Engineering Applications of Artificial Intelligence 62 (2017) 152–163

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Instance labeling in semi-supervised learning with meaning values of words

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

[a] Engineering Faculty, Department of Computer Engineering, Marmara University, Turkey
[b] Engineering Faculty, Department of Computer Engineering, Yıldız Technical University, Turkey

ARTICLE INFO

Keywords:
Text classification
Semantic kernel
Semi-supervised learning
Instance labeling
Helmholtz principle

ABSTRACT

In supervised learning systems; only labeled samples are used for building a classifier that is then used to predict the class labels of the unlabeled samples. However, obtaining labeled data is very expensive, time consuming and difficult in real-life practical situations as labeling a data set requires the effort of a human expert. On the other side, unlabeled data are often plentiful which makes it relatively inexpensive and easier to obtain. Semi-Supervised Learning methods strive to utilize this plentiful source of unlabeled examples to increase the learning capacity of the classifier particularly when amount of labeled examples are restricted. Since SSL techniques usually reach higher accuracy and require less human effort, they attract a substantial amount of attention both in practical applications and theoretical research. A novel semi-supervised methodology is offered in this study. This algorithm utilizes a new method to predict the class labels of unlabeled examples in a corpus and incorporate them into the training set to build a better classifier. The approach presented here depends on a meaning calculation, which computes the words' meaning scores in the scope of classes. Meaning computation is constructed on the Helmholtz principle and utilized to various applications in the field of text mining like feature extraction, information retrieval and document summarization. Nevertheless, according to the literature, ILBOM is the first work which uses meaning calculation in a semi-supervised way to construct a semantic smoothing kernel for Support Vector Machines (SVM). Evaluation of the proposed methodology is done by performing various experiments on standard textual datasets. ILBOM's experimental results are compared with three baseline algorithms including SVM using linear kernel which is one of the most frequently used algorithms in text classification field. Experimental results show that labeling unlabeled instances based on meaning scores of words to augment the training set is valuable, and increases the classification accuracy on previously unseen test instances significantly.

1. Introduction

Text categorization is a popular task whose aim is to label documents according to predefined class labels. There is a big amount of textual data collected on the internet especially on social networks, microblogging sites, blogs, forums, news, etc. This tremendous amount of texts continues to enlarge by the contributions of millions of people every day. Automatically processing and extracting meaning from these great amounts of documents is one of the main difficulties not only for research platforms but also for commercial platforms. The text classification plays a very important role in several popular and widely used applications such as document filtering, sentiment classification, information extraction, summarization and question answering. It is also significant to remember that, one of these applications is likely to be a part of a very important military, health and security engineering problem in real world cases. Nevertheless a very big portion of the accumulated data consists of unlabeled samples.

Bag of Words (BOW) is traditional representation methodology of unstructured textual data in the literature. Each of these terms in the same document represents an independent dimension in a vector space (Salton and Yang, 1973). There is no order of terms in BOW feature demonstration. Also, a bag is able to be demonstrated as a vector as well as a group of bags is able to be demonstrated as a matrix. The rows of this matrix represent the documents and columns of this matrix represent the corresponding term frequencies of these documents; which is called Vector Space Model (VSM). This approach mainly emphasizes the frequency of terms. The BOW methodology makes the representation of words simpler in documents by disregarding the following semantic and syntactic relations between words in natural language: 1.) It assumes independency between words, since it ignores the semantic connections among words. This will be an important problem especially for the documents which include multi-word expressions. 2.) It processes polysemous words like a particular unit. For example, word "bank" could have two distinct meanings according

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

http://dx.doi.org/10.1016/j.engappai.2017.04.003
Received 30 November 2016; Received in revised form 24 February 2017; Accepted 3 April 2017
Available online 19 April 2017

to the context it appears; one is a financial institution and the other is a river side (Wang and Domeniconi, 2008). 3.) It maps synonymous terms into completely different entities (Salton and Yang, 1973). Each class of texts has two forms of vocabulary: i)"core" vocabulary that is related to the theme of that class, ii) "general" vocabulary which may have almost identical distributions in distinct classes like stop words as Steinbach et al. (2000) analyze and discuss. Therefore, two unlike documents which cover completely distinct topics and belong to different classes may have several general terms in common as well as may have high similarity value according to their BOW feature demonstration.

An expected output of accurate and efficient text classification algorithms is to label unlabeled textual materials based on specified classes that comprise of identical textual materials. On account of accomplishing this goal, there are various classification methods which based on distance or similarity measures. These similarity measures compare pairs of documents and compute their similarities. It is also known that vector space demonstration of texts results sparsity and high dimensionality. This is a very big difficulty especially when there are numerous class labels however an inadequate training data. Hence it is critical that a successful and accurate text classifier should scale well with the large number of classes and features under the circumstances of restricted training data. However, rather preferably, terms in documents convey semantic information, i.e., the sense carried by the words of the textual materials. Therefore, a perfect text classification system should be able to take advantage of this semantic information.

