Neurocomputing 000 (2017) 1–15

Contents lists available at ScienceDirect

Neurocomputing

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

On expressiveness and uncertainty awareness in rule-based classification for data streams

Thien Le[a], Frederic Stahl[a,*], Mohamed Medhat Gaber[b], João Bártolo Gomes[c], Giuseppe Di Fatta[a]

[a] Department of Computer Science, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AY, United Kingdom
[b] School of Computing and Digital Technology, Birmingham City University, Curzon St, Birmingham B4 7XG, England, United Kingdom
[c] DataRobot, 1 International Place, 5th Floor, Boston, MA 02110, United States

ARTICLE INFO

Article history:
Received 25 February 2016
Revised 24 May 2017
Accepted 26 May 2017
Available online xxx

Keywords:
Data Stream Mining
Big Data Analytics
Classification
Expressiveness
Abstaining
Modular classification rule induction

ABSTRACT

Mining data streams is a core element of Big Data Analytics. It represents the *velocity* of large datasets, which is one of the four aspects of Big Data, the other three being *volume, variety* and *veracity*. As data streams in, models are constructed using data mining techniques tailored towards continuous and fast model update. The *Hoeffding Inequality* has been among the most successful approaches in learning theory for data streams. In this context, it is typically used to provide a statistical bound for the number of examples needed in each step of an incremental learning process. It has been applied to both classification and clustering problems. Despite the success of the Hoeffding Tree classifier and other data stream mining methods, such models fall short of explaining how their results (i.e., classifications) are reached (*black boxing*). The expressiveness of decision models in data streams is an area of research that has attracted less attention, despite its paramount of practical importance. In this paper, we address this issue, adopting *Hoeffding Inequality* as an upper bound to build decision rules which can help decision makers with informed predictions (*white boxing*). We termed our novel method *Hoeffding Rules* with respect to the use of the *Hoeffding Inequality* in the method, for estimating whether an induced rule from a smaller sample would be of the same quality as a rule induced from a larger sample. The new method brings in a number of novel contributions including handling uncertainty through abstaining, dealing with continuous data through Gaussian statistical modelling, and an experimentally proven fast algorithm. We conducted a thorough experimental study using benchmark datasets, showing the efficiency and expressiveness of the proposed technique when compared with the state-of-the-art.

1. Introduction

One problem the research area of 'Big Data Analytics' is concerned with is the analysis of high velocity data, also known as streaming data [1,2], that challenge our computational resources. The analysis of these fast streaming data in real-time is also known as the emerging area of Data Stream Mining (DSM) [2,3]. One important data mining technique, and in turn DSM category of techniques is classification. Traditional data mining builds its classification models on static batch training sets allowing several iterations over the data. This is different in DSM as the classification model needs to be induced in a linear or sublinear time complexity [4]. Furthermore, DSM classification techniques need to allow dynamic adaptation to concept drifts as the data streams in [4]. Applications of DSM classification techniques are manifold and comprise for example monitoring the stock market from handheld devices [5], real-time monitoring of a fleet of vehicles [6], real-time sensing of data in the chemical process industry using soft-sensors [7], sentiment analysis using real-time micro-bogging data such as twitter data [8], to mention a few.

The challenge of data stream classification lies in the need of the classifier to adapt in real-time to concept drifts, which is significantly more challenging if the data stream is of high velocity. Many data stream classification techniques are based on the 'Top Down Induction of Decision Trees', also known as the 'divide-and-conquer' approach [9], such as [10,11]. However, the decision tree format is also a major weakness and often requires irrelevant information to be available to perform a classification task [12]. Moreover, adaptation of the trees is harder compared with rules when

* Corresponding author.
E-mail addresses: t.d.le@pgr.reading.ac.uk (T. Le), f.t.stahl@reading.ac.uk, Frederic.T.Stahl@gmail.com (F. Stahl), mohamed.gaber@bcu.ac.uk (M.M. Gaber), joao@datarobot.com (J.B. Gomes), g.difatta@reading.ac.uk (G.D. Fatta).

http://dx.doi.org/10.1016/j.neucom.2017.05.081

a change occurs, this could be a disadvantage for real-time applications.

