A STUDY WITH MULTI-WORD FEATURES IN TEXT CLASSIFICATION

Wen Zhang[1], Taketoshi Yoshida[1], Xijin Tang[2]

[1] School of Knowledge Science, Japan Advanced Institute of Science and
Technology, 1-1 Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen, yoshida}@jaist.ac.jp
[2] Institute of Systems Science, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100080, P.R.China
xjtang@amss.ac.cn

ABSTRACT

We carried out a series of experiments on text classification using multi-word features. A hand-crafted method was proposed to extract the multi-words from text data set and two different strategies were developed to normalize the multi-words into two different versions of multi-word features. After the texts were represented respectively using these two different multi-word features, text classification was conducted in contrast to examine the effectiveness of these two strategies. Also the linear and nonlinear polynomial kernel of support vector machine (SVM) was compared on the performance of text classification task.

Keywords: text classification, multi-word features, feature selection, SVM

INTRODUCTION

Automated text classification utilizes a supervised learning method to assign predefined category labels to new documents based on the likelihood suggested by a trained set of labels and documents. During the process of transforming the unstructured text into structured data as numerical vectors for the data mining methods, bag of words (BOW) (Chidanad et al., 1994) is often used to represent the text with single words obtained from the given text data set. As a simple and intuitive method, BOW method makes the representation and learning easy and highly efficient. But an obvious disadvantage of BOW method is that it ignores the ordering and composing of words occurring in the text which are used to describe a concept, not merely a mixture of single words, and this kind of concept usually can provide more evidential information for text classification. For example, key points as a collection of keywords and utterances are used for knowledge creation in (Liu and Tang, 2003). From the view of Ontology, this kind of "synset" can be regarded as concepts that contain a set of words, each of which has a sense that names the concept in WordNet (http://wordnet.princeton.edu/). Also the phenomena as small world for the keywords co-occurrence from a document was demonstrated in (Matsuo et al., 2001) which is a convincing evidence for the existence of common concepts among the documents and the clusters of words in small world network can be treated as some instances of concepts. With the disadvantage of BOW method and the advocating of concepts in documents, it is reasonable to conjecture that adopting ordering and composing information in text representation purposely might improve the text classification performance. Generally, multi-word features are not found too frequently in a text data set, but when they do occur they are often highly predictive. Based on this motivation, the method of multi-word features (Nitin et al., 2005) was proposed in this paper, and the effectiveness of this method was examined with text classification using multi-word features.

Recently, a lot of research has been undertaken in the text mining field with the expectation to enhance BOW by both linguistic characteristics and logic character-istics of words in text. The ordering and position of a word in document was considered as the background relation between text categories for classification in (Cohen et al., 1995). The concept of ontology was introduced in (Hotho et al., 2003) to represent text, and different strategies to make use of ontologies, the linguistic relationships between words and concepts, are discussed. However, these studies are all based on single word representation and concentrate on the hidden relationships between single words in the aspect of linguistics or logic. In this paper, multi-word features, i.e., a group of consecutive words were proposed for text representation and their effectiveness was examined with the text classification task. Two strategies were developed to post-process the extracted multi-words into multi-word features (multi-word features are the multi-words after post-processing) to represent the documents, and their text classification performances were compared. Also, the linear kernel and nonlinear kernel of support vector machine (SVM) were compared in this paper on text classification task.

DATA SELECTION

The Reuters-21578 data set (http://www.research.att.com/~lewis) was selected as our experiment data. It appeared as Reuters-22173 in 1991 and was indexed with 135 categories by personnel from Reuters Ltd in 1996. By our statistics, it contains in total 19403 valid texts with average 5.4 sentences for each text. For convenience, the texts from 4 categories, “grain”, “crude”, “trade” and “interest” were selected as our target data set, on the condition that the number of sentences for each text in these categories is between 4 and 7. With this method, 252 texts from “grain”, 208 texts from “crude”, 133 texts from “interest” and 171 texts from “trade” were assigned as our target data set and 2/3 of these texts from each category were used as training data and 1/3 of them as test data by random sampling. Thus, the text collection for the experiment was performed and all the processing described in this paper was conducted on this data set.

DATA PREPROCESSING

The purpose of this section is to explain the processing of the texts selected in Section 2 into standard format and to extract multi-word features from the selected texts used as training data. The usually adopted preprocessing methods in text mining area were employed, such as stop word elimination, stemming, sentence boundary determination. Furthermore, a hand-crafted method for multi-word extraction and post-processing of the multi-words into normalized features were included in this section as well. The following sections give the details of each procedure.

