An Efficient Feature Selection using Multi-Criteria in Text Categorization for Naïve Bayes Classifier

SON DOAN
Graduate School of Information Science
Japan Advance Institute of Science and Technology
Asahidai 1-1, Tatsunokuchi, Ishikawa 923-1292,
JAPAN

SUSUMU HORIGUCHI
Graduate School of Information Science
Tohoku University, Aoba 09, Sendai, 980-8579
JAPAN

Abstract: - Feature selection is one of the most interesting problems in machine learning in general and text categorization in particular. Previous researches in feature selection often focus on choosing appropriate measument to evaluate features. This seems to be good for structured data but rather difficult to text, a non-structured data. Our main contribution in this paper is to propose a new approach of feature selection based on multi-criteria ranking of features. A new model for feature selection is propose; based on a threshold value for each criterion, a new procedure for feature selection is proposed and applied to a text categorization. Experiments show that the proposed model outperforms performances in compare to conventional feature selection methods.

Key-Words: - feature selection, text categorization, text mining

1 Introduction

Text categorization is defined as the problem of assigning a natural document into one or more predefined classes [6],[12]. One of the most interesting issues recently in text categorization is feature selection problem. Feature selection plays a very important role in data mining in general and text categorization in particular. Theoretically, feature selection is shown as the NP-hard problem [1] and many solutions based on search heuristics are proposed such as [3],[5],[7].

Feature selection problem is aggravated by text data due to the non-structured format in the form of raw text or its semi-structured format in the forms of email or Web pages. In addition, the large amount of terms in text documents leads difficult to construct a classifier.

Text data itself has properties of natural language in human sense such as semantics, syntax, thesaurus, etc. Feature selection in text categorization is equivalent to the questions: *what is the feature and which features should be chosen from set of text documents with respect to the category of documents ?* Some results from previous researches [4],[6] showed that a phrase did not much affect the performance of text categorization, a feature hereafter is treated as a term, not a phrase, which is extracted from a given corpus (a set of documents). The question now remains that which terms should be chosen with respect to the category of documents. A term itself has several criteria characterizing its qualitative amount in a document as well as a corpus. Based on those multi-criteria, we propose a new approach and a model to the feature selection in text categorization by selecting features. We also show that, by experiments, using some criteria in feature selection can achieve better performance in text categorization compared to using only one criterion.

This paper is organized as follows. Section 2 briefly introduces related work. Section 3 proposes a general model for feature selection and a procedure for feature selection based on multi-criteria ranking. Experimental results are shown in Section 4. Section 5 draws some conclusions and outlines future work.

2 Related Work

Text categorization consists of two main steps : pre-processing and classifier building. Pre-processing includes tasks such as feature extraction, feature selection, and document representation. The output of the first step is the input for the second in which machine learning algorithms can be used for classification purpose.

There are two common model in text representation, the vector space model and the probabolistic model. A document will be represented as a vector of features in the vector space model [10] or a ``bag-of-words'' in the probabilistic model; features are the components in a vector or a ``word''. Therefore, feature selection plays a very important role in later steps and affects the performance of the whole system.

Two most common approaches in feature selection are the filter and the wrapper approaches [3],[5],[9]. In the wrapper approach, the subset of features is chosen based on the accuracy of classifiers. Technically, the wrapper method is relatively difficult to implement, especially with a large amount of data. Instead, the filtering approach is usually chosen because it is easily understood and for its independent classifiers.

