Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 155–165

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Information gain and divergence-based feature selection for machine learning-based text categorization ☆

Changki Lee *, Gary Geunbae Lee *

Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31 Hyoja dong, Nam Gu, Pohang 790-784, Korea (South)

Received 2 February 2004; accepted 11 August 2004

Abstract

Most previous works of feature selection emphasized only the reduction of high dimensionality of the feature space. But in cases where many features are highly redundant with each other, we must utilize other means, for example, more complex dependence models such as Bayesian network classifiers. In this paper, we introduce a new information gain and divergence-based feature selection method for statistical machine learning-based text categorization without relying on more complex dependence models. Our feature selection method strives to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization. Empirical results are given on a number of dataset, showing that our feature selection method is more effective than Koller and Sahami's method [Koller, D., & Sahami, M. (1996). Toward optimal feature selection. In *Proceedings of ICML-96, 13th international conference on machine learning*], which is one of greedy feature selection methods, and conventional information gain which is commonly used in feature selection for text categorization. Moreover, our feature selection method sometimes produces more improvements of conventional machine learning algorithms over support vector machines which are known to give the best classification accuracy.

Keywords: Text categorization; Feature selection; Information gain and divergence-based feature selection

☆ This work was supported by the Brain Korea 21 Project (Ministry of Education) in 2003. The short and preliminary version of this paper was published in HLT/NAACL2004.

* Corresponding authors. Tel.: +82 54 279 5581; fax: +82 54 279 2299 (C. Lee), tel.: +82 54 279 2254; fax: +82 54 279 2299 (G.G. Lee).

E-mail addresses: leeck@postech.ac.kr (C. Lee), gblee@postech.ac.kr (G.G. Lee).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.08.006

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 (Yang & Liu, 1999).

A major characteristic, or difficulty, of text categorization problems is the high dimensionality of the feature space (Joachims, 1998). The native feature space consists of the unique terms (words or phrases) 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 (Lewis & Ringuette, 1994; Schutze, Hull, & Pedersen, 1995; Yang & Pedersen, 1997). 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 (Yang & Pedersen, 1997). They also discovered that IG and CHI scores of a term are strongly correlated.

Another major characteristic of text categorization problems is the high level of feature redundancy (Joachims, 2001). While there are generally many different features relevant to a classification task, often several such cues occur in one document, and these cues are partly redundant. Naive Bayes, which is a popular learning algorithm, is commonly justified using assumptions of conditional independence or linked dependence (Cooper, 1991). However, theses assumptions are generally accepted to be false for text. To remove these violations, more complex dependence models such as Bayesian network classifiers have been developed (Sahami, 1998), but they require complex models by trading efficiency.

Most previous works of feature selection emphasized only the reduction of high dimensionality of the feature space (Lewis & Ringuette, 1994; Schutze et al., 1995; Yang & Pedersen, 1997). 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 the previously selected features. MMR provides precisely such functionality (Carbonell & Goldstein, 1998). 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 (Koller & Sahami, 1996; Pietra, Pietra, & Lafferty, 1997). Pietra et al. proposed a method for incrementally constructing random field (Pietra et al., 1997). Their method builds increasingly complex fields to approximate the empirical distribution of a set of training examples by allowing potential functions, or features, which are supported by increasingly large sub-graphs. Each feature is assigned a weight, and the weights are trained to minimize the Kullback–Leibler divergence between the field and the empirical distribution of the training data. Features are incrementally added to the field using a top–down greedy algorithm, with the intent of capturing the salient properties of the empirical samples 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 (Koller & Sahami, 1996). 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 that of support vector machine (SVM) in Section 4. SVM is a new learning method introduced by Vapnik et al. (Vapnik, 1998). Previous works show that SVM consistently achieves good performance on text categorization tasks, outperforming existing methods substantially and significantly (Joachims, 1998, 2001). With its ability to generalize well in high dimensional feature spaces and high level of feature redundancy, SVM is known that it does not help much with sophisticated feature selection (Joachims, 2001).

