Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **2**, Number **1**, January **2017**

doi:10.3934/bdia.2017009
pp. **69–75**

MULTIPLE-INSTANCE LEARNING FOR TEXT CATEGORIZATION BASED ON SEMANTIC REPRESENTATION

Jian-Bing Zhang*, Yi-Xin Sun and De-Chuan Zhan

National Key Laboratory for Novel Software Technology
Nanjing University, China

Abstract. Text categorization is the fundamental bricks of other related researches in NLP. Up to now, researchers have proposed many effective text categorization methods and gained well performance. However, these methods are generally based on the raw features or low level features, e.g., tf or tfidf, while neglecting the semantic structures between words. Complex semantic information can influence the precision of text categorization. In this paper, we propose a new method to handle the semantic correlations between different words and text features from the representations and the learning schemes. We represent the document as multiple instances based on word2vec. Experiments validate the effectiveness of proposed method compared with those state-of-the-art text categorization methods.

1. **Introduction.** Nowadays, with the rapid growth of information on the Internet, it has become more and more difficult for us to search the specific information we need. Since most of the informations are text information, text categorization has become one of the key techniques for handling and organizing these information. Text categorization is a basic NLP task. It assigns a document to one or more predefined categories. Researchers had proposed many approaches to deal with text categorization and most of these approaches are relying on a simple document representation in a word-based input space.

Traditional word-based document representations such as the well known VSM and LSA are widely used for extracting representative features of documents. But these traditional methods are all based on the frequency of words or tf-idf weighting. They disregard semantic information of words. While the semantics of document also plays a very important role in text categorization. Besides, with increase of content of documents, the dimension of vectors will increase quickly and lead to high computational-complexity.

In this paper, we propose a new method to represent a document as several vectors based on the word2vec, and use Multiple-Instance learning model for the final categorization operation[1]. Word2vec, introduced by Mikolov, is an efficient method for learning highquality vector representations of words from large amounts of unstructured text data. The learned vectors explicitly encode many linguistic regularities and patterns. So it is considered as a perfect estimation of word representations in vector space. It can also achieve large improvements in accuracy at much lower computational cost. Multiple-Instance learning is an extension of the

2010 *Mathematics Subject Classification.* 97R40.

Key words and phrases. Text categorization, text representation, Multiple-Instance learning, mi-SVM, word2vec.

standard supervised learning settings. Multiple-Instance learning was coined by Dietterich et al. in the context of drug activity prediction[5]. In Multiple-Instance learning, the train set is composed by bags and each bag consists of one or more instances. As we know, an article often contains several subtopics and each paragraph will represent relatively independent subtopic. But not all these subtopics are consistent with the topic of categorization, so these uncorrelated subtopics may be the noise for text categorization. But the computer can't distinguish whether the subtopic of the paragraph in the article is helpful to categorization beforehand. Fortunately, Multiple-Instance learning is exactly proposed under such situation. So in our method, we represent each paragraph as a vector and each document is represented as several vectors. Then we use mi-SVM for categorization.

The rest of this paper is organized as follows. In Section 2, we will discuss word2vec and Multiple-Instances learning. In Section 3, we will introduce the detail of our method for text categorization and the next section is our experiment. Finally, the conclusion will be shown in Section 5.

2. **Related work.** In this paper, we propose a new method for text categorization. In our method, we represent a document as several vector with the same dimension. Then, we put these vectors into the Multiple-Instance learning model to train a classifier and test the precision of the classifier.

Text categorization has experienced a long research history. But the task is still facing some challenges. So far, researchers have proposed so many methods for text representation such as n-gram, TF-IDF, LSA etc [3, 8, 7] and text categorization such as Decision Trees, SVM, Bayesian Classifiers and Neural Network Classifiers[11, 6, 9]. All of these methods are able to obtain well performance in text categorization. But, with the rapid growth of text on the Internet, if we still use the traditional representation method, the dimension of the vector will be very large. Fortunately, word2vec, an open source tool provided by Google, can help us solve the problem of dimension disaster. Word2vec can represent word or phrase as a vector effectively by predefining the dimension of the vector[12]. Besides, word2vec can represent the semantic meaning of words or phrases which appear in documents. So in our methods, instead of using word-based representation methods such as TFIDF, n-gram etc, we use word2vec to representation documents.

Multiple instance learning was first proposed by Dietterich et al. to deal with the problem of drug activity prediction[5]. Up to new researchers have done many researches on Multiple-Instance learning and proposed many algorithms such as Diverse Density[10], Citation-kNN[14], ID3-MI[4], BP-MIP[15][16], MI SVMs[2] to solve problems about Multiple-Instance learning and gain well performance in related fields. However, there is little application of Multiple-Instance learning in text categorization. Nowadays, the scale of text data is becoming bigger and bigger and the semantic expression in the text becomes more and more complex. If we can separate the relevant and irrelevant contents according to the topic of categorization in the text and take advantage of these relevant contents, we will achieve better performance on text categorization.