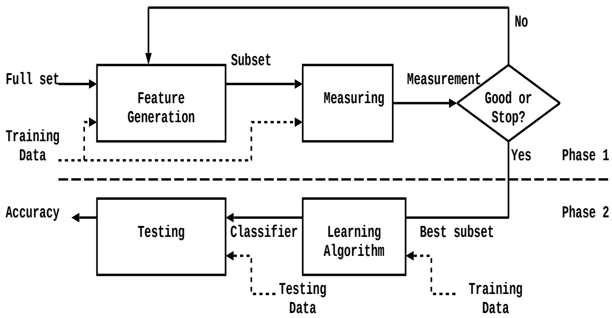

Figure 1. The framework of the filter model

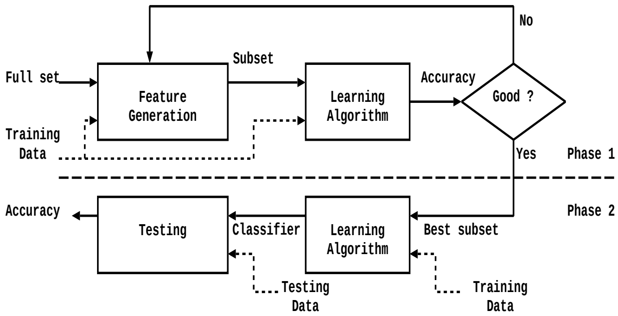

Figure 2. The framework of the wrapper model

Two models of feature selection are shown in Figure 1 and Figure 2 respectively.

The filter approach, as its name implies, chooses a subset of features by filtering based on the scores which were assigned by a specific weighting method. In text categorization, the filter approach is often used and features are selected by one of these following criteria [9],[11],[13].

1. Document frequency criterion: Features are selected by their frequencies in document, with a threshold.
2. Class-based criterion: Select features based on their frequency in a class.
3. Information gain measure: Given a set of categories $\mathbf{C}=\{c_i\}_{i=1}^{m}$ the information gain of term x is given by [11],[13]:

IG(x)=

$$\sum_{i=1}^{m} P(ci)\log P(c_i) + P(x)\sum_{i=1}^{m} P(c_i \mid x)\log(c_i \mid x)$$

$$+ P(x)\sum_{i=1}^{m} P(c_i \mid \bar{x})\log(c_i \mid \bar{x}). \qquad (1)$$

4. Mutual information measure: Mutual information of term x in class is given by [11],[13].

$$MI(x) = \sum_{i=1}^{m} \log \frac{P(x \wedge c_i)}{P(x).P(c_i)} \qquad (2)$$

There are also other measures for feature selection, for example, chi-square and odd-ratio ... [9],[11],[13]. Among these measures, mutual information measure is the most common measure used recently [2],[11],[13],([12]). For this reason we use mutual information measure as the baseline feature selection method in this paper.

3 Feature Selection Based on Multi-Criteria Ranking

The feature selection problem in text categorization can be stated as follows: Given a set **X** consisting of n features $x_1, x_2, \ldots x_n$, the problem in feature selection is to choose the optimal subset **S** of **X** ($\|\mathbf{S}\| \ll \|\mathbf{X}\|$) with highest effectiveness for the system.

To solve this problem, our basic idea is to filter features based on a procedure of multi-criteria ranking for terms. Each feature, according to a criterion, will be weighted with a term weight; thus, with t criteria, we will have t ways of ordering features. The feature selection problem can be expressed as follows :

*Choose a proper subset of **X**, given a set of criteria $\theta_1, \theta_2, \ldots \theta_t$, within which each criterion determines a ranking of **X**.*

Formally, for each criterion θ_i, we have a set of preference $X_i=\{x_{(i)1}, \ldots, x_{(i)n}\}$ where $x_{(i)j} \in \mathbf{X}$. Thus, we have t sets of X_i.

After ranking according to a multiple criteria as above, for each criterion θ_i, we select a subset $\mathbf{S_i}$ of **X** based on a threshold τ_i. Then the set of selected terms is defined by

$$S = \bigcup_{i=1}^{t} S_i$$

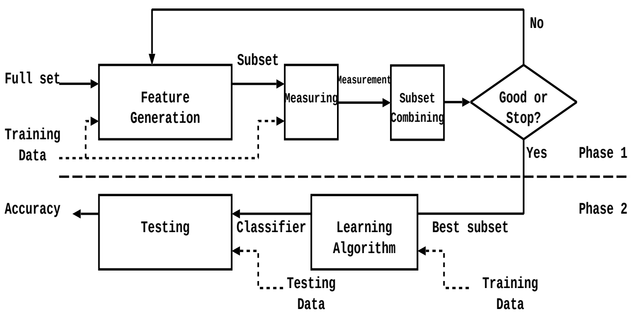

Figure 3. The proposed model for feature selection

The framework of the model for feature selection is shown in Figure 3. In the proposed model subset is evaluated by a measurement. For each measurement we have a ranked list of terms, the optimal feature subset is obtained by a filter choosing from threshol values of each criterion.

