MMR-based feature selection for text categorization

Changki Lee
Dept. of Computer Science and Engineering
Pohang University of Science & Technology
San 31, Hyoja-Dong, Pohang,
790-784, South Korea
phone: +82-54-279-5581
leeck@postech.ac.kr

Gary Geunbae Lee
Dept. of Computer Science and Engineering
Pohang University of Science & Technology
San 31, Hyoja-Dong, Pohang,
790-784, South Korea
phone: +82-54-279-5581
gblee@postech.ac.kr

Abstract

We introduce a new method of feature selection for text categorization. Our MMR-based feature selection method strives to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization. Empirical results show that MMR-based feature selection is more effective than Koller & Sahami's method, which is one of greedy feature selection methods, and conventional information gain which is commonly used in feature selection for text categorization. Moreover, MMR-based feature selection sometimes produces some improvements of conventional machine learning algorithms over SVM which is known to give the best classification accuracy.

1 Introduction

Text categorization is the problem of automatically assigning predefined categories to free text documents. A growing number of statistical classification methods and machine learning techniques have been applied to text categorization in recent years [9].

A major characteristic, or difficulty, of text categorization problems is the high dimensionality of the feature space [10]. The native feature space consists of the unique terms that occur in documents, which can be tens or hundreds of thousands of terms for even a moderate-sized text collection. This is prohibitively high for many machine learning algorithms. If we reduce the set of features considered by the algorithm, we can serve two purposes. We can considerably decrease the running time of the learning algorithm, and we can increase the accuracy of the resulting model. In this line, a number of researches have recently addressed the issue of feature subset selection [2][4][8]. Yang and Pederson found information gain (IG) and chi-square test (CHI) most effective in aggressive term removal without losing categorization accuracy in their experiments [8].

Another major characteristic of text categorization problems is the high level of feature redundancy [11]. While there are generally many different features relevant to classification task, often several such cues occur in one document. These cues are partly redundant. Naïve Bayes, which is a popular learning algorithm, is commonly justified using assumptions of conditional independence or linked dependence [12]. However, theses assumptions are generally accepted to be false for text. To remove these violations, more complex dependence models have been developed [13].

Most previous works of feature selection emphasized only the reduction of high dimensionality of the feature space [2][4][8]. The most popular feature selection method is IG. IG works well with texts and has often been used. IG looks at each feature in isolation and measures how important it is for the prediction of the correct class label. In cases where all features are not redundant with each other, IG is very appropriate. But in cases where many features are highly redundant with each other, we must utilize other means, for example, more complex dependence models.

In this paper, for the high dimensionality of the feature space and the high level of feature redundancy, we propose a new feature selection method which selects each feature according to a combined criterion of information gain and novelty of information. The latter measures the degree of dissimilarity between the feature being considered and previously selected features. Maximal Marginal Relevance (MMR) provides precisely such functionality [5]. So we propose MMR-based feature selection method which strives to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization.

In machine learning field, some greedy methods that add or subtract a single feature at a time have been developed for feature selection [3][14]. S. Della Pietra et al. proposed a method for incrementally constructing random field [14]. Their method builds increasingly complex fields to approximate the empirical distribution of a set of training examples by allowing features. Features are incrementally added to the field using a top-down greedy algorithm, with the intent of capturing the

salient properties of the empirical sample while allowing generalization to new configurations. However the method is not simple, and this is problematic both computationally and statistically in large-scale problems.

Koller and Sahami proposed another greedy feature selection method which provides a mechanism for eliminating features whose predictive information with respect to the class is subsumed by the other features [3]. This method is also based on the Kullback-Leibler divergence to minimize the amount of predictive information lost during feature elimination.

In order to compare the performances of our method and greedy feature selection methods, we implemented Koller and Sahami's method, and empirically tested it in section 4.

We also compared the performance of conventional machine learning algorithms using our feature selection method with Support Vector Machine (SVM) using all features in section 4. Previous works show that SVM consistently achieves good performance on text categorization tasks, outperforming existing methods substantially and significantly [10][11]. With its ability to generalize well in high dimensional feature spaces and high level of feature redundancy, SVM is known that it does not need any feature selection [11].

