D-039

A Comparison of Automatic Document Classification Methods for Vietnamese Language

Nguyen GIANG SON Shigeru OYANAGI Katsuhiro YAMAZAKI Masatoshi KAMIHARAKO

Graduate School of Science and Engineering
Ritsumeikan University, Biwako-Kusatsu Campus Noji Higashi
1 chome, 1-1 Kusatsu, 525-8577 Shiga-ken, JAPAN

Abstract *Document classification, an important area of Text Mining, has many interesting applications such as email spam filtering. There have been many researches on document classification methods but mainly for English. Hence, we conducted a comparison of some well-known document classification methods including Decision Tree (DT), Naïve Bayes classifiers, and Support Vector Machines (SVM) for Vietnamese language. The results showed that SVMs still got the highest performance whereas the performance of DT classifiers was not so high and the Naïve Bayes classifiers were quite competitive.*

1 Introduction

Automatic document classification or categorization (DC) is a task done by a computer to assign automatically categories to documents. It has many important applications like email spam filtering, document organization, document routing, and so on. An increasing number of Machine Learning approaches have been well studied but they mainly are applied for English language. Therefore in this research we focus on applying those methods including the Decision Tree, Naïve Bayes classifiers, and the SVM classifier for Vietnamese document categorization to understand their performance on the language.

The remaining of this article is structured as following. Section 2 describes a brief background about classifiers, their uses, and evaluation methods. Section 3 displays experiments and theirs results. The last section 4 is our conclusion and future work.

2 Classifiers and Performance Evaluation

2.1 Classifiers

2.1.1 Decision Tree

DT is one of the most widely used methods in supervised learning. The decision tree model is built by recursively splitting the training set based on a locally optimal criterion (the goodness function – Information Gain) until all or most of the records belonging to each of the leaf nodes bear the same class label. After that the tree is pruned to avoid overfit problem. We used the standard C4.5 (J4.8 from the Weka tool in [6]) version with default parameters for our experiments.

2.1.2 Naïve Bayes Classifiers

Naïve Bayes classifiers are based mainly on Bayes assumption, which is that words of a document are independent. Although this assumption does not fit the real world, it has been show to produce very good performance. The formula to identify a label of a document:

$$\arg\max_{c_k} P(c_k \mid x) = \arg\max_{c_k} P(c_k)P(x \mid c_k)$$

Conditional probabilities of words with a given class are calculated by using specific event models. There are two event models, Bernoulli event model (BayesB using binary weighting scheme) and multinomial event model (BayesM using term frequency weight). Both types of Naïve Bayes classifiers were used in our experiments.

2.1.3 SVM classifiers

Main idea of SVMs is to solve the binary linear separable problem by finding the linear separating hyper-plane which maximizes the margin, the optimal separating hyper-plane.

For solving nonlinear separable problems, kernel functions are used to transform problem space to linear separable derived feature space. We applied the Linear SVM, because it has been proved very efficient for solving DC problems. A Java version of SVM [5] was implemented for the experiments.

2.2 Performance Measures

Our experiments adopt commonly used performance measures, including the micro-recall, micro-precision, and micro F_1 measure to evaluate classifiers on multi-class problems. To make a comparison for classifiers, they are trained with various training data sets and tested on the same test dataset.

3. Experiments and Evaluation

3.1 News Dataset Preparation

Unlike in English there is no standard Vietnamese corpus for testing DC. In our experiments, news datasets were crawled automatically from a Vietnamese news website (http://dantri.com.vn). Then properties of the news articles like title or headline, label, and body were extracted from their web pages by removing HTML tags. In theses experiments, the news with total 5000 articles is divided into 4 training datasets with various sizes ranging from 1000 to 4000 and a test dataset with size 1000. The number of labels/categories of news datasets is 11.

3.2 Feature Extraction and Selection

To build input feature vectors, documents were first tokenized by using the newest Vietnamese word segmentation program, JVnsegmenter as in [4] (this technique has high performance, highest F_1-measure 94.09). Then, the stop words were removed from the input vectors. There is no word stemming steps for processing Vietnamese texts. Lastly, to reduce dimension of feature space, X^2 technique were used. X^2 method has been proved that it has well performance on TC as in [3]. The size of input feature vector is 3000, which is in medium size.