The remainder of this paper is organized as follows. In Section 2, we describe the information gain and divergence-based feature selection. Section 3 presents in-depth experiments, discussions and the results. Section 4 concludes the research.

2. Information gain and divergence-based feature selection

In this section, we describe the maximal marginal relevance (MMR) and the MMR-based feature selection.

2.1. 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 the previously selected documents. In document retrieval and summarization, marginal relevance should be maximized, hence the method is labeled as MMR (Carbonell & Goldstein, 1998).

$$\mathrm{MMR} = \arg\max_{D_i \in R\backslash S}\left[\lambda \cdot \mathrm{Sim}_1(D_i, Q) - (1-\lambda)\max_{D_j \in S}\mathrm{Sim}_2(D_i, D_j)\right]$$

where $C = \{D_1, \ldots, D_i, \ldots\}$ is a document collection (or document stream); Q is a query or user profile; $R = \mathrm{IR}(C, Q, \theta)$, 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 a degree of match or number of documents); S is the subset of documents in R which is already selected; $R\backslash S$ 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 a relevance ranking between documents (passages) and a query; and Sim_2 can be the same as Sim_1 or a different metric.

Given the above definition, MMR computes incrementally the standard relevance-ranked list when the parameter $\lambda = 1$, and computes a maximal diversity ranking among the documents in R when $\lambda = 0$. For intermediate values of λ in the interval $[0, 1]$, a linear combination of both criteria should be optimized.

2.2. MMR-based feature selection

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

$$\text{MMR\_FS} = \arg\max_{D_i \in R \setminus S} \left[\lambda \cdot \text{IG}(w_i; C) - (1-\lambda) \max_{w_j \in S} \text{IGpair}(w_i; w_j | C) \right]$$

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, $R \setminus S$ 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:

$$\begin{aligned} \text{IG}(w_i; C) &= -\sum_k p(C_k) \log p(C_k) + p(w_i) \sum_k p(C_k|w_i) \log p(C_k|w_i) \\ &\quad + p(\bar{w}_i) \sum_k p(C_k|\bar{w}_i) \log p(C_k|\bar{w}_i) \\ \text{IGpair}(w_i; w_j|C) &= -\sum_k p(C_k) \log p(C_k) + p(w_{i,j}) \sum_k p(C_k|w_{i,j}) \log p(C_k|w_{i,j}) \\ &\quad + p(\bar{w}_{i,j}) \sum_k p(C_k|\bar{w}_{i,j}) \log p(C_k|\bar{w}_{i,j}) \end{aligned}$$

where $p(w_i)$ is the probability that word w_i occur, $\bar{w}_i$ means that word w_i does not occur, $p(C_k)$ is the probability of the kth class value, $p(C_k|w_i)$ is the conditional probability of the kth class value given that w_i occur, $p(w_{i,j})$ is the probability that w_i and w_j co-occur, and $\bar{w}_{i,j}$ means that w_i and w_j does not co-occur but w_i or w_j can occur (i.e. $p(\bar{w}_{i,j}) = 1 - p(w_{i,j})$).

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

3. Experiments

In order to compare the performance of MMR_FS 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 (Joachims, 1997)) and Maximum Entropy using Rainbow (McCallum & Kachites, 1996). Rainbow is an executable program that does document classification. It provides Naive Bayes, TFIDF/Rocchio, Probabilistic Indexing (PrTFIDF), K-nearest neighbor, Maximum Entropy and SVM.

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

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

For all datasets, we did not remove stopwords. We performed 10-fold cross validation for all datasets. MMR_FS 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 11λ values (i.e. $0, 0.1, 0.2, \ldots, 1$) and selected the best λ values.

3.1. Datasets

3.1.1. 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 (Koller & Sahami, 1996), we construct two subsets from Reuter corpus. The first, Reuters1, consists of articles on the topic 'coffee', 'iron-steel', and 'livestock'. These topics are not likely to have many meaningful overlapping words. The second set, Reuters2, contains articles on 'reserves', 'gold', and 'gnp', which are likely to have many similar words used in different contexts across topics. We also construct Reuters3 which consists of articles on the most frequent topic 'earn', 'acq', and 'money-fx'.

