Knowledge-Based Systems 108 (2016) 50–64

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A new hybrid semi-supervised algorithm for text classification with class-based semantics

Berna Altınel, Murat Can Ganiz*

Department of Computer Engineering, Marmara University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 15 November 2015
Revised 13 June 2016
Accepted 14 June 2016
Available online 15 June 2016

Keywords:
Semantics
Semi-supervised classification
Text classification
Semantic smoothing kernel
Class-based transformations

ABSTRACT

Vector Space Models (VSM) are commonly used in language processing to represent certain aspects of natural language semantics. Semantics of VSM comes from the distributional hypothesis, which states that words that occur in similar contexts usually have similar meanings. In our previous work, we proposed novel semantic smoothing kernels based on classspecific transformations. These kernels use class-term matrices, which can be considered as a new type of VSM. By using the class as the context, these methods can extract class specific semantics by making use of word distributions both in documents and in different classes. In this study, we adapt two of these semantic classification approaches to build a novel and high performance semi-supervised text classification algorithm. These approaches include Helmholtz principle based calculation of term meanings in the context of classes for initial classification and a supervised term weighting based semantic kernel with Support Vector Machines (SVM) for the final classification model. The approach used in the first phase is especially good at learning with very small datasets, while the approach in the second phase is specifically good at eliminating noise in a relatively large and noisy training sets when building a classification model. Overall, as a semantic semi-supervised learning algorithm, our approach can effectively utilize abundant source of unlabeled instances to improve the classification accuracy significantly especially when the amount of labeled instances are limited.

1. Introduction

Vector Space Models (VSM) are commonly used language processing to represent some aspects of natural language semantics. In this model, documents are simply represented as points in vector space and closeness of two points is proportional to their semantic similarity. There are several categories of VSM including word-context, pair-pattern and term-document matrices [5]. As pointed out in [5], semantics of VSM comes from the distributional hypothesis, which states that words that occur in similar contexts usually have similar meanings [1]. One of our main motivations is to use a class of documents as the context. In our previous work, we proposed novel semantic smoothing kernels based on class specific transformations [2,3]. Since these kernels use class labels explicitly, they are strictly supervised. Furthermore, they can be considered as a new type of VSM consist of term-class matrices specific to the class labeled documents. By using the class as the context, these matrices can extract class specific semantics by making use of word distributions both in documents and in different classes. These semantic kernels can be integrated into supervised classifiers i.e. SVM for text classification and they could be able to outperform baseline classifiers using document based traditional VSM. In this study, by combining one of these supervised semantic kernel based classification algorithm [3] with class meanings based methodology to classify unlabeled documents which is inspired from [4], we propose a novel and high performance semi-supervised text classification algorithm.

In machine learning applications there are two conventional strategies; supervised learning and unsupervised learning. Traditional supervised learning algorithms need a set of sufficient labeled data as training set to build the classification model, which will be used to predict the class memberships of the unlabeled examples. On the other hand, unsupervised learning, solely based on unlabeled samples, doesn't need any labeled data to learn a model. So as to train a classifier, it attempts to discover the indirect structure of unlabeled data [6]. There have been massive amounts of accumulated data on the web, particularly on blogs, forums, social networks and continue to increase day by day without any doubt. But unfortunately most of the available data does not have pre-assigned labels which limit their use in several

* Corresponding author.
E-mail addresses: berna.altinel@marmara.edu.tr (B. Altınel), murat.ganiz@marmara.edu.tr (M.C. Ganiz).

http://dx.doi.org/10.1016/j.knosys.2016.06.021
0950-7051/

practical machine learning application fields such as text classification, sentiment recognition, speech recognition. Moreover, generally it can be time-consuming, tedious and expensive to assign labels to them manually. Most importantly, learning a classifier with only a few labeled training data may not generate sufficient performance. In situations where labeled data is inadequate, many algorithms have suggested exploiting and utilizing the unlabeled data to support to learning process for better classification. SSL approaches utilize not only labeled data but also unlabeled data to increase the classification accuracy. Recently, SSL has become popular and gained increased attention of both academic and commercial platforms as a new machine learning strategy. SSL is different from two ordinary classification approaches by the usage of unlabeled data to mitigate the effect of insufficient labeled data on classifier accuracy. Many SSL systems have been offered in the past years, like co-training [7], self-training [8,9] graph-based methods [10], semi-supervised support vector machines [6], EM with generative mixture models [11], transductive support vector machines [12].