Term weighting schemes were used for TC classifiers as following. TFxIDF term weighting was used for DT C4.5. Inverse Document Frequency was used for SVM because we saw that it had better accuracy than that of TFxIDF. For Naïve Bayes classifiers, it depends on the event models (term frequency or binary).

3.3 Experimental Results and Discussion

Figure 1, 2, and 3 describe the four classifier's performance on the test corpus in terms of micro averaged *precision*, *recall*, and F_1 measure.

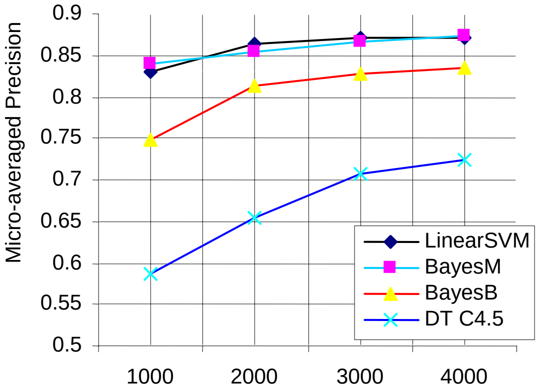

Figure 1. Micro-averaged *precision* curves

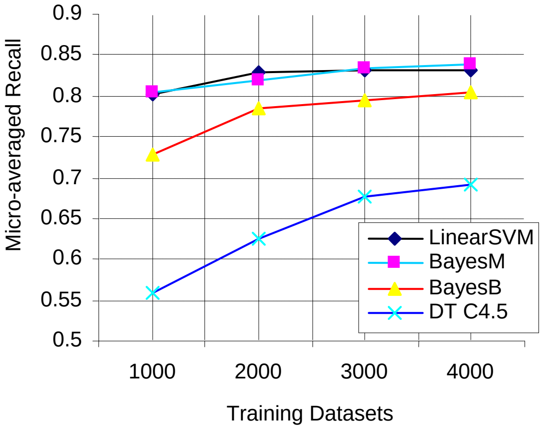

Figure 2. Micro-averaged *recall* curves

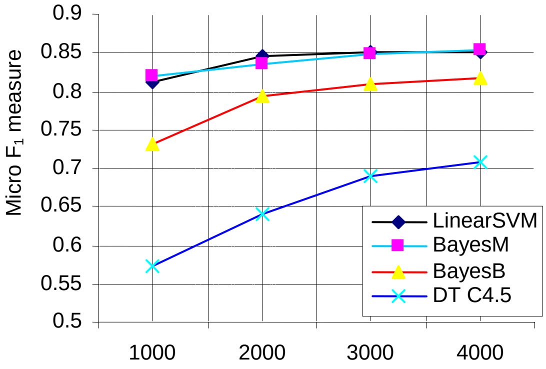

Figure 3. Micro F_1 measure of classifiers

The LinearSVM and BayesM exhibited similar good performance. In contrast the DT C4.5 displayed poor performance. Perhaps it is not suitable for high dimension problem with spare data. Although BayesB was lower than BayesM, its performance was quite good. Ranks of classifiers remained the same in both micro precision and recall measures.

4 Conclusion and future work

We have tested TC problem with well-known classifiers applied for Vietnamese text corpus to invest their performance. In general, except DT, the three remaining classifiers produced good performance (score of 0.80 or above). This is initial researches about TC in Vietnamese. To have a complete view on performance of TC methods for Vietnamese, the comparison could be executed by advanced testing conditions like various kinds of news sources, large size of corpora, and changing from small to large dimension of feature vectors.

References

[1] Tom Mitchell. Machine Learning. *McGraw Hill*. 1996.

[2] Fabrizio Sebastiani. Machine learning in automated text categorization. In *ACM Computing Surveys*. 2002

[3] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412-420, Nashville, US, 1997.

[4] Cam-Tu Nguyen and Xuan-Hieu Phan, "JVnSegmenter: A Java-based Vietnamese Word Segmentation Tool", *http://jvnsegmenter.sourceforge.net/*, 2007.

[5] Java version of SVM: *http://www.csie.ntu.edu.tw/~cjlin/libsvm/*

[6] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, *Morgan Kaufmann*, San Francisco, 2005.