Stop words elimination

Stop words, or stopwords, is a name given to words which are filtered out prior to the processing of natural language data (text). They are generally regarded as 'functional words' which do not carry meaning. In this research, we obtained the stop words from USPTO (United States Patent and Trademark Office), which contains about 100 'functional words' in English (http://ftp.uspto.gov/patft/help/stopword.htm). By our observation concerning the occurring positions of the stop words in a sentence, three kinds of full matching of stop words are programmed to eliminate the stop words as “stop word + white space” for the beginning position, “white space + stop word + punctuation” for the end position and “white space + stop word + white space” for the middle position while “+” means “followed by”.

Sentence boundary determination

In order to extract the multi-words from texts, it is necessary to break up the full text into separate sentences. Sentence boundary determination is essentially the problem of deciding which instances of a period followed by white space are sentence delimiters (full stop) and which are not. In this research, sentence boundary determination is also conducted with a hand-crafted program. The punctuation marks “?” and “!” are regarded as the end of a sentence, and many other rules were made to identify a period as a delimiter for a sentence, especially on distinguishing it from decimal point.

Basic hypothesis for Multi-word

Many methods could be applied to extract the multi-words from text, such as the frequency approach, correlation approach and mutual information approach, etc. However, a newly programmed approach was proposed in this research to extract the multi-words from training texts. A basic hypothesis was conceived with a multi-word, that if a multi-word appears in a text and has the power enough to discriminate the category of this text from others, it should occur more than once in all the texts of this category. For example, "money market" is a significant multi-word feature for the category "interest", so we assumed that it is impossible for it to occur only once in all the texts of the "interest" category. Otherwise, it could not be regarded as a multi-word with the power of discrimination, which can be used to distinguish its category from other categories.

Multi-word extraction

The usually adopted preprocessing methods in text mining area were employed, such as stop word elimination, stemming, sentence boundary determination. Furthermore, a hand-crafted method for multi-word extraction and post-processing of the multi-words into normalized features were included in this section as well. The following sections give the details of each procedure. Based on the above hypothesis, the multi-words were extracted by the comparison between any two sentences in the same category so to find out the same consecutive matching in both sentences. Our multi-word extraction algorithm was followed below.

```
Input:
$s_1$, the first sentence
$s_2$, the second sentence
Output:
Multi-word extracted from $s_1$ and $s_2$.
Procedure:
$s_1 = \{w_1,w_2,\ldots,w_n\}$ $s_2 = \{w_1',w_2',\ldots,w_m'\}$ k=0
for each word $w_i$ in $s_1$
        for each word $w_j$ in $s_2$
            while($w_i$ == $w_j$)
                k++
            end while
            if k>1
            combine the words from $w_i$ to $w_{i+k}$ into a multi-word
            End if
        End for
End for
```

After the processing of multi-word extraction, 468 multi-words were obtained from training data "grain", 407 multi-words from training data "crude", 366 multi-words from training data "trade", and 273 multi-words from training data "interest".

Multiword post-processing with two strategies

More often than not, the multi-words overlapped each other such as "U.S. agriculture department", "U.S. agriculture" and "agriculture department". In this research, we develop two types of strategies to post-process the extracted multi-words.

Strategy 1: also named "decomposition strategy". With this strategy, if a short multi-word is included in a long multiword, the long multi-word will be eliminated from the multi-word feature set.

Strategy 2: also named “combination strategy”. With this strategy, if a short multi-word is included in a long multiword, the short multi-word will be eliminated from the multi-word feature set.

With the decomposition strategy, total 1514 multi-words obtained in Section 3.4 were decomposed into 984 multi-word features and with the combination strategy, the extracted multi-words was combined into 1037 multi-word features.

TEXT CLASSIFICATION WITH MULTI-WORD FEATURES

Text representation with multi-word features

With Strategy 1, only the frequency of the multi-word feature in a text was considered, because long multi-word features were decomposed into short ones and texts were represented in vector space model.

With Strategy 2, the texts were represented with Boolean model, because long multi-word features include more single words than short ones, and these features do not frequently occur in a text.

However, the occurrence of a long multi-word feature in a text can not be easily determined with simple full matching. Instead, two indicators were introduced to determine whether a long multi-word feature occurred in a given text. One is the ratio of single words of the multi-word feature (multi-word feature comprises single words) occurring in the text, and the other is the minimum distance that these single words occurred in a given text. Following is the algorithm designed to calculate these two indicators.

```
Input:
$D = \{s_1, s_2, \ldots, s_n\}$ // D is a text in selected text collection and $s_i$ is the ith sentence in D;
$W = \{w_1, w_2, \ldots, w_m\}$ // W is a multi-word feature and $w_i$ is the ith single word in W;
Output:
r --- ratio of W occurred in D;
l --- minimum length of W occurring in D;
Procedure:
$S = \bigcup_{i=1}^{k} s_i$ ;
$L = \varnothing$;
For each $w_i$ in W
    If $w_i$ exists in S
        $L = L \cup \{w_i\}$
    End if
End for
$r = |L|/|W|$;
For each $w_i$ in L
    $L_j = \{a_{j,k} \mid a_{j,k}$ is the $kth$ position for $w_j$ occurred in S$\}$;
End for
// the total number of $L_j$ is $|L|$;
For each $a_{ik}$ in $L_1$
    $L_k' = \{a_{ik}\}$;
```

```
        While $|L_k'| < |L|$
                next = $|L_k'| + 1$;
                Find $a_{next,*}$ in $L_{next}$ which has the minimum difference with anyone element in $L_k'$
                $L_k' = L_k' \cup \{ a_{next,*} \}$
        End while
End for
For each $L_k'$ where k = 1 to $|L_1|$
        $b_k$ = maximum in $L_k'$ – minimum in $L_k'$
End for
l = minimum $b_j$ (j=1,…,$|L_1|$)
```

