Neurocomputing 000 (2018) 1–7

Contents lists available at ScienceDirect

Neurocomputing

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

Brief Papers

Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning

Han-joon Kim[a,*], Jiyun Kim[a], Jinseog Kim[b], Pureum Lim[a]

[a] School of Electrical and Computer Engineering, University of Seoul, Korea
[b] Department of Applied Statistics, Dongguk University, Korea

ARTICLE INFO

Article history:
Received 7 May 2017
Revised 27 March 2018
Accepted 3 July 2018
Available online xxx

Communicated by Dr. Tie-Yan Liu

Keywords:
Text classification
Naïve Bayes learning
Tensor space
Wikipedia
Semantic features

ABSTRACT

This paper suggests a novel way of dramatically improving the Naïve Bayes text classifier with our semantic tensor space model for document representation. In our work, we intend to achieve a perfect text classification with the semantic Naïve Bayes learning that incorporates the semantic concept features into term feature statistics; for this, the Naïve Bayes learning is semantically augmented under the tensor space model where the 'concept' space is regarded as an independent space equated with the 'term' and 'document' spaces, and it is produced with concept-level informative Wikipedia pages associated with a given document corpus. Through extensive experiments using three popular document corpora including *Reuters-21578, 20Newsgroups*, and *OHSUMED* corpora, we prove that the proposed method not only has superiority over the recent deep learning-based classification methods but also shows nearly perfect classification performance.

1. Introduction

Text classification is to automatically assign an unknown textual document to its appropriate one or more classes. Nowadays, the most popular approach towards text classification is to use machine learning techniques that inductively build a classification model of pre-defined classes from a training set of labeled documents. The machine learning methods for text classification include Naïve Bayes [1], *k*-nearest neighbors [2], decision trees [3], support vector machine (SVM) [4], and the recent deep learning methods such as convolutional neural network (CNN) [5] and recurrent neural network (RNN) [6]. In our work, we focus on improving the Naïve Bayes learning algorithm because it is simple yet accurate technique in spite of its wrong independence assumption. More importantly, the Naïve Bayes learning algorithm has a number of superior advantages compared with other learning algorithms in constructing the *operational* text classification system even though it is old.

Basically, machine learning algorithms for text classification should effectively deal with the *curse-of-dimensionality* problem since textual data have a huge number of term features. The Naïve Bayes algorithm is less sensitive than other learning algorithms in terms of overcoming the problem. Moreover, it is very easy to incrementally update its classification model due to its simplicity; when new documents are given as training data, the current term feature statistics are easily updated and additional feature evaluation is immediately carried out without re-processing the past training data. This characteristic is essential in the case where the document corpus is highly evolutionary. Besides, the Naïve Bayes learning does not require a complex generalization process unlike support vector machine, decision trees, and deep learning-based algorithms; it has only to calculate the feature statistics per class with a single pass over training documents.

Because of the above advantages, there have been many studies to improve the Naïve Bayes text classifier in different aspects, and one promising approach is to enrich the representation of textual documents with external or internal semantic features [7–10]. In this paper, we propose a semantic Naïve Bayes text classifier that is based upon a semantic tensor space model proposed in our previous research. In [11], we have proposed a document representation model conforming to the definition of the 'concept' in the formal concept analysis framework [12]. The model represents a single document as not a vector but a matrix (i.e., 2nd-order tensor) that reflects the relationship between term features and semantic features within a document. To realize this semantically enriched text model, we employ the Wikipedia encyclopedia as an external knowledge source, and a concept-level informative Wikipedia page is defined as a 'single' semantic concept.

The rest of the paper is organized as follows. Section 2 describes the related work. In Section 3, we discusses the traditional

* Corresponding author.
E-mail address: khj@uos.ac.kr (H.-j. Kim).

https://doi.org/10.1016/j.neucom.2018.07.002
0925-2312/

Naïve Bayes learning framework for text classification. Section 4 introduce the proposed semantic Naïve Bayes text classification with a tensor based document representation model. Section 5 describes our experimental setup and results. Lastly, we conclude our paper in Section 6.

2. Related work

2.1. Semantic Naïve Bayes classification

The traditional Naïve Bayes text classifier is learned as a generative model that fits the distribution of the document instances given a class label, and it makes a strong assumption that the term features within a document are conditionally independent given a class label. As mentioned earlier, it can be improved by enriching document representation with external (or internal) semantic features since the semantic features can mitigate the strong independence assumption. As good quality external knowledge such as Wikipedia (http://en.wikipedia.org/), WordNet (http://wordnet.princeton.edu/), and Minimal Recursion Semantics [13] have been built, these semantic resources have been thus utilized to enhance the Naïve Bayes algorithm [9,10,14,15]. To enhance the text classifier, the semantic features associated with terms occurring in documents are extracted from the external knowledge, which are used to augment the initial set of features. In addition, without the help of external knowledge, internal (or latent) semantic features can be derived through singular value decomposition (SVD) [7]. Also, in [8] as a study most similar to our work, Jing et al. proposed a semantic Naïve Bayes classification method that incorporates inherent semantic information, which is obtained by applying latent topic models from training documents without external knowledge. The problem of these related studies is that new semantic features are only added without being distinguished from original term features, and the learning framework itself has not changed.