The remainder of this paper is organized as follows. In section 2, we describe the Maximal Marginal Relevance, and in section 3, we describe the MMR-based feature selection. Section 4 presents the in-depth experiments and the results. Section 5 concludes the research.

2 Maximal Marginal Relevance

Most modern IR search engines produce a ranked list of retrieved documents ordered by declining relevance to the user's query. In contrast, the need for 'relevant novelty' was motivated as a potentially superior criterion. A first approximation to relevant novelty is to measure the relevance and the novelty independently and provide a linear combination as the metric.

The linear combination is called 'marginal relevance' - i.e. a document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents. In document retrieval and summarization, marginal relevance is strived to maximize, hence the method is labeled 'Maximal Marginal Relevance' (MMR) [5].

$$MMR = \underset{D_i \in R \setminus S}{Arg\max}\left[\lambda \cdot Sim_1(D_i, Q) - (1-\lambda)\max_{D_j \in S} Sim_2(D_i, D_j)\right]$$

where $C=\{D1,...,Di,...\}$ is a document collection (or document stream); Q is a query or user profile; $R = IR(C, Q, \theta)$, i.e., the ranked list of documents retrieved by an IR system, given C and Q and a relevance threshold θ, below which it will not retrieve documents (θ can be degree of match or number of documents); S is the subset of documents in R which is already selected; $R \setminus S$ is the set difference, i.e. the set of as yet unselected documents in R; Sim_1 is the similarity metric used in document retrieval and relevance ranking between documents (passages) and a query; and Sim_2 can be the same as Sim_1 or a different metric.

3 MMR-based Feature Selection

We propose a MMR-based feature selection which selects each feature according to a combined criterion of information gain and novelty of information. We define MMR-based feature selection as follows:

$$MMR\_FS = \underset{w_i \in R \setminus S}{Arg\max}\left[\lambda \cdot IG(w_i; C) - (1-\lambda)\max_{w_j \in S} IGpair(w_i; w_j \mid C)\right]$$

where C is the set of class labels, R is the set of candidate features, S is the subset of features in R which was already selected, R\S is the set difference, i.e. the set of as yet unselected features in R, IG is the information gain scores, and IGpair is the information gain scores of co-occurrence of the word (feature) pairs. IG and IGpair are defined as follows:

$$\begin{aligned} IG(w_i; C) = &-\sum_k p(C_k)\log p(C_k) \\ &+ p(w_i)\sum_k p(C_k \mid w_i)\log p(C_k \mid w_i) \\ &+ p(\overline{w}_i)\sum_k p(C_k \mid \overline{w}_i)\log p(C_k \mid \overline{w}_i) \end{aligned}$$

$$\begin{aligned} IGpair(w_i; w_j \mid C) = &-\sum_k p(C_k)\log p(C_k) \\ &+ p(w_{i,j})\sum_k p(C_k \mid w_{i,j})\log p(C_k \mid w_{i,j}) \\ &+ p(\overline{w}_{i,j})\sum_k p(C_k \mid \overline{w}_{i,j})\log p(C_k \mid \overline{w}_{i,j}) \end{aligned}$$

where $p(w_i)$ is the probability that word w_i occurred, $\overline{w}_i$ means that word w_i doesn't occur, $p(C_k)$ is the probability of the k-th class value, $p(C_k|w_i)$ is the conditional probability of the k-th class value given that w_i occurred, $p(w_{i,j})$ is the probability that w_i and w_j co-occurred, and $\overline{w}_i$ means that w_i and w_j doesn't co-occur but w_i or w_j can occur (i.e. $p(\overline{w}_{i,j}) = 1 - p(w_{i,j})$).

Given the above definition, MMR_FS computes incrementally the information gain scores when the parameter λ =1, and computes a maximal diversity among the features in R when λ=0. For intermediate values of λ in the interval [0,1], a linear combination of both criteria is optimized.

4 Experiments

In order to compare the performance of MMR-based feature selection method with conventional IG and

greedy feature selection method (Koller & Sahami's method, labeled 'Greedy'), we evaluated the three feature selection methods with four different learning algorithms: naive Bayes, TFIDF/Rocchio, Probabilistic Indexing (PrTFIDF [7]) and Maximum Entropy using Rainbow [6].

