Engineering Applications of Artificial Intelligence 43 (2015) 54–66

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A corpus-based semantic kernel for text classification by using meaning values of terms

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

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

ARTICLE INFO

Article history:
Received 21 November 2014
Received in revised form
19 March 2015
Accepted 30 March 2015
Available online 29 April 2015

Keywords:
Support vector machines
Text classification
Semantic kernel
Meaning
Higher-order relations

ABSTRACT

Text categorization plays a crucial role in both academic and commercial platforms due to the growing demand for automatic organization of documents. Kernel-based classification algorithms such as Support Vector Machines (SVM) have become highly popular in the task of text mining. This is mainly due to their relatively high classification accuracy on several application domains as well as their ability to handle high dimensional and sparse data which is the prohibitive characteristics of textual data representation. Recently, there is an increased interest in the exploitation of background knowledge such as ontologies and corpus-based statistical knowledge in text categorization. It has been shown that, by replacing the standard kernel functions such as linear kernel with customized kernel functions which take advantage of this background knowledge, it is possible to increase the performance of SVM in the text classification domain. Based on this, we propose a novel semantic smoothing kernel for SVM. The suggested approach is based on a meaning measure, which calculates the meaningfulness of the terms in the context of classes. The documents vectors are smoothed based on these meaning values of the terms in the context of classes. Since we efficiently make use of the class information in the smoothing process, it can be considered a supervised smoothing kernel. The meaning measure is based on the Helmholtz principle from Gestalt theory and has previously been applied to several text mining applications such as document summarization and feature extraction. However, to the best of our knowledge, ours is the first study to use meaning measure in a supervised setting to build a semantic kernel for SVM. We evaluated the proposed approach by conducting a large number of experiments on well-known textual datasets and present results with respect to different experimental conditions. We compare our results with traditional kernels used in SVM such as linear kernel as well as with several corpus-based semantic kernels. Our results show that classification performance of the proposed approach outperforms other kernels.

1. Introduction

Text categorization plays a significantly important role in recent years with the rapid growth of textual information on the web, especially on social networks, blogs and forums. This enormous data increases by the contribution of millions of people every day. Automatically processing these increasing amounts of textual data is an important problem. Text classification can be defined as automatically organizing documents into predetermined categories. Several text categorization algorithms depend on distance or similarity measures which compare pairs of text documents. For this reason similarity measures play a critical role in document classification. Apart from the other, structured data types, the textual data includes semantic information, i.e., the sense conveyed by the words of the documents. Therefore, classification algorithms should utilize semantic information in order to achieve better results.

In the domain of text classification, documents are typically represented by terms (words and/or similar tokens) and their frequencies. This representation approach is one of the most common one and it is called Bag of Words (BOW) feature representation. In this representation, each term constitutes a dimension in a vector space, independent of other terms in the same document (Salton and Yang, 1973). The BOW approach is very simple and commonly used; yet, it has a number of restrictions. Its main limitation is that it assumes independency between

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

http://dx.doi.org/10.1016/j.engappai.2015.03.015
0952-1976/

terms, since the documents in BOW model are represented with their terms ignoring their position in the document or their semantic or syntactic connections between other words. Therefore it clearly turns a blind eye to the multi-word expressions by breaking them apart. Furthermore, it treats polysemous words (i.e., words with multiple meanings) as a single entity. For instance the term "organ" may have the sense of a body-part when it appears in a context related to biological structure, or the sense of a musical instrument when it appears in a context that refers to music. Additionally, it maps synonymous words into different components; as mentioned by Wang and Domeniconi (2008). In principle, as Steinbach et al. (2000) analyze and argue, each class has two types of vocabulary: one is "core" vocabulary which are closely related to the subject of that class, the other type is "general" vocabulary those may have similar distributions on different classes. So, two documents from different classes may share many general words and can be considered similar in the BOW representation.

In order to address these problems several methods have been proposed which use a measure of relatedness between term on Word Sense Disambiguation (WSD), Text Classification and Information Retrieval domains. Semantic relatedness computations fundamentally can be categorized into three such as knowledge-based systems, statistical approaches and hybrid methods which combine both ontology-based and statistical information (Nasir et al., 2013). Knowledge-based systems use a thesaurus or ontology to enhance the representation of terms by taking advantage of semantic relatedness among terms, for examples see (Bloehdorn et al., 2006), (Budanitsky and Hirst, 2006), (Lee et al., 1993), (Luo et al., 2011), (Nasir et al., 2013), (Scott and Matwin, 1998), (Siolas and d'Alché-Buc, 2000), and (Wang and Domeniconi, 2008). For instance in (Bloehdorn et al., 2006), (Siolas and d'Alché-Buc, 2000) the distance between words in WordNet (Miller et al., 1993) is used to capture semantic similarity between English words. The study in (Bloehdorn et al., 2006) uses super-concept declaration with different distance measures between words from WordNet such as Inverted Path Length (IPL), Wu-Palmer Measure, Resnik Measure and Lin Measure. A recent study of this kind can be found in (Zhang, 2013), which uses HowNet as a Chinese semantic knowledge-base. The second type of semantic relatedness computations between terms are corpus-based systems in which some statistical analysis based on the relations of terms in the set of training documents is performed in order to reveal latent similarities between them (Zhang et al., 2012). One of the famous corpus-based systems is Latent Semantics Analysis (LSA) (Deerwester et al., 1990) that partially solves the synonymy problem. Finally, approaches of the last category are called hybrid since they combine the information acquired both from the ontology and the statistical analysis of the corpus (Nasir et al., 2013), (Altınel et al., 2014a). There is a recent survey in (Zhang et al., 2012) about these studies.