Although the threshold of these indicators should be set carefully and used jointly according to the practical application, the threshold for the first indicator was simply set as 0.5 and latter one is as the 1.75 times length of the given multi-word feature. That is, if more than half number of the total words of multi-word feature occurred in the given text and the minimum length of these words occurred in the given text is no more than 1.75 times length of this multi-word feature, this multi-word feature would be regarded as occurrence in the given text. Otherwise, it would be regarded as absence from the given text.

Feature selection with Information gain

Information gain (IG) is usually employed as a term goodness criterion in the field of machine learning (Quinlan, 1986). It is reported in (Yang et al., 1997) that 98% removal of unique terms yields text classification accuracy up to 89.2% with the Reuters-22173 data set. After the text representation in the previous section, IG value of each feature was calculated based on the entropy of classes and feature values, also the features were sorted in an ascending order according to their IG values. The text classification experiment was designed varying the removal percentages of the sorted features as shown in Table 1. The motivation for us to do this is that we wanted to examine the effectiveness of IG selection method with multi-word features.

Table 1. The percentage of high IG value features corresponding to each designed test

Test No.	1	2	3	4	5	6	7	8	9	10
Removal Percentage (%)	0	50	70	75	80	85	90	95	98	99

Learning with SVM

SVM is a classifier derived from a statistical learning theory by Vapnik and Chervonenkis, and it was first introduced in 1995 (Mulier, 1999). Some published results in (Apte et al., 1998; Yang et al., 1999) reported that it can obtain better performance than other learning methods in text classification tasks. In this research, we carried out the text classification task with SVM using linear and nonlinear kernel, respectively, in order to compare the performance between them. To simplify, the $(u*v)^1$ was used as the linear kernel, and the polynomial kernel $(u*v+1)^2$ was used as the nonlinear kernel for the task. Also, the different representation strategies introduced in Section 4.1 are combined in the experiments. In detail, four types of experiments were designed, as shown in Table 2. The motivation for us to devise these experiments is that we also wanted to compare the effect of representation strategy and kernel types on the classification, besides the single comparison of strategies or kernels.

Table 2. The experiments designed with two kinds of text representation strategies and kernels.

Experiment No.	1	2	3	4
Representation strategies	Strategy 1	Strategy 2	Strategy 1	Strategy 2
Kernel type	Linear	Linear	Nonlinear	Nonlinear

EXPERIMENTAL RESULTS

According to the designs in Section 4.3, we carried out all the experiments with the help of libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The results of these experiments were as follows. It should be noticed here that each type of experiment was carried out at 10 different removal percentages of features specified in Section 4.2. Table 3 demonstrated the text classification accuracy of each classifier at different percentage of removal feature.

Table 3. The accuracy of the designed experiments on four-class text classification task

R(%) EX	I 00	50	70	75	80	85	90	95	98	99
1	0.7835	0.7462	0.7873	0.7798	0.7910	0.7649	0.7798	0.7798	0.7089	0.5373
2	0.8694	0.8656	0.8358	0.8395	0.8209	0.7948	0.8172	0.7649	0.7500	0.7350
3	0.5186	0.5858	0.7089	0.7089	0.7201	0.7351	0.7500	0.7723	0.7201	0.5037
4	0.8134	0.8208	0.7835	0.8283	0.8097	0.7985	0.8022	0.7686	0.7500	0.7388

From the comparison of the results of experiment No.1 and experiment No.2, it can be seen that on average Strategy 2 outperforms Strategy 1 with the linear kernel, except for the result at 95% removal of total features. On nonlinear kernel of experiment No.3 and No.4, we can confidently deduce that a better performance was obtained with Strategy 2 than Strategy 1. From the pair comparison of the results of experiment No.1 and No.3, we can draw the conclusion that the performance of linear kernel was better than that of the nonlinear kernel on Strategy 1, and also the IG method exhibited its effectiveness in the feature selection as it kept the performance stable when more and more features were removed from the feature set. Although it also can be seen from comparison between experiment No.2 and No.4 that the performance of linear kernel is better than that of nonlinear kernel with Strategy 2, the robustness of Strategy 2 is not very good because its performance declined dramatically when more and more features were removed from the feature set.

