Contents lists available at ScienceDirect

Expert Systems With Applications

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

Ensemble multi-label text categorization based on rotation forest and latent semantic indexing

Haytham Elghazel*, Alex Aussem, Ouadie Gharroudi, Wafa Saadaoui

University of Lyon, Université Lyon 1, LIRIS UMR CNRS 5205, F-69622, France

ARTICLE INFO

Article history:
Received 27 October 2015
Revised 22 March 2016
Accepted 23 March 2016
Available online 24 March 2016

Keywords:
Multi-label classification
Text categorization
Ensemble learning
Rotation forest
Content analysis and indexing

ABSTRACT

Text categorization has gained increasing popularity in the last years due the explosive growth of multimedia documents. As a document can be associated with multiple non-exclusive categories simultaneously (e.g., Virus, Health, Sports, and Olympic Games), text categorization provides many opportunities for developing novel multi-label learning approaches devoted specifically to textual data. In this paper, we propose an *ensemble* multi-label classification method for text categorization based on four key ideas: (1) performing Latent Semantic Indexing based on distinct orthogonal projections on lower-dimensional spaces of concepts; (2) random splitting of the vocabulary; (3) document bootstrapping; and (4) the use of BoosTexter as a powerful multi-label base learner for text categorization to simultaneously encourage diversity and individual accuracy in the committee. Diversity of the ensemble is promoted through random splits of the vocabulary that leads to different orthogonal projections on lower-dimensional latent concept spaces. Accuracy of the committee members is promoted through the underlying latent semantic structure uncovered in the text. The combination of both rotation-based ensemble construction and Latent Semantic Indexing projection is shown to bring about significant improvements in terms of *Average Precision, Coverage, Ranking loss* and *One error* compared to five state-of-the-art approaches across 14 real-word textual data sets covering a wide variety of topics including health, education, business, science and arts.

1. Introduction

Document categorization is the assignment of documents to one or more predefined classes or categories based on their similarity to the conceptual content of the categories. It is a fundamental issue in natural language processing, including information retrieval, information extraction, and text mining. The document categorization problem may be cast as a multi-label classification (MLC) problem.

Formally, MLC amounts to finding a mapping from a space of features (the words) to a space of labels (the categories). Given a multi-label training set $\mathcal{D}$, the goal of multi-label learning is to find a function which is able to map any unseen example to its proper set of labels. Multi-label learning also emerged in other challenging applications such as gene function classification, spam filtering, sentiment analysis and semantic annotation of images (Boutell, Luo, Shen, & Brown, 2004; Clare & King, 2001; Zhang & Zhou, 2006) to cite a few. More recently, the MLC problem, viewed from a statistical perspective, attracted a great deal of interest in the machine learning community (Dembczynski, Waegeman, Cheng, & Hüllermeier, 2012; Gasse, Aussem, & Elghazel, 2015; Madjarov, Kocev, Gjorgjevikj, & Dzeroski, 2012; Zhang & Zhou, 2013) and several approaches have been proposed (see for instance (Madjarov et al., 2012; Zhang & Zhou, 2013) for a comparative review).

Research on multi-label learning was initially motivated by the difficulty of concept ambiguity encountered in text categorization, where each document may belong to one or more topics (labels) simultaneously. However, the existence of an underlying latent structure in textual data capture calls for MLC approaches that are specifically tailored to document categorization. Therefore, developing powerful and scalable MLC approaches devoted to document categorization is still an important issue in the field of machine learning.

In the last decade, there has been a great deal of research focused on ensemble MLC methods in order to improve the robustness and the generalization ability of single MLC learners which suffer from severe limitations in the presence of high-dimensional data, noisy, or imbalanced data. To achieve higher prediction accuracy than individual classifiers, it is crucial that the ensemble consists of highly accurate classifiers which at the same time disagree as much as possible.

* Corresponding author. Tel.: +33426234465.
E-mail addresses: haytham.elghazel@liris.cnrs.fr (H. Elghazel), alexandre.aussem@liris.cnrs.fr (A. Aussem), ouadie.gharroudi@liris.cnrs.fr (O. Gharroudi), wafa.saadaoui@univ-lyon2.fr (W. Saadaoui).

http://dx.doi.org/10.1016/j.eswa.2016.03.041
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

With this motivation in mind, we present a novel ensemble multi-label text categorization algorithm, termed Multi Label Rotation Forest (MLRF), based on a combination of Rotation Forest and Latent Semantic Indexing. The combination of both paradigms brings about significant benefits. On the one hand, Rotation Forest (Rodríguez, Kuncheva, & Alonso, 2006) is a one of the most powerful ensemble methods for binary classification problems as shown in a number of recent extensive experimental studies (Bibimoune, Elghazel, & Aussem, 2013; Kuncheva & Rodríguez, 2007) over a wide range of data sets. On the other hand, Latent semantic indexing (LSI) is an efficient indexing and retrieval method that uses a rank-reduced singular value decomposition (SVD) to identify patterns in the relationships between the words (or terms) and the (latent) concepts. The key idea is to apply the LSI on small random subsets of the vocabulary in order to build a collection of training sets with distinct samples and concept representations. The goal is to encourage simultaneously individual accuracy and diversity within the ensemble. Diversity is promoted through the different splits of the set of words that lead to different orthogonal projections on lower-dimensional subspaces, namely the space of concepts. Accuracy is promoted through the underlying latent semantic structure in the text uncovered by LSI. The LSI also reduces noise and other undesirable artifacts of the original space.

