. Next, in Section 5, it is compared with its previous version, IADEM-2, and various state-of-the-art algorithms based on decision trees for mining data streams: VFDT [4], HAT [15], and OnlineTree2 [11]. The comparison is based on common performance measures (accuracy, model size and processing time) and uses various synthetic and real-world datasets. The experimental results show that IADEM-3 is competitive with the state-of-the-art algorithms: it is more robust to false detections and can recover more quickly from concept drifts. In consequence, IADEM-3 often reaches higher levels of accuracy, maintaining a suitable model size and processing time. Finally, Section 6 presents the main conclusions from this study.

2. Definitions and types of change

A data stream is commonly defined as a very large (or possibly infinite) sequence $S = (\vec{a}_1, c_1); (\vec{a}_2, c_2); \ldots$ of *examples* or *instances*$(\vec{a}_i, c_i)$ that arrive over time, where $\vec{a}_i \in \overrightarrow{\mathcal{A}}$ is a vector in which each component is called *attribute* and $c_i \in \mathcal{C}$ is its corresponding *class label*, named *class*, taken from a finite set $\mathcal{C}$ of the possible classes. Assuming the existence of a target function $f(\vec{a}_i) = c_i$, the incremental learning task is to obtain a model $\hat{f}$ that approximates *f*, so that $\hat{f}$ maximizes the prediction accuracy. Often it is also assumed that the examples are regulated by a probability density function $P(\overrightarrow{\mathcal{A}}, \mathcal{C})$. *Concept* refers to the whole distribution $P(\overrightarrow{\mathcal{A}}, \mathcal{C})$ of the problem at a certain point in time. Therefore, a change in the whole distribution of the problem $P(\overrightarrow{\mathcal{A}}, \mathcal{C})$ (also known as context; [18]) is called *concept change* or concept drift. Two major changes are usually considered: *virtual drift* [19,20] when only the prior probability $P(\overrightarrow{\mathcal{A}})$ changes; and *real change* [21] when the conditional probability $P(\mathcal{C}|\overrightarrow{\mathcal{A}})$ changes, either with or without change in $P(\overrightarrow{\mathcal{A}})$ [2].

Another important categorization refers to the *extent of change* (also called *drift rate*). For simplicity, let us consider that a concept drift involves two different concepts: initial concept (P_I) and final concept (P_F). Depending on how different P_I and P_F are, we can say that the extent of change is *total* or *partial*. A total change implies that the distributions of P_I and P_F do not share anything. This kind of change is usually simulated with artificial data, as in real data the change is usually partial.

The speed of change, namely the time used to change from P_I to P_F (t_{change}), is commonly divided into *abrupt change* ($t_{change} = 0$) and *gradual change* ($t_{change} > 0$). When the change is gradual we can consider two additional types of change. If the change is made progressively, by means of intermediate concepts that are slightly different from each other (and different from P_I and P_F), we have a *continuous* gradual change. If both concepts P_I and P_F exist and their presence alternates, varying the frequency during the transition (with higher presence of P_I at the beginning and higher presence of P_F at the end), we have a *discrete* gradual change. Another