If we are to include the semantic features to improve the text classifier, then considering the dependence between the term and the concept would contribute significantly to its performance. In our work, the traditional Naïve Bayes learning framework is enhanced so as to reflect the dependence between term and semantic concept, and to estimate the degree of dependence, our previous tensor-based document representation using Wikipedia [11] is effectively used. In [11], we attempted to improve the accuracy of text clustering by suggesting a Frobenius norm-based similarity function between documents expressed in a *term-by-concept* matrix without probabilistically revealing dependencies between terms and concepts.

2.2. Document representation for text classification

In terms of document representation, the Naïve Bayes classifier uses the Bag of Word (BoW) model that only considers frequencies of terms occurring in a class [15]. The key to improving the classifier is to solve the 'loss of term senses' problem of the BoW model and incorporate the enhanced representation model into the Naïve Bayes learning framework. In our work, text classification is done based on a semantic tensor space model, in which a document is represented by a *term-by-concept* matrix (i.e., 2nd-order tensor), and a document corpus is thus represented as a 3rd-order *document-by-term-by-concept* tensor (See Fig. 1 (c)) [11]. The 'concept' space is regarded as an independent space equated with the 'term' and 'document' spaces, and it is produced with concept-level informative Wikipedia pages associated with a given document corpus where a Wikipedia page is defined as a single concept. The important thing is that the semantic features should encompass the correct meanings of terms in a document to improve the Naïve Bayes classifier.

Actually, our document representation model is related to the studies on mapping documents or terms onto a concept space. As such an earlier approach, latent semantic indexing, which is a variant of classical vector space, attempted to produce a concept space for document indexing by capturing the latent concepts hidden in documents [7]. During a past decade, several studies on deriving correct meanings of words through Wikipedia pages have significantly contributed to improve text mining algorithm [16–19]. In [17], to semantically represent document with Wikipedia pages, significant terms in the document are identified and their meaning are represented in terms of Wikipedia-based concepts. Furthermore, Boubacar and Niu [16,18] attempted to improve the performance of text clustering by enriching document representation with concept-level Wikipedia pages. Similarly, Wang et al. [19] proposed a way of improving the performance of text classification by expanding the vector space model with semantic relations such as synonymy, hyponymy, and associative relations derived from Wikipedia. Most of the related studies is to express a document itself as a Wikipedia-based concept vector or to do a simple combination of a word vector and a concept vector. In contrast, our approach is to generate a Wikipedia-based concept vector at the individual term level, resulting in a single document being represented as a *term-by-concept* matrix. This matrix representation contains a dependency between the term and the concept, which can help to greatly improve the Naïve Bayes algorithm. Previous studies did not take into account the dependency of Wikipedia-based concepts and words. Also, with the matrix representation, a document can be expressed as a 1st-order tensor (i.e., a vector) by summing all the components of each row or column; in other words, a document can be represented by a concept vector as well as a term vector. In this context, we can say that our representation model is a generalization of the models suggested in the other studies such as [16–19].

3. Naïve Bayes learning framework

The Naïve Bayes (NB) text classifier produces its classification model as a result of learning (estimation) process based on the Naïve Bayes learning algorithm which belongs to a family of probabilistic classifiers based on the Bayes theorem. The estimated classification model consists of two kinds of parameters: the term probability estimates $\hat{\theta}_{t|c}$, and the class prior probabilities $\hat{\theta}_c$; that is, the classification model $\hat{\theta}_{NB} = (\hat{\theta}_{t|c}, \hat{\theta}_c)$. Each parameter can be estimated according to maximum a posteriori (MAP) estimation. For classifying a given document, Naïve Bayes learning system estimates the posterior probability of each class via Bayes rule; that is, $Pr(c|d) = \frac{Pr(c)\cdot Pr(d|c)}{Pr(d)}$, where $Pr(c|d)$ is the probability that a document d belongs to the class c in a set of classes C, $Pr(c)$ is the class prior probability that any random document from the document corpus belongs to the class c, $Pr(d|c)$ is the probability that a randomly chosen document from documents in the class c is the document d, and $Pr(d)$ is the probability that a randomly chosen document from the whole corpus is the document d.[1] The document d is then assigned to a class $argmax_{c\in C} Pr(c|d)$ $(=\Phi_{\hat{\theta}_{NB}}(d))$ with the highest posterior probability. Here, in the context of the Naïve Bayes, the document d is represented by a bag of words $(t_1, t_2, \ldots, t_{|d|})$, where multiple occurrences of words are preserved. Moreover, the Naïve Bayes assumes that the terms in a document are mutually independent and the probability of term occurrence is independent of position within the document given a class. This assumption allows simplifying the classification function

[1] Throughout the paper, we use the standard notational shorthand for random variables; that is, $Pr(X = x)$ is simply written as $Pr(x)$.

Please cite this article as: H.-j. Kim et al., Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.07.002