There is a question raising from framework: how to choose the threshold values for criteria. This seems to be turn out to the another problem with multiple constrants. In this paper we investigate only the effects of multi-criteria on feature selection in text categorizarion problem.

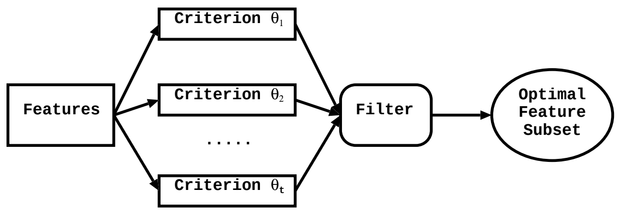

Figure 4: The framework for subset combining process based on multi-criteria.

Procedure EFS(X: Original feature set, S-Optimal feartureset, τ_1, .. τ_1 – threshold values)
Begin
For i:=1 to t **loop**
    $S_i \leftarrow \varnothing$;
    Step 1. Ranking all features based on criterion θ_i;
    Step 2. Add the first features based on τ_i to S_i;
End loop;
    $S \leftarrow S_1 \cup .. \cup S_t$;
Return S;
End

Figure 5. The EFS procedure for selecting the optimal feature subset.

The framework for subset combining process based on multi-criteria is shown in Figure 4 and the procedure describing our approach is depicted as in Figure 5.

Naïve Bayes Classifier

After the pre-processing step a document is represented by features and these features are inputs for the second text categorization step, classifier building. Among existing machine learning techniques is the naive Bayes which is one of the most common techniques used in text categorization and is viewed as the baseline method until now [11],[12]. In this paper we use the naive Bayes algorithm as the standard algorithm for the classifier.

The naive Bayes algorithm can be briefly described as follows.

Given m classes $\mathbf{C}=\{c_1, c_2, \ldots c_m\}$, with a document d’, our problem is to build a classifier σ that can assign the document d’ to a class.

Without loss of generality, suppose a document d' consisting of terms x_1, x_2,…x_n. The naive Bayes algorithm calculates the probability of a class belonging to each document by the formulation

$$P(c_i \mid d') \propto P(d' \mid c_i)P(c_i) = P((x_1,...x_n) \mid c_i)P(c_i)$$

$$= \prod_{j=1}^{n} P(x_j \mid c_i)P(c_i). \qquad (3)$$

Thus, the class of document d' is calculated by the following formula,

$$\sigma(d') = \arg\max_{i\in[1..m]} P(c_i \mid d). \qquad (4)$$

4 Experiments

4.1 Real-world Data Set

Table 1: Details of top 10 categories of Reuters21578 data set.

Category	#training docs	#testing docs
Earn	2,877	1,083
Acq	1,650	719
Money-fx	538	179
Grain	433	149
Crude	389	189
Trade	368	117
Interest	347	131
Ship	197	89
Wheat	212	71
Corn	181	56
Total	**7,769**	**3,019**

Table 2: The contingency table for a category c_i

Category c_i	Human assign YES	Human assign NO
Classifier predict YES	a_i	b_i
Classifier predict NO	c_i	d_i

To examine our proposed method, we used a standard text data set Reuters-21578 for our problem. Reuters has been viewed as the standard data for text categorization community until now. There are various versions of Reuters, of which Reuter-21578 is the most common used [11],[13]. The top 10 categories were chosen for implementation; they are described in Table 1.

Reuters-21578 data set is preprocessed by removing common words such as *the*, *a*, *an*, etc in the stop list, words are stemmed by the Porter algorithm. After preprocessing, the number of vocabulary is 19,791 words.