Semantic text classification groups the documents into meaningful classes. In these kinds of classifiers, semantic connections among the words and the documents are taken into consideration. The texts which are semantically correlated to each other are classified with the same class label while the texts which are semantically unconnected are classified with different class labels. Semantic classification algorithms can also help in detecting the subject of a class. Semantic classification methodologies focuses on meanings of the terms and therefore the semantic approach mostly uses a dictionary or statistical calculations extracted from the corpus to build the classifier and then classify the test instances.

Advantages of semantic text classification over traditional text classification could be listed as follows:

- Semantic text classification algorithms help in information and relationship detection among words of the texts.
- Semantic text classification algorithms can contribute semantically relating the classes to one another.
- Semantic text classification approach can give the opportunity to extract the latent relationships between words and documents.
- Semantic text classification algorithms can generate meaningful keywords for the existing classes.
- Common text classification methods have poor capabilities in explaining to users why a certain result is achieved because traditional text classification algorithms cannot relate semantically to nearby terms. As well, they cannot explain how the result clusters are related to one another. But on the other side, the good news is that semantic text classification algorithms have the capability to locate the instances semantically, explain and analyze the classification results.
- Traditional text classification methods focus on only syntax that produces poor classification results. So, semantic understanding of text is necessary to improve progress of the efficiency and accuracy of classification.
- Synonymy is a term or phrase that means exactly or nearly the same as another word or phrase in the same language even though they are written differently. Polysemy is the ability for a term or phrase to have more than one meaning. Many languages have several synonyms. For instance "peak-summit", "minuscule-minute" pairs are synonyms in English. There are also many polysemous terms in English. For example, the verb "to get" can mean "procure", "understand" (I get it), etc. Traditional text classifiers cannot make use of semantic approaches and they only concentrate on syntax in a document. Thus, they ignore the semantic connections between words and documents and they evaluate a word as it is independent from its context. Conversely, semantic text classification algorithms have the opportunity to handle synonymy and polysemy better than traditional text classification algorithms since they take advantages of semantic connections between words. Consequently, semantic approaches make semantic classification algorithms assess and interpret a word within its context.

In machine learning applications, especially in the field of text classification there are two conventional strategies; supervised learning and unsupervised learning. A sufficient amount of labeled data is required as training corpus to build the classifier in conventional supervised classification methods, which will be helpful to guess the class labels of the unlabeled instances. Conversely, unsupervised learning, only depends on unlabeled instances, and doesn't require class labels to build a classifier so; they attempt to explore the latent composition of unlabeled data to train a model (Zhu, 2005). Unfortunately most of the huge amount of accumulated data on the web is unlabeled. This restrict their usage in numerous machine learning applications like speech recognition, sentiment recognition and text classification. Moreover, assigning labels to them manually is expensive, tedious and time-consuming. Furthermore, to train a classifier with very little labeled data possibly will not yield adequate classification accuracy. Semi-supervised Learning (SSL) algorithms take advantages of both labeled and unlabeled instances to improve the classification performance. A lot of SSL algorithms have been suggested in the former decades, like co-training (Blum and Mitchell, 1998), self-training (Rosenberg, 2005; Yarowsky, 1995), graph-based methods (Zhu, 2005), semi-supervised support vector machines (Zhu, 2005), Estimation-Maximization (EM) with generative mixture models (Nigam et al., 2000), transductive support vector machines (Chapelle and Zien, 2005).

It is known that Latent Semantic Indexing (LSI) utilizes latent higher-order structure between terms and documents (Kontostathis and Pottenger, 2006). Higher-order relations in LSI get "hidden semantics". The LSI algorithm (Deerwester, 1990) is a very popular and commonly-used technique in the fields of text mining and information retrieval. There are several LSI-based classifiers. For instance, in (Zelikovitz and Hirsh, 2004) the authors propose an LSI-based k-Nearest Neighborhood (LSI k-NN) algorithm in a semi-supervised setting for short text classification which is one of the simple uses of LSI in text classification. In this work, the authors use the k-Nearest Neighborhood (k-NN) algorithm that is based on calculating similarities or distance between training instances and a test instance in the transformed LSI space. They set the number of neighbors to 30 and use the noisy-or operator. A similar approach is used in a supervised setting to build an LSI-based k-NN algorithm as one of the baseline algorithms in (Ganiz et al., 2011). In this study, the number of neighbors is set to 25, and the dimension parameter (k) of the LSI algorithm is optimized.

In a recent study (Altınel et al., 2015), a novel supervised semantic smoothing kernel for SVM is offered: Class Meaning Kernel (CMK). CMK uses Helmholtz principle (Balinsky et al., 2010, 2011a, 2011b, 2011c) for smoothening a document's words in BOW demonstration. Evaluation of CMK on experimental data reveals significant improvement in classification accuracy over linear kernel. This is very important since linear kernel is a benchmark algorithm for text classification field.

Inspired by the benefits of CMK over linear kernel, and concentrated on the truth that there is inadequate labeled samples in actual world cases, a non-iterative semi-supervised version of CMK is built, which is named Instance Labeling Based on Meaning (ILBOM). This