3.1.2. WebKB

This dataset 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 seven categories: 'course', 'department', 'faculty', 'project', 'staff', 'student' and 'other'. Following (McCallum & Nigam, 1998), we discard the categories 'other', 'department' and 'staff'. The remaining part of the corpus contains 4199 documents in four categories.

Since the web pages are in HTML format, they contain much non-textual information. We filtered out MIME headers and HTML tags.

3.2. Experimental results

Fig. 1 displays the performance curves for four different machine learning algorithms on Reuters1 after term selection using MMR_FS (number of features is 25). When the parameter $\lambda = 0.5$, all machine learning algorithms have best performance and significant improvements compared to the conventional information

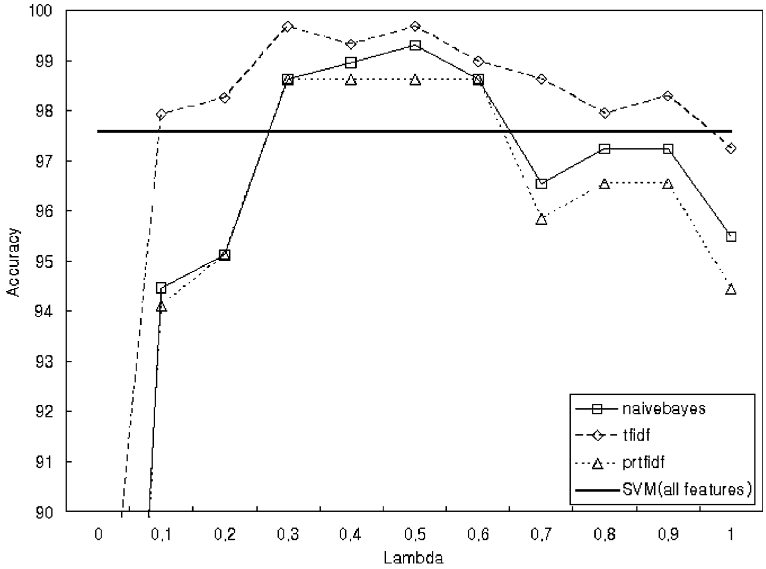


Fig. 1. MMR-based feature selection for three machine learning algorithm on Reuters1 (#features = 25).

gain (i.e. $\lambda = 1$) with a statistical significance at the 99% level. Moreover, all machine learning algorithms have improvements compared to SVM using all features.

Figs. 2–4 display the performance curves for Naive Bayes, TFIDF, and PrTFIDF on Reuters1 using three feature selection methods (MMR_FS with $\lambda = 0.5$, Greedy, and conventional IG), all features with no feature selection (5595 terms), and SVM using all features, respectively. As seen from these figures, MMR_FS is more effective than Greedy and IG, and moreover, produces improvements of conventional