Classification of texts requires special techniques to transform unstructured text into structured format required for classification algorithms; usually a vector of features. In this domain, documents are usually symbolized by terms and their corresponding frequencies. This kind of representation methodology is actually the most popular one in the literature and it is named as Bag of Words (BOW) feature demonstration. BOW is known as the simplest feature representation method where each term comprises a dimension in a vector space, being independent of other terms in the same document [13]. A bag is mathematically similar to set with the difference that there can be duplicate values. As in sets, the order of words is lost in bag representation. Similarly, a bag can be represented as a vector and a group of bags also can be represented as a matrix where the rows are documents and columns are term frequencies. This also called Vector Space Model (VSM). Although, the VSM and BOW approach is very simple and commonly used, they have several limitations as discussed in [2]. One of the restrictions is the assumption of independency between terms. Documents are represented only with their term frequencies, disregarding their position in the document or their semantic or syntactic links between other words. This is a big problem since it clearly ignores the multi-word expressions by separating them. Moreover; it cannot handle polysemous words (i.e. words with multiple meanings) since it treats them as a single entity. Furthermore, it maps synonymous words into different components; as discussed in [14]. In principle, as Steinbach et al. [15] mention, each class has two kinds of vocabulary: one is “core” vocabulary which are closely correlated to the theme of that class, the other type is “general” vocabulary those may have similar distributions on different classes. Consequently, two documents from different classes may commonly have many general words and can be categorized as similar in the BOW representation.

In our recent study in [2], we offer a novel method for constructing a supervised semantic smoothing kernel for SVM, which we name Class Meaning Kernel (CMK). The proposed method smoothens the words of a document in BOW demonstration by class-based meaning values of terms. The meaning scores of words are calculated based on the Helmholtz principle from Gestalt theory [16–19] in the scope of classes. According to our experimental results, CMK is superior to the traditional kernels such as linear kernel, polynomial kernel and RBF kernel. In one of our previous studies, we suggest a new classifier for textual data, named as Supervised Meaning Classifier (SMC) [4]. The SMC classifier uses meaning calculations, which is based on Helmholtz principle from Gestalt Theory. In SMC, meaningfulness of terms in the scope of classes are calculated and used for classification of a document. According to the experiment results this new SMC classifier outperforms Multinomial Naïve Bayes (MNB) and SVM specifically on insufficient training data.

In another recent study of ours [3], we offer a novel approach for constructing a supervised semantic kernel for SVM, which we name Class Weighting Kernel (CWK). The proposed method smoothens a document in BOW representation by using class-based weights of words. The weights of terms are calculated based on a term weighting approach that is designed as a part of a feature extraction algorithm is presented in [20,21]. According to our experimental results the classification performance of CWK is higher than the classification performance of the other commonly used kernels (i.e. linear kernel, polynomial kernel and RBF kernel).

Inspired by the advantages of CMK, CWK and SMC, and motivated by the fact that there are insufficient labeled data in real life practical applications, we build a new hybrid semi-supervised form of CWK and SMC, which is called Hybrid Class Semantics Classifier (HCSC). More precisely, in this article we suggest a novel non-iterative semi-supervised methodology that uses class-based meaning values and weights of terms. The suggested approach utilizes both labeled and unlabeled data in order to build a classifier. First it smoothens the terms of the labeled documents in BOW demonstration (document vector represented by term frequencies) by class-based meanings of words as the same way it is done in CMK [2] and SMC [4]. Then, it attempts to find suitable labels for unlabeled instances. It succeeded this labeling process by weighting the terms of the unlabeled documents in BOW representation with the help of meaning calculations [4]. Following this, HCSC combines the original labeled data with newly classified unlabeled data. Finally, CWK is applied on this enlarged labeled dataset in order to predict the labels of the samples in the test dataset. The smoothing process in HCSC increases the significance of important (i.e. meaningful) words specific to a particular class while reducing the importance of general terms those have a similar distribution in all classes. Since this method is used in the transformation phase of a kernel function from input space into a feature space, it considerably decreases the effects of above mentioned disadvantages of BOW. We note that HCSC advances the accuracy of SVM compare to the linear kernel by giving more significance to the class specific concepts, those may possibly be synonymous or very closely associated in the scope of a class. The HCSC uses a semantic smoothing matrix in the transformation of the original space into the feature space. This semantic smoothing mechanism maps the similar documents to close positions in the feature space of SVM if they are written using semantically nearby sets of terms on the same subject. The main novelty of our approach is the use of this meaning information in the both labeling of unlabeled data and smoothing process of the semantic kernel. The meanings of words are calculated based on the Helmholtz principle from Gestalt theory [16–19] in the scope of both classes and documents as in [2]. HCSC directly incorporates class information to the semantic kernel for labeled instances. Additionally it also incorporates unlabeled instances in the training process. Therefore, it can be considered as a semi-supervised approach.

We performed a number of experiments on several document datasets with numerous different labeled/unlabeled/test splits of the corpus. According to our experimental results HCSC broadly outperforms the performance of baseline algorithms.

The first gain of our proposed solution is the classification capability of HCSC. To show the performance difference and robustness of HCSC we perform some experiments on various textual datasets with different labeled/unlabeled portions of the dataset. Experimental results demonstrate that HCSC exceeds the performance of baselines including the semi-supervised form of linear kernel. In linear kernel the inner product between two document vectors is