Classification of Text Documents

Y. H. LI AND A. K. JAIN

Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA
Email: liyongho,jain@cps.msu.edu

The exponential growth of the internet has led to a great deal of interest in developing useful and efficient tools and software to assist users in searching the Web. Document retrieval, categorization, routing and filtering can all be formulated as classification problems. However, the complexity of natural languages and the extremely high dimensionality of the feature space of documents have made this classification problem very difficult. We investigate four different methods for document classification: the naive Bayes classifier, the nearest neighbour classifier, decision trees and a subspace method. These were applied to seven-class Yahoo news groups (business, entertainment, health, international, politics, sports and technology) individually and in combination. We studied three classifier combination approaches: simple voting, dynamic classifier selection and adaptive classifier combination. Our experimental results indicate that the naive Bayes classifier and the subspace method outperform the other two classifiers on our data sets. Combinations of multiple classifiers did not always improve the classification accuracy compared to the best individual classifier. Among the three different combination approaches, our adaptive classifier combination method introduced here performed the best. The best classification accuracy that we are able to achieve on this seven-class problem is approximately 83%, which is comparable to the performance of other similar studies. However, the classification problem considered here is more difficult because the pattern classes used in our experiments have a large overlap of words in their corresponding documents.

1. INTRODUCTION

The World Wide Web (WWW) is a widely distributed and dynamic information gallery. A pessimistic estimate is that, in the last half of 1996, the www consisted of more than 60 million documents on 12 million hosts and 600,000 servers, up from 9 million hosts and 250,000 servers at the beginning of the year, and these numbers are increasing every day. These Web documents contain rich textual information, but the rapid growth of the internet has made it increasingly difficult for users to locate the relevant information quickly on the Web. This has led to a great deal of interest in developing useful and efficient tools and software to assist users in searching on the Web.

Search engines were designed to reduce the effort and information overload on the Web. Commercial search engines, such as Yahoo, HotBot, InfoSeek, WebCrawler and Lycos, etc. are examples of tools that construct indices and find information requested by a user. However, it is not uncommon that search engine queries often return some sites that have little to do with user interests. This has led to the development of intelligent agents which are playing an important role in making the internet more usable [1, 2, 3].

Document retrieval [4], categorization [5], routing [6] and filtering systems (agents) [1, 2] are often based on text classification. A typical classification problem can be stated as follows: given a set of labelled examples belonging to two or more classes (training data), we classify a new test sample to a class with the highest similarity. Document retrieval, routing and filtering systems, can often be viewed as a two-class classification problem where a document is labelled as relevant or non-relevant. User feedback provides a set of training examples with positive and negative labels. A document is presented to the user if it is classified as the relevant class. In document categorization, which is the topic of classification of USENET news groups [5], we already have human indexed training data available. A classifier is used to automatically determine to which news-group a new document should be posted.

Text classification presents many challenges and difficulties. First, it is difficult to capture high-level semantics and abstract concepts of natural languages just from a few key words. For instance, there are many ways to represent similar concepts (e.g. agent, softbot, robot, or bot) and the same word can represent different meanings (e.g. bank can be either related to a finance problem or a river). Furthermore, semantic analysis, which is a major step in designing a natural language information retrieval system, is not well understood, although there are some techniques that have been successfully applied to limited domains [7]. Second, high dimensionality (thousands of features) and variable length, content and quality are the characteristics of a huge number of documents on the Web. These place both efficiency and accuracy demands on classification systems.

A number of methods have been discussed in the literature for document classification. These include the naive Bayes classifier [8], decision trees [9], nearest neighbour classifier [5], linear discriminant analysis (LDA) [6], logistic regression [10] and neural networks [10]. Lewis and Ringutte [9] compared their ProBayes method and a decision tree classifier (using IND package) on two data sets (Reuters newswire benchmark and MUC-3) with different numbers of features. They showed that the maximum effectiveness was reached for both algorithms when the term (feature) selection was based on collection frequency and mutual information. Pazzani *et al.* [11] developed a software agent that learns to rate pages on the www based on user judgement. They also compared three different algorithms: the Bayesian classifier, decision tree (ID3) and nearest neighbour with a binary feature vector on the two categories of user preferences (hot-list and cold-list). Their experiments used only a small number of (20 to 120) training examples. The empirical results indicated that ID3 was not suited to their problem and the nearest neighbour classifier worked well over other methods when presented with a large number of examples. Schutze *et al.* [10] have empirically analysed how feature selection affects the three statistical classification techniques (LDA, logistic regression and neural networks) for the routing (two-class) problems. They used optimal term selection (χ^2 measure) and latent semantic indexing (LSI) to reduce the number of features. Their experimental results showed that features based on LSI are more effective for techniques such as LDA and logistic regression, whereas neural network based classification performs well with both the feature selection methods.

In this paper, we report on our experience using news data from the Yahoo web site, which are categorized into seven groups (business, entertainment, health, international, politics, sports and technology), each item of news is indexed manually by human experts (i.e. we have labelled training examples and 'true class labels' of testing samples). Among these news groups, there are some categories that have large overlaps (such as international and politics news groups). Meanwhile, there are classes that are better separated (e.g. health and sports). So, the data is well suited for testing the classification algorithms. We apply four different classification methods: the naive Bayes classifier, nearest neighbour classifier, decision trees and subspace method to the text document classification individually. Combinations of different classifiers by simple voting, adaptive classifier selection and our own adaptive classifier combination approaches were also investigated. Due to the fact that document classification often requires working in high-dimensional feature space and with sparse data, we also study the effect of feature reduction on these classification algorithms. Term-grouping in the subspace method is explored, together with two other dimensionality reduction schemes: optimal term selection and principal component analysis (PCA).