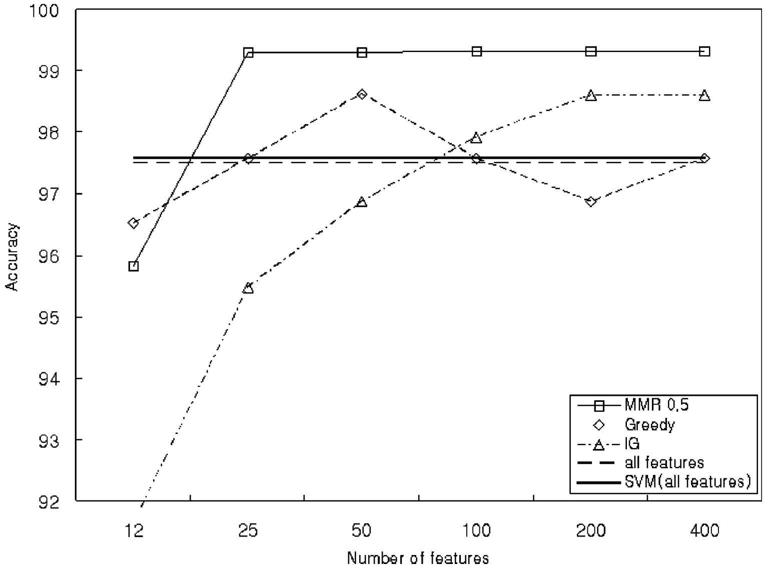


Fig. 2. Naive Bayes on Reuters1.

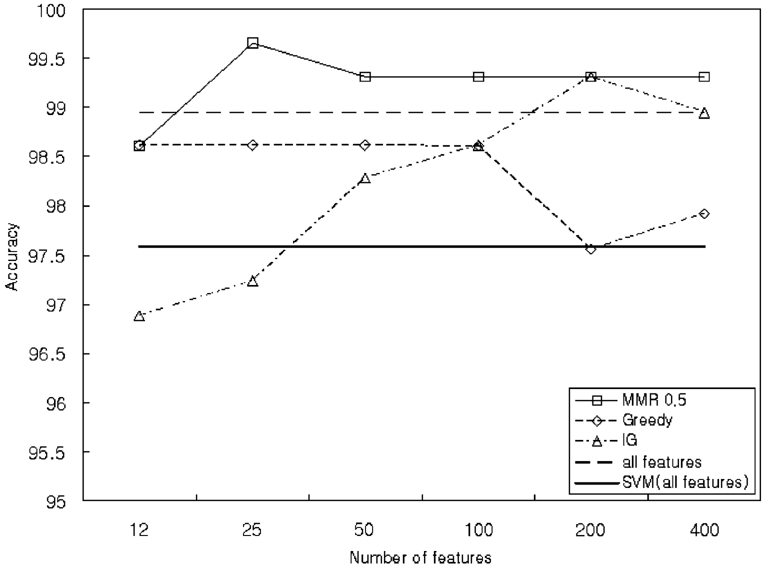


Fig. 3. TFIDF on Reuters1.

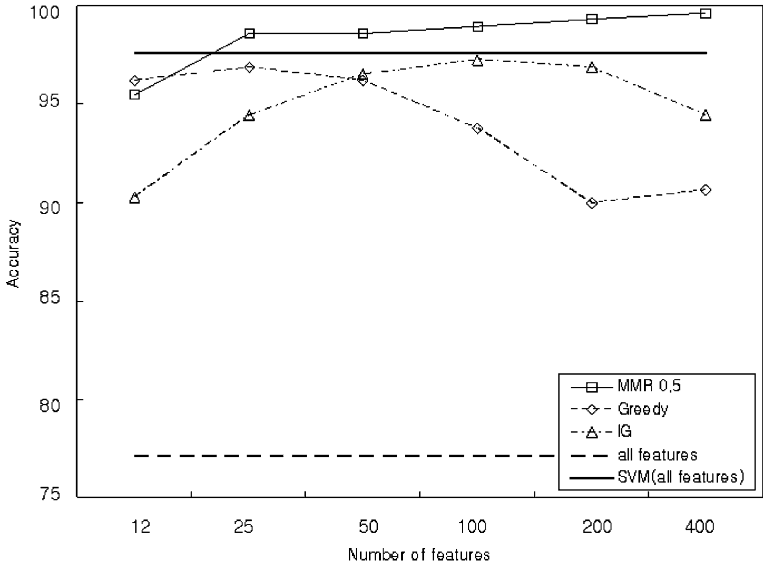


Fig. 4. PrTFIDF on Reuters1.

machine learning algorithms over SVM which is known to give the best classification accuracy. MMR_FS can reduce the feature vocabulary from 5595 terms to 25–100 with best performance in accuracy. For example, the vocabulary is reduced from 5595 terms to 100 (98.2% reduction), and the accuracy is improved from 97.50% to 99.31% in Naive Bayes (the micro-averaged F1 is improved from 97.59 to 99.31).