3. **Our method.** In our method, we first represent each word as a K-dimensional digital vector, then, we will represent the document as several vectors. After that, we use mi-SVM to test the performance of our method.

3.1. **Document segmentation.** According to the assumption of Multiple-Instance learning, we can know that the training set comprises labeled bags that are composed of unlabeled instances and our task is to predict the label of unseen bags[5].

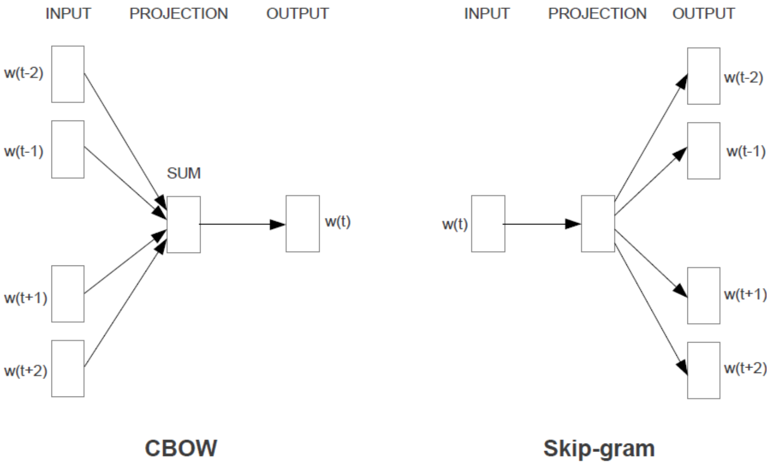


Figure 1. The structure of Bag-of-Words and Skip-Gram

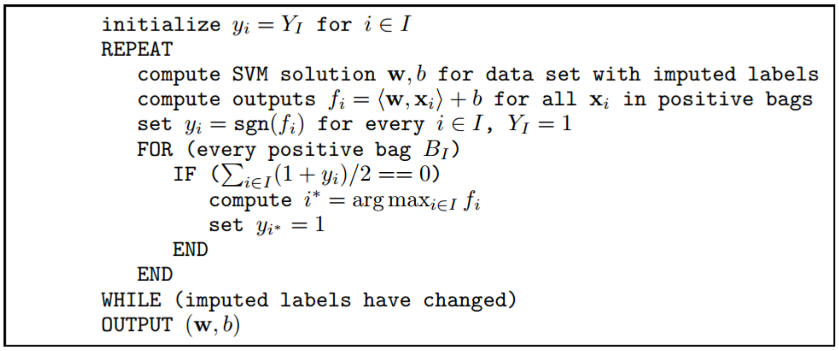
So, in our task, we should take some methods to represent each document in the same form as bag so that we can predict the category of a document by predicting the label of the bag corresponding to it. In order to represent the document as a bag, the most important thing is document segmentation. By dividing the document into several segments, our goal is to separate these contents that are relevant to the topic of categorization from these irrelevant contents.

In this paper, we choose to segment the text by paragraph, which is more in line with human writing habit. Depending on the experience of writing, we can know that each paragraph in the text will express relatively consistent topic and the topics of different paragraph are relatively independent to some extent. So we think that segmenting the text based on the paragraph can be a simple but effective way.

3.2. **Document representation.** Text is a more abstract way of expression, so we want to classify the text in a way that is more similar to human thinking rather than only using word-based features. Traditional word-based text representation only reflects the features of word distribution, it overlooks the semantic features of text. But the semantic features may be more important in text categorization.

In this paper, we choose to use word2vec to represent the text. Word2vec has two novel model architectures for computing continuous vector representations of words from very large datasets[13]. The first architecture is Continuous Bag-of-Words Model(CBOW) which predicts the current word based on the context and the other one is Continuous Skip-gram Model which predicts surrounding words given the current word. Word2vec simplifies the context processing to vector processing in a K-dimensional vector space. Structures of CBOW model and Continuous Skip-gram Model are shown in the Fig1.

We choose the CBOW architecture to train the word embedding and the dimension we set for the vector is 300. After getting the pre-trained word embedding, we can represent each word in the text as a 300 dimension vector and we can get the representation of each paragraph by adding all the vectors which its related



```
initialize $y_i = Y_I$ for $i \in I$
REPEAT
    compute SVM solution $\mathbf{w}, b$ for data set with imputed labels
    compute outputs $f_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b$ for all $\mathbf{x}_i$ in positive bags
    set $y_i = \text{sgn}(f_i)$ for every $i \in I$, $Y_I = 1$
    FOR (every positive bag $B_I$)
        IF ($\sum_{i \in I}(1 + y_i)/2 == 0$)
            compute $i^* = \arg\max_{i \in I} f_i$
            set $y_{i^*} = 1$
        END
    END
WHILE (imputed labels have changed)
OUTPUT $(\mathbf{w}, b)$
```

Figure 2. Pseudo-code for mi-SVM

word appears in the paragraph. Then, we can represent each paragraph as a vector and these vectors are called "instance" mentioned above. So in our method, each document will be represented as several vectors.

