2nd International Conference on Autonomous Robots and Agents
December 13-15, 2004 Palmerston North, New Zealand

An Agent-Based Approach to Feature Selection in Text Categorization

Son Doan and Susumu Horiguchi
Graduate School of Information Science
Japan Advance Institute of Science and Technology
Asahidai 1-1, Tatsunokuchi, Ishikawa 923-1292, Japan.
{s-doan, hori}@jaist.ac.jp

Abstract

This paper considers the problem of feature selection in text categorization. The problem of feature selection is viewed as the problem of choosing a subset of features based on agent-based approach in which each category can be considered as an agent. A procedure for feature selection is proposed based on choosing appropriate threshold values for each agent and applied to a text categorization problem. Experiments dealing with Reuters-21578, the English benchmark data and the naive Bayes algorithm show that the proposed approach outperforms performances in compare to conventional feature selection methods.

Keywords: feature selection, text categorization, agent-based approach, text mining

1 Introduction

Text categorization is defined as the problem of assigning a text document into one or more predefined categories. It has wide applications such as email filtering, information organization, document routing and so on.

One of the most interesting issues in machine learning in general and text categorization in particular is feature selection which selects "good" features for a classifier. This problem is also proven as the NP-hard [1], almost existing solutions based on search heuristics or prior knowledge of a system [2],[3],[4]. Two common approaches to feature selection are wrapper and filter models. Wrapper model uses a classifier as the heuristic information in feature selection, otherwise, filter model selected features only based on information of features. Both models are based on a classical view of a system which uses inductive principles to extract features from the system.

Agent approach has been a new direction to artificial intelligence recently and it has many applications in both theory and practice [5]. In this paper, we introduce a new agent-based approach to feature selection in text categorization problem. We consider each category as an agent instead of a set of targets in conventional approaches and environment is whole corpus. Features are then ranked by criteria given by agents in the system. By a feature combining strategy we can obtain a subset of features. Experimental results showed that our approach outperformed conventional feature selection method, including performance measures BEP, F_1 and ROC curves.

This paper is organized as follows. Section 2 introduces our agent-based approach in feature selection problem and a procedure for feature selection is also introduced. Section 3 presents an application of our approach to text categorization problem. Expriments are in Section 4. Conclusions are drawn in Section 5.

2 An Agent-Based Approach to Feature Selection

The feature selection problem in general can be stated as follows: Given a set $\mathscr{X}$ consisting of n features $x_1, x_2, \ldots x_n$, the problem in feature selection is to choose the optimal subset S of $\mathscr{X}$ ($||S|| \ll ||\mathscr{X}||$) with highest effectiveness for the system. This problem is also a basic problem in data mining in general and text categorization in particular [2],[6],[7],[8],[9].

Let us consider the problem of text categorization: Given a set of categories $\mathscr{C} = \{c_1, \ldots, c_k\}$, the purpose of text categorization is to learn a classifier which is capable of classifying properly with a new document as an input. All information to learn this classifier must be known in a training process.

From a viewpoint of the agent approach, each category can be considered as an agent with an environment consisting of information of features (features can treated as terms, index terms, phrases, etc). The feature selection problem now is considered as the problem of learning features that characterizing each category (this problem also was generalized in an agent model for concept learning [10]).

Features in the environment can be ranked by each agent (category) c_i. For k category we have k ranking of features as follows

$$\begin{array}{ll} \text{Agent } c_1: & x_{\sigma_1(1)} \preceq \ldots \preceq x_{\sigma_1(N)} \\ \ldots & \ldots \\ \text{Agent } c_k: & x_{\sigma_k(1)} \preceq \ldots \preceq x_{\sigma_k(N)} \end{array}$$

where σ_i is a permutation of the set $\{1,\ldots,N\}$, and $\preceq$ is the order relation based on each agent.

Related to relation between agents, we also have an ranking of features based on mutual information measure [7],[8],[9].

Let H be entropy as follows,

$$H(X) = \sum_{x \in X} p(x) \log p(x), \tag{1}$$

where X is a random variable with distribution function $p(x)$

Entropy is also called self-information for one random variable in a system. Relative entropy $I(X,Y)$ called mutual information between two random variables is defined by:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{2}$$

where X, Y are two random variables with distribution functions $p(x)$ and $p(y)$ and $p(x,y)$ is a joint distribution function.