Fig. 5 shows the same performance curves on Reuters2 after term selection using MMR_FS (number of features is 400). In this dataset, most machine learning algorithms also have best performance and significant

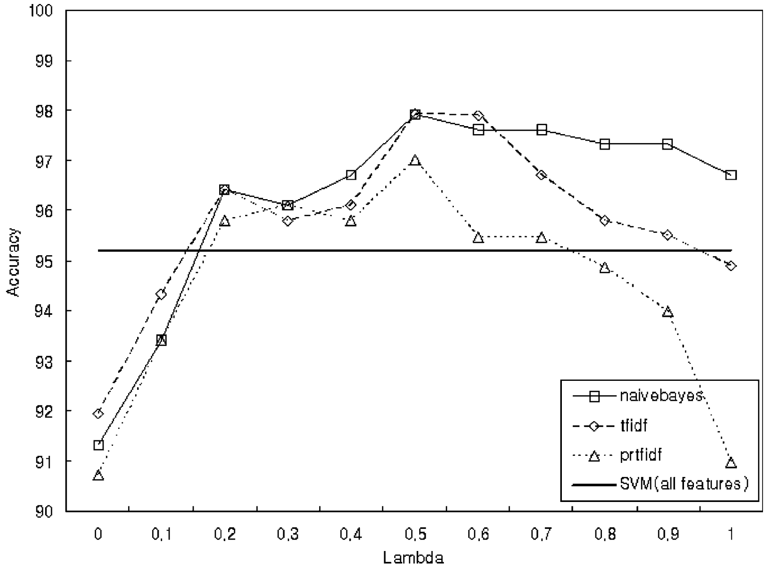


Fig. 5. MMR-based feature selection for three machine learning algorithms on Reuters2 (#features = 400).

improvements compared to conventional IG (with a statistical significance at the 90% level) and SVM using all features, when the parameter $\lambda = 0.5$.

Figs. 6–8 show the performance curves for Naive Bayes. TFIDF, and PrTFIDF on Reuters2 using the three feature selection methods, all features (5478 terms) and SVM using all features, respectively. In these figures, MMR_FS is also more effective than greedy method and IG, and moreover, produces improve-

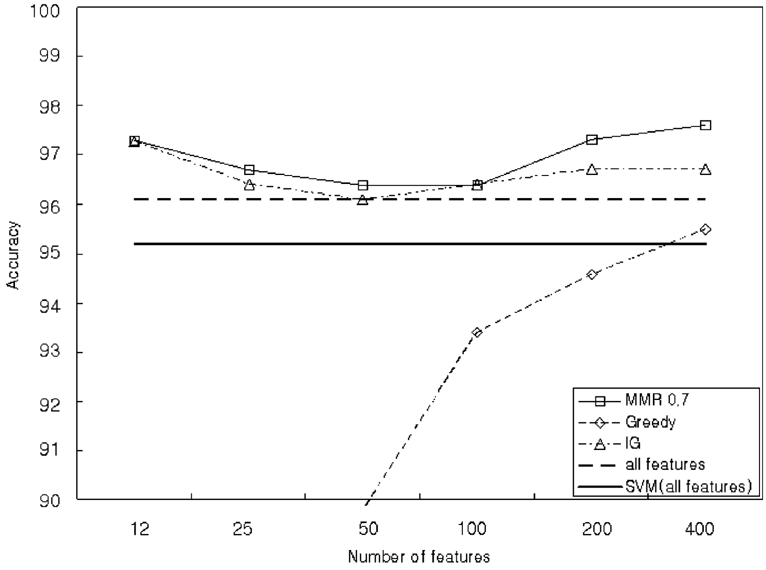


Fig. 6. Naive Bayes on Reuters2.