3.3. **Document categorization.** In our method, we represent each document as several instances(vector) and these instances have no labels, only the bag(document) has label. However, traditional supervised methods require each instance must have a label, so we can't use these traditional methods to train the model. $mi - SVM$ proposed in [2] is an extension of the Support Vector Machine(SVM).In Multiple-Instance learning, we have an assumption that an example is positive if and only if one or more of its instances are positive. According to the assumption of Multiple-Instance learning and the describing in [2], we can describe the relation between instance labels y_i and bag labels Y_I as following form:

$$\sum_{i \in I} \frac{y_i + 1}{2} \geq 1, \forall I \quad \text{s.t.} \quad Y_I = 1, \text{and} \quad y_i = -1, \forall I \quad \text{s.t.} \quad Y_I = -1 \tag{1}$$

In our method, we will still use this assumption. And according to this assumption, we can formulate our optimization goal as follows:

$$\begin{aligned} &\min_{\{y_i\}} \min_{\omega, b, \varepsilon} \frac{1}{2}|\omega|^2 + C \sum_i \varepsilon_i \\ &\text{s.t.} \quad \forall i : y_i(\langle \omega, x_i \rangle + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, y_i \in \{-1, 1\}, \text{ and (1) hold} \end{aligned} \tag{2}$$

Based on this optimization goal, we are looking for a separating linear discriminant such that there is at least one instance from every positive bag in the positive halfspace, while all instances belonging to negative bags are in the negative halfspace. So in our task, we are looking for a classification hyperplane such that if a document belongs to label, at least one paragraph in this document can be classified as this label.The pseudo-code of mi-SVM is given in Fig2[2].

4. **Experiments.** In this section, we will introduce in detail about our experiments.

4.1. **Datasets.** In order to evaluate the performance of our method, we choose the sougouC corpus and 20newsgroup as the dataset in our experiments. SougouC corpus is a chinese corpus and 20newsgroup is an english corpus. We choose 5 categorys of SougouC as dataset and the dataset contains 40000 documents marked

Table 1. Results of experiments on sougouC

Model	car	finance	IT	health	sport
SVM + TF-IDF	0.8473	0.8420	0.8363	0.8326	0.8737
SVM + Word2vec	0.9303	0.8571	0.8755	0.9163	0.9828
mi-SVM + Word2vec	**0.9599**	**0.8904**	**0.8943**	**0.9325**	**0.9842**

with 5 categorys. These 5 categorys are car, finance, IT, health and sport. Each category contains 8000 documents. 20newsgroup corpus contains 18828 documents labeled with 20 labels. We use 10-fold cross-validation to test the precision of method on both chinese and english corpora. These two corpora are enough to evaluate the performance and obtain objective results.

4.2. **Designs of experiments.** In this section, we will introduce the detail of our experimental design. In order to use sougouC corpus and 20newsgroup corpus to train word2vec, we should first deal with the problem of word segmentation. In our experiment, we use IKAnalyzer, a word segmentation toolkit in java, to deal with the chinese word segmentation and use lucene to deal with the english word segmentation. In the process of word segmentation, we get rid of those stop words. After that, we organize these words in one document and all of words in the same document are in the same line and the order of the words are not changed. Then we put the document into word2vec toolkit to gain the map of words and vectors. In this step, we use the Continuous Bag-of-words architecture. We set the dimension of a vector as 300 and set the size of the window as 8. After about 8 hours, we can get the map of words and vectors. In order to compare our method with the traditional method, we should represent each document in corpus as one instance and multiple instances. In our experiment, we just add all the vectors of words in the same group to gain the representation vector. For the traditional method, a document is not segmented to several segments so that each document is represented as one vector. While in our method, each document will be divided into several segments, so, the document will be represented as multiple instances.

To verify the performance of our method, we design two group text categorization experiments on Chinese and english corpora. In each experiment group, we set three groups of experiments. For traditional methods, we choose tf-idf and SVM for experiments and the result of this experiment is seen as our baseline. In order to compare performance of semantic representation, we repeat the experiment mentioned above, but instead of using tf-idf, we use word2vec to represent the text. For our method, we separate each document into several segments and use word2vec to represent these segments, then, mi-SVM will be used for text categorization. All of these experiments are using the strategy of 10-fold cross-validation to get the ultimate performance of each method. The results of traditional methods and our method are presented in the next section.