If X is a feature with distribution function $p(x)$ and Y is an agent c_{k+1} with distribution function $p(y)$. Then, for each agent c_i, we also have another ranking of features,

$$\text{Agent } c_{k+1}: \quad x_{\sigma_{k+1}(1)} \preceq \ldots \preceq x_{\sigma_{k+1}(N)}$$

After ranked $\mathscr{X}$ according to a multiple criteria as above, for each agent c_i, we select a subset S_i of $\mathscr{X}$ based on a threshold τ_i. Thus, the set of selected terms is defined by

$$S = \bigcup_{i=1}^{k} S_i \cup S_{k+1} \tag{3}$$

Algorithmically, the process of feature selection is depicted as in Figure 1.

Procedure ABS($\mathscr{X}$ - original feature set, S-optimal feature set, $\tau_1,\ldots,\tau_k,\tau_{k+1}$ - threshold values)

for i =1 **to** k **loop**
    $S_i \leftarrow$ ø;
    Step 1. Ranking all features based on agent c_i;
    Step 2. Choose the first features based on τ_i;
    Step 3. **Return** S_i;
end loop;
    $S \leftarrow S_1 \cup S_2 \cup \ldots \cup S_k$;
    Ranking features based on mutual information
    Choose the first features based on τ_{k+1};
    $S \leftarrow S \cup S_{k+1}$;

Return S

Figure 1: The ABS (Agent-Based Selection) procedure for selecting the optimal feature set

3 Applications in Text Categorization

Text categorization is defined as the problem of assigning a document into one or more predefined categories. Like supervised learning, text categorization has two main steps : pre-processing and classifier building. Pre-processing includes tasks such as feature extraction, feature selection and document representation. After pre-processing, a document will be represented as a vector of features in vector space model or a "bag-of-words" in probabilistic model; features are the components in a vector or a term. Therefore, feature selection plays a very important role in steps later and affects the performance of the whole system.

Features in text categorization usually are selected by one of following measures: document frequency, term frequency, class-based frequency, mutual information, gain entropy, chi-square etc [2],[6],[8],[7],[9]. Between those measures, mutual information measure is one of the most common methods recently [7],[6],[8]. For this reason we use mutual information measure as the baseline method to compare to our proposed method in this paper,

After the pre-processing step, a document is represented by features and these features are inputs for the second step, classifier building. Several machine learning techniques are used for building a classifier, for example, decision tree, neural network, perception, naive Bayes algorithm, genetic algorithms, etc. One of the most common techniques used in text categorization is Naive Bayes and it is viewed as the baseline method for the classifier [9],[6]. In this paper, we use this algorithm as the baseline algorithm for the classifier.

Naive Bayes algorithm

The naive Bayes algorithm is based on probabilistic model, in which each document can be represented as a bag-of-word. It means that just only words existing in documents can be chosen for document representation. Without loss of generality, suppose that a document d' consisting of terms $t_1,t_2,\ldots t_n$. The naive Bayes algorithm calculates the probability of a class belonging to each document with the assumption of independent

variables (attributes). The formulation is based on the Bayes theorem and is given by:

$$P(c_i|d') \propto P(d'|c_i)P(c_i) = P((t_1,t_2,\ldots,t_n)|c_i)P(c_i) = \prod_{j=1}^{k} P(t_j|c_i)P(c_i). \quad (4)$$

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

$$\sigma(d') = \arg\max_{i\in[1\ldots k]} P(c_i|d). \quad (5)$$

4 Experiments

To examine our proposed method, we used a standard text data set Reuters-21578 for our problem [1]. Reuters-21578 is standard data set for text categorization research. The original version was created by Lewis in SGML format and it is preprocessed as some version, for instance, Reuters-21450, Reuters-21173, Reuters-21578, etc, in which Reuters-21578 was seen as the benchmark data set in text categorization community. Reuters-21578, also called the ModApte version, has 21,578 documents and includes 90 categories; we used the subset containing 7,769 documents for training and 3,019 documents for testing. Several documents belong to more than one class: there are 1,192 documents in training data set and 437 documents in testing data set. Within the 90 categories, the top 10 categories are most often used as the standard data set. Top ten categories of this data are selected and 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 was 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. In the baseline method, the number of vocabulary is chosen was 2,000 terms ($\simeq 1/10$ vocabulary), this number is often used for mutual information measure [11].

