Weighted Kernel Model for Text Categorization

Lei Zhang **Debbie Zhang** **Simeon J. Simoff** **John Debenham**

Faculty of Information Technology
University of Technology, Sydney
PO Box 123 Broadway NSW 2007 Australia
Email{leizhang, debbiez, simeon, debenham}@it.uts.edu.au

Abstract

Traditional bag-of-words model and recent word-sequence kernel are two well-known techniques in the field of text categorization. Bag-of-words representation neglects the word order, which could result in less computation accuracy for some types of documents. Word-sequence kernel takes into account word order, but does not include all information of the word frequency. A weighted kernel model that combines these two models was proposed by the authors [1]. This paper is focused on the optimization of the weighting parameters, which are functions of word frequency. Experiments have been conducted with Reuter's database and show that the new weighted kernel achieves better classification accuracy.

Keywords: Bag-of-words Kernel, Word-sequence Kernel, Weighted Kernel Model, Text Categorization

1 Introduction

Text categorization is the task of assigning documents into predefined categories (classes), specified by their topics. For example, the documents might be news items and the classes might be national news, sports news and business news. Documents are classified based on their content[2]. As documents are characterized by the words that appear in each document, they are firstly transformed into a representation that is suitable for the classification task. Then learning algorithms are applied to perform the classification task. Automated text categorization has been successfully demonstrated in many applications [3] [4].

One of the widely used representation of documents is known as the bag-of-words model [5], which is a set of words contained in the documents. Bag-of-words is based on both frequency of a word in a document and the corpus. However, bag-of-words model has many shortcomings. Particularly, it ignores both syntax and semantics of the documents. Without considering the word position, the information of the sequence of words is lost.

Lodhi proposed the use of string kernels [6], which was the first significant departure from the bag-of-words model. In string kernels, the features are not word frequencies or related expansions, but the extent to which all possible ordered subsequences of characters are presented in the document. Cancedda [7] proposed the use of string kernel with sequences of words rather than characters, known as the word-sequence kernel. The word-sequence kernel has several advantages, in particular it is more computationally efficient and it ties in closely with standard linguistic pre-processing techniques. Although word-sequence kernel takes into account word positions, it does not include the information about word frequency. This issue will be discussed further in Section 3.

To make both the word frequency and position information available for the learning algorithms, a combined weighted kernel model was proposed [1]. However, not every combination of the bag-of-words approach and word-sequence kernel approach will result in improved computational accuracy. This is because these two kernels, which represent the similarity between documents respectively, have different contribution to construct the new kernel. Moreover simply combining these two kernels do not satisfy valid kernel conditions [8].

This paper is based on our previous work of combining kernels. It emphasizes on the optimization of the weighting parameters, which is critical to the categorization accuracy. The rest of this paper is organized as follows: In Section 2, the basic idea of bag-of-words kernel is reviewed. The word-sequence kernel and issues of this kernel are presented in Section 3. The detail implementation of proposed new kernel model and algorithm for determining the weighting parameter are described in Section 4. Section 5 presents the experimental results using the Reuters data set, followed by the conclusions in Section 6.

2 Bag-of-words Kernel

Bag-of-words model is the traditional approach for representing a document as a term vector. A bag is a set of a dictionary. As repeated elements are allowed, this representation takes into account not only the presence of a word but also its frequency [9]. A document is represented by a row vector

$$\phi(d) = [tf(t_1, d), tf(t_2, d), ..., tf(t_N, d)] \in \mathbb{R}^N \quad (1)$$

where $tf(t_i, d)$ is the frequency of the term t_i in the document d. Hence, a document is mapped into a space of dimensionality N being the size of the dictionary. Each entry records how many times a particular term is used in the document. The vector space kernel or bag-of-words kernel is given by the following definition

$$k(d_1, d_2) = \langle\phi(d_1), \phi(d_2)\rangle = \sum_{j=1}^{N} tf(t_j, d_1)tf(t_j, d_2) \quad (2)$$