2. FEATURE REPRESENTATION

We adopt the commonly used 'bag-of-words' [12] document representation scheme (vector space model), in which we ignore the structure of a document and the order of words in the document. The feature vectors represent the words observed in the documents. The *word-list* $\mathcal{W} = (w_1, \ldots, w_d)$ in the training set consists of all the distinct words (also called *terms*) that appear in the training samples after removing the *stopwords* [13] (those words which are not helpful for retrieval, such as 'the', 'some' and 'of') and the low-frequency words (which only occur once in the training examples). Typically, there can be thousands of features in document classification (the number of commonly used English words is about 50,000). Given a document $\mathcal{D}$, its feature (term) vector is represented by $\mathcal{T} = (t_1, \ldots, t_d)$ constructed from $\mathcal{W}$. The value of each component of $\mathcal{T}$ could be either binary (a value of 1 indicates whether the corresponding word appeared in the document) or an integer representing the number of times the corresponding word was observed. In this paper we use binary representation. Frequency representation is also used in the nearest neighbour classifier and the subspace method classifier to calculate the weight for each term. One training example of the *business* news group after Web document parsing and *stopwords* and low-frequency words removal is also shown in Appendix A. The value of d (total number of different terms) is a function of the training data. For our training data, $d = 4724$.

3. CLASSIFICATION ALGORITHMS

In the following sections, we briefly describe the naive Bayes classifier, nearest neighbour and decision tree classification methods used in our study and introduce our use of the subspace method for document classification.

3.1. Naive Bayes classifier

The naive Bayes classifier [12] has been successfully used in the Rainbow text classification system [8]. Let $\mathcal{C} = (c_1, \ldots, c_m)$ be m document classes. Given a new unlabelled document $\mathcal{D}$ and its corresponding word-list $\bar{\mathcal{W}} = (w_1, \ldots, w_{d'})$ (defined in the same way as the word-list for the training set), the naive Bayes approach assigns $\mathcal{D}$ to a class c^*_{NB} as follows:

$$c^*_{\text{NB}} = \text{argmax}_{c_j \in \mathcal{C}} P(c_j) \prod_{i=1}^{d'} P(w_i|c_j), \qquad (1)$$

where $P(c_j)$ is the *a priori* probability of class c_j and $P(w_i|c_j)$ is the conditional probability of word w_i given class c_j. The underlying assumption of the naive Bayes approach is that for a given class c_j, the probabilities of words occurring in a document are independent of each other.

When the size of the training set is small, the relative frequency estimates of probabilities, $P(w_i|c_j)$, will not be reasonable; if a word never appears in the given training data, its relative frequency estimate will be zero. Instead, we applied the Laplace law of succession [14] to estimate

$P(w_i|c_j)$. The estimate of the probability $P(w_i|c_j)$ is given as:

$$P(w_i|c_j) = \frac{n_{ij} + 1}{n_j + k_j}, \tag{2}$$

where n_j is the total number of words in class c_j, n_{ij} is the number of occurrences of word w_i in class c_j and k_j is the vocabulary size of class c_j. This is the result of the Bayesian estimation with a uniform prior assumption, i.e. probabilities of the occurrence of words appearing in class c_j are equally likely.

3.2. Nearest neighbour classifier

The nearest neighbour decision rule assigns the new unlabelled document $\mathcal{D}$ to the document class c_j if the training pattern closest to $\mathcal{D}$ is from class c_j. We use the TF–IDF (TF is the term frequency in a document and IDF is the inverse document frequency) weighting scheme and use the cosine similarity [15] instead of Euclidean distance to measure the similarity of the two documents. Given two documents $\mathcal{D}_1$ and $\mathcal{D}_2$, their corresponding weighted feature vectors are $\mathcal{T}_1 = (t_{1i}\delta_i 1)_{i=1}^d$ and $\mathcal{T}_2 = (t_{2i}\delta_i 2)_{i=1}^d$, where δ_{ki} is the weight of word w_i in document k (TF-IDF). The similarity between $\mathcal{D}_1$ and $\mathcal{D}_2$ is then defined as:

$$S(\mathcal{D}_1, \mathcal{D}_2) = \frac{\mathcal{T}_1^{\mathrm{T}}\mathcal{T}_2}{\|\mathcal{T}_1\|\|\mathcal{T}_2\|}, \tag{3}$$

where $\|\cdot\|$ denotes the norm of the vector.

3.3. Decision tree classifier

Decision trees are one of the most widely used inductive learning methods. Their robustness to noisy data and their capability to learn disjunctive expressions seem suitable for document classification. One of the most well known decision tree algorithms is ID3 [16] and its successor C4.5 [17] and C5.[1] It is a top-down method which recursively constructs a decision tree classifier. At each level of the tree, ID3 selects the attribute that has the highest *information gain* [12]. For our experiments, we chose the C5 decision tree package since it has many nice features over its predecessor ID3 and C4.5. For example, the *rulesets* used in C5 are more accurate, faster and require less memory.[2] Furthermore, *adaptive boosting* [18] is incorporated into the software. The basic idea of boosting is to generate n ($n > 1$, n is specified by the user) classifiers (either decision trees or rule sets) instead of one. The ith classifier is constructed by examining the errors made by the $(i - 1)$th classifier. When a new document is to be classified, a voting scheme based on n classifiers is used to determine the final class of the document.

3.4. Subspace method

The subspace model [19] decomposes a given feature space into m subregions of lower dimensionality (subspaces), where each region is a representative feature space for its corresponding pattern class $c_i, i = 1, \ldots, m$. A test document is classified based on a comparison of its compressed representation in each feature space with that of different classes.