The main contribution of this paper is an investigation of the extent to which MRLF is powerful compared to state-of-the-art methods. Extensive experiments are conducted on various benchmark text categorization multi-label data sets. Our results demonstrate that the proposed method enjoys significant advantages compared to other methods.

The rest of the paper is organized as follow: Section 2 reviews recent studies on ensemble learning and multi-label learning methods with special emphasis on the multi-label document categorization methods; Section 3 introduces our multi-label classification method for text categorization; Experiments using relevant benchmark text categorization data sets are presented in Section 4; finally, Section 5 concludes the study and identifies some future research directions.

2. Related work

In this section, we review of the Rotation Forest algorithm and the standard MLC methods with special emphasis on the MLC methods devoted to document categorization.

2.1. Rotation Forest

The idea of exploiting ensemble learning to improve multi-label classification has received an increasing attention in the last few years. Dietterich (2000) states that “A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse”. Many methods have been proposed to generate accurate, yet diverse, sets of models. Bagging (Breiman, 1996), boosting (Freund & Shapire, 1996), Random Subspaces (Ho, 1998), Random Forest (Breiman, 2001) and Rotation Forest (Rodríguez et al., 2006) are the most popular examples of this methodology. We assume the reader is familiar with these techniques. We point the reader to (Zhou & Zhou, 2012) for a brief review.

Proposed by Rodríguez et al. (2006), Rotation Forest is another successful ensemble classifier generation technique in which the training set for each base classifier is formed by applying PCA (Principal Component Analysis) to rotate the original attribute axes. Specifically, to create the training data for a base classifier, the attribute set of data is randomly split into K subsets and PCA is applied to each subset. All principal components are retained in order to preserve the variability information in the data. Thus, K axis rotations take place to form the new attributes for a base classifier. The main idea of Rotation Forest is to simultaneously encourage diversity and individual accuracy within the ensemble: diversity is promoted through doing feature extraction for each base classifier and accuracy is sought by keeping all principal components and also using the whole data set to train each base classifier. Recent extensive experimental studies Kuncheva and Rodríguez (2007) and Bibimoune et al. (2013) over a wide range of benchmark and real data sets has demonstrated the effectiveness of Rotation Forest in the context of binary classification problems compared to a variety of well-known ensemble methods, including Adaboost, Rotation Forest and Rotboost (Zhang & Zhang, 2008).

2.2. Multi-label learning

As mentioned in the Introduction, the issue of learning from multi-label data has recently attracted significant attention over the last years (Dembczynski et al., 2012; Madjarov et al., 2012; Zhang & Zhou, 2013). While, many efficient approaches have been proposed, their theoretical underpinnings remain weak, at least, as compared to the rather complete theory of binary classification learning. In fact, the benefit of exploiting label dependence is closely dependent on the loss function as discussed in Dembczynski et al. (2012).

Most MLC methods intend to exploit, in one way or the other, dependencies between the class labels. Basically, these methods can be summarized into three categories: (a) algorithm adaptation methods, (b) problem transformation methods, and (c) *ensemble methods* (Madjarov et al., 2012). We discuss each category in what follows.

2.2.1. Algorithm adaptation methods

These methods extend specific learning algorithms (like decision trees, SVM, neural networks and k-nearest neighbors) to handle multi-label data directly. Clare et al. adapted the entropy function in the C4.5 decision tree algorithm to handle the multi-label data (Clare & King, 2001) by summing the label entropies. PCT, in Blockeel, Raedt, and Ramon (1998), is another algorithm adaptation decision tree capable of predicting multiple target attributes at once. The induction process in PCT uses the sum of the Gini indices throughout all labels to identify the best separation at each node. Elisseeff and Weston presented in (Elisseeff, 2005) a ranking approach based SVM to handle multi-label data. They propose to use the average fraction of incorrectly ordered pairs of labels as a cost function. Crammer and Singer propose to use neural network approach called BP-MLL in (Zhang & Zhou, 2006) which is an adaptation of back-propagation in the multi-label setting. The important modification of the algorithm is the use of function error that take considers multiple labels. From the popular k-Nearest Neighbors (k-NN), various multi-label learning have been proposed. Zhang and Zhou proposed in (Zhang, 2007) a lazy learning approach (MLkNN). Their model is similar to the traditional k-NN algorithm. Although, the determination of labels for a new test instance is different. The algorithm use prior and posterior probabilities of each label among the k-NN. More recently, based on the variable precision neighborhood rough sets, two multi-label classification approaches named MLRS and MLRS-LC were proposed in (Yu, Pedrycz, & Miao, 2014). Both approaches consider the aspects of correlation among the labels and uncertainty in the mapping between the feature space and the label space to improve the quality of multi-label classification. MLRS and MLRS-LC respectively provide a global and local view at the label correlation. Although, a series of experiments reported in (Yu et al., 2014) have shown that both approaches perform well in several domains, their performances depend on the nature of the data. As a nearest-neighbors-based method, they have been recognized as having poor perfor-