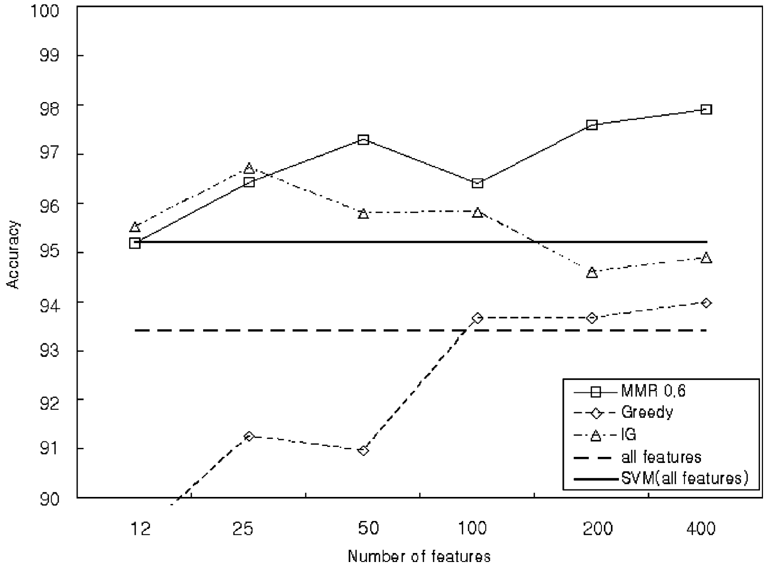


Fig. 7. TFIDF on Reuters2.

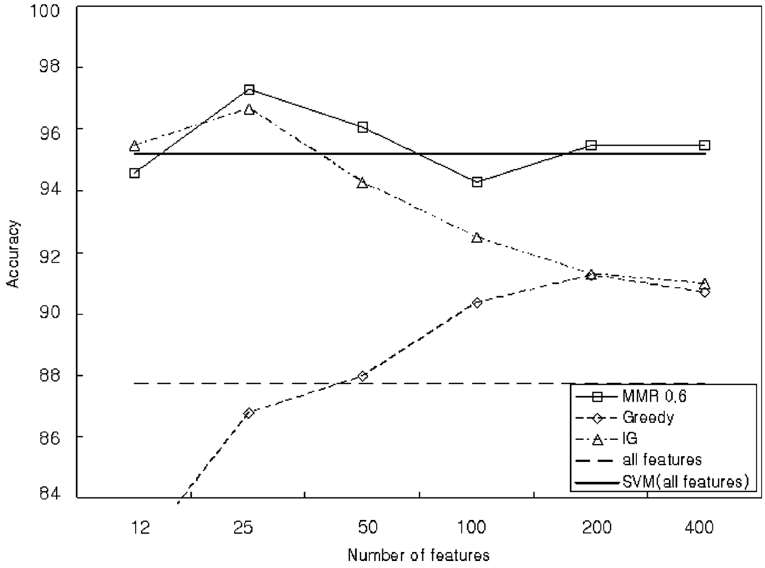
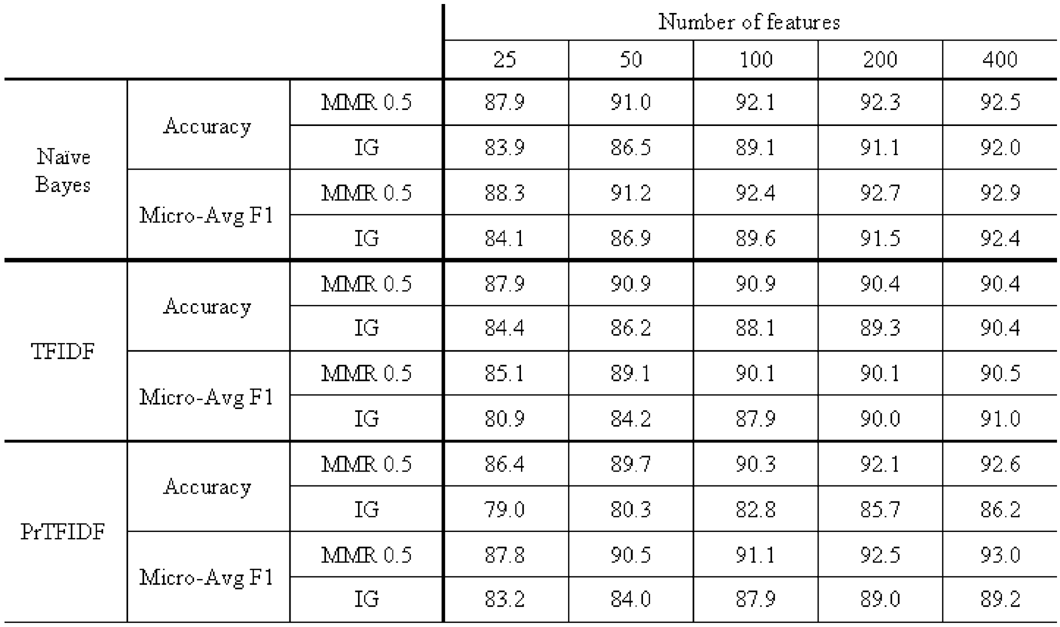