To compare our method with the baseline method and all term method, we selected features as the same as the baseline for S_{k+1} in the ABS procedure. Each agent is treated equivalently, therefore $\tau_1 = \ldots = \tau_k$ (in this case, $k = 10$) and the threshold values τ_i are chosen as 100 and 200 respectively. That is, $\tau_i = 100, i \in \{1..k\}$ (we called this case ABS-100) and $\tau_i = 200, i \in \{1..k\}$ (we called this case ABS-200). The number of terms in the ABS-100 is 2,314 terms, and the number of terms in the ABS-200 is 2,619 terms. To compare to baseline methods, we chose the threshold value for mutual information is the same as the baseline method, that is $\tau_{k+1} = 2,000$. Experiments are executed in SunOS 5.8 operating system, Perl, sed, awk, C programming languages.

[1] This dataset can be obtained from http://www.daviddlewis.com/resources/testcollection/reuters21578

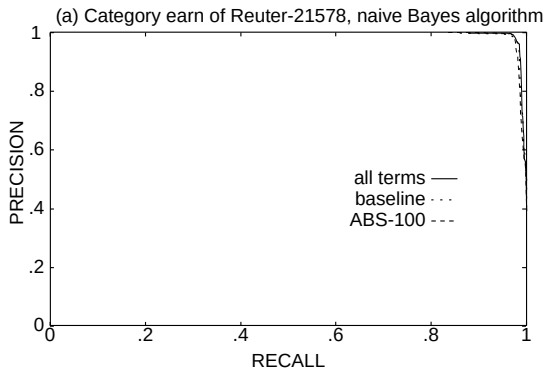

Figure 2: The ROC curves for category `earn` with Reuters-21578 corpus

Table 1 shows the performances BEP and F_1 of text categorization. It is obvious that our proposed method (ABS-100 and ABS-200 have outperformed both all term method and baseline method. The results of all term and baseline methods also the same as the results obtained for feature selection in [11] on Reuters-21578 data set. Further, compared to all term and baseline method, microaveraging of BEP and F_1 were effectively increased. These results also suggests that if thresholds are chosen appropriately, a good performance of the system can be achieved. In this case ABS-100 has the best performance in all experimental results.

Figure 2 describes the ROC curves of the category `earn` and Figure 3 describes the ROC curves of the category `acq`. These two categories have highest number of training documents in Reuters-21578 data set, $2,877$ and $1,650$ documents respectively. It is obvious that the curves show very high performances; the P and R measures are approximately 1.0. In Table 1, the BEP and F_1 measures of the category `earn` are higher when using all term method than for either the ABS-100 or the baseline methods. However, category `acq`, when using all term method achieves higher performances than the baseline method, 96.45% vs. 96.04% with BEP measure and 96.48% vs. 96.21% with F_1; our proposed the ABS-100 method shows better results than either of the other.

In Table 1 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 181 documents respectively. The results show the advantages of using feature selection with our ABS-100 method, with BEP rising from

Table 1: The *BEP* and F_1 of Reuters-21578

Category	*BEP*				F_1			
	all terms	baseline	ABS-100	ABS-200	all terms	baseline	ABS-100	ABS-200
Earn	97.65	97.47	97.43	97.38	98.10	97.91	98.04	98.04
Acq	96.45	96.04	96.60	96.66	96.48	96.21	96.67	96.67
Money-fx	76.54	75.98	76.54	76.19	76.92	75.98	76.54	76.30
Grain	50.34	49.49	51.04	51.50	59.76	54.42	57.47	57.41
Crude	80.00	78.09	78.51	78.51	81.40	79.67	79.43	79.43
Trade	79.15	84.62	84.12	84.12	82.59	85.59	85.60	85.60
Interest	72.52	68.96	70.23	70.23	73.00	73.83	73.38	73.38
Ship	62.92	60.00	59.55	59.55	67.00	67.58	68.75	68.96
Wheat	31.76	40.85	41.13	39.72	39.82	49.24	48.39	47.83
Corn	33.93	35.40	37.50	37.50	35.89	46.40	44.02	44.30
macro ave	68.13	68.69	**69.26**	69.14	71.10	72.68	**72.83**	72.79
micro ave	72.31	74.54	**74.55**	74.55	73.34	73.86	**74.06**	74.03