We apply this model to document classification as follows. Suppose we have m document classes $\mathcal{C} = (c_k)_{k=1}^m$. Class c_k is represented by a subspace $\mathcal{L}_k$ of cardinality d_k. Let $\mathcal{T} = (t_i)_{i=1}^d$ denote the term-vector in the original d-dimensional feature space, corresponding to the word-list of the training set $\mathcal{W} = (w_i)_{i=1}^d$. Let the word-list of the subspace $\mathcal{L}_k$ be denoted $\mathcal{W}_k = (w_i^k)_{i=1}^{d_k}$, where w_i^k are the words observed in class c_k. Given a vector $\mathcal{T}$ in the original feature space, the weighted projection Π_k of vector $\mathcal{T}$ on the subspace $\mathcal{L}_k$ is defined as:

$$\boldsymbol{\mathcal{T}}_k = \Pi_k(\mathcal{T}) = \mathcal{H}_k\mathcal{T}, \tag{4}$$

where $\mathcal{H}_k = (h_{ij})_{d_k \times d}$ is a $(d_k \times d)$ matrix and the ith row corresponds to the ith component of the word-list $\mathcal{W}_k$ in the subspace $\mathcal{L}_k$, while the jth column is the jth component of the word-list $\mathcal{W}$ in the original feature space. The elements h_{ij} are calculated as follows:

$$h_{ij} = \begin{cases} \delta_j^k, & \text{when the term } w_i^k \text{ is the} \\ & \text{same as the term } w_j \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where δ_j^k is the weight of term w_j^k in subspace $\mathcal{L}_k$. We define δ_j^k as:

$$\delta_j^k = \frac{\mathrm{CLASSFREQ}_{jk}}{log_2(\mathrm{DOCFREQ}_j + 1)}, \tag{6}$$

where $\mathrm{CLASSFREQ}_{jk}$ denotes the ratio of the number of documents in which the term w_j occurred in c_k to the number of documents in c_k and $\mathrm{DOCFREQ}_j$ represents the ratio of the number of documents in all classes in which the term w_j occurred to the size of the training samples.

The Euclidean vector norm of $\mathcal{T}_k$ is $\|\mathcal{T}_k\| = \sqrt{\mathcal{T}_k{}^{\mathrm{T}}\mathcal{T}_k}$. For a new document $\mathcal{D}$, the subspace decision rule classifies $\mathcal{D}$ to the class on whose subspace its term-vector $\mathcal{T}$ has the largest projection in terms of the Euclidean vector norm.

4. COMBINATION OF MULTIPLE CLASSIFIERS (CMC)

A number of researchers have shown that combining different classifiers can improve the classification accuracy [20, 21, 22]. Larkey and Croft [21] applied weighted linear combinations of different classifiers to the medical document domain, where the weights were assigned by the user. Another CMC approach is dynamic classifier selection (DCS) [20, 22], where a single classifier is selected which has the highest local accuracy in small regions of feature space surrounding the test sample presented to the system. We investigated three different combination approaches: simple voting, DCS and our own approach of adaptive classifier combination (ACC).

[1] http://www.rulequest.com/see5-unix.html

[2] http://www.rulequest.com/see5-comparison.html

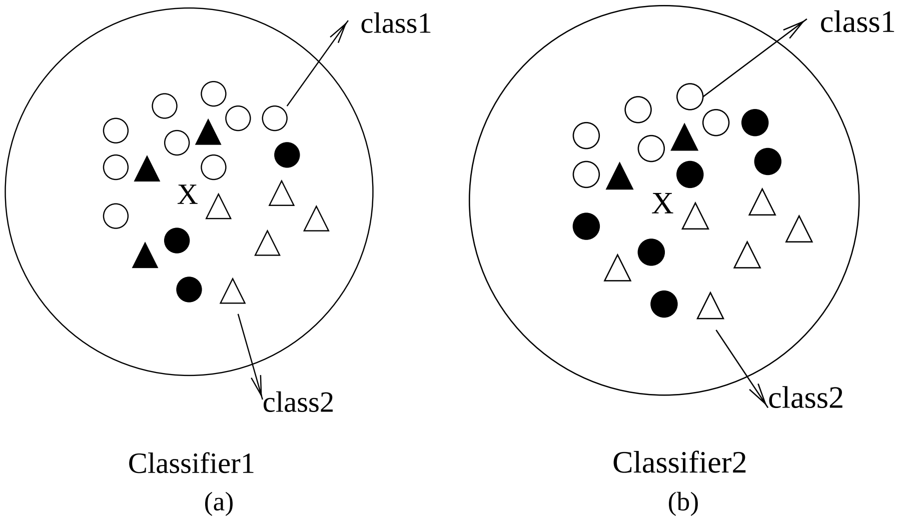

FIGURE 1. Illustration of *DCS* and *ACC* methods. Darkened patterns are misclassified using leave-one-out method. (a) *DCS* method: $Pc_{(\text{classifier1})} = 14/20$, $Pc_{(\text{classifier2})} = 12/20$, classifier1 should be chosen, X → class1; (b) *ACC* method: $Pc_{(\text{class1})} = 9/12 + 6/12 = 5/4$, $Pc_{(\text{class2})} = 5/8 + 6/8 = 11/8$, X → class2.

4.1. Simple voting

For each test document, we classify it to class c_i, where a majority of the classifiers individually assign the test document to class c_i.

