Information Sciences 357 (2016) 125–143

Contents lists available at ScienceDirect

Information Sciences

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

Dynamic Clustering Forest: An ensemble framework to efficiently classify textual data stream with concept drift☆

Ge Song[a,b], Yunming Ye[b,*], Haijun Zhang[b], Xiaofei Xu[b], Raymond Y.K. Lau[c], Feng Liu[b]

[a] College of Mathematics And Informatics, South China Agricultural University, Guangzhou, China
[b] Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[c] Department of Information Systems, City University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 11 October 2014
Revised 10 January 2016
Accepted 28 March 2016
Available online 12 April 2016

Keywords:
Clustering tree
Ensemble learning
Concept drift
Textual stream

ABSTRACT

Textual stream mining with the presence of concept drift is a very challenging research problem. Under a realistic textual stream environment, it often involves a large number of instances characterized by a high-dimensional feature space. Accordingly, it is computationally complex to detect concept drift. In this paper, we present a novel ensemble model named, Dynamic Clustering Forest (DCF), for textual stream classification with the presence of concept drift. The proposed DCF ensemble model is constructed based on a number of Clustering Trees (CTs). In particular, the DCF model is underpinned by two novel strategies: (1) an adaptive ensemble strategy to dynamically choose the discriminative CTs according to the inherent property of a data stream, (2) a dual voting strategy that takes into account both credibility and accuracy of a classifier. Based on the standard measure of Mean Square Error (MSE), our theoretical analysis demonstrates the merits of the proposed DCF model. Moreover, based on five synthetic textual streams and three real-world textual streams, the results of our empirical tests confirm that the proposed DCF model outperforms other state-of-the-art classification methods in most of the high-dimensional textual streams.

1. Introduction

The recent trend in delivering emails, publishing blogs, establishing chatting rooms and forums on the Internet have led to the generation of a huge number of dynamic textual streams. The underlying characteristics of these textual streams pose some serious challenges of effectively classifying the dynamic textual streams. First, the concepts embedded in a data stream will change over time. This characteristic is referred to as concept drift, which requires the adaptation of a classifier with respect to the up-to-dated concepts. For instance, the topical interest of a reader may change over time after s/he has read a large number of online news with diversified topics on the Internet. This phenomenon motivates us to develop adaptive learning models to capture readers' evolving topical interests. Second, a textual stream usually consists of a large number of objects (instances), and these objects are characterized by a high-dimensional feature space (e.g., the news topics referred

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author. Tel.: +86 75526033008.
E-mail address: yeyunming@hit.edu.cn, yym@hitsz.edu.cn (Y. Ye).

http://dx.doi.org/10.1016/j.ins.2016.03.043
0020-0255/

to in a textual stream are described by a large vocabulary). Unfortunately, most of the existing data stream classification methods fail to tackle textual streams due to their high-dimensional feature spaces [26].

Many models have been proposed to deal with textual streams. Among them, the ensemble classification method that aims at combining the predictions of individual classifiers to form a final classification decision is a promising approach [5]. Under the ensemble framework, each chunk of a textual stream is treated as a sub-model to train a classifier, and then these well-trained sub-models are combined to predict the labels of incoming instances. However, existing ensemble methods cannot provide a satisfactory solution to tackle the task of textual stream classification because of the following open questions [31]:

1. How to adaptively select a suitable number of sub-models such that outdated concepts can be removed?
2. How to combine sub-models to generate an optimal global prediction?

In this paper, we propose a new ensemble approach named, Dynamic Clustering Forest (DCF), for the classification of textual streams. DCF aims at effectively combining a number of Clustering Trees (CTs) [28] by using a principled way. It is worth noting that CT is a clustering-based classification algorithm [18]. CT is suitable for classifying large, high dimensional, and sparse textual data with many classes [18]. For the proposed DCF model, we employ CT as a sub-classifier for classifying textual streams. First, we train a set of CTs by using some sequential chunks of a textual stream. Then, we dynamically select a suitable number of CTs and combine these CTs to accomplish the optimal classification performance. During the combination process, we exploit two ensemble strategies: an adaptive ensemble strategy and a dual voting strategy. For the adaptive ensemble strategy, a threshold, which is defined according to the accuracy weight of a CT, is applied to determine whether the CT is too "old" to classify a new concept. The threshold is estimated by using the average (or minimum) prediction accuracy of each CT in the model, rather than using the random prediction accuracy employed by most of the existing models [12,29]. For the dual voting strategy, a credibility weight is introduced to each testing sample. This credibility weight enables us to determine whether a CT is "credible" enough to classify the testing sample. The estimation of the credibility weight relies on the similarity between the testing sample and the centroid of the cluster that this testing sample belongs to.

To verify if the proposed strategies are sound or not, we conducted a theoretical analysis of the DCF model in terms of the standard measure of Mean Square Error (MSE). Moreover, we performed empirical tests against the DCF model by comparing the performance of the DCF model with that of seven state-of-the-art ensemble models available on the Massive Online Analysis (MOA) platform [6] in classifying several synthetic and real-world textual streams. Our experimental results confirm that the DCF model demonstrates promising performance in terms of average accuracy and plotting accuracy for high-dimensional textual stream classification tasks.

The rest of the paper is organized as follows. In Section 2, we discuss the existing research work which is related to our study. We then introduce the preliminaries and the background information about textual stream classification with concept drift in Section 3. In Section 4, we present an overview of the proposed DCF framework, followed by the illustration of the proposed adaptive ensemble strategy and the dual voting strategy of the DCF model in Section 5. In Section 6, we report a theoretical analysis of the inherent properties of the proposed DCF model, followed by a description of the empirical tests against the DCF model and some state-of-the-art ensemble models in Section 7. We then discuss the characteristics of the proposed algorithm by referring to our experimental results in Section 8. Finally, we offer concluding remarks and suggest future directions of our research work.

2. Related work

Our approach is related to two main research areas, namely data stream classification with concept drift and text mining. We briefly describe related research work in the following sub-sections.

2.1. Data streams with concept drift

Large amount of data and drifting concepts are two main features of a data stream. These inherent challenges of data streams lead to the development of two common methods to classify data streams: an incremental mining (IM) method, and an ensemble learning (EL) method.

The IM method revises and refines a single model continuously when new data arrive [29]. However, most existing IM algorithms are not efficient because the corresponding models must be updated frequently. In addition, these algorithms are not effective to handle drifting concepts, especially for recurring drifting concepts.

EL approach [10,31] is another promising approach for data stream mining. We summarize three popular ensemble strategies according to existing ensemble methods for data stream classification with concept drift reported in literature[15].

(1) *Ensemble approach based on resampling and adaptive sliding window (Adwin)*: Resampling is a technique that reuses (or selects) data, adaptively reweights and combines sub-models to improve classification performance. Some popular resampling methods include bagging and boosting [3]. But traditional resampling methods cannot deal with dynamic data stream classification. To solve the drifting concept problem, Bifet et al. [3] have proposed a method named, Adwin, which constructs a sliding window with varying size to choose a suitable amount of training data for learning