Knowledge-Based Systems 89 (2015) 265–277

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A novel semantic smoothing kernel for text classification with class-based weighting

Berna Altınel [a,*], Banu Diri [b], Murat Can Ganiz [c]

[a] Computer Engineering Department, Marmara University, Istanbul, Turkey
[b] Computer Engineering Department, Yıldız Technical University, Istanbul, Turkey
[c] Computer Engineering Department, Doğuş University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 24 December 2014
Received in revised form 28 May 2015
Accepted 10 July 2015
Available online 17 July 2015

Keywords:
Support vector machines
Text classification
Semantic kernel
Semantic smoothing kernel
Class-based term weighting

ABSTRACT

In this study, we propose a novel methodology to build a semantic smoothing kernel to use with Support Vector Machines (SVM) for text classification. The suggested approach is based on two key concepts; class-based term weighting and changing the orthogonality of vector space. A class-based term weighting methodology is used for transformation of documents from the original space to the feature space. This class-based weighting basically groups terms based on their importance for each class and consequently smooths the representation of documents. This is accomplished by changing the orthogonality of the Vector Space Model (VSM) with introducing class-based dependencies between terms. As a result, on the extreme case, two documents can be seen as similar even if they do not share any terms but their terms are similarly weighted for a particular class. The resulting semantic kernel can directly make use of class information in extracting semantic information between terms, therefore it can be considered as a supervised kernel. For our experimental evaluation, we analyze the performance of the suggested kernel with a large number of experiments on benchmark textual datasets and present results with respect to varying experimental conditions. To the best of our knowledge, this is the first study to use class-based term weighting in order to build a supervised semantic kernel for SVM. We compare our results with kernels that are commonly used in SVM such as linear kernel, polynomial kernel, Radial Basis Function (RBF) kernel and with several corpus-based semantic kernels. According to our experimental results the proposed method favorably improves classification accuracy over linear kernel and several corpus-based semantic kernels in terms of both accuracy and speed.

1. Introduction

In recent years, with the ever accumulating online information on the Internet and social media, text categorization has become one of the key techniques for organizing and handling textual data. Automatically processing these huge amounts of textual data is an essential problem. Text classification can be defined as the utilization of a supervised learning methodology to assign predefined class labels to documents using a model learned from labels of the documents in the training set. An important requirement of efficient and accurate text classification systems is to organize documents into pre-determined categories that contain similar documents. In order to achieve this goal, several classification algorithms depend on similarity or distance measures that compare pairs of text documents. It is also known that vector space representation of textual documents yields high dimensionality and related to this; sparsity. This is especially a problem when there are a large number of category labels but limited amount of training data. It is thus crucial that a good text classification algorithm should scale well with the increasing number of features and classes. Most importantly, words in textual data carry semantic information, i.e., the sense carried by the terms of the documents. Consequently, an ideal text classification algorithm should be able to make use of this semantic information.

Bag-of-words (BOW) feature representation is well accepted as the fundamental approach in the domain of text classification. In BOW approach documents are characterized by the frequencies of individual words or terms and each term represents a dimension in a vector space independent of other terms in the same document [1]. It basically focuses on the frequency of words. The BOW approach over simplifies the representation of terms in documents by ignoring the several different syntactic or semantic

* Corresponding author.
E-mail addresses: berna.altinel@marmara.edu.tr (B. Altınel), banu@ce.yildiz.edu.tr (B. Diri), mcganiz@dogus.edu.tr (M.C. Ganiz).

http://dx.doi.org/10.1016/j.knosys.2015.07.008
0950-7051/

relations between terms in natural language, e.g. it treats polysemous words (words with multiple meanings) as a single entity. For instance the term "bank" may have different meanings like financial institution or a river side based on the context it appears. Additionally, the BOW feature representation maps synonymous words into different components [2]. In principle, as Steinbach et al. [3] investigate, each class of documents has two kinds of vocabulary: one is "core" vocabulary which are intimately associated to the topic of that class, the other type is "general" vocabulary (e.g. stop words) those may have similar distributions on different classes. Therefore, two different documents from different classes may share many general words and will have high similarity based on their BOW representations.

To address the above mentioned weaknesses of BOW model, several methods are proposed in the fields of word sense disambiguation, text classification and Information Retrieval (IR). These methods which enhance the representation of documents can be categorized as domain knowledge based systems, statistical approaches, hybrid methods, word sequence enhanced systems and linguistic enriched methods. These studies are discussed in Section 2.