4.2. Dynamic classifier selection (DCS)

We implemented a version of DCS described in [20, 22]. For a test document $\mathcal{D}$, we use the k-nearest neighbour approach to find the neighbourhood of $\mathcal{D}$ and the 'leave-one-out' method [23] is applied on the training data to find the local accuracy in the neighbourhood of $\mathcal{D}$. We used the 'soft' measure [22] of the local accuracy, where the weight of each neighbour of $\mathcal{D}$ is the cosine similarity measure between $\mathcal{D}$ and that neighbour. See Figure 1a for the illustration.

4.3. Adaptive classifier combination (ACC)

Instead of selecting the best classifier with the highest local accuracy for a test document, we assign the document to class c_i, which is the class identified by the classifier that has the highest local accuracy among all the classifiers. See Figure 1b for the illustration. The outline of our algorithm, given n classifiers, is described as follows.

(1) For a test document $\mathcal{D}$, find the neighbourhood of $\mathcal{D}$, $\mathcal{NB}(\mathcal{D}) = (x_1, \ldots, x_k)$, $x_i \in \mathit{Training\_Set}$, using the k-nearest neighbour algorithm.

(2) Denote the classification results for $\mathcal{D}$ by n classifiers as $\bar{\mathcal{C}} = (\bar{c_1}, \ldots, \bar{c_n})$, $\bar{c_j} \in \{c_1, \ldots, c_m\}$.

(3) For each class $\bar{c_j} \in \bar{\mathcal{C}}$, calculate $\mathrm{Acc}^j_{\mathrm{loc}} = \sum_{s=1}^{n} \sum_{i=1}^{k} W_i P_s(\bar{c_j}|x_i \in \bar{c_j})$, where $P_s(\bar{c_j}|x_i \in \bar{c_j})$ is the local accuracy of neighbourhood patterns x_i, i.e. the *a posteriori* probability that x_i belongs to the class $\bar{c_j}$. The local accuracy of each x_i can be obtained by using the 'leave-one-out' method on the training data, and W_i is the cosine similarity measure between pattern x_i and $\mathcal{D}$.

(4) Classify $\mathcal{D}$ to class c_κ, where $\kappa = \mathrm{argmax}_j(\mathrm{Acc}^j_{\mathrm{loc}})$.

5. DIMENSIONALITY REDUCTION

Document classification is often characterized by the high dimensionality (thousands of features) of the associated feature space and a relatively small number of training samples. We must, therefore, guard against the potential problems of 'curse of dimensionality' [24]. We study three feature dimensionality reduction approaches: feature selection, feature extraction and the proposed term grouping.

5.1. Feature selection

One way to reduce the number of features in document classification is to select a subset of the best terms from the entire feature space. In this paper, we use the *individual best features* approach, where terms are sorted by their weights and the top n terms with the highest weights are selected. We use mutual information as suggested in [25] to assign weights to the terms. Other weighting schemes have been used in best-term subset selection [15, 10]. Traditional feature selection approaches in pattern recognition, such as sequential forward/backward selection and 'plus l-take away *r*' selection [26] may perform better, but they are often much more expensive in terms of computational cost. This is a major consideration in high dimensional feature spaces encountered in document classification problems.

5.2. Feature extraction

Another method of dimensionality reduction is to map original measurements into a more effective lower dimensional subspace. Each new feature is a combination of the original features. Principal component analysis (PCA) (also called the Karhunen–Loève transform) is a commonly used linear projection method [27]. Using the vector space representation scheme in document classification, let $\mathcal{T}_1, \ldots, \mathcal{T}_N$ denote the N d-dimensional training vectors, while their normalized vectors with zero-mean are denoted as $\mathcal{T}_1^*, \ldots, \mathcal{T}_N^*$. Let the p basis vectors, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ be a set of orthonormal vectors that best describe the distribution of documents in the p-dimensional subspace (eigenspace), $p \leq d$. The k^{th} eigenvector, $\boldsymbol{e}_k, k = 1, \ldots, p$, is computed such that

$$\lambda_k = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{e}_k^{\mathrm{T}}\mathcal{T}_i^*)^2 \tag{7}$$

is maximum, subject to

$$\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{e}_j = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The value λ_k is the kth largest eigenvalue of the covariance matrix Σ which can be estimated using the training samples by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{T}_i^*\mathcal{T}_i^{*\mathrm{T}}. \tag{9}$$

The vector $\boldsymbol{e}_k$ is the kth eigenvector of the covariance matrix $\hat{\Sigma}$ corresponding to λ_k.

With the p-dimensional eigenspace defined, training vectors, $\mathcal{T}_1^*, \ldots, \mathcal{T}_N^*$, can be represented as a set of p-dimensional feature vectors, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p$:

$$\boldsymbol{\xi}_k = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\mathcal{T}}_i^*, \quad i = 1, \ldots, N, \tag{10}$$

where $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p)$.

The sum of the first p eigenvalues is the 'variance' retained in the subspace, while the sum of all the d eigenvalues is the 'variance' in the original pattern space [27]. We can choose p such that $(\sum_{i=1}^{p}\lambda_i / \sum_{i=1}^{d}\lambda_i) \geq \nu$, where ν is a user-specified value representing the desired 'variance' retained in the p-dimensional subspace.