From the above analysis of Table 3, it can be summarized that Strategy 2 has better performance than Strategy 1, and linear kernel also has a little better performance than the nonlinear kernel in text classification. The effectiveness of the IG method for feature selection was demonstrated, because the overall performance of classification is kept stable when more and more low IG value features were removed from the applied feature set. Further, it can also be deduced that the effect of the employment of different strategies on the text classification task is more significant than that of using different kernels.

CONCLUDING REMARKS

Multi-word feature is a newly practical method for feature selection in text mining. In this paper, a hand-crafted method was proposed to extract the multi-words in the text based on our hypothesis that a multi-word cannot occur only once in the texts of its category. With this method, the multi-words are extracted from the texts of the same category. In order to normalize the initial multi-words into standard multi-word features, two strategies were developed: the first is the decomposition strategy and the second is combination strategy. Next, the texts in the selected data set were represented with these two types of different multi-word features, respectively. Then, IG method was employed to evaluate the importance of the features for text classification. The motivation of the feature evaluation was that we wanted to examine the effectiveness of IG method, and the robustness of the multi-word text classification, when the applied features are varied. Finally, the text classification was carried out with SVM in both linear and nonlinear kernels, and the results are compared on not only the different kernels but also the different strategies. The experimental results demonstrated that in multi-word text classification, Strategy 2 outperforms Strategy 1, and linear kernel outperforms nonlinear kernel with SVM. However, it also appeared that Strategy 2 has poorer robustness than Strategy 1 when the low IG value features are removed from the applied feature set, and its performance also declined dramatically. Nevertheless, IG method was proved an effective approach for feature selection in most of the cases, because it kept the classification performance stable when the low IG value features were removed from the applied feature set gradually.

Although the experiment results have provided us with some clues on text classification with multi-word features, a generalized conclusion was not obtained from this examination because of the lack of theoretical proof. To be frank, our work is an initial step, and more examination and investigation should be undertaken for more convincing work

One of the promising directions in the text mining field concerns predictive pattern discovery from large amounts of documents. In order to achieve this goal, many kinds of work are involved in this field such as algorithm optimization, linguistics and machine learning. As for our further research, we would like to develop more precise algorithms (Smith et al., 1981) for multi-word extraction, and use the linguistics in multiword extraction, instead of only literal extraction from texts. Also, the name entity will be considered seriously in extraction, as well as the adoption of support from a third-party dictionary. Another aspect that we also should advance is the improvement of the learning method such as SVM for multi-class classification. Despite the fact that the basic disciplines of the learning methods are well established, the performance of classification will be improved if more processes are refined according to our practical research.

ACKNOWLEDGMENTS

This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the “Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project” and partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001.

REFERENCES

Apte, C., F. Damerau., Weiss, S.M. (1998). Text Mining with Decision Trees and Decision Rules, in Conference on Automated Learning and Discovery, Carnegie-Mellon University, June 1998.

Chidanand, A., Fred, D., Sholom, M.W. (1994).Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3):233-251.

Cohen, W. W. (1995). Learning to classify English text with ILP methods. In: Proceeding of 5th International Workshop on Inductive Logic Programming. 3-24.

Hotho, A., Staab, S., Stumme, G. (2003). Ontologies improve text document clustering. In Proceeding of IEEE International Conference on Data Mining (ICDM03), 541-544.

Liu, Y. J., Tang, X. J. (2003) A Visualized Augmented Tool for Knowledge Association in Idea Generation, in Knowledge and Systems Sciences: Toward Meta-Synthetic Support for Decision Making (the proceedings of the Fourth International Symposium on Knowledge and Systems Sciences, Lecture Notes in Decision Sciences 3, Global-Link Publishers, Hong Kong, 2003, 19-24.

Mulier,F.(1999). Vapnik-Chervonenkis (VC) Learning Theory and Its Application. IEEE Trans on Neural Networks. 10(5) 5-7.

Matsuo, Y., Oshawa, Y., Ishizuka, M.(2001) Key-world: Extracting keywords from a Document as a small world. K. P. in Jantke and A. Shinohara (Eds): DS 2001, LANI 2226, Springer-Verlag, Berlin Heideberg, 271-281

Nitin, I., Fred, J. D., Zhang, T (2005). Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer Science and Business Media, Inc. 15-37.

Quinlan, J.R.(1986) Induction of decision trees. Machine Learning, 1(1):81-106.

Smith, T.F., Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology, 147, (1981) 195-197.

Yang, Y.M., Pedersen, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning, 412-420.

Yang, Y.M., Lin, X. (1999). A re-examination of text categorization methods. In: Proceedings on the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley,C A, 42-49.