Another issue in the BOW model is how to weight a term. There are different weighting approaches to assign appropriate weights to the terms to improve the classification performance including binary, Term Frequency (TF), Term Frequency–Inverse Document Frequency (TF–IDF) [4,5], Gain Ratio, Information Gain (IG), Odds Ratio, [6,7], Term Frequency–Relevance Frequency (TF–RF) [8] and Term Frequency–Inverse Class Frequency (TF–ICF) [9,10]. These methods are summarized in Section 2.3.

In our previous studies, we introduced a number of corpus-based semantic kernels: Higher-Order Semantic Kernel (HOSK) [11], Iterative Higher-Order Semantic Kernel (IHOSK) [12], Higher-Order Term Kernel (HOTK) [13] and Class Meaning Kernel (CMK) [15] for SVM. In those studies, we extend the traditional linear kernel (i.e. a dot product between document vectors) for text classification by embedding higher-order relations between terms and documents into the kernel. In the CMK, we employed meaningfulness calculations for terms and used these values to build a semantic kernel. These methods are discussed in Section 2.2.

In this study, we propose a novel approach for building a semantic smoothing kernel which makes use of the class-based term weights to improve the performance of SVM especially for text classification. The proposed approach is called Class Weighting Kernel (CWK). This class-based weighting basically groups terms based on their importance for each class. Consequently it smooths the representation of documents which changes the orthogonality of the vector space model by introducing class-based dependencies between terms. As a result, on the extreme case, two documents can be seen as similar even if they do not share any terms but their terms are similarly weighted for a particular class.

The suggested approach smooths the terms of a document in BOW representation with a methodology includes the calculation of the terms' discriminating power for each class in the training set. This in turn increases the importance of core or in other words significant terms specific to a particular class while reducing the importance of general terms that have a similar distribution in all classes. Since this approach is used in the transformation phase of a kernel function from input space into a feature space, it considerably reduces the effects of above mentioned disadvantages of BOW. We observe that CWK improves the accuracy of SVM compare to the linear kernel by increasing the importance of class specific concepts, which can be synonymous or very closely related in the context of a class. The CWK 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 nearby positions in the feature space of SVM if they are written using semantically closer sets of terms on the same topic/class.

The first advantage of our suggested solution is the capability of CWK to perform much better than standard kernels in terms of classification accuracy. To demonstrate performance improvements, we conduct several experiments on varied benchmark datasets with several different test environments. According to our experimental results CWK exceeds the performance of linear kernel which is one the state of the art kernels for text classification [14,22]. Additionally, experimental results show that CWK is superior to both polynomial kernel and RBF kernel. In linear kernel, the inner product between two document vectors is used as kernel function, which utilizes the information about shared terms in these two documents. However, CWK can take advantage of class-based weighting of terms; therefore it extends the context from a single document to a class of documents. In this way, semantic relation between two terms is composed of class-based weights of these terms for all classes. So, if the two terms are significant terms in the same class then the semantic relatedness value between them will be higher.

The second advantage of CWK is about its execution time. We evaluate CWK by comparing it to the traditional kernels as well as the corpus-based kernels of HOSK, IHOSK, HOTK [11–13] and CMK [15] by using several benchmark datasets. The CWK outperforms other corpus-based semantic kernels in many cases in terms of accuracy with less execution time.

The third advantage of the proposed approach is about its simplicity and independency of the outside semantic sources such as WordNet. As a result it can be applied to any domain without adjustments or parameter optimizations. To show this wide applicability of our kernel we present results with different experimental settings, such as: (i) several datasets from different domains such as newsgroups postings and movie reviews classified into sentiments, (ii) different training portions of the dataset in order to observe the effect of sparsity, and (iii) varying values of misclassification cost (C) parameter of the SVM.

The other benefit of CWK is that it also forms a foundation that can easily be combined with other term-based semantic similarity methods such as unsupervised semantic similarity measures. It is also possible to combine with similarities between terms derived from a semantic source like WordNet or Wikipedia.

The experimental results show significant improvements in the classification accuracy when class-based term weighting kernel is used in SVM, compared to the performance of the two types of previously mentioned baselines; linear kernel and our former semantic kernels like IHOSK [12] and HOTK [13]. To the best of our knowledge, class-dependent weighting is used in the transformation phase of SVM for the first time in the literature and give significant insights on the class-based semantic smoothing of terms in a kernel for text classification.

The current work extends beyond our previous studies on semantic kernels in [11–13]. The main contributions of this work, which distinguish it from our former works, can be summarized in the following:

- Embedding class-based term weights which reflect the importance of terms on classes, into a semantic kernel to smooth the representation of the text documents.
- Building a kernel with the capability of reaching higher accuracy in compare to linear kernel and our previous semantic kernels [11–13].
- Designing a smoothing mechanism in a kernel which works faster than our previous semantic kernels of IHOSK [12] and HOTK [13].