The value of N in equation 1 is related to the length of the documents. Excessive number of irrelevant words and terms will not only increase the computational cost but also decrease the accuracy of the classification. Therefore the most frequent words and

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at Australiasian Data Mining Conference (AusDM 2006), Sydney, December 2006. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 61. Peter Christen, Paul Kennedy, Jiuyong Li, Simeon Simoff and Graham Williams, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

terms are usually selected in order to construct the bag-of-words kernel. As mentioned in Section 1, this technique only takes into account the word frequencies, ignoring the information on word positions. In many language modeling applications, such as speech recognition and short message classification, word order is extremely important. Furthermore, it is likely that word order can assist in topic inference. For example, consider the following two sentences

“The interest rate goes up, US dollar goes down.”
“The interest rate goes down, US dollar goes up.”

These two sentences [10] have exactly the same words and word frequencies. It is the different word order that results in opposite meanings, which cannot be distinguished by the bag-of-words method. Therefore the string kernel and word-sequence kernel were introduced to tackle with the word order issue.

3 Word-sequence Kernels

In string kernels, the features are not word frequencies. The document is represented by all possible ordered subsequences of characters. However, this method is computationally expensive for long documents. More recently, Cancedda et al. extended the string kernel to word-sequence kernel, where all possible sequences of words are used instead of sequences of characters. This novel way to compute the document similarity based on matching subsequence has outperformed the string kernel in many applications [11] [12].

Following the definition of Lodhi [6], let Σ be a finite alphabet set. A string is a finite sequence of characters from Σ, including the empty sequence. For strings s and t, l_s denotes the length of the string s, where $s = s_1 ... s_{l_s}$. The string $s[i : j]$ is the substring $s_i ... s_j$ of s. u is a subsequence of s, if there exist indices $\mathbf{i} = (i_1, ..., i_{l_u})$, with $1 \leq i_1 < ... < i_{l_u} \leq l$, such that u = $s[\mathbf{i}]$. The length l($\mathbf{i}$) of the subsequence in s is $i_{l_u} - i_1$ +1. Σ^n denotes the set of all finite strings of length n, and Σ^* is the set of all strings $\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$. The feature mapping ϕ for a string s is given by defining the u coordinate $\phi_u(s)$ for each $u \in \Sigma^n$.

$$\phi_u(s) = \sum_{\mathbf{i}:u=s[\mathbf{i}]} \lambda^{l(\mathbf{i})} \qquad (3)$$

where λ is the decay factor. These features measure the number of occurrences of subsequences in the string s weighting them according to their lengths. Hence, the inner product of the feature vectors for two strings s and t gives a sum over all common subsequences weighted according to their frequency of occurrence and lengths

$$K_n(s,t) = \sum_{u \in \Sigma^n} \langle \phi_u(s) \cdot \phi_u(t) \rangle = \sum_{u \in \Sigma^n} \sum_{\mathbf{i}:u=s[\mathbf{i}]} \sum_{\mathbf{j}:u=t[\mathbf{j}]} \lambda^{l(\mathbf{i})+l(\mathbf{j})} \qquad (4)$$

Following the example of Lodhi, we examine different values of λ when n = 2, 3. We can compute the similarity between the following two sentences:

K(”science is organized knowledge”, ”wisdom is organized life”)

The similarity with $\lambda = 0.5$ was calculated by Lodhi, which were 0.580 when n equals to 2 and 0.478 when n equals to 3. We examine this algorithm by choosing different values of the λ as shown in the following table.

We could conclude that the string kernel algorithm is not much influenced by the value of λ. Therefore, the median value 0.5 was chosen for λ for the rest of experiments in this paper.

