Knowledge-Based Systems 83 (2015) 176–189

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Term-weighting learning via genetic programming for text classification

Hugo Jair Escalante [a,*], Mauricio A. García-Limón [a], Alicia Morales-Reyes [a], Mario Graff [b], Manuel Montes-y-Gómez [a], Eduardo F. Morales [a], José Martínez-Carranza [a]

[a] Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro 1, Puebla 72840, Mexico
[b] INFOTEC – Centro de Investigación e Innovación, en Tecnologías de la Información y Comunicación, Cátedras CONACyT, Aguascalientes, Mexico

ARTICLE INFO

Article history:
Received 16 December 2014
Received in revised form 16 March 2015
Accepted 19 March 2015
Available online 28 March 2015

Keywords:
Term-weighting learning
Genetic programming
Text mining
Representation learning
Bag of words

ABSTRACT

This paper describes a novel approach to learning term-weighting schemes (TWSs) in the context of text classification. In text mining a TWS determines the way in which documents will be represented in a vector space model, before applying a classifier. Whereas acceptable performance has been obtained with standard TWSs (e.g., Boolean and term-frequency schemes), the definition of TWSs has been traditionally an art. Further, it is still a difficult task to determine what is the best TWS for a particular problem and it is not clear yet, whether better schemes, than those currently available, can be generated by combining known TWS. We propose in this article a genetic program that aims at learning effective TWSs that can improve the performance of current schemes in text classification. The genetic program learns how to combine a set of basic units to give rise to discriminative TWSs. We report an extensive experimental study comprising data sets from thematic and non-thematic text classification as well as from image classification. Our study shows the validity of the proposed method; in fact, we show that TWSs learned with the genetic program outperform traditional schemes and other TWSs proposed in recent works. Further, we show that TWSs learned from a specific domain can be effectively used for other tasks.

1. Introduction

Text classification (TC) is the task of associating documents with predefined categories that are related to their content. TC is an important and active research field because of the large number of digital documents available and the consequent need to organize them. The TC problem has been approached with pattern classification methods, where documents are represented as numerical vectors and standard classifiers (e.g., naïve Bayes and support vector machines) are applied [35]. This type of representation is known as the vector space model (VSM) [34]. Under the VSM one assumes a document is a point in a *N*-dimensional space and documents that are closer in that space are similar to each other [41]. Among the different instances of the VSM, perhaps the most used model is the bag-of-words (BOW) representation. In the BOW it is assumed that the content of a document can be determined by the (orderless) set of terms[1] it contains. Documents are represented as points in the vocabulary space, that is, a document is represented by a numerical vector of length equal to the number of different terms in the vocabulary (the set of all different terms in the document collection). The elements of the vector specify how important the corresponding terms are for describing the semantics or the content of the document. BOW is the most used document representation in both TC and information retrieval. In fact, the BOW representation has been successfully adopted for processing other media besides text, including, images [7], videos [37], speech signals [38], and time series [43] among others.

A crucial component of TC systems using the BOW representation is the so called term-weighting scheme (TWS), which is in charge of determining how relevant a term is for describing the content of a document [20,4,26,11]. Traditional TWSs are term-frequency (*TF*), where the importance of a term in a document is given by its frequency of occurrence in the document; Boolean (*B*), where the importance of a term in a document is either 1, when the term appears in the document or 0, when the term does not appear; and term-frequency inverse-document-frequency (*TF-IDF*), where the importance of a term for a document is determined by its occurrence frequency times the inverse frequency of the term across the corpus (i.e., frequent terms in the corpus, as prepositions and articles, receive a low weight). Although, TC is a widely studied topic with very important developments in the last two

* Corresponding author.
E-mail addresses: hugojair@inaoep.mx (H.J. Escalante), mauricio.garcia.cs@gmail.com (M.A. García-Limón), a.morales@inaoep.mx (A. Morales-Reyes), mgraffg@gmail.com (M. Graff), mmontesg@inaoep.mx (M. Montes-y-Gómez), emorales@inaoep.mx (E.F. Morales), carranza@inaoep.mx (J. Martínez-Carranza).

[1] A term is any basic unit by which documents are formed, for instance, terms could be words, phrases, and sequences (n-grams) of words or characters.

http://dx.doi.org/10.1016/j.knosys.2015.03.025
0950-7051/

decades [35,20], it is somewhat surprising that little attention has been paid to the development of new TWSs to better represent the content of documents for TC. In fact, it is quite common in TC systems that researchers use one or two common TWSs (e.g., *B*, *TF* or *TF-IDF*) and put more effort in other processes, like feature selection [21,44], or the learning process itself [1,2,14]. Although all of the phases in TC are equally important, we think that by putting more emphasis on defining or learning effective TWSs we can achieve substantial improvements in TC performance.

This paper introduces a novel approach to learning TWS for TC tasks. A genetic program is proposed in which a set of primitives and basic TWSs are combined through arithmetic operators in order to generate alternative schemes that can improve the performance of a classifier. Genetic programming is an evolutionary algorithm in which a population of programs is evolved [27], where programs encode solutions to complex problems, in this work programs encode TWSs. The underlying hypothesis of our proposed method is that an evolutionary algorithm can learn TWSs of comparable or even better performance than those proposed so far in the literature.