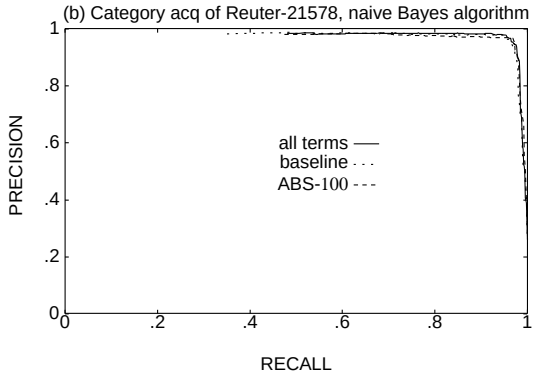

Figure 3: The ROC curves for category `acq` with Reuters-21578 corpus

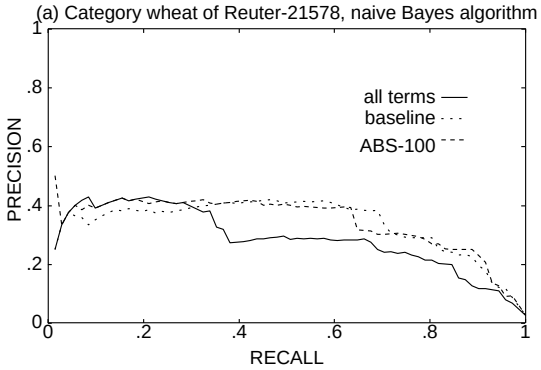

Figure 4: The ROC curves for category `earn` with Reuters-21578 corpus

31.78% when using all term method to 41.13% with the ABS-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 1, the F_1 measures in categories `wheat` and `corn` show that the proposed method is also higher than the all term method. However, our F_1 results are lower than the baseline method. ROC curves of `wheat` and `corn` are shown in Figure 4 and Figure 5.

To illustrate the relative relationship of *BEP*, F_1 and *ROC*, Figure 6 shows the ROC curves for `ship` which the *BEP* measure is lower but the F_1 measure is higher compared to all term method and the baseline.

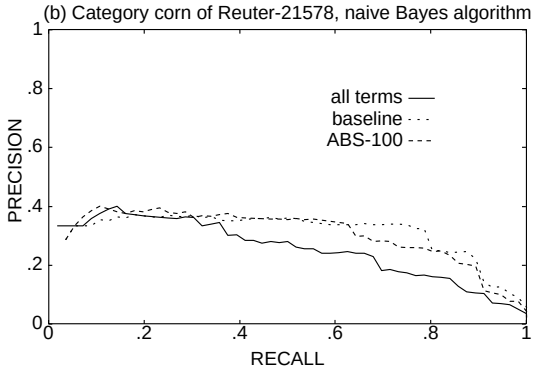

Figure 5: The ROC curves for category `acq` with Reuters-21578 corpus

5 Conclusions

This paper proposed a feature selection approach based on the agent-based approach in text categorization problem. Experimental results shows the following advantages compared to both all term and baseline methods:

1. The proposed method outperformed the performance, including F_1, *BEP* measures and *ROC* curves over baseline methods.

2. The proposed method has better performance, especially for macroaveraging measures.

The adaptation of agents in environments and threshold values for feature selection will be investigated in the future.

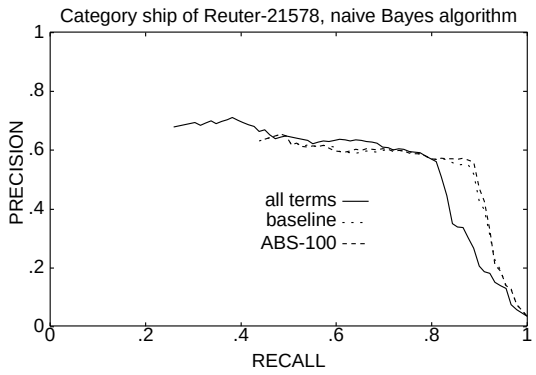

Figure 6: The ROC curves for category `ship` with Reuters-21578 corpus

6 Acknowledgments

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

References

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

[2] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic, 1998.

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

[4] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[5] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2002.

[6] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*, 1:69–90, 1999.

[7] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *Proceeding of the 14th International Conference on Machine Learning (ICML97)*, pages 412–420, 1997.

[8] D. Mladenic. Feature subset selection in text learning. In *Proc of European Conference on Machine Learning(ECML)*, pages 95–100, 1998.

[9] F. Sebastiani. Machine learning in automated text categorization. *ACM computing survey*, 34(1):1–47, 2002.

[10] C. Thornton. Why concept learning is a good idea. In A. Clark and P. Millican, editors, *Connectionism, concepts, and folk psychology (The legacy of Alan Turing)*, volume 2, pages 181–194. Oxford Clarendon press, 2003.

[11] F. Debole and F. Sebastiani. An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and Technology (JASIST)*, 2004. Forthcoming.