We also compared the performance of conventional machine learning algorithms using our feature selection method and SVM using all features.

MMR-based feature selection and greedy feature selection method (Koller & Sahami's method) requires quadratic time with respect to the number of features. To reduce this complexity, for each data set, we first selected 1000 features using IG, and then we applied MMR-based feature selection and greedy feature selection method to the selected 1000 features.

For all datasets, we did not remove stopwords. The results reported on all dataset are averaged over 10 times of different test/training splits. A random subset of 20% of the data considered in an experiment was used for testing (i.e. we used Rainbow's '--test-set=0.2' and '--test=10' options), because Rainbow does not support 10-fold cross validation.

MMR-based feature selection method needs to tune for λ. It appears that a tuning method based on held-out data is needed here. We tested our method using 11 λ values (i.e. 0, 0.1, 0.2, …, 1) and selected the best λ value.

4.1 Reuters-21578

The Reuters-21578 corpus contains 21578 articles taken from the Reuters newswire. Each article is typically designated into one or more semantic categories such as 'earn', 'trade', 'corn' etc., where the total number of categories is 114.

Following [3], we constructed a subset from Reuter corpus. The subset is comprised of articles on the topic 'coffee', 'iron-steel', and 'livestock'.

4.2 WebKB

This data set contains WWW-pages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base (WebKb) project of the CMU text learning group. The 8282 pages were manually classified into 7 categories: 'course', 'department', 'faculty', 'project', 'staff', 'student' and 'other'. Following [1], we discarded the categories 'other', 'department' and 'staff'. The remaining part of the corpus contains 4199 documents in four categories.

4.3 Experimental Results

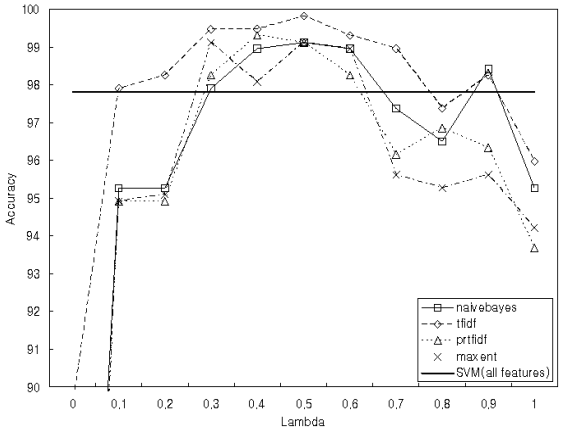

Figure 1. MMR feature selection for four machine learning algorithms on Reuters (#features=25)

Figure 1 displays the performance curves for four different machine learning algorithms on the subset of Reuters after term selection using MMR-based feature selection (number of features is 25). When the parameter λ=0.5, most machine learning algorithms have best performance and significant improvements compared to conventional information gain (i.e. λ=1) and SVM using all features.

Table 1. WebKB.

		Number of features						
		12	25	50	100	200	400	All features
SVM	-	-	-	-	-	-	-	**92.37**
Naïve Bayes	MMR 0.7	78.75	82.31	86.22	89.19	**90.49**	90.10	85.26
	Greedy	81.17	83.34	**87.22**	86.98	87.03	85.89	
	IG	79.31	83.54	83.46	**87.71**	84.96	84.65	
TFIDF	MMR 0.5	84.99	87.56	87.45	88.47	87.35	**89.00**	89.76
	Greedy	78.31	82.36	85.63	**86.87**	86.85	85.96	
	IG	83.62	82.03	83.21	**86.67**	86.42	86.17	
PrTFIDF	MMR 0.6	71.00	79.02	81.30	**82.47**	81.01	81.92	61.75
	Greedy	66.10	74.46	**74.56**	68.26	64.72	61.32	
	IG	**72.16**	70.06	67.25	68.36	62.21	59.21	
Maximum Entropy	MMR 0.8	78.88	83.30	86.25	89.76	90.87	**92.05**	89.34
	Greedy	74.33	83.72	87.90	90.63	90.92	**91.50**	
	IG	78.82	84.85	86.81	90.73	**91.61**	91.32	