Fig. 8. PrTFIDF on Reuters2.

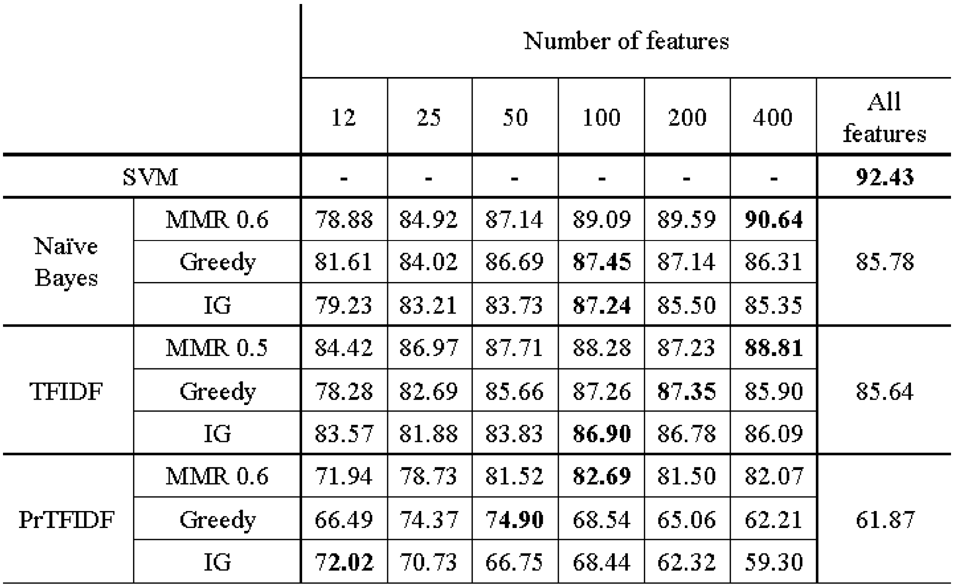
ments over SVM. MMR_FS can reduce the vocabulary from 5478 terms to 200–400 with best performance in accuracy for the corpora even with many similar overlapping terms such as Reuters2.

Fig. 9 shows the performance of three machine learning algorithms on Reuters3 using MMR-based feature selection method. In this dataset, MMR_FS has significant improvements compared to IG. Using



			Number of features				
			25	50	100	200	400
Naïve Bayes	Accuracy	MMR 0.5	87.9	91.0	92.1	92.3	92.5
		IG	83.9	86.5	89.1	91.1	92.0
	Micro-Avg F1	MMR 0.5	88.3	91.2	92.4	92.7	92.9
		IG	84.1	86.9	89.6	91.5	92.4
TFIDF	Accuracy	MMR 0.5	87.9	90.9	90.9	90.4	90.4
		IG	84.4	86.2	88.1	89.3	90.4
	Micro-Avg F1	MMR 0.5	85.1	89.1	90.1	90.1	90.5
		IG	80.9	84.2	87.9	90.0	91.0
PrTFIDF	Accuracy	MMR 0.5	86.4	89.7	90.3	92.1	92.6
		IG	79.0	80.3	82.8	85.7	86.2
	Micro-Avg F1	MMR 0.5	87.8	90.5	91.1	92.5	93.0
		IG	83.2	84.0	87.9	89.0	89.2

Fig. 9. MMR-based feature selection for three machine learning algorithms on Reuters3.



