Information Sciences 406–407 (2017) 87–101

Contents lists available at ScienceDirect

Information Sciences

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

Combining dissimilarity spaces for text categorization

Roberto H.W. Pinheiro*, George D.C. Cavalcanti, Ing Ren Tsang

Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jornalista Anibal Fernandes s/n, Cidade Universitária 50740-560, Recife, PE, Brazil

ARTICLE INFO

Article history:
Received 20 June 2016
Revised 29 March 2017
Accepted 9 April 2017
Available online 12 April 2017

Keywords:
Text categorization
Multiple classifier system
Dissimilarity representation

ABSTRACT

Text categorization systems are designed to classify documents into a fixed number of predefined categories. Bag-of-words is one of the most used approaches to represent a document. However, it generates high-dimensional sparse data matrix with a high feature-to-instance ratio. An aggressive feature selection can alleviate these drawbacks, but such selection degrades the classifier's performance. In this paper, we propose an approach for text categorization based on Dissimilarity Representation and multiple classifier systems. The proposed system, Combined Dissimilarity Spaces (CoDiS), is composed of multiple classifiers trained on data from different dissimilarity spaces. Each dissimilarity space is a transformation of the original space that reduces the dimensionality, feature-to-instance ratio, and sparseness. Experiments using forty-seven text categorization databases show that CoDiS presents a better performance in comparison to literature systems.

1. Introduction

Text categorization, also known as text classification or topic spotting, aims at classifying a document into one or more previously known categories. A common way to represent a document as a feature vector is using Bag-of-Words (BoW), in which each feature is a term that appears in any document of the *Corpus* (database of documents). BoW was first introduced for text retrieval [17] but now is the most used approach to represent a document in text categorization problems. However, BoW has three well-known drawbacks: i) High-dimensionality [22]: Each document is represented by a feature vector, in which each feature is a term that appears in any document of the *Corpus*. Since all terms of the *Corpus* are considered as features, it is common to have tens of thousands of features; ii) High feature-to-instance ratio [12]: the number of features (tens of thousands) is usually much higher than the number of documents, or instances, in the Corpus; iii) Sparse data matrix [37]: The value for the features of the terms that do not appear in the document is zero. Since each document has only a small percentage of all existing terms, the data matrix is sparse.

Feature selection algorithms reduce the dimensionality of the feature vectors by removing features based on heuristics. This strategy is commonly used to address the drawbacks of BoW. To determine which features should be discarded, different criteria are employed, such as feature evaluation functions [6,11,39] which generate a score per feature. In this case, features with a score greater than a threshold is considered relevant. There are other alternatives to select features such as the proposal of Mitra et al. [25] that uses structural similarity; ALOFT [30] that relies on the idea that each document should contribute to the feature vector; and, the proposal of Maldonado et al. [24] that is based on loss functions. Consequently, feature selection algorithms reduce the dimensionality and, consequently, the feature-to-instance ratio; but, sparseness is

* Corresponding author.
E-mail addresses: rhwp@cin.ufpe.br (R.H.W. Pinheiro), gdcc@cin.ufpe.br (G.D.C. Cavalcanti), tir@cin.ufpe.br (I.R. Tsang).

http://dx.doi.org/10.1016/j.ins.2017.04.025
0020-0255/

Table 1
Related works.

Method	Feature	Sparseness	Ensemble	Summary
Prabowo and Thelwall [33]	✓		✓	Combines different types of classifiers
Ozgur and Gungor [27]	✓			Bag of Words extension based on pruning concepts
Xia et al. [46]			✓	Integrates different types of classifiers, features sets and combination approachs
Pinheiro et al. [30]	✓			Feature selection method, called ALOFT
De Silva et al. [7]	✓		✓	Combines different feature representation with different types of classifiers
Wang et al. [41]	✓		✓	Compares ensemble methods
Wu et al. [45]	✓		✓	Hybrid (SVM and Random Forest) method for imbalanced text data
Jun et al. [18]	✓	✓		Document support vector clustering with dimension reduction
Pinheiro et al. [32]	✓	✓		Cosine representation with prototype selection
Zhang and He [49]		✓	✓	Two classifier ensemble with enriched Bag-of-Words
Altinel et al. [1]	✓	✓		Class weighting kernel
Onan et al. [26]	✓		✓	Statistical keyword extraction
Proposed method	✓	✓	✓	Combines different dissimilarity spaces

slightly diminished. An aggressive feature selection is required to address these three drawbacks. However, removing many features can increase the classification error because of information loss [47].

