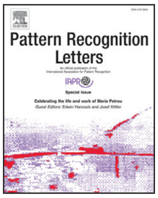
Pattern Recognition Letters 000 (2016) 1–12

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



# Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs[☆]

Thiago de Paulo Faleiros[*], Rafael Geraldeli Rossi, Alneu de Andrade Lopes

*Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo – Campus de São Carlos, Caixa Postal 668, 13560-970 São Carlos SP, Brazil*

ARTICLE INFO

*Article history:*
Available online xxx

*Keywords:*
Text classification
Transductive learning
Graph-based learning
Text mining
Label propagation
Bipartite graphs

ABSTRACT

Transductive classification is an useful way to classify a collection of unlabeled textual documents when only a small fraction of this collection can be manually labeled. Graph-based algorithms have aroused considerable interests in recent years to perform transductive classification since the graph-based representation facilitates label propagation through the graph edges. In a bipartite graph representation, nodes represent objects of two types, here documents and terms, and the edges between documents and terms represent the occurrences of the terms in the documents. In this context, the label propagation is performed from documents to terms and then from terms to documents iteratively. In this paper we propose a new graph-based transductive algorithm that use the bipartite graph structure to associate the available class information of labeled documents and then propagate these class information to assign labels for unlabeled documents. By associating the class information to edges linking documents to terms we guarantee that a single term can propagate different class information to its distinct neighbors. We also demonstrated that the proposed method surpasses the algorithms for transductive classification based on vector space model or graphs when only a small number of labeled documents is available.

## 1. Introduction

The most common way of storing information is in textual format. In fact, a huge amount of data produced and stored every day are textual documents. In this scenario, automated techniques to help classify textual documents is one of the most important tasks to organize, manage, and extract knowledge from these data and still remain as worthwhile research topics for machine learning and data mining communities.

The task of text classification is usually carried out by inductive learning algorithms which aim to induce a model to classify new or unseen documents. A considerable number of labeled documents is necessary to create an accurate classification model. However, a consistent set of labeled documents to induce a classification model is not available in most of the real applications, since the labeling task be an expensive, time consuming and expert dependent task. Thus, a more practical way of approaching text classification in real applications is to employ methods which make use of a small set of labeled documents with unlabeled documents (a large set is usually available).

Transductive approaches are widely used when labeled training data are insufficient and the goal is to classify a known set of documents. In this case, they make use of unlabeled data to improve classification performance [3,7,11,12,19]. Transductive classification directly estimates the labels of unlabeled instances without creating a model to classify new texts.

In addition to the classifier problem, another issue of the text classification task is the data representation. The traditional representation of textual documents is the vector space model (VSM). Nevertheless, more expressive representations, such as homogeneous or heterogeneous graphs may be employed. In a homogeneous graph, only link between objects of the same type is allowed, therefore it contains only document-document or term-term relations. In a heterogeneous graph, only link between pair of objects of different types is allowed. An intuitive way to represent a collection of documents is by creating a bipartite graph where vertices correspond to documents and terms, and edges represent the occurrence of the word in the document. Eventually, a weight can be assigned to the edge according to the frequency of the word in the document.

To describe algorithms over a graph representation has several advantages, since graph representation: (1) avoids sparsity and ensures low memory consumption; (2) enables easy description of operations for the inclusion of the topological structure of a dataset; (3) enables an optimal description of the topological

[☆] This paper has been recommended for acceptance by Cheng-Lin Liu.

[*] Corresponding author. Tel.: +551633739678.
*E-mail address:* thiagopf@icmc.usp.br (T. de Paulo Faleiros).

http://dx.doi.org/10.1016/j.patrec.2016.04.006
0167-8655/© 2016 Elsevier B.V. All rights reserved.

structure of a dataset; (4) provides local and global statistics of the datasets structure; and (5) allows extracting patterns which are not extracted by algorithms based on vector-space model [6].

Graph-based algorithms are mostly used in a label propagation schema in which some labeled objects propagate their labels to other objects through the graph structure to perform transductive classification [4,21,32,34]. Appropriate use of the richness of information conveyed by these graphs can lead to label propagation algorithms that using just few labeled examples surpass the classification performance of inductive classifiers using a large number of labeled examples [19].

Here, we propose an algorithm that uses a bipartite heterogeneous graph representation. The rationale behind the proposed graph-based semi-supervised algorithm is to associate class information to each vertex and edge. The class information is a *l*-dimensional vector, where *l* is the number of classes. Our algorithm is an iterative propagation procedure, in which the class information related to a vertex influences their neighbors' labels until convergence, by using a label propagation schema. In our proposal, and due to the characteristic of bipartite heterogeneous graph, the documents propagate their class information to edges, and edges propagate their class information back to documents and terms. Traditional label propagation algorithms optimize the class information vectors considering each dimension independently whereas our propagation algorithm optimizes the divergence between closely related vertices considering all dimensions of the class information vectors. The divergence between class information is optimized by maximizing the generalized Kullback–Leibler divergence [17]. The proposed algorithm, named TPBG (Transductive Propagation in Bipartite Graph) obtains better classification performance than state-of-the-art transductive algorithms based on vector space or graphs models when only a small number of labeled documents is available.

