Expert Systems with Applications 41 (2014) 5431–5450

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Efficient classification of multi-labeled text streams by clashing

Ricardo Ñanculef [a,b,*], Ilias Flaounas [b], Nello Cristianini [b]

[a] Department of Informatics, Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile
[b] Intelligent Systems Laboratory, University of Bristol, MVB, Woodland Rd, Bristol, BS8 1UB, UK

ARTICLE INFO

Keywords:
Text classification
Data streams
Multi-label classification
Feature hashing
Massive data mining

ABSTRACT

We present a method for the classification of multi-labeled text documents explicitly designed for data stream applications that require to process a virtually infinite sequence of data using constant memory and constant processing time.

Our method is composed of an online procedure used to efficiently map text into a low-dimensional feature space and a partition of this space into a set of regions for which the system extracts and keeps statistics used to predict multi-label text annotations. Documents are fed into the system as a sequence of words, mapped to a region of the partition, and annotated using the statistics computed from the labeled instances colliding in the same region. This approach is referred to as *clashing*.

We illustrate the method in real-world text data, comparing the results with those obtained using other text classifiers. In addition, we provide an analysis about the effect of the representation space dimensionality on the predictive performance of the system. Our results show that the online embedding indeed approximates the geometry of the full corpus-wise TF and TF-IDF space. The model obtains competitive F measures with respect to the most accurate methods, using significantly fewer computational resources. In addition, the method achieves a higher macro-averaged F measure than methods with similar running time. Furthermore, the system is able to learn faster than the other methods from partially labeled streams.

1. Introduction

The efficient analysis of massive datasets is one of the main challenges in modern machine learning and data mining applications (Hand, 2013; Rajaraman & Ullman, 2012; Wu, Zhu, Wu, & Ding, 2014). Usually in these scenarios, data is being generated continuously, arriving to the system in the form of a fast and virtually infinite data stream (Aggarwal, 2007; Bifet, 2013). Examples include the stream of messages exchanged on a social network or the stream of daily stories generated by different news outlets. The challenge for a mining system designed to work under this setting is being ready to predict at any time, learning constantly from new observations, but using limited computational resources i.e. bounded memory and constant processing time.

In this paper, we present a simple and efficient method to classify large data streams of documents: one of the most common type of user generated data. Document classification is one of the most frequent and important problems in textual data analysis, with applications from information retrieval and spam filtering to content personalization and natural language processing. The task is that of learning a mechanism from data to automatically annotate documents with thematic categories or labels from a given set (Aggarwal & Zhai, 2012; Sebastiani, 2002). Since documents can be associated with multiple non-exclusive categories (e.g. *politics*, *economics* and *international affairs*) simultaneously, this task is one of the most common examples of multi-label classification (Tsoumakas, Katakis, & Vlahavas, 2010).

The novelty and contribution of this work is in addressing explicitly and simultaneously text representation and multi-label classification in streaming environments with limited computational resources. As we explain below, this is a challenging setting because of the large number of possible features arising in textual domains and the traditional batch setting for dimensionality reduction and multi-label classifier design (Read, Bifet, Holmes, & Pfahringer, 2012).

1.1. Context of this research

In this section we motivate the settings of this research, discussing current work in the field, and stressing the novel aspects of our approach.

* Corresponding author at: Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile. Tel.: +56 32 2654256.
E-mail addresses: jnancu@inf.utfsm.cl (R. Ñanculef), ilias.flaounas@bristol.ac.uk (I. Flaounas), nello.cristianini@bristol.ac.uk (N. Cristianini).

http://dx.doi.org/10.1016/j.eswa.2014.02.017
0957-4174/

1.1.1. Text representation

Different methods to classify documents have been investigated in the last years (Aggarwal & Zhai, 2012; Sebastiani, 2002). A fundamental component of these systems is the way to represent text into an amenable form for machine learning algorithms. This representation is commonly obtained by selecting a set of indexing terms suitable to capture document content (the vocabulary) and a weighting scheme assigning values to the dimensions of the feature vector spanned them (Joachims, 2002; Zhang, Yoshida, & Tang, 2011). The bag of words (BOW) indexing model is the most widely used text representation method in current research (Lan, Tan, Su, & Lu, 2009; Ren & Sohrab, 2013). In this model, each possible word in the set of known texts corresponds to a dimension of the feature space used to embed documents. Along with BOW, the TF-IDF weighting scheme is usually applied to obtain the final representation of a document (Lan et al., 2009; Aggarwal & Zhai, 2012). TF-IDF is proportional to the number of times a particular word appeared in a document and inversely proportional to the number of documents containing the word. Despite its widespread acceptance among practitioners, this approach for text representation has some drawbacks that recently have started to be addressed by researchers in the field. First, documents are treated as collections of unordered words. A number of authors have thus investigated the use of longer indexing units, linguistically or statistically meaningful for content identification (Zhang et al., 2011). They include k-grams (Caropreso, Matwin, & Sebastiani, 2001), frequent word sequences (Li, Chung, & Holt, 2008) and frequent word sets (Zhang, Yoshida, Tang, & Wang, 2010; Zhang et al., 2011). Unfortunately, in text categorization problems, these methods have shown improvements somewhat disappointing (Li, Xiong, Zhang, Liu, & Li, 2011; Zhang et al., 2011). Second, TF-IDF does not exploit the co-occurrence of terms and categories in the weighting process. Therefore, methods capable to exploit information about the different distribution of terms among the documents of each class have been focus of increasing interest in the last years (Guan, Zhou, & Guo, 2009; Lan et al., 2009; Luo, Chen, & Xiong, 2011; Ren & Sohrab, 2013; Wang & Zhang, 2013) and is shaping up as an important direction of research. However, as regards the scalability of text classification systems, the most important drawback of BOW and TF-IDF is the high-dimensionality of the resulting representation space.