5.3. Term grouping in subspace

The occurrences of different words in documents are usually not independent; there are correlations between words in a group of documents. Wulfekuhler and Punch [28] applied K-means clustering within each document category to find clusters of words for that document class. To capture the correlation of all word-pairs, we construct a bigram matrix for each document class k. Let $\mathcal{W}_k = (w_1^*, \ldots, w_{d_k}^*)$ be a d_k-dimensional word-list in subspace $\mathcal{L}_k$. The bigram matrix $\mathcal{B}_k = (b_{ij})_{d_k \times d_k}$ is a $(d_k \times d_k)$ matrix with b_{ij} representing the number of documents from class c_k in which both the term w_i^* and term w_j^* jointly appeared. A complete-link hierarchical clustering algorithm is applied to the proximity matrix $\mathcal{B}_k$. The resulting dendrogram (tree) is then cut into p term-groups, where p is a user-specified parameter. Let the term-groups be denoted as $\mathcal{T}_k^* = (h_1, \ldots, h_p)$, where $h_i \bigcap h_j = \emptyset, i \neq j$ and $h_i = (w_{i1}^*, \ldots, w_{in_i}^*)$, $w_{ij}^* \in \mathcal{W}_k$. Given a vector $\mathcal{T}_k$ in subspace $\mathcal{L}_k$, the projection of $\mathcal{T}_k$ to p-dimensional space, $\Xi = (\xi_1, \ldots, \xi_p)$ is defined as

$$\xi_i = (\mathcal{T}_k^{\mathrm{T}}\mathcal{H}_i)\boldsymbol{I}, \tag{11}$$

where $\mathcal{H}_i = (h_{uv})_{n_i \times d_k}$ is defined (as in Section 3.4) as

$$h_{uv} = \begin{cases} 1 & \text{when the term } w_{iu}^* \text{ is the} \\ & \text{same as the term } w_v^* \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

and $\boldsymbol{I}$ is the n_i-dimensional unit vector.

6. EXPERIMENTAL RESULTS

The data used in our experiments are the news items downloaded from the Yahoo news group. We preprocess the HTML news items by (i) document parsing (remove headers and tags in the HTML files) and (ii) removing *stopwords* and low-frequency words as mentioned earlier. We have used a total of 814 documents belonging to seven different classes (business (B), entertainment (E), health (H), international (I), politics (P), sports (S) and technology (T)) for training and two test data sets (news items at different time intervals, see Table 1). There are 265 distinct terms (only observed in at most one of the classes) in the training data and 75 terms are common among all the categories. On average, a document contains about 120 words.

Using the 814 training documents, we compared the four classification algorithms (naive Bayes classifier (NB), nearest neighbour classifier (NN), decision tree classifier (DT) and the subspace classifier (SS)) on our two test data sets. Table 2 shows a comparison of the recognition rates (precision) using these four classification algorithms individually. The experimental results show that all four classification algorithms perform reasonably well; the naive Bayes approach performs the best on test data set1, but the subspace method outperforms all others on test data set2. Confusion matrices of the classification results using NB on test set1 and using SS on test set2 are shown in Tables 3 and 4, respectively.

Results of multiple classifier combinations using different combination approaches are summarized in Table 5. We set $k = 20$ (neighbourhood size) in our experiments. Note that for these two test data sets, there was no significant improvement by using a combination of classifiers. This shows that the performance of classifier combinations is data dependent.

We used the PCA method to project the original feature space onto a lower dimensional subspace. We retained 90% of the variance with about 500 eigenvectors. Since the derived features obtained by the PCA method are linear combinations of the original features, we lose information about individual words in the document, so the naive Bayes method cannot be used on this reduced feature set. We applied the nearest neighbour algorithm and the decision tree

TABLE 1. Training and testing data.

	Categories	B	E	H	I	P	S	T
Training data	No. of documents	130	133	91	110	130	130	90
	Total no. of terms	1848	2045	1213	1974	2070	1659	1364
Test data set1	No. of documents	110	111	79	80	110	111	79
	Total no. of terms	2155	2583	1535	1999	2439	1952	1618
Test data set2	No. of documents	100	101	78	70	101	101	70
	Total no. of terms	2046	2834	1803	2604	2070	1974	1689

TABLE 2. Comparison of the four classification algorithms (NB, NN, DT and SS).

		NB	NN	DT	SS
Test data set1	No. of misclassifications	115	165	178	139
	Recognition rate (%)	**83.1**	75.7	73.8	79.6
Test data set2	No. of misclassifications	125	179	144	111
	Recognition rate (%)	79.87	71.18	76.8	**82.13**

TABLE 3. Confusion matrix of the classification results using the NB classifier on test set1.

	B	E	H	I	P	S	T	Recognition rate (%)
B	81	1	2	0	7	0	19	73.6
E	1	89	1	8	3	2	7	80.2
H	0	0	79	0	0	0	0	100.0
I	1	1	0	53	25	0	0	66.3
P	5	0	1	10	91	0	3	82.7
S	2	0	1	1	3	102	2	91.9
T	7	0	2	0	0	0	70	88.6

TABLE 4. Confusion matrix of the classification results using the SS method on test set2.

	B	E	H	I	P	S	T	Recognition rate (%)
B	62	2	0	0	6	2	22	68.0
E	0	84	3	6	0	6	2	83.2
H	1	0	72	0	4	1	0	92.3
I	5	0	2	50	12	0	1	71.4
P	12	1	7	10	69	1	1	68.3
S	0	0	0	0	0	101	0	100.0
T	3	0	0	0	1	0	66	94.3

TABLE 5. Classification accuracies of combinations of multiple classifiers.

Combination of classifiers	Combination approach	Testing data set1 (%)	Testing data set2 (%)
NB, SS, NN	Simple voting	80.29	81.96
	DCS	80.00	77.13
	ACC	82.21	82.45
NB, SS	DCS	80.44	79.87
	ACC	**83.24**	**82.93**
Best Individual Classifier		83.1	82.13

TABLE 6. Comparison of the two classification algorithms (NN and DT) before and after using PCA feature reduction technique on test set1.