The main contributions of this paper are threefold: (1) to bring the advantages of bipartite graph representation and iterative propagation to the semi-supervised transductive learning process; (2) to propose an algorithm which surpasses the classification accuracy of state-of-the-art algorithms based on vector space model or graphs when considering a small number of labeled documents; (3) to carry out a comprehensive comparative evaluation of our proposal.

In the experimental evaluation we also present the behavior of the algorithms for a different range of labeled documents. The results showed that our algorithm returns consistent results, which makes the method a competitive alternative and a new exploratory possibility to state-of-the-art of semi-supervised algorithms.

The remainder of this paper is organized as follows. Section 2 presents related works about transductive classification. Section 3 presents details on the proposed algorithm for transductive classification of texts using bipartite graphs. Section 4 presents details of the experimental evaluation and the results. Finally, Section 5 presents the conclusions and points to future work.

## 2. Related works

Differently from supervised inductive classification, which aims to learn a model from labeled examples, the goal of transductive learning is to predict the class labels of the given labeled and unlabeled examples. In the context of document classification, transductive learning assigns weights for each dimension of the class information vector of each document and the documents are classified considering these weights.

In general, there are two ways to represent text collections to perform transductive learning: vector-space model and graph based representations. In the vector-space model, documents are represented as vectors and each dimension correspond to term of the document collection. The values in the vectors are based on the frequency of a term, such as binary weights, term frequency (tf) or term frequency-inverse document frequency (tf-idf). In the graph based representation, the objects corresponding to documents or terms are represented as vertices, and the relationship between pairs of objects are represented by edges. Different types of objects and different relations can be used to generate a graph-based representation. Documents can be connected according to “explicit relations” as hyperlinks or citations [15,25], or considering similarity [2]. Terms can be connected by precedence in text [1], if they present syntactic/semantic relationship [24], or if they co-occur in text collection or in pieces of texts as sentences/windows [13,16,27,29]. A combination of different types of objects is also used. In this case, documents and terms generate a bipartite graph where terms are connected to documents in which they occur [8,19,21].

### 2.1. Transductive learning on vector space model

The first proposals on transductive learning for text classification considered text collections represented by vector space model [5,11,14,30]. Perhaps, the most natural way to perform transductive learning is through Self-Training. Self-Training assumes that the most confident classifications are correct and re-induce the model by adding these new labeled instances to the training set.

Support Vector Machines (SVM) are one of the most popular classification algorithms used in machine learning. Its transductive version, Transductive Support Vector Machine (TSVM), have been used for text classification [11]. TSVM considers labeled and unlabeled documents to obtain a maximal margin hyperplane. The coefficients of a hyperplane correspond to the class information of terms. Based on the assumption that the classes are well-separated, the hyperplane with maximal margin will fall into a low density region. When this assumption does not hold, the TSVM classification algorithm is not accurate.

Transductive learning can also be performed by a probabilistic model. In [14] is presented a probabilistic framework which uses unlabeled data to improve a text classifier. A Expectation Maximization (EM) algorithm based in Multinomial Naive Bayes is used to estimate maximum a posteriori probability. The EM algorithm performs two steps. In the E-Step, the naive Bayes parameters, $\theta$, is used to estimate the component membership of each unlabeled document. In the M-Step, the parameter $\theta$ is re-estimated using all the documents and is established the class information for terms. EM classification is not accurate if the generative assumption is violated.

### 2.2. Transductive learning on graph

Here we define a graph as a triple $G = (\mathcal{V}, \mathcal{E}, f)$, where $\mathcal{V}$ is a set of vertices, $\mathcal{E}$ is a set of edges, and $f$ is a mapping which associate an edge to a real number, *i.e.* $f : \mathcal{E} \to \mathbb{R}$. To simplify notation we denote $f(e_{j,\,i})$ as $f_{j,\,i}$ for $e_{j,i} \in \mathcal{E}$. When $\mathcal{V}$ is compounded by a single type of objects, the graph is called homogeneous graph. When $\mathcal{V}$ is compounded by $h$ different types of objects, *i.e.*, $\mathcal{V} = \{V_1 \cup \ldots \cup V_h\}$, the graph is called heterogeneous graph [10]. To create a graph in a textual context, the vertices can be associated to documents, terms, pieces of a text, sentences or paragraphs, and all these objects can be combined in pairs to describe an edge. Usually, homogeneous networks are created considering explicit relations between pairs of documents [15,25], or considering similarity metric between documents [2]. With respect to heterogeneous graph in textual context, terms can be connected to documents [8,20,21] or sentence [28] in which they occur.

Please cite this article as: T. de Paulo Faleiros et al., Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.04.006