Expert Systems With Applications 55 (2016) 532–545

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A unified pipeline for online feature selection and classification

Veronica Bolon-Canedo*, Diego Fernández-Francos, Diego Peteiro-Barral, Amparo Alonso-Betanzos, Bertha Guijarro-Berdiñas, Noelia Sánchez-Maroño

Departamento de Computación, Universidade da Coruña, Campus de Elviña S/N, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 16 December 2014
Revised 27 January 2016
Accepted 15 February 2016
Available online 27 February 2016

Keywords:
Machine learning
Online learning
Classification
Feature selection
Discretization

ABSTRACT

With the advent of Big Data, data is being collected at an unprecedented fast pace, and it needs to be processed in a short time. To deal with data streams that flow continuously, classical batch learning algorithms cannot be applied and it is necessary to employ online approaches. Online learning consists of continuously revising and refining a model by incorporating new data as they arrive, and it allows important problems such as concept drift or management of extremely high-dimensional datasets to be solved. In this paper, we present a unified pipeline for online learning which covers online discretization, feature selection and classification. Three classical methods—the k-means discretizer, the χ^2 filter and a one-layer artificial neural network—have been reimplemented to be able to tackle online data, showing promising results on both synthetic and real datasets.

1. Introduction

During the last years and with increasing frequency, real-time production systems generated tremendous amount of data at unprecedented rates, such as network event logs, telephone call records or sensoring and surveillance video streams. To deal with data streams that flow continuously, classical batch learning algorithms cannot be applied and it is necessary to employ online approaches. Online data mining consists of continuously revise and refine a model by incorporating new data as they arrive (Wang, Fan, Yu, & Han, 2003). Note that any online method is inherently incremental. This type of learning has been applied in fields such as classification of textual data streams, financial data analysis, credit card fraud protection, traffic monitoring and predicting customer behavior (Elwell & Polikar, 2009; Katakis, Tsoumakas, & Vlahavas, 2006; Wang et al., 2003).

Most of these applications present a great challenge for machine learning researches due to the high amount of data available. Theoretically, it could seem logical that having more features could lead to better results, but this is not always the case due to the so-called *curse of dimensionality* (Bellman, 1966). This phenomenon happens when the dimensionality increases and the time required by the machine learning algorithm to train the data increases exponentially. To overcome these problems, feature selection is a well-known dimensionality reduction technique. *Feature selection* consists of selecting the relevant features and discarding the irrelevant ones to obtain a subset of features that describes properly the problem with a minimum degradation of performance (Guyon, Gunn, Nikravesh, & Zadeh, 2006).

A special case of feature selection is known as *online feature selection* (Glocer, Eads, & Theiler, 2005; Nguyen, Wu, & Mukherjee, 2015; Perkins & Theiler, 2003; Wang, Zhao, Hoi, & Jin, 2014; Wu, Yu, Wang, & Ding, 2010), which can be very useful, being one of the most interesting when a *concept drift* appears. This phenomenon is present in situations where the underlying data distribution changes. These changes make the model built on old data inconsistent with the new data, and regular updating of the model is necessary (Tsymbal, 2004). Applied to feature selection, a concept drift may cause that the subset of relevant features changes over the time. In other words, as time goes by, different sets of features become important for classification and some totally new features with high predictive power may appear. Online feature selection has been faced mostly individually, i.e., by selecting features previously in a single step independent of the online machine learning step, or performing online feature selection without performing online classification afterwards. Notice that after an online feature selection process, where the set of relevant features changes across the time, the classification algorithm has to be capable of updating its model according not only to new samples but also to new features, limiting the alternatives available capable of coping with both requirements.

* Corresponding author. Tel.: +34 981167000.
E-mail addresses: vbolon@udc.es, veronica.bolon@udc.es (V. Bolon-Canedo), dfernandezf@udc.es (D. Fernández-Francos), dpeteiro@udc.es (D. Peteiro-Barral), ciamparo@udc.es (A. Alonso-Betanzos), cibertha@udc.es (B. Guijarro-Berdiñas), nsanchez@udc.es (N. Sánchez-Maroño).

http://dx.doi.org/10.1016/j.eswa.2016.02.035
0957-4174/

Therefore, in this work we propose a method that covers both online feature selection and online learning. Our proposal includes an algorithm that performs online feature selection and classification at the same time, by modifying a classical feature selection algorithm and introducing a novel implementation for a classification training algorithm. Among the different feature selection methods available, we chose a representative of so-called *filter methods* (Guyon et al., 2006) since they are known for being fast, simple, classifier-independent and having a low computational cost (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013). Specifically, we reimplemented the χ^2 metric (Liu & Setiono, 1995), chosen because of its simplicity and effectiveness, as well as having some characteristics that make it inherently incremental. However, this filter requires data to be discrete, and thus, well-known *k*-means discretizer (MacQueen et al., 1967; Tou & González, 1977; Ventura & Martinez, 1995) was also adapted to make it incremental.