	NN without PCA	NN with PCA	DT without PCA	DT with PCA
No. of misclassifications	165	166	178	151
Recognition rate (%)	75.7	75.6	73.8	77.8

classifier using the reduced feature space. Table 6 shows a comparison of the two classification algorithms (nearest neighbour and decision tree) before and after using the PCA feature reduction technique on test set1. We can see that the performance of the DT classifier is improved by using the PCA feature extraction strategy, while the performance of the NN classifier is not affected very much.

We apply our term-grouping technique to the subspace method. A total of 30 term groups were chosen in our experiment. A comparison of the recognition rates before and after using the term-grouping technique on data set1 is shown in Table 7. It shows that the performance of the SS method improved marginally with the term-grouping technique.

We also used mutual information to weigh the words appearing in the training documents; a subset of the words with the highest weights was selected. We compared the four classifiers (NB, NN, DT and SS) using this feature selection technique. Figure 2 shows the recognition rate of

TABLE 7. Comparison of using the subspace method with or without the term-grouping feature reduction technique on test set1.

	SS without term-grouping	SS with term-grouping
No. of misclassifications	139	137
Recognition rate (%)	79.6	79.9

TABLE 8. Comparison of the four classification algorithms (NB, NN, DT, and SS) on test set1 for the reduced 5-class problem.

	NB	NN	DT	SS
No. of misclassifications	67	95	128	101
Recognition rate (%)	90.15	86.03	81.18	85.15

the four classification algorithms with different sized feature subsets of test set1. From this figure, we can see that a small number of features is not suitable for the NB classifier; the performance of the NB and SS methods tends to improve when the number of features is increased, while NN and DT classifiers have some fluctuations.

7. CONCLUSIONS AND DISCUSSION

We have applied four different classification methods (NB, NN, DT and SS) to the problem of document categorization. These methods were evaluated individually and when used in combination. Since document classification involves high-dimensional feature space, we also studied the effect of different feature reduction techniques (the individual best feature selection, PCA and term grouping) on the performance of these classifiers. The seven classes of Yahoo news items used in our experiments have a large overlap of words in their documents (e.g. in 2,948 total words, there are 1,096 common words between *international* and *politics* news categories, 744 out of 2,468 words are common between *business* and *technology* news groups), so this is a difficult classification problem. We can make the following observations based on our experimental results.

(1) All the four classifiers perform reasonably well on our data sets. Weiss *et al.* [5] reported that the accuracy of human judgement of 1000 messages on 10 USENET news groups (misc.health.diabetes, sci.math.num-analysis, dc.politics, rec.food.restaurants, alt.tv.seinfeld, comp.sys.ibm.pc.games.sports, rec.arts.comics.dc.universe, sci.military.naval, talk.philosophy.misc and humanities.lit.authors.shakespeare) is about 85%. NB and SS classifiers work better than NN and DT methods, but the performance of NB and SS is data dependent. Most of the misclassifications are between *international* and *politics* categories and between *business* and *technology* document classes which inherently have a large overlap of terms. If we combine *international* and *politics* news groups and combine *business* and *technology* news groups, the performance of all four classification algorithms on the resulting five-class problem improved by an average of 7%. Classification accuracies on test set1 for the four classifiers on this five-class problem are shown in Table 8.

(2) The naive Bayes method works well on our news data set, despite the 'independence' assumption which is not always satisfied in document classification. This agrees with observations of Domingos and Pazzani [29]: 'attribute dependence is not a good predictor of the naive Bayes classifier's differential performance versus approaches that can take it into account'.

(3) The simple SS method performs best on one test data set and outperforms the NN and DT without dimensionality reduction. It works extremely well on the two most separable classes, *health* and *sports*, but not quite so good for classes with a large overlap (e.g. between *business* and *technology* news groups).

(4) Combinations of multiple classifiers do not always improve the classification accuracy. The adaptive classifier combination introduced here worked better than the simple voting and the dynamic classifier selection approaches on our two test data sets.

(5) The 'curse of dimensionality' and overfitting do not seem to be a problem for NB, NN and SS classifiers. The performance of DT improved significantly when incorporating the boosting technique (average improvement is about 13%). With feature extraction (PCA), the DT performs better (the recognition rate increases by 4%). Note that PCA does not incorporate category information in dimensionality reduction, so the linear discriminant analysis (LDA) will work better than PCA for classification problems [27]. However, LDA requires a huge amount of memory space when used for high dimensional feature vectors.

(6) There is no significant 'peak' in classification performance observed in our experiments with feature selection. In particular, the performance of NB improves as the number of features increase (which is different from the results obtained in [9]). Additionally, NB does not use a small number of features effectively (which is also different from the observation in [9] that 10 features for Reuters news groups performed the best). This indicates that our data set is less separable than the Reuters newswire.

(7) The problem with feature selection is that the small number of selected words may not generalize well to new documents. However, the advantage of dimensionality reduction is not only to improve the recognition rate (eliminate the problem of overfitting), but the reduced number of features leads to lower time and space complexities. The term-grouping method reduces the feature dimensionality and overcomes the generalization problem of feature selection, while maintaining the performance of the classifier.