Traditional TWSs combine term-importance and term-document-importance factors to generate TWSs. For instance in *TF-IDF*, *TF* and *IDF* are term-document-importance and term-importance factors, respectively. Term-document weights are referred as local factors, because they account for the occurrence of a term in a document (locally). On the other hand, term-relevance weights are considered global factors, as they account for the importance of a term across the corpus (globally). It is noteworthy that the actual factors that define a TWS and the combination strategy itself have been determined manually. Herein we explore the suitability of learning these TWSs automatically, by providing a genetic program with a pool of TWSs' building blocks with the goal of evolving a TWS that maximizes the classification performance for a TC classifier. We report experimental results in many TC collections that comprise both: thematic and non-thematic TC problems. Throughout extensive experimentation we show that the proposed approach is very competitive, learning very effective TWSs that outperform most of the schemes proposed so far. We evaluate the performance of the proposed approach under different settings and analyze the characteristics of the learned TWSs. Additionally, we evaluate the generalization capabilities of the learned TWSs and even show that a TWS learned from text can be used to effectively represent images under the BOW formulation.

The rest of this document is organized as follows. Next section formally introduces the TC task and describes common TWSs. Section 3 reviews related work on TWSs. Section 4 introduces the proposed method. Section 5 describes the experimental settings adopted in this work and reports results of experiments that aim at evaluating different aspects of the proposed approach. Section 6 presents the conclusions derived from this paper and outlines future research directions.

2. Text classification with the Bag of words

The most studied TC problem is the so called thematic TC (or simply text categorization) [35], which means that classes are associated to different themes or topics (e.g., classifying news into "*Sports*" vs. "*Politics*" categories). In this problem, the sole occurrence of certain terms may be enough to determine the topic of a document; for example, the occurrence of words/terms *"Basketball"*, *"Goal"*, *"Ball"*, and *"Football"* in a document is strong evidence that the document is about *"Sports"*. Of course, there are more complex scenarios for thematic TC, for example, distinguishing documents about sports news into the categories: *"Soccer"* vs. *"NFL"*. Non-thematic TC, on the other hand, deals with the problem of associating documents with labels that are not related to their topics. Non-thematic TC includes the problems of authorship attribution [39], opinion mining and sentiment analysis [32], authorship verification [25], author profiling [24], among several others [33,23]. In all of these problems, the thematic content is of no interest, nevertheless, it is common to adopt standard TWSs for representing documents in non-thematic TC as well (e.g., BOW using character n-grams or part-of-speech tags [39]).

It is noteworthy that the BOW representation has even trespassed the boundaries of the text media. Nowadays, images [7], videos [37], audio [38], and other types of data [43] are represented throughout analogies to the BOW. In non-textual data, a codebook is first defined/learned and then the straight BOW formulation is adopted. In image classification, for example, visual descriptors extracted from images are clustered and the centers of the clusters are considered as visual words [7,45]. Images are then represented by numerical vectors (i.e., a VSM) that indicate the relevance of visual words for representing the images. Interestingly, in other media than text (e.g., video, images) it is standard to use only the *TF* TWS, hence motivating the study on the effectiveness of alternative TWSs in non-textual tasks. Accordingly, in this work we also perform preliminary experiments on learning TWSs for a standard computer vision problem [19].

TC is a problem that has been approached mostly as a supervised learning task, where the goal is to learn a model capable of associating documents to categories [35,20,1]. Consider a data set of labeled documents $\mathcal{D} = (\mathbf{x}_i, y_i)_{\{1,\ldots,N\}}$ with N pairs of documents ($\mathbf{x}_i$) and their classes (y_i) associated to a TC problem; where we assume $\mathbf{x}_i \in \mathbb{R}^p$ (i.e., a VSM) and $y_i \in C = \{1, \ldots K\}$, for a problem with K-classes. The goal of TC is to learn a function $f : \mathbb{R}^p \rightarrow C$ from $\mathcal{D}$ that can be used to make predictions for documents with unknown labels, the so called test set: $\mathcal{T} = \{\mathbf{x}_1^T, \ldots, \mathbf{x}_M^T\}$. Under the BOW formulation, the dimensionality of documents' representation, p, is defined as $p = |V|$, where V is the vocabulary (i.e., the set all the different terms/words that appear in a corpus). Hence, each document d_i is represented by a numerical vector $\mathbf{x}_i = \langle x_{i,1} \ldots, x_{i,|V|} \rangle$, where each element $x_{i,j}, j = 1, \ldots, |V|$, of $\mathbf{x}_i$ indicates how relevant word t_j is for describing the content of d_i, and where the value of $x_{i,j}$ is determined by the TWS.

Many TWSs have been proposed so far, including unsupervised [35,34,20] and supervised schemes [11,26], see Section 3. Unsupervised TWSs are the most used ones, they were firstly proposed for information retrieval tasks and latter adopted for TC [35,34]. Unsupervised schemes rely on term frequency statistics and measurements that do not take into account any label information. For instance, under the Boolean (B) scheme $x_{i,j} = 1$ *iff* term t_j appears in document i and 0 otherwise; while in the term-frequency (*TF*) scheme, $x_{i,j} = \#(d_i, t_j)$, where $\#(d_i, t_j)$ accounts for the times term t_j appears in document d_i. On the other hand, supervised TWSs aim at incorporating discriminative information into the representation of documents [11]. For example in the *TF-IG* scheme, $x_{i,j} = \#(d_i, t_j) \times IG(t_j)$, is the product of the TF TWS for term t_j and document d_i (a local factor) with the information gain of term t_j ($IG(t_j)$, global factor). In this way, the discrimination power of each term is taken into account for the document representation; in this case through the information gain value [44]. It is important to emphasize that most TWSs combine information from both term-importance (global) and term-document-importance (local) factors (see Section 3), for instance, in the *TF-IG* scheme, IG is a term-importance factor, whereas TF is a term-document-importance factor.

Although acceptable performance has been reported with existing TWS, it is still an art determining the adequate TWS for a particular data set; as a result, mostly unsupervised TWSs (e.g., *B, TF*