The last step of our proposed online pipeline requires an incremental classifier, however, those available in the literature are incremental in the instance space, but not in the feature space. Up to the authors' knowledge, a complete pipeline as the one introduced here has not been presented elsewhere. In fact, the popular machine learning tool Weka (Hall et al., 2009) provides methods able to receive new instances, but they do not support different sets of features, perhaps with different sizes, in each iteration. Thus, an online training algorithm for one-layer artificial neural networks ANNs is also introduced in this paper, which continuously adapts the input layer to those features, that remind might vary in number, selected at each time. In order to achieve this, we are presenting a new implementation of our previously proposed algorithm (Fontenla-Romero, Guijarro-Berdiñas, Pérez-Sánchez, & Alonso-Betanzos, 2010), which reaches a minimum error in a few epochs of training and exhibits a higher speed when compared to other classical methods. Moreover, the structure of this algorithm makes it suitable for a dynamic input space, as happens when selecting features on-line. In this research, we propose a novel implementation, which continuously adapts the size of the input layer to those features selected at each time. In summary, the contribution of this paper consists of introducing an approach that uses three components together conforming a pipeline: (a) an online discretizer, (b) an online filter, and (c) an online learning algorithm; that will be applied to either online or large data.

The rest of the paper is organized as follows: Section 2 presents the state of the art in the field of online machine learning, Section 3 describes the method proposed in this research, Section 4 describes the experimental settings, Section 5 shows the experimental results, Section 6 is focused on a case study about the influence of the order of occurrence of the samples (data order) on the performance of the pipeline, finally, Section 7 presents the discussion and conclusions.

2. Background

Online learning has become a trending area in the last few years since it allows to solve important problems such as concept drift or managing extremely high-dimensional datasets. For this reason, advances in this field have recently appeared. However, online feature selection has not evolve in line with online learning. Zhang, Ruan, and Tan (2011) proposed an incremental computation feature subset selection algorithm which, originated from Boolean matrix technique, selects useful features for the given data objective efficiently. Nevertheless, the efficiency of the feature selection method has not been tested with an incremental machine learning algorithm. Keerthika and Priya (2015) examined various feature reduction techniques for intrusion detection, where training data arrive in a sequential manner from a real time application. Katakis et al. (2006) mentioned the idea of a dynamic feature space. The features that are selected based on an initial collection of training documents are the ones that are subsequently considered by the learner during the operation of the system. However, these features may vary over time and in some applications an initial training set is not available. In the approach presented inhere, we are interested in flexible feature selection methods able to modify the selected subset of features as new training samples arrive, in both subset size and specific features selected. It is also desirable that these methods can be executed in a dynamic feature space that would be empty at the beginning and add features when new information arrives (e.g., documents in their text categorization application). Katakis et al. (2006) applied incremental feature selection combined with what they called a feature based learning algorithm to deal with online learning in high-dimensional data streams. This framework is applied to a special case of concept drift inherent to textual data streams, which is the appearance of new predictive words over time. The problem with this approach is that they assume that features have discrete values. Perkins, Lacker, and Theiler (2003) presented a novel and flexible approach, called grafting, which treats the selection of suitable features as an integral part of learning a predictor in a regularized learning framework. To make it suitable for large problems, grafting operates in an incremental iterative fashion, gradually building up a feature set while training a predictor model using gradient descent. Perkins and Theiler (2003) tackle the problem in which, instead of all features being available from the start, features arrive one at a time. Online Feature Selection (OFS) assumes that, for any reason, is not affordable to wait until all features have arrived before learning begins, therefore one needs to derive a mapping function *f* from the inputs to the outputs that is as "good as possible" using a subset of just the features seen so far. By Wu et al. (2010), a promising alternative method, Online Streaming Feature Selection (OSFS), to online select strongly relevant and non-redundant features is presented. Glocer et al. (2005) demonstrated the power of OFS in the image processing domain by applying it to the problem of edge detection. Mao, Yuan, Wu, Qu, and Li (2014) proposed a real-time compressive tracking algorithm based on online feature selection to address the problems of drifting and tracking lost caused by changes in the appearance of the tracked object. The discriminating features selected are then integrated to construct a classifier to carry out the tracking process. And Nguyen et al. (2015) presented an online unsupervised feature selection method for background suppression in video sequences, that allows them to prune the feature set avoiding any combinatorial search.

Finally, some other researches have been found in the literature comprising online feature selection and classification. Kalkan and Çetisli (2011) presented an online learning algorithm for feature extraction and classification, implemented for impact acoustics signals to sort hazelnut kernels. Levi and Ullman (2010) proposed to classify images by ongoing feature selection. However, their approach only uses at each stage a small subset of the training data. Carvalho and Cohen (2006) performed online feature selection based on the weights assigned to each input of the classifiers. Note, however, that this method is highly dependent on the classifier. Another method that is strongly dependent on the classifier was presented by Wang et al. (2014). They addressed two different tasks of OFS: (1) learning with full input, where the learner is allowed to access all the features to decide the active ones, and (2) learning with partial input, where only a limited number of features is allowed to be accessed for each instance by the learner. In a recent work, Roy (2015) proposed an interesting algorithm for streaming big data and for highly parallel implementation on Apache Spark based on Kohonen networks. It examines some streaming data to select the features with a high