In our previous studies, we proposed several corpus-based semantic kernels such as Higher-Order Semantic Kernel (HOSK) (Altınel et al., 2013), Iterative Higher-Order Semantic Kernel (IHOSK) (Altınel et al., 2014a) and Higher-Order Term Kernel (HOTK) (Altınel et al., 2014b) for SVM. In these studies, we showed significant improvements on classification performance over traditional kernels of SVM such as linear kernel, polynomial kernel and RBF kernel by taking advantage of higher-order relations between terms and documents. For instance, the HOSK is based on higher-order relations between the documents. The IHOSK is similar to the HOSK since they both propose a semantic kernel for SVM by using higher-order relations. However, IHOSK makes use of the higher-order paths between both the documents and the terms iteratively. Therefore, although, the performance of IHOSK is superior, its complexity is significantly higher than other higher-order kernels. A simplified model, the HOTK, uses higher-order paths between terms. In this sense, it is similar to the previously proposed term-based higher-order learning algorithms Higher-Order Naïve Bayes (HONB) (Ganiz et al., 2009) and Higher-Order Smoothing (HOS) (Poyraz et al., 2012, 2014).

In this article, we propose a novel approach for building a semantic kernel for SVM, which we name Class Meaning Kernel (CMK). The suggested approach smoothes the terms of a document in BOW representation (document vector represented by term frequencies) by class-based meaning values of terms. This in turn, increases the importance of significant or in other words meaningful terms for each class while reducing the importance of general terms which are not useful for discriminating the classes. This approach reduces the above mentioned disadvantages of BOW and improves the prediction abilities in comparison with standard linear kernels by increasing the importance of class specific concepts which can be synonymous or closely related in the context of a class. The main novelty of our approach is the use of this class specific information in the smoothing process of the semantic kernel. The meaning values of terms are calculated according to the Helmholtz principle from Gestalt theory (Balinsky et al., 2010, 2011a, 2011b, 2011c) in the context of classes.

We conducted several experiments on various document datasets with several different evaluation parameters especially in terms of the training set amount. Our experimental results show that CMK widely outperforms the performance of the other kernels such as linear kernel, polynomial kernel and RBF kernel. Please note that SVM with linear kernel is accepted as one the best performing algorithms for text classification and it virtually become de-facto standard in this domain. In linear kernel, the inner product between two document vectors is used as kernel function, which includes information about only the terms that these documents share. This approach can be considered as first-order method since its context or scope consists of a single document only. However, CMK can make use of meaning values of terms through classes. In this case semantic relation between two terms is composed of corresponding class-based meaning values of these terms for all classes. So if these two terms are important terms in the same class then the resulting semantic relatedness value will be higher. In contrast to the other semantic kernels that make use of WordNet or Wikipedia[1] in an unsupervised fashion, CMK directly incorporates class information to the semantic kernel. Therefore, it can be considered a supervised semantic kernel.

One of the important advantages of the proposed approach is its relatively low complexity. The CMK is a less complex and more flexible approach than the background knowledge-based approaches, since CMK does not require the processing of a large external knowledge base such as Wikipedia or WordNet. Furthermore, since CMK is constructed from corpus based statistics it is always up to date. Similarly, it does not have any coverage problem as the semantic relations between terms are specific to the domain of the corpus. This leads to another advantage of CMK: it can easily be combined with background knowledge-based systems that are using Wikipedia or WordNet. As a result, CMK outperforms other similar approaches in most of the cases both in terms of accuracy and execution time as can be seen from our experimental results.

The remainder of the paper is organized as follows: The background information with the related work including SVM, semantic kernels, and meaningfulness calculation summarized in Section 2. Section 3 presents and analyzes the proposed kernel for text classification algorithm. Experimental setup is described in Section 4, the corresponding experiment results including some discussion points are given in Section 5. Finally, we conclude the paper in Section 6 and provide a discussion on some probable future extension points of the current work.

[1] http://www.wikipedia.org/