In our experiments, we chose two standard methods in feature selection, all terms (that is method containing all terms in vocabulary) and feature selection based on mutual information measure. For easily understanding later, we called the first case all term method and the second case the baseline method. Thus, the number of all term method is 19,791; with the baseline method, the number of vocabulary is chosen was 2,000 terms (≈1/10 vocabulary).

To compare our method with the baseline method and all term method, we used two criteria, the mutual information and class-based frequency. A threshold for mutual information was τ_1=2000 and two thresholds for class-based frequency measure are selected, τ_2=100 and, τ_2=200, respectively. That is, parameters in the EFS procedure are t=2, τ_1=2000, τ_2=100 and t=2, τ_1=2000, τ_2=200. We called the first case in the our proposed method the EFS-100 method and the second the EFS-200 method. The number of terms in the EFS-100 is 2,314 terms, and the number of terms in the EFS-200 is 2,619 terms. Experiments are executed in SunOS 5.8 operating system, Perl, sed, awk, C programming languages and *libbow* library [8].

4.2 Performance Measures

Two basic measures in text categorization are precision P and recall R. They are expressed mathematically by a contingency table in Table.

$$P = \frac{a_i}{a_i + b_i} \quad and \quad R = \frac{a_i}{a_i + c_i} \qquad (5).$$

To evaluate the performances of whole categorization system, the macro-averaging and micro-averaging P and R are used

$$macro-P = \sum_{i=1}^{k} \frac{P_i}{k} \quad and \quad macro-R = \sum_{i=1}^{k} \frac{R_i}{k} \quad (6).$$

Micro-averaging of P and R are calculated by,

$$micro - P = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} (a_i + b_i)} \quad and$$

$$micro - P = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} (a_i + c_i)} \qquad (7).$$

F_1 is defined by,

$$F_1 = \frac{2PR}{P + R} \qquad (8).$$

BEP measure is calculated by interpolation between two points P and R, that is the point where P=R. It is often calcualted by taking the average of P and R. The macro- and micro-F_1 and BEP are calculated by replacing P and R

Table 3: BEP performances of Reuters-21578

Category	All terms	Baseline	EFS-100	EFS-200
Earn	97.65	97.47	97.43	97.38
Acq	96.45	96.04	96.60	96.66
Money-fx	76.54	75.98	76.54	76.19
Grain	50.34	49.49	51.04	51.50
Crude	80.00	78.09	78.51	78.51
Trade	79.15	84.62	84.12	84.12
Interest	72.52	68.96	70.23	70.23
Ship	69.92	60.00	59.55	59.55
Wheat	31.76	40.85	41.13	39.72
Corn	33.93	35.40	37.50	37.50
Macro ave	68.13	68.69	**69.26**	69.14
Micro ave	72.31	74.54	**74.55**	74.55

Table 4: F_1 performances of Reuters-21578

Category	All terms	Baseline	EFS-100	EFS-200
Earn	98.10	97.91	98.04	98.04
Acq	96.48	96.21	96.67	96.67
Money-fx	76.92	75.98	76.54	76.30
Grain	59.76	54.42	57.47	57.41
Crude	81.40	79.67	79.43	79.43
Trade	82.59	85.59	85.60	85.60
Interest	73.00	73.83	73.38	73.38
Ship	67.00	67.58	68.75	68.96
Wheat	39.82	49.24	48.39	47.83
Corn	35.89	46.40	44.02	44.30
Macro ave	71.10	72.68	**72.83**	72.79
Micro ave	73.34	73.86	**74.06**	74.03

with the corresponding macro and micro of P and R in Equation 6 and 7.

Two macro-averaging and micro-averaging of BEP and F_1 are viewed as the whole performances of text categorization systems.

4.3 Experimental Results

The macro and micro averages F_1 and BEP are considered as the system performances in text categorization. Table 2 shows the results of BEP and table 3 shows the results of F_1. Results indicated that both two proposed methods the EFS-100 and the EFS-200 had higher performances than the baseline and the all term methods.