4.3. **Results.** In our experiments, we contrast traditional methods and our method. The experimental results are shown in Table 1 and Table 2.

Compared with pictures, text is a more abstract way of information expression for human. So, the more important attribute of text may be semantics rather than word frequency, tf-idf, etc. and classifying the text based on the semantics is more similar to human thinking. For these reasons, we choose to use the word embedding

Table 2. Results of experiments on 20newsgroup

Model	SVM+tf-idf	SVM+Word2vec	mi-SVM+Word2vec
Average	0.8508	0.8421	**0.8619**

to represent text and verifying the performance of it through the experiment. From the result of experiment, we can see that for Chinese corpus, the performances of using word embedding on five categories are better than using tfidf. If we look at the content of the document in Chinese corpus, we can also judge the category of the document very easily. But for 20newsgroup, the performance of word embedding doesn't outperform tfidf, to find the cause of this situation, we choose some documents in 20newsgroup randomly to check their contents. We find that every document has headers and most of the documents contain lots of content which are unrelated to category and we can see these contents as noises. All of these noises will influence the performance of categorization. The results of our method also support our opinion.

When we compare our method to traditional methods, the results of experiments show that our method outperforms the traditional methods for most of the labels. It shows that, in a certain degree, our method can separate the useful information and noise. According to the assumption of Multiple-Instance learning, if one or more of document's instances are positive, we will label the document as positive. So if document's useful information and noise are separated, we will get a more accurate result. For example, if a document is represented as 5 instances and 4 of them are noises. In our method, classifier can still judge which label the document belongs to accurately according the remaining instance. But if we encode these information into only one instance, it will be difficult for classifier to judge document's label because most of the informations are noises.

5. **Conclusion.** In this paper, we propose a new method to represent a document as several vectors and use Multiple-Instance learning method to get the accuracy of categorization and evaluate the performance of categorization. Our datasets are chinese corpus called sougouC and english corpus called 20newsgroup. We use word2vec to get the document represents vectors. As we can see in our experiment, our method can get higher accuracy in document categorization than other traditional method such as SVM. But, there are also some shortcomings in our method. For example, the time we cost in training the mi-SVM model is too long. Besides, the accuracy of text categorization is influenced by the result of word segmentation.

In the future, we will do more experiments on more datasets. And we should also improve the method we represent the document rather than simply separate the document by paragraph.

REFERENCES

[1] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, *Artificial Intelligence*, **201** (2013), 81–105.

[2] S. Andrews, I. Tsochantaridis and T. Hofmann, Support vector machines for multiple-instance learning, *Advances in Neural Information Processing Systems*, **15** (2002), 561–568.

[3] W. B. Cavnar, J. M. Trenkle, et al., N-gram-based text categorization, *Ann Arbor MI*, **48113** (1994), 161–175.

[4] Y. Chevaleyre and J. D. Zucker, Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. application to the mutagenesis problem, In *Biennial*

Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, (2001), 204–214.

[5] T. G. Dietterich, R. H. Lathrop and T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, *Artificial Intelligence*, **89** (1997), 31–71.

[6] S. Dumais, Using svms for text categorization, *IEEE Expert*, **13** (1998), 21–23.

[7] N. Ishii, T. Murai, T. Yamada and Y. Bao, Text classification by combining grouping, lsa and knn, In *Ieee/acis International Conference on Computer and Information Science and Ieee/acis International Workshop on Component-Based Software Engineering,software Architecture and Reuse*, (2006), 148–154.

[8] Q. Kuang and X. Xu, Improvement and application of tfidf method based on text classification, In *International Conference on Internet Technology and Applications*, (2010), 1–4.

[9] S. Lai, L. Xu, K. Liu and J. Zhao, Recurrent convolutional neural networks for text classification, In *AAAI*, (2015), 2267–2273.

[10] O. Maron and T. Lozano-Pérez, A framework for multiple-instance learning, *Advances in Neural Information Processing Systems*, **200** (1998), 570–576.

[11] A. Mccallum and K. Nigam, A comparison of event models for naive bayes text classification, *In AAAI-98 Workshop On Learning For Text Categorization*, **62** (2009), 41–48.

[12] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, Computer Science, 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, **26** (2013), 3111–3119.

[14] J. Wang and J. D. Zucker, Solving multiple-instance problem: A lazy learning approach, *Proc.international Conf.on Machine Learning*, (2000), 1119–1126.

[15] M. L. Zhang and Z. H. Zhou, Improve multi-instance neural networks through feature selection, *Neural Processing Letters*, **19** (2004), 1–10.

[16] Z. H. Zhou and M. L. Zhang, Neural networks for multi-instance learning, In *International Conference on Intelligent Information Technology*, 2002.

E-mail address: zjb@nju.edu.cn
E-mail address: sunyx@nlp.nju.edu.cn
E-mail address: zhandc@nju.edu.cn