The dimensionality of TF-IDF matches the size of the vocabulary i.e. the number of distinct terms across the entire dataset. This "curse of dimensionality" in traditional text representation brings about huge memory requirements and huge computation since most classification models scale linearly or super linearly in the dimensionality of the feature set size. In data stream scenarios, the problem is still worse, because the word distribution (required to compute TF and IDF) is not known beforehand and both the vocabulary and the corpus is constantly growing. Recently, various feature selection techniques to reduce dimensionality have been studied and compared in text domains (Forman, 2003; Fragoudis, Meretakis, & Likothanassis, 2005; Spolaôr & Tsoumakas, 2013; Yang, Liu, Zhu, Liu, & Zhang, 2012). Most of these methods correspond to filter approaches, that is, methods selecting features from general characteristics of the training data regardless of the learning algorithm (Spolaôr & Tsoumakas, 2013). Although wrapping methods, using the classifier to determine the quality of selected features, usually outperform filter methods, they tend to be prohibitively expensive on large-scale datasets (Spolaôr & Tsoumakas, 2013; Wang, Zhao, Hoi, & Jin, 2013; Yang et al., 2012) and thus more simple and efficient methods such as Information Gain, Chi-Squared and Bi-normal Separation (Forman, 2003) are preferred in practice. Unfortunately, despite the increasing importance of data stream scenarios, most studies on feature selection are restricted to the batch setting, that is, the selection task is conducted off-line and all the features and training instances are supposed to be known a priori. Thus, weighting schemes like TF-IDF and feature selection methods used to reduce dimensionality, need to be updated at a corpus level which demands lots of computation and requires the storage of a large amount of training data. A fully online feature selection method has been recently presented in (Wang et al., 2013). However, it is focused on a binary single label classification with a perceptron classifier. Indeed, research on multi-label feature selection is still very scarce. Feature selection for text datasets often applies traditional filter methods focusing on a single label and then uses some aggregation strategy to obtain a decision. It is well known that this approach can neglect strongly predictive features for unfrequent labels in unbalanced categorization problems. Therefore, both multi label and online feature selection are still topics that need to be studied in the field.

In this paper we investigate a fully online *embedding* method for approximating TF-IDF using constant memory and time. In contrast to most dimensionality reduction approaches, we do not apply a reduction method on the original TF-IDF representation but directly on the sequence of words contained in a document. The method is built on the ideas of count min sketching (Cormode, 2012) and feature hashing (Shi, Petterson, Langford, Smola, & Strehl, 2009; Shi et al., 2009), methods introduced to estimate data stream distributions and high dimensional dot products respectively. Therefore, our method is related also to other data-oblivious embedding techniques like random projections (RP) (Achlioptas, 2003) and fast Johnson Lindenstrauss transforms (FJL) (Ailon & Chazelle, 2010), for which there has been significant interest in the community. In contrast to these techniques, the tools we use do not rely on the application of cumbersome projection matrices to data but rely on simple hashing functions which can be directly applied to words. Random projections for text representation have been applied to text representation in Lin and Gunopulos (2003) and DeBarr and Wechsler (2012). Up to our knowledge, the computationally efficient versions of RP described in Ailon and Chazelle (2010) still have not been explicitly studied in text domains. Formerly, Baena-Garcia, Carmona-Cejudo, Castillo, and Morales-Bueno (2011) has proposed using the count min sketch to allow the efficient computation of IDF for massive streams of documents, studying the similarity between the ranking of the exact TF-IDF values and that of the approximate values obtained from approximate IDF. However, this algorithm works with exact TF and authors do not assess (theoretically or empirically) the effects of this approximation on document classification tasks. Approximating TF using with fewer dimensions is important in text categorization problems, because classification algorithms usually rely on the computation of metrics which scale linearly in the number of dimensions of the representation space. In this work, we study the approximate computation of both TF and IDF studying both theoretically and empirically the quality of the approximation. Recently, a generic approach for mining massive data using sketches has been suggested in Gupta, Agnihotri, and Saha (2013). However it relies on low rank matrix multiplications and does not focus on text representation or text categorization in online environments.

1.1.2. Multi-label text classification

Most classification methods studied in machine learning are devised to deal with single label assignments i.e. a data item belongs to one and only one class of the set of possible categories. Therefore, multi-label classification methods for problems arising in areas like text categorization, image annotation and protein function classification are of increasing research interest in the last years (Jiang, Tsai, & Lee, 2012; Madjarov, Kocev, Gjorgjevikj, & Dzeroski, 2012; Montañes et al., 2014; Tsoumakas et al., 2010; Yu, Pedrycz, & Miao, 2014). Multi-label classification is usually approached using either a problem transformation approach, where the problem is decomposed into several classic classification tasks, or by directly