The macro average BEP or the EFS-100 is 69.26% vs. 68.13% when using the all term method and 68.69% when using the baseline method. In case of the EFS-200, the macro average BEP is 69.14%; it is higher than both baseline and the all term methods but lower than the EFS-100. The micro average BEP for both proposed methods is the same (74.55%). It is also not different from that for the baseline method (74.54%) but higher than the all term method (72.74%).

Similarly, the macro and micro averages of F_1 for both proposed methods are higher than the baseline and the all term methods. The macro averages F_1 are 72.83% for the EFS-100 and 72.79% for the EFS-200 respectively, vs. 71.10% for all term method and 72.68% for the baseline method. The micro averages F_1 are 74.06% for the EFS-100 and 74.03% for the EFS-200 while they are 73.86% and 73.34% for the baseline method and the all term method respectively.

In summary, our proposed method outperformed the baseline method and the all term method, especially for macroaveraging measures. Furthermore, the results also showed that the EFS-100 has better performance than the EFS-200, it has been suggested that appropriate parameters τ_1, τ_2 for our proposed method can be tuned for achieving better performance.

In Table 2 we also see the BEP measures for the two categories, corn and wheat. Two these categories have smallest number of training documents in Reuters-21578, with 212 and 181documents respectively. The results show the advantages of using feature selection with our EFS-100 method, with BEP rising from 31.78% when using all term method to 41.13% with the EFS-100 method for wheat, and from 33.93% to 37.50% for corn. Compared to the baseline method, the results show that the performance improved from 40.85% to 41.13% for wheat and from 35.40% to 37.50% for corn, In Table 3, the F_1 measures in categories wheat and corn show that the proposed method is also higher than the all term method.

5 Conclusions

This paper proposed a novel feature selection approach based on the multi-criteria ranking of features in text categorization problem. A new general framework for feature selection was proposed and applied to Reuters-21578 data set.

Experimental results shows the following advantages:

1. The proposed approach has shown that using multi-criteria to feature selection is promising approach.
2. The proposed approach outperformed the all term method and the baseline method in terms of F_1 and BEP measures.

Acknowledgments

This work was supported in partly the Grand-in-Aid of Scientific Research, JSPS, Japan.

References:

[1] Amaldi, E. & Kann, V. (1998), "On the approximation of minimizing non zero variables or unsatisfied relations in linear systems", *Theoretical Computer Science* (209), 237-260.

[2] Baker, L. & McCallum, A. (1998), "Distributional clustering of words for text classification", *in* `Proc of SIGIR-98', pp. 96-103.

[3] Blum, A. & Langley, P. (1997), "Selection of relevant features and examples in machine learning", *Artificial Intelligence* 97(1-2), 245-271.

[4] Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998), "Inductive learning algorithms and representations for text categorization", *in* `Proceeding of the 1998 ACM 7th International Conference on Information and Knowledge Management', pp. 148-155.

[5] Kohavi, R. & John, G. (1997), "Wrappers for feature subset selection", *Artificial Intelligence* 97(1-2), 273-324.

[6] Lewis, D. (1991), "Representation and Learning in Information Retrieval", PhD thesis, Graduate School of the University of Massachusetts.

[7] Liu, H. & Motoda, H. (1998), *"Feature Selection for Knowledge Discovery and Data Mining"*, Kluwer Academic.

[8] McCallum, A. K. (1996), "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering". http://www.cs.cmu.edu/~mccallum/bow.

[9] Mladenic, D. (1998), "Feature subset selection in text learning", *in* `Proc of European Conference on Machine Learning(ECML)', pp. 95-100.

[10] Salton, G., Wong, A. & Yang, C. (1975), "A vector space model for automatic indexing", *Communications of the ACM* 18(11), 613-620.

[11] Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM computing survey* 34(1), 1-47.

[12] Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Information Retrieval Journal* 1, 69-90.

[13] Yang, Y. & Pedersen, J. (1997), "A comparative study on feature selection in text categorization", *in* `Proceeding of the 14th International Conference on Machine Learning (ICML97)', pp. 412-420.