Table 1 shows the performance of four machine learning algorithms on WebKB using three feature selection methods and all features (41763 terms). In this data set, again MMR-based feature selection has best performance and significant improvements compared to greedy method and IG. Using MMR-based feature selection, for example, the vocabulary is reduced from 41763 terms to 200 (a 99.5% reduction), and the accuracy is improved from 85.26% to 90.49% in Naïve Bayes. Using greedy method and IG, however, the accuracy is improved from 85.26% to about 87% in Naïve

Bayes. PrTFIDF is most sensitive to feature selection method. Using MMR-based feature selection the best accuracy is 82.47%. Using greedy method and IG, however, the best accuracy is only 72~74%. In this dataset, however, MMR-based feature selection does not produce improvements of conventional machine learning algorithms over SVM.

The observation in Reuters and WebKB are highly consistent. MMR-based feature selection is consistently more effective than greedy method and IG on two data sets, and sometimes produces improvements even over the best SVM.

5 Conclusion

In this paper, we proposed a MMR-based feature selection method which strives to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization.

We carried out extensive experiments to verify the proposed method. Based on the experiment results, we can verify that MMR-based feature selection is more effective than Koller & Sahami's method, which is one kind of greedy methods, and conventional information gain which is commonly used in feature selection for text categorization. Besides, MMR-based feature selection method sometimes produces improvements of conventional machine learning algorithms over SVM which is known to give the best classification accuracy.

A disadvantage in using MMR-based feature selection is that the computational cost of computing the pairwise information gain (i.e. IGpair) is quadratic time with respect to the number of features. To reduce this computational cost, we can use MMR-based feature selection method on the reduced feature set resulting from IG as our experiments in section 4. Another drawback of our method is the need to tune for λ. It appears that a tuning method based on held-out data is needed here

References

[1] Andrew Mccallum and Kamal Nigam. 1998. *A Comparison of Event Models for Naive Bayes Text Classification.* In AAAI-98 Workshop on Learning for Text Categorization.

[2] David D. Lewis and Marc Ringuette. 1994. *A Comparison of Two Learning Algorithms for Text Categorization.* In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval.

[3] Daphne Koller and Mehran Sahami. 1996. *Toward Optimal Feature Selection.* In Proceedings of ICML-96, 13th International Conference on Machine Learning.

[4] Hinrich Schütze and David A. Hull, and Jan O. Pedersen. 1995. *A Comparison of Classifiers and Document Representations for the Routing Problem.* In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[5] Jaime Carbonell and Jade Goldstein. 1998. *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries.* In Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval.

[6] McCallum and Andrew Kachites. 1996. *Bow: A toolkit for statistical language modelling, text retrieval, classification and clustering.* http://www.cs.cmu.edu/~mccallum/bow.

[7] Thorsten Joachims. 1997. *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.* In Proceedings of ICML-97, 14th International Conference on Machine Learning.

[8] Yiming Yang and Jan O. Pedersen. 1997. *A Comparative Study on Feature Selection in Text Categorization.* In Proceedings of ICML-97, 14th International Conference on Machine Learning.

[9] Yiming Yang and Xin Liu. 1999. *A re-examination of text categorization methods.* In Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval.

[10] Thorsten Joachims. 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* In Proceedings of ECML-98, 10th European Conference on Machine Learning.

[11] Thorsten Joachims. 2001. *A Statistical Learning Model of Text Classification for Support Vector Machines.* In Proceedings of the 24th ACM-SIGIR International Conference on Research and Development in Information Retrieval.

[12] William S. Cooper. 1991. *Some Inconsistencies and Misnomers in Probabilistic Information Retrieval.* In Proceedings of the 14th ACM SIGIR International Conference on Research and Development in Information Retrieval.

[13] Mehran Sahami. 1998. *Using Machine Learning to Improve Information Access.* PhD thesis, Stanford University.

[14] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. *Inducing Features of Random Fields.* IEEE Transactions on Pattern Analysis and Machine Intelligence.