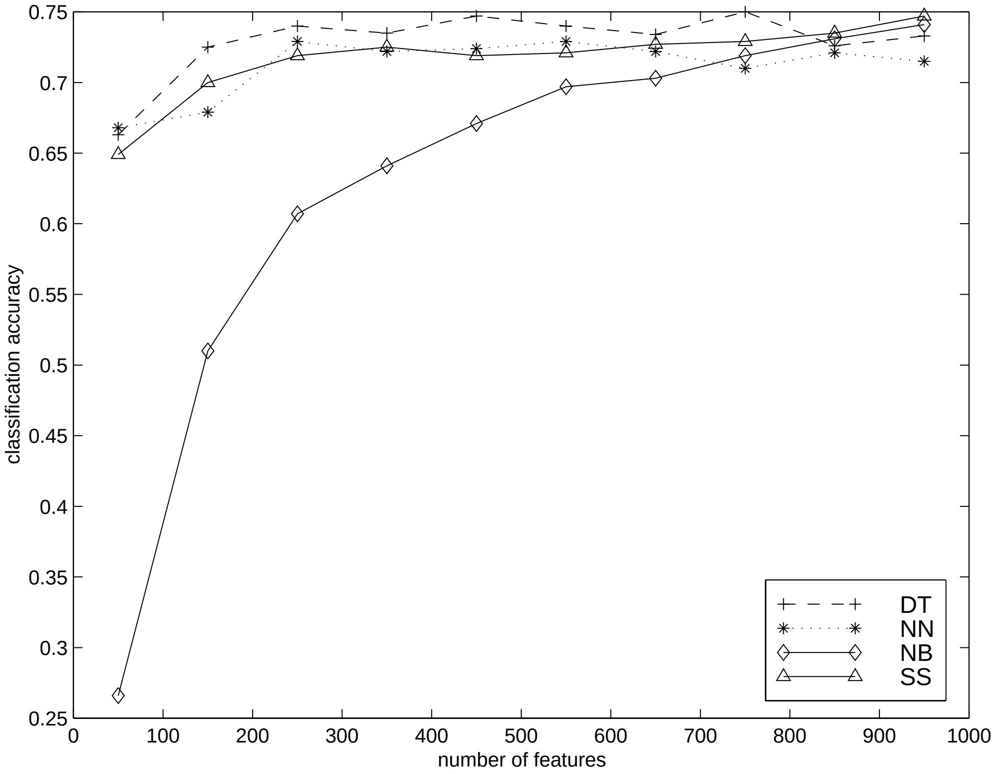

FIGURE 2. Accuracy of classification algorithms with different numbers of features on test set1.

REFERENCES

[1] Yan, T. and Molina, H. (1995) SIFT—a tool for wide-area information dissemination. In *Proc. 1995 USENIX Technical Conf.*, pp. 177–186.

[2] Lang, K. (1995) NewsWeeder: learning to filter netnews. In *Proc. 12th Int. Conf. on Machine Learning*, pp. 331–339.

[3] Joachims, T., Freitag, D. and Mitchell, T. (1997) WebWatcher: a tour guide for the world wide web. In *Proc. IJCAI97*.

[4] Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P. (1997) Keyword detection, navigation, and annotation in hierarchical text. In *Proc. 23rd VLDB*, pp. 446–455.

[5] Weiss, S., Kasif, S. and Brill, E. (1996) Text classification in USENET newsgroup: a progress report. In *AAAI Spring Symp. on Machine Learning in Information Access Technical Papers*, Palo Alto, March 1996.

[6] Hull, D., Pedersen, J. and Schutze, H. (1996) Document routing as statistical classification. In *AAAI Spring Symp. on Machine Learning in Information Access Technical Papers*, Palo Alto, March 1996.

[7] Faloutsos, C. and Oard, D. (1995) *A Survey of Information Retrieval and Filtering Methods*. Technical Report CS-TR-3541, University of Maryland.

[8] *Gentle Introduction to RainBow*. URL: http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/gentle_introduction.html.

[9] Lewis, D. D. and Ringutte, M. (1994) A comparison of two learning algorithms for text categorization. In *Third Annual Symp. on Document Analysis and Information Retrieval*, Las Vegas, NV, pp. 81–93.

[10] Schutze, H., Hull, D. and Pedersen, J. (1995) A comparison of classifiers and document representations for the routing problem. In *Proc. SIGIR*, pp. 229–237.

[11] Pazzani, M., Muramatsu, J. and Billsus, D. (1996) Syskill & Webert: identifying interesting web sites. In *AAAI Spring Symp. on Machine Learning in Information Access Technical Papers*, Palo Alto, March 1996.

[12] Mitchell, T. (1997) *Machine Learning*. McGraw-Hill, New York.

[13] Fox, C. (1992) Lexical analysis and stoplist. In Frakes, W. and Baeza-Yates, R. (eds), *Information Retrieval Data Structures and Algorithms*, pp. 102–130. Prentice Hall, Englewood Cliffs, NJ.

[14] Ristad, E. (1995) *A Natural Law of Succession*. Technical Report CS-TR-495-95, Princeton University.

[15] Salton, G. and McGill, M. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

[16] Quinlan, J. (1986) Induction of decision trees. *Machine Learning*, **1**, 81–106.

[17] Quinlan, J. (1993) *C4.5:Programs for Machine Learning*. Morgan Kaufmann, San Matteo, CA.

[18] Freund, F. and Schapire, R. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. Second European Conference on Computational Learning Theory*, pp. 23–37.

[19] Oja, E. (1983) *Subspace Methods of Pattern Recognition*. Wiley, New York.

[20] Woods, K., Kegelmeyer, W. and Bowyer Jr, K. (1997) Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. PAMI*, **19**, 405–410.

[21] Larkey. S. and Croft, W. (1996) Combining classifiers in text classification. In *Proc. SIGIR*, pp. 81–93.

[22] Giacinto, G. and Roli, F. (1997) Adaptive selection of image classifiers. In *Proc. ICIAP (Springer Verlag Lecture Notes in CS Vol. 1310)*, pp. 38–45.