		Number of features						
		12	25	50	100	200	400	All features
SVM		-	-	-	-	-	-	**92.43**
Naïve Bayes	MMR 0.6	78.88	84.92	87.14	89.09	89.59	**90.64**	85.78
	Greedy	81.61	84.02	86.69	**87.45**	87.14	86.31	
	IG	79.23	83.21	83.73	**87.24**	85.50	85.35	
TFIDF	MMR 0.5	84.42	86.97	87.71	88.28	87.23	**88.81**	85.64
	Greedy	78.28	82.69	85.66	87.26	**87.35**	85.90	
	IG	83.57	81.88	83.83	**86.90**	86.78	86.09	
PrTFIDF	MMR 0.6	71.94	78.73	81.52	**82.69**	81.50	82.07	61.87
	Greedy	66.49	74.37	**74.90**	68.54	65.06	62.21	
	IG	**72.02**	70.73	66.75	68.44	62.32	59.30	

Fig. 10. WebKB.

MMR_FS, for example, the accuracy is improved from 86.5% to 91.0% and the micro-averaged F1 is improved from 86.9 to 91.2 in Naive Bayes (number of feature is 50).

Fig. 10 shows the performance of three machine learning algorithms on WebKB also using the same three feature selection methods and all features (41763 terms). In this dataset, again MMR_FS has the best performance and significant improvements compared to Greedy and IG. Using MMR_FS, for example, the vocabulary is reduced from 41763 terms to 200 (99.5% reduction), and the accuracy is improved from 85.78% to 90.64% in Naive Bayes. Using Greedy and IG, however, the accuracy is improved from 85.78% to about 87% in Naive Bayes. PrTFIDF is most sensitive to feature selection method. Using MMR_FS the best accuracy is 82.69%. Using Greedy and IG, however, the best accuracy is only 72–74%. In this dataset, however, MMR_FS does not produce improvements of conventional machine learning algorithms over SVM.

The observation in Reuters and WebKB are highly consistent. MMF_FS method is consistently more effective than Greedy and IG method on two datasets, and sometimes produces improvements of simple conventional classifiers even over state-of-the-art SVM.

4. Conclusion

In this paper, we proposed an information gain and divergence-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. The experiments were performed using three different machine learning algorithms on both Reuters and WebKB dataset. Based on the experiment results, we can verify that our MMR-based feature selection is more effective than Koller and 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 simple conventional machine learning algorithms over SVM which are 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 the 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 systematic tuning method based on held-out data is needed in the future.

References

Carbonell, J., & Goldstein, J. (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*.

Cooper, W. S. (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*.

Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML-97, 14th international conference on machine learning*.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European conference on machine learning*.

Joachims, T. (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*.

Koller, D., & Sahami, M. (1996). Toward optimal feature selection. In *Proceedings of ICML-96, 13th international conference on machine learning*.

Lewis, D. D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*.

McCallum, & Kachites, A. (1996). Bow: a toolkit for statistical language modelling, text retrieval, classification and clustering. Available from <http://www.cs.cmu.edu/mccallum/bow>.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*.

Pietra, S. D., Pietra, & V. D., Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sahami, M. (1998). *Using machine learning to improve information access.* PhD thesis, Stanford University.

Schutze, H., Hull, D. A., & Pedersen, J. O. (1995). Toward optimal feature selection. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Chichester, GB: Wiley.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th international conference on machine learning*.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd ACM-SIGIR international conference on research and development in information retrieval*.