The here presented work proposes the *Hoeffding Rules* data stream classifier that is based on modular classification rules instead of trees. Hoeffding Rules can easily be assimilated by humans, and at the same time does not require unnecessary information to be available for the classification tasks. Rule induction from data streams can be traced back to the *Very Fast Decision Rules (VFDR)* [13] and *eRules* data stream classifiers [9] for numerical data. *eRules* induces modular classification rules from data streams, but requires extensive parameter tuning by the user in order to achieve adequate classification accuracy including the setting of the window size. Noting this drawback that affects the accuracy, if the parameters are not set correctly, a statistical measure that automatically tunes the parameters is desirable. Addressing this issue, Hoeffding Rules adjusts these parameters dynamically with very little input required by the user. The here presented Hoeffding Rules algorithm is based on the *Prism* [12] rule induction approach using a sliding window [14,15]. However, this window for buffering training data is adjusted dynamically by making use of the Hoeffding Inequality [16]. One important property of Hoeffding Rules compared with the popular Hoeffding Tree data stream classification approach [10] is, that Hoeffding Rules can be configured to abstain from classifying an unseen data instance when it is uncertain about its class label. In addition, our approach is computationally efficient and hence is suitable for real-time requirements. An important strength of the proposed technique is the high expressiveness of the rules. Thus, having the rules as the representation of the output can help users in making timely and informed decisions. Output expressiveness increases trust in data stream analytics which is one of the challenges facing adaptive learning systems [17]. To address the expressiveness issue for offline black box machine learning models, the new algorithm *Local Interpretable Model-Agnostic Explanations (LIME)* has been proposed in [18]. The method generates a short explanation for each new classified or regressed instance out of a predictive model, in a form that is interpretable by humans (can be expressed as rules, in a way). The work has attracted a great deal of media attention, and has emphasised the need for expressive models. Model trust has been further emphasised. This work and many other follow-up research papers have been the result of experimental work that showed some serious flaws in deep learning models (a highly accurate black box approach) [19]. The work showed that miss-classification by deep learning models of some images – due to added noise to these images – can occur to surprisingly very obvious examples to humans. Again, model interpretability and trust have been emphasised as an important area of research.

The *utility of expressiveness* is introduced in this paper to refer to the cost of expressiveness when comparing the accuracy of two methods. As accuracy has been the dominating measure of interest in comparing classifiers in both static and streaming environments, it is evident that real-time decision making based on streaming models still suffers from the issue of trust [17]. To address this issue, the user is able to determine an accuracy loss band (ζ), such that the model can be expressive enough to grant trust, and at the same time the accuracy can be tolerated at ($-\zeta\%$) of any other best performing classifier which is less expressive (can be a total black box). We argue that such a new measure will open the door for more trustful white box models. In many applications (e.g., surveillance, medical diagnosis, terrorism detection), decisions need to be based on clear arguments. In such applications, having a trustful model with a competent accuracy can be much more appreciated than having a highly accurate model that does not convey any reasoning about its decision. Other examples of applications that require convincing arguments can be found in [18].

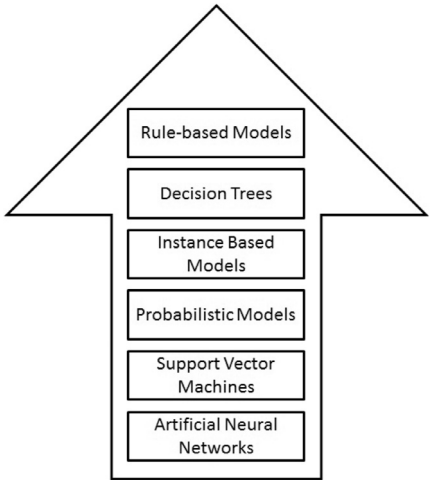

Fig. 1. Hierarchy of output expressiveness.

This paper is organised as follows: Section 2 highlights works related to the Hoeffding Rules approach. Section 3 highlights our dynamic rule induction and adaptation approach for data streams. An experimental evaluation and discussion is presented in Section 4. Concluding remarks are provided in Section 5.

2. Related work

The velocity aspect of the Big Data trend is the main driver of work done for over a decade in the area of data stream mining – long before the Big Data term was coined. Among the proposed techniques in the area come a long list of classification techniques. Approaches to data stream classification varied from approximation to ensemble learning. Two motives stimulated such developments: (1) fast model construction addressing the high speed nature of data streams; and (2) change in the underlying distribution of the data, in what has been commonly known as concept drift.

Hoeffding Inequality [16] has found its way from the statistical literature in the 60s of the last century to make an impact in data stream mining, having a number of techniques, mostly in classification, with notable success. The *Hoeffding bound* is a statistical upper bound on the probability that the sum of a random variable deviates from its expected value. The basic *Hoeffding bound* has been extended and adopted in successfully developing a number of streaming techniques that were termed collectively as *Very Fast Machine Learning* (VFML) [20].

Earlier work on data stream mining addressed the aforementioned issues. However, the end user perspective has been greatly missing, and hence the user's trust in such systems was frequently questioned. This issue has been discussed in a position paper by Zliobaite et al [17].

In this paper we address this issue, attempting to provide the end user with the most expressive knowledge representation for data stream classification, i.e., rules. We argue that rules can provide the users with informative decisions that enhance the trust in streaming systems. Fig. 1 shows a hierarchy of output expressiveness, with rule-based models being at the top of all of the other classification techniques.

2.1. Rule induction from data streams

FLORA is a family of algorithms for data stream rule induction that adjusts its window size dynamically using a heuristic based on the predictive accuracy and concept descriptions. The most recent FLORA algorithm, FLORA4, addresses the issue of concept drift. It can use previous concept descriptions in situations

Please cite this article as: T. Le et al., On expressiveness and uncertainty awareness in rule-based classification for data streams, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.05.081