	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
kernel (n = 2)	0.557	0.580	0.620
kernel (n = 3)	0.483	0.478	0.483

Table 1: Result by using string kernel

	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
BOW	1	1	1
n=2	0.874	0.837	0.825
n=3	0.492	0.487	0.558
BOW + n=2	0.999	0.992	0.994
BOW + n=3	0.999	0.993	0.878
BOW + n=2 + n=3	0.999	0.985	0.865

Table 2: Result by simply combination of two kernels

However string kernel and word-sequence kernel do not include the information of word frequency. Considering the term ”kernel methods” and ”kernel model”, if we choose word-sequence kernel with $n = 2$, the similarity is zero. However, in many situation, these two terms are considered the same.

4 Weighted and Combined Kernel Model

Bag-of-words representation could be considered as a special case of word-sequence kernel. Consider the length $l = 0$, $\lambda^l = \lambda^0 = 1$. When $l = 0$, word-sequence kernel only contains the information of word itself, which is essentially the representation of bag-of-word.

However, the simply combination of $l = 0$ with word-sequence kernels is not a feasible approach. Let us consider the interest rate example in Section 2. We examine the bag-of-words and word-sequence kernels with n=2, 3, and combined n = 2, 3 with bag-of-word. The similarity of these two sentences is shown in the following table.

Simply combination of bag-of-word and word-sequence kernel will result in less computational accuracy as the entries of the element of these two kernel are in different scale. For example, when λ equals 0.5 and n equals 3, the similarity is 0.487, which makes sense while 0.99 is not properly. That is because the two sentences in the ”interest rate” example are two related sentences, but the meaning is different. Given a high value near one is not properly, neither given a zero value. Therefore a value in the middle is very ideal. This is also where the idea come of choosing the parameter based on the word frequency information.

Moreover simply combine these two kernel wouldn’t result in a new valid kernel, because the similarity of the same document may be greater than one.

Here we propose an approach to combine the word-sequence kernel and bag-of-word kernel.

$$K_{combined} = (1-\lambda)*K_{BOW}+\lambda*K_{Word-sequence} \qquad (5)$$

The λ in the above formula is no longer the decay factor in Lodhi’s word-sequence kernel. The parameter λ is now for balancing the contribution of the word frequency and word order information. It may be fixed, or determined from the data. Fixed parameters are more likely rely on the domain knowledge, which is different according to different projects. In this paper, we propose a technique to determine the parameter from the data by using the word frequency information.

Since the common understanding of a binary classification problem, the more words occurs in both

	BOW + n=2	BOW + n=3	BOW + n=2,3
$\lambda = 0.5$	0.862	0.695	0.792

Table 3: Result by using new combined kernel

Frequent Word	BOW	Word-sequence	Weighted Kernel
50	80 %	80%	85 %
100	80 %	80%	85 %

Table 4: Results on C15 and C22 data based on 100 Documents

Frequent Word	BOW	Word-sequence	Weighted Kernel
50	70%	60%	80 %
100	75%	60%	85 %

Table 5: Results on C21 and C22 data based on 100 Documents

classes, the less important the bag-of-words is. The ideal situation for bag-of-words would be no words occurs in two classes. In such a situation there is no need for word order information, and bag-of-words itself would accurately classify these two classes. And only the situation of many words occurs in both classes, we need give more attention to the word order. Therefore the parameter is determined by how many words occurs in both classes defined as follows

$$\lambda = n/N \tag{6}$$

where n is the number of words occurs in both classes for two classes classification or all classes for multi-classification. N_i is the sum of words in class i, and N is the average number of words of all classes defined by $N = (\sum_{i=1}^{c} N_i)/c$, where c is the number of classes. This new approach contain both word information and word sequence information, and does not require switching between bag-of-word kernel and word-sequence kernel. Compute the interest rate example by using the new kernel, we have the result as shown in Table 3.

There may be other techniques to determine the value of the parameter. The parameter could also be influenced by the domain knowledge by human beings. In this paper, the proposed algorithm for λ has been demonstrated successful in the example and the experiment presented in the next section.

