Information Processing and Management 52 (2016) 217–257

Contents lists available at ScienceDirect

Information Processing and Management

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

Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts

Rafael Geraldeli Rossi*, Alneu de Andrade Lopes, Solange Oliveira Rezende

Institute of Mathematics and Computer Science, University of São Paulo, Brazil

ARTICLE INFO

Article history:
Received 28 July 2014
Revised 22 April 2015
Accepted 6 July 2015
Available online 6 November 2015

Keywords:
Text classification
Transductive learning
Graph-based learning
Text mining
Label propagation
Bipartite heterogeneous network

ABSTRACT

Transductive classification is a useful way to classify texts when labeled training examples are insufficient. Several algorithms to perform transductive classification considering text collections represented in a vector space model have been proposed. However, the use of these algorithms is unfeasible in practical applications due to the independence assumption among instances or terms and the drawbacks of these algorithms. Network-based algorithms come up to avoid the drawbacks of the algorithms based on vector space model and to improve transductive classification. Networks are mostly used for label propagation, in which some labeled objects propagate their labels to other objects through the network connections. Bipartite networks are useful to represent text collections as networks and perform label propagation. The generation of this type of network avoids requirements such as collections with hyperlinks or citations, computation of similarities among all texts in the collection, as well as the setup of a number of parameters. In a bipartite heterogeneous network, objects correspond to documents and terms, and the connections are given by the occurrences of terms in documents. The label propagation is performed from documents to terms and then from terms to documents iteratively. Nevertheless, instead of using terms just as means of label propagation, in this article we propose the use of the bipartite network structure to define the relevance scores of terms for classes through an optimization process and then propagate these relevance scores to define labels for unlabeled documents. The new document labels are used to redefine the relevance scores of terms which consequently redefine the labels of unlabeled documents in an iterative process. We demonstrated that the proposed approach surpasses the algorithms for transductive classification based on vector space model or networks. Moreover, we demonstrated that the proposed algorithm effectively makes use of unlabeled documents to improve classification and it is faster than other transductive algorithms.

1. Introduction

Text automatic classification (TAC) is one of the most important tasks to manage, retrieve and extract knowledge from a huge number of textual documents (Manning, Raghavan, & Schütze, 2008; Nedjah, Mourelle, Kacprzyk, Frana, & de Souza, 2008; Berry & Castellanos, 2008; Li, Zhu, & Ogihara, 2008; He & Zhou, 2011; Uysal & Gunal, 2014). TAC automatically assigns a predefined category to a textual document.

* Corresponding author at: Department of Computer Science, Institute of Mathematics and Computer Science, University of São Paulo, Brazil. Tel.: +55 (16) 3373 9646; fax: +55 (16) 3373 9751.
E-mail address: ragero@icmc.usp.br (R.G. Rossi).

http://dx.doi.org/10.1016/j.ipm.2015.07.004
0306-4573/

Generally TAC is carried out by using inductive leaning algorithms (Weiss, Indurkhya, & Zhang, 2012; Sebastiani, 2002), which induce classification models to classify new or unseen texts. Usually a large number of labeled documents are necessary to induce an accurate classification model. Nevertheless, labeling texts is usually expensive and time consuming. Thus, a more practical approach is to employ methods which make use of the plenty of unlabeled texts available to perform and improve TAC.

Transductive approaches are widely used when labeled training data are insufficient. In this case, they make use of unlabeled data to improve classification performance (Kong, Ng, & Zhou, 2013; Chapelle, Schölkopf, & Zien, 2006; Belkin, Niyogi, & Sindhwani, 2006; Joachims,1999). Transductive classification directly estimates the labels of unlabeled instances without creating a model to classify new texts. Several algorithms considering texts represented in a vector space model have been developed to perform transductive classification such as Self-Training (Yarowsky,1995), Co-Training (Blum & Mitchell, 1998), Expectation Maximization (EM) (Nigam, McCallum, Thrun, & Mitchell, 2000), and Transductive Support Vector Machines (TSVM) (Joachims,1999). However, the use of these algorithms is unfeasible in practical applications due to the assumptions of these algorithms about the data distribution and computational cost. Moreover, the assumption that instances or terms are independent also impairs their classification performances.