Here we propose an approach, called Combined Dissimilarity Spaces (CoDiS), for text categorization. CoDiS is a multiple classifier system in which each classifier is trained using a different Dissimilarity Representation [28] that transforms the feature vectors to a new low-dimensional representation. In this representation, each document is represented by a dissimilarity vector composed of distances to all documents that belongs to a representation set (a subset of the training set). Since the representation set contains documents from different categories, the dissimilarity between documents of the same category should be small, and the distance between documents in different categories should be significant, this property increases the discrimination between the categories [29].

Multiple classifier system (MCS) relies on the assumption that combining classifiers can lead to an improvement in the accuracy rate and usually performs better than individual classifiers [50]. Gangeh et al. [12] proposed an MCS that presented better performance in text categorization than state-of-the-art individual classifiers, such as Support Vector Machines (SVM). They used a classifier generation method called Random Subspace (RSS) [15] in which each classifier is trained with a random subset of the original features. Thus, each classifier can deal with a different part of the feature space, and the final answer is given by the combination of the classifiers responses. In the literature, there are other works [4,48] that successfully used Random Subspace to deal with high-dimensional problems.

The proposed approach, CoDiS, uses Dissimilarity Representation and multiple classifier systems to improve the accuracy rate while diminishes the drawbacks of BoW. We claim that the Dissimilarity Representation entails less information loss compared to an aggressive feature selection because feature selection discards features, whereas all terms are built in the calculation of the distance between documents in the Dissimilarity Representation. As a consequence of the dimensionality reduction promoted by the Dissimilarity Representation and the transformation of the original features to distances, all three BoW drawbacks are diminished. Multiple classifiers is a more robust solution than a single classifier to deal with different databases [48]. CoDiS also avoids searching for the best representation set since multiple classifiers can overcome local optima by combining different initial conditions [9].

Through a set of comprehensive experiments on 47 databases, we show that the proposed method can considerably reduce the dimensionality while improving the recognition rates. The results reached by our method compare favorably to other ensemble methods using three statistical tests.

Our contributions and findings can be summarized as follows: i) A new Text Categorization system, CoDiS; ii) Dissimilarity Representation can benefit from multiple classifiers; iii) Dissimilarity Representation can reduce sparseness and dimensionality with a good compromise between performance and reduction; iv) Euclidean distance obtains better results than cosine similarity for CoDiS because Dissimilarity Representation with Euclidean distance is able to generate diverse classifiers; v) Random prototype selection methods are more adequate than prototype selection algorithms to produce a diverse ensemble in the CoDiS architecture.

This paper is organized as follows: The next section reviews the literature focusing on works that deal with a high number of features, the sparseness of data and that use ensemble in Text Categorization. Section 3 reviews the Dissimilarity Representation, which is a transformation capable of reducing the drawbacks of Bag-of-Words. Section 4 presents the proposed system, called Combined Dissimilarity Spaces (CoDiS), a Text Categorization system that combines different dissimilarity spaces. Section 5 describes the methodology of the experiments, presents preliminary experiments and the analysis of the final experimental results. Finally, Section 6 concludes this paper and describes some future works.

2. Related works

In this section, we discuss some important Text Categorization works available in the literature. The articles showed in Table 1 focus on three aspects: i) feature reduction (column Feature) which indicates the use of feature selection, feature extraction, or other feature transformation; ii) sparseness treatment (column Sparseness) which indicates that the work