[23] Duda, R. and Hart, P. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

[24] Jain, A. and Chandrasekaran, B. (1982) Dimensionality and sample size considerations in pattern recognition practice. In Krishnaiah, P. and Kanal, L. (eds), *Handbook of Statistics*, pp. 835–855. North-Holland, Amsterdam.

[25] Mladenic, D. (1996) *Personal WebWatcher: Design and Implementation*. Technical Report IJS-DP-7472, Carnegie Mellon University.

[26] Jain, A. and Zongker, D. (1997) Feature selection: evaluation, application and small sample performance. *IEEE Trans. PAMI*, **19**, 153–157.

[27] Jain, A. and Dubes, R. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

[28] Wulfekuhler, M. and Punch, W. (1997) Finding salient features for personal web page categories. In *Sixth International World Wide Web Conference*, Santa Clara, CA.

[29] Domingos, P. and Pazzani, M. (1996) Beyond independence: conditions for the optimality of the simple bayesian classifier. In *Proc. MLC*, pp. 105–112.

APPENDIX A. EXAMPLE OF A TRAINING SAMPLE

We show one of the training examples from Yahoo news groups after Web document parsing and their word list (used to construct the feature vector) after *stopwords* and low-frequency words removal (Figure 3).

APPENDIX B. EXAMPLES OF MISCLASSIFICATION

We show some examples of Yahoo news items that were misclassified (Figures 4, 5, 6 and 7). By looking at the contents of these news items, these misclassifications do not appear to be unreasonable.

Minimum Wage Rises - Nearly 7	minimum wage rises _ nearly _
million Americans are getting a	million americans ___ getting _
raise on this Labor Day. The	raise __ ____ labor day_ ___
federal minimum wage is rising to	federal minimum wage __ rising __
$5.15 an hour. Fast food workers,	_____ __ hour_ fast food workers_
retail clerks, gas station	retail clerks_ gas station
attendants and others will be	attendants ___ ______ ____ __
earning 40 cents an hour more when	earning __ cents __ hour ____ ____
they report to work as the second	____ report __ ____ __ ___ ______
phase of the hike goes into effect.	phase __ ___ hike goes ____ effect_
It was first raised to $4.75 last	__ ___ _____ raised __ _____ ____
Oct. 1. According to a report to	oct_ __ according __ _ report __
be issued tomorrow by the Economic	__ issued tomorrow __ ___ economic
Policy Institute, most of the 6.8	policy institute_ ____ __ ___ ___
million workers affected by the	million workers affected __ ___
minimum wage hike are women who	minimum wage hike ___ women ___
work in the service sector. The	____ __ ___ service sector_ ___
Washington, D.C.-based think tank’s	washington_ _____based _____ tank__
study found that in 18 states, more	study found ____ __ __ _______ ____
than 10 percent of the work force	____ __ percent __ ___ ____ force
will be affected by the minimum	____ __ affected __ ___ minimum
wage increase.	wage increase_
(a)	(b)

FIGURE 3. An example of the *business* news group: (a) a training sample; (b) extracted word list.

Yeltsin Drops Hint on Third Term–Eleven months after life-saving heart surgery, President Boris Yeltsin is hinting that he might consider running for a third term. On a visit to the city of Nizhny Novgorod, the 66-year-old president was asked Thursday if there was any chance he would run again in the year 2000. Yeltsin replied: “My friends and colleagues have forbidden me from talking about this.” A month ago, he flatly denied that he would seek re-election. Whether Yeltsin is really willing or able, legally or physically, to lead Russia into the 21st century, no one can say. The 1993 Russian constitution limits presidents to two terms.

FIGURE 4. An example of the *international* news item that was misclassified into the *politics* news group.

WorldCom Makes Bid for MCI–WorldCom, the nation’s fourth largest long-distance telephone company, made an unsolicited bid of $30 billion today to buy rival MCI, topping an offer from British Telecommunications. WorldCom, a little known but fast growing communications company, says it would pay $41.50 in Worldcom stock for each share of MCI in a deal that Wall Street analysts believe delivered a stunning blow to British Telecom’s global expansion. MCI shares traded as high as $36.19, up $6.81, on Nasdaq, where it’s the most active share listed. An MCI spokesman says it’s “premature for us to comment” on the WorldCom bid.

FIGURE 5. An example of the *technology* news item that was misclassified into the *business* news group.

Knight-Ridder Unit is Sold–M.A.I.D. a money-losing British electronic information provider, said today it’s acquiring Knight-Ridder Information in a $420 million deal that will create the world’s largest online information company. KR Information, part of U.S. newspaper publisher Knight-Ridder, owns huge databases containing the equivalent of six billion pages of text. Miami-based Knight-Ridder, publisher of the Philadelphia Inquirer, recently sold its real-time news operation as part of a withdrawal from the electronic information business.

FIGURE 6. An example of the *business* news item that was misclassified into the *technology* news group.

Boyz II Men Tops Charts–The latest album by Boyz II Men debuted at No. 1 on the pop album charts, but its first week of sales were sharply lower than those of the group’s previous release, according to data issued Wednesday. The quartet’s “Evolution” sold 211,000 copies in the week that ended Sept. 28, while its “II” had opened with more than 302,000 copies in September 1994. Last week’s No. 1 album, Mariah Carey’s “Butterfly,” flew down to third. Holding steady at second place was LeAnn Rimes’ “You Light Up My Life.” Country duo Brooks & Dunn’s “The Greatest Hits Collection” rose three places to No. 4 in its second week, while Rapper Master P.’s “Ghetto D” slipped a notch to No. 5.

FIGURE 7. An example of the *entertainment* news item that was misclassified into the *business* news group.