Network-based algorithms came up to avoid the drawbacks of the algorithms based on vector space model and to improve transductive classification. Networks are mostly used for label propagation, in which some labeled objects propagate their labels to other objects through the network connections to perform transductive classification (Zhu & Goldberg, 2009; Rossi, Lopes, & Rezende, 2014; Subramanya & Bilmes, 2008; Zhou, Bousquet, Lal, Weston, & Schölkopf, 2004). Label propagation using just few labeled examples can obtain higher classification performance than inductive classification using a large number of labeled examples for TAC (Rossi, Lopes, & Rezende, 2014). Moreover, the use of networks to model text collections allows extracting patterns which are not extracted by algorithms based on vector-space model (VSM) (Breve, Zhao, Quiles, Pedrycz, & Liu, 2012).

Text collections are modeled as networks using homogeneous or heterogeneous networks. Homogeneous networks contain objects of a single type and heterogeneous networks are compounded by objects of different types. Document homogeneous networks have been used to model text collections as networks for label propagation (Jebara, Wang, & Chang, 2009; Kim, Pantel, Duan, & Gaffney, 2009; Subramanya & Bilmes, 2008; Wang & Zhang, 2006; Castillo, Donato, Gionis, Murdock, & Silvestri, 2007; Zhou et al., 2004; Zhu, Ghahramani, & Lafferty, 2003). In such networks, documents propagate their labels directly to other documents. Documents are connected according to hyperlinks, citations or similarities. The use of just hyperlinks and citations to build document networks reduces the quality of classification (Angelova & Weikum, 2006) and limits the application domains. On the other hand, documents wired considering similarity have been applied since they model any type of text collections and improve the classification quality (Angelova & Weikum, 2006). However, computing similarities poses a high computational cost, and the parameters such as minimum similarity or number of neighbors, significantly impact the classification accuracy (de Sousa, Rezende, & Batista, 2013).

Bipartite networks have come up as an alternative to model text collections as networks (Rossi, Faleiros, Lopes, & Rezende, 2012; Rossi, Lopes, Faleiros, & Rezende, 2014; Rossi, Lopes, & Rezende, 2014), in which objects correspond to documents and terms. Terms are linked to documents in which they are present. This network is easily generated, since there is no need to set parameters or compute similarities. Moreover, it has provided promising results for text classification (Rossi et al., 2012; Rossi, Lopes, Faleiros, et al., 2014: Rossi, Lopes, & Rezende, 2014). In such networks, documents propagate their labels to terms and then the terms propagate their labels to documents.

Instead of using the bipartite network structure just as means to propagate labels, this structure can be used to set the relevance scores of terms for classes, i.e., how much the presence of a term in a document increases or decreases the probability of a document belonging to a class. In (Rossi et al., 2012; Rossi, Lopes, Faleiros, et al., 2014) the relevance scores of terms for classes are induced using the bipartite network structure. These relevance scores were used to classify new/unseen documents, providing accuracies higher than state-of-the-art algorithms. However, scenarios with only few labeled documents impair the induction of term scores and consequently the classification accuracy.

In this paper we propose an algorithm to set the relevance scores of terms for classes considering labeled and unlabeled documents represented in a bipartite heterogeneous network. The relevance scores are obtained through an optimization process considering the current labels of the documents. The obtained relevance scores are propagated to define the new labels to unlabeled documents. Optimization and label propagation are repeated iteratively until converge, i.e., until the labels assigned to unlabeled documents do not change. The proposed algorithm, named TCBHN (*Transductive Classification based on Bipartite Heterogeneous Network*) obtains better classification performance and is faster than transductive algorithms based on vector space model or networks.

The main contributions of this article are fivefold:

- We propose a transductive classification algorithm which effectively makes use of unlabeled data to improve text classification.
- We propose a scalable transductive classification algorithm which makes use of bipartite networks to perform transductive classification.