5 Experiment

Experimental studies have been carried out to compare the performance of bag-of-words kernel, word-sequence kernel and proposed weighted kernel approach. The Reuters News Data Sets, which are frequently used as benchmarks for classification algorithms, was used in this paper for the experiments. The Reuters 21578 collection is a set of 21,578 short (average 200 words in length) news items, largely financially related, that have been pre-classified manually into 118 categories.

The experiments were conducted using 100 documents from three news group: C15 (performance group), C22 (new products/services group) and C21 (products/services group). The first set of experiments used C15 and C22 data, while the second set of experiments used C21 and C22. The second set of data is more difficult to classify than the first set since data sets C21 and C22 are closely related. This is confirmed by the experimental results by using the bag-of-word kernel and word-sequence kernel separately. However, as shown in Tables 4 and 5, the combined and weighted kernel achieves similar results.

50 and 100 frequent keywords were chosen for the bag-of-word kernel. For the word-sequence kernel, frequent words sequences were used in instead of a full list of words. The first column shows the number of selected frequent words. The second column shows the bag-of-word classification result. The third column shows the result of simple combination of the bag-of-word kernel and word-sequence kernel. The last column shows the result of the proposed combined and weighted kernel approach presented in section 4.

The above results show that classification based on bag-of-word model is better than word-sequence kernel in C21 and C22 group. This implies that the word-sequence kernel does not include the bag-of-word information. Although there are many identical keywords in C21 and C22, there is little information on the keyword sequence. Because keywords are not always occur in the same order in a sentence. Therefore word-sequence kernel alone does not reveal any feature representing the documents.

The bag-of-word kernel works better for the C15 and C22 data sets. This is because these two groups are not very close and have not many keywords in common. While the simple combination of the above two approaches results in poorer accuracy in all experiments, the proposed new kernel produces far better results.

6 Conclusion

Bag-of-words model and word-sequence kernel are two important techniques applied in the field of text categorization. Combining word frequency and word order is taking the advantage of both bag-of-words kernel and word-sequence kernel. The parameter based on the word frequency information makes the weighted kernel valid and high computational accuracy. Experiments was conducted with Reuter's database and show the new weighted kernel achieves better classification accuracy.

References

[1] Zhang, L., Zhang, D., Simoff, J.S.: Combined kernel approach for text categrization (submitted). In: the 19th ACS Australian Joint Conference on Artificial Intelligence, Sydney, Australia (2006)

[2] Sebastiani, F.: Machine learning in automated text categorisation. ACM Computing Surveys **34** (2002) 1–47

[3] Amasyali, M., Yildirim, T.: Automatic text categorization of news articles. In: Signal Processing and Communications Applications Conference, 2004. Proceedings of the IEEE 12th, Turkish (2004) 224 – 226

[4] Basu, A., Walters, C., Shepherd, M.: Support vector machines for text categorization. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Hawaii (2003)

[5] Shawe-Taylor, J.: Kernel Methods for Pattern Analysis. University of Cambridge, Cambridge (2004)

[6] Lodhi, H.: Text classification using string kernels. Journal of Machine Learning Research **2** (2002) 419–444

[7] Cancedda, N.: Word-sequence kernels. Journal of Machine Learning Research **3** (2003) 1059–1082

[8] Schlkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, Massachusetts (2002)

[9] Jalam, R., Teytaud, O.: Kernel-based text categorisation. In: International Joint Conference on Neural Networks. Volume 3., Washington, DC (2001) 1891 – 1896

[10] Zhang, D., Simoff, J.S., Debenham, J.: Exchange Rate Modelling for e-Negotiators. Springer, Computational Intelligence (2006)

[11] Sato, K.: Extracting word sequence correspondences with support vector machines. In: the 19th international conference on Computational linguistics. (2002)

[12] Li, Y.: Text document clustering based on frequent word sequences. In: the 14th ACM international conference on Information and knowledge management. (2005)