Classification of Text Documents Based on Minimum System Entropy

Raghu Krishnapuram KRAGHURA@IN.IBM.COM
Krishna P Chitrapura KCHITRAP@IN.IBM.COM
Sachindra Joshi JSACHIND@IN.IBM.COM
IBM India Research Lab,
Block-I, Indian Institute of Technology,
Hauz Khas, New Delhi, INDIA

Abstract

In this paper, we describe a new approach to classification of text documents based on the minimization of system entropy, i.e., the overall uncertainty associated with the joint distribution of words and labels in the collection. The classification algorithm assigns a class label to a new document in such a way that its insertion into the system results in the maximum decrease (or least increase) in system entropy. We provide insights into the minimum system entropy criterion, and establish connections to traditional naive Bayes approaches. Experimental results indicate that the algorithm performs well in terms of classification accuracy. It is less sensitive to feature selection and more scalable when compared with SVM.

1. Introduction

With the phenomenal increase in information content on the Internet and intranets, the need for intelligent tools to organize the information into taxonomies or ontologies has become acute. Among the plethora of techniques for classification of text documents that have been suggested in the literature, some important ones are naive Bayes (Lewis, 1998; McCallum & Nigam, 1998; Friedman et al., 1997; Langley et al., 1992), support vector machines (Joachims, 1998; S. T. Dumais & Sahami, 1998), rule-learning (Cohen & Singer, 1996; Slattery & Craven, 1998), k-nearest neighbor (Yang, 1999), and maximum entropy (Nigam et al., 1999). While some techniques (Robertson & Spark-Jones, 1976; Larkey & Croft, 1996; Koller & Sahami, 1997) assume the multi-variate Bernoulli model that uses a binary vector representation for each document, other techniques (Lewis & Gale, 1994; Kalt & Croft, 1996; Joachims, 1997; Li & Yamanishi, 1997; Nigam et al., 1998; McCallum et al., 1998) assume a multinomial model that accounts for multiple occurrences of a feature in a document. A study done by McCallum and Nigam (1998) shows that multinomial model typically outperforms the multi-variate Bernoulli model.

In this paper, we propose a new approach to classification of text documents based on the concept of 'system entropy'. We view a collection of labelled documents as being generated by a joint distribution of two random variables, one representing the class labels and the other the words in the vocabulary. We define the entropy of such a system, i.e., collection of documents. Since the entropy of a well organized system should be low, we argue that a new document should be inserted into the system in such a way that the updated entropy (after insertion) is as low as possible, i.e., the insertion should result in the maximum decrease (or least increase) of the entropy of the system. We refer to the resulting classification algorithm as the minimum system entropy (MSE) classification algorithm.

The rest of the paper is organized as follows. In Section 2, we define the entropy of a collection of labelled documents and present our approach to classification. In Section 3, we derive an approximate expression for the change in entropy when a new document is inserted into a particular class. Based on the approximation, we show that the MSE criterion and the naive Bayes criterion are related under some assumptions. In Section 4, we present experimental results that compare the performance of the MSE classifier with the naive Bayes classifier and SVM on several standard data sets. The experiments show that MSE does not require feature selection and has other advantages such as scalability. We also present a study on the effect of feature selection and the size of the training data set on the proposed algorithm. Finally, Section 5 contains the summary and conclusions.

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

2. The Minimum Entropy Criterion

2.1. Joint Entropy of Two Random Variables

Let A be a random variable that assumes values a in the set $\mathcal{A}$, and let B be a random variable that assumes values b in the set $\mathcal{B}$. Let $p(a,b)$ denote the joint probability that $A = a$ and $B = b$. The joint entropy, which is a measure of the uncertainly associated with the joint distribution of the pair of random variables A and B, is given by

$$E(A,B) = -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a,b). \qquad (1)$$

Similarly, the conditional entropy $E(A|B)$, which is a measure of the conditional distribution of A given B is defined as

$$\begin{aligned} E(A|B) &= -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a|b) \\ &= -\sum_{b\in\mathcal{B}} P(b)\sum_{a\in\mathcal{A}} p(a|b)\log p(a|b), \qquad (2) \end{aligned}$$

where the last equality follows from Bayes rule, i.e., $p(a,b) = p(a|b)P(b) = p(b|a)P(a)$. Here $P(a)$ and $P(b)$ are the marginal probabilities of a and b. It is easy to verify that (1) can be written as

$$E(A,B) = E(A) + E(B|A) = E(B) + E(A|B), \quad (3)$$

where $E(A) = -\sum_{a\in\mathcal{A}} P(a)\log P(a)$ and $E(B) = -\sum_{b\in\mathcal{B}} P(b)\log P(b)$ are the marginal entropies of A and B.

2.2. Entropy of a Collection of Labelled Documents

Let $D = \{D_1, D_2, \ldots, D_n\}$ denote a set of n documents where each document has been assigned a class label from the label set $\mathcal{C} = \{C_1, C_2, ..., C_c\}$. Let $n_i, i = 1,\ldots,c$, be the number of documents with the label C_i, with $\sum_{i=1}^{c} n_i = n$. We assume that the documents are modeled by the word set (vocabulary) $\mathcal{W} = \{w_1, w_2, \ldots, w_d\}$. Each document D_i is represented by a d-dimensional vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$, where x_{ij} represents the frequency of occurrence of word w_j in document D_i. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ represent the set of feature vectors corresponding to the document set. In addition, we let $k_{ij} = \sum_{\mathbf{x}_i\in C_i} x_{ij}$ denote the frequency of occurrence of word w_j in class C_i, and $k_j = \sum_{i=1}^{c} k_{ij}$ denote the overall frequency of occurrence of word w_j. We define K_i to be $\sum_{j=1}^{d} k_{ij}$. If we think of each class C_i as a bag of words, then K_i represents the size of bag C_i. We also define $K = \sum_{i=1}^{c} K_i = \sum_{j=1}^{d} k_j$, and estimate the prior probability P_i of bag C_i as K_i/K. Finally, we let $p_j = k_j/K, j = 1,\ldots,d$, and $p_{ij} = k_{ij}/K_i, i = 1,\ldots,c; j = 1,\ldots,d$, denote the probability of w_j occurring in X and C_i respectively.

If we view the collection of documents as being generated by a joint distribution of two random variables W and C that assume values in $\mathcal{W}$ and $\mathcal{C}$ respectively, from (1) and (3), the entropy $E(W,C)$ of the collection becomes:

$$E(W,C) = E(C) + \sum_{k=1}^{c} P_k E(W_k), \qquad (4)$$

where $E(C) = -\sum_{k=1}^{c} P_k \log P_k$, and $E(W_k) = -\sum_{j=1}^{d} p_{kj}\log p_{kj}$. In (4), $E(C)$ is a measure of uncertainty arising out of assigning class labels to documents and $E(W_k)$ is a measure of uncertainty arising out of the distribution of the words within class C_k. We refer to $E(W,C)$ as the system entropy.

2.3. Approach to Classification

Let $\mathbf{x} = [x_1, x_2, \ldots, x_d]$, be the new vector to be labelled, and let $L(\mathbf{x})$ denotes its length. The idea is to assign it the label that gives rise to the maximum decrease in the overall entropy of the system. It is to be noted that when we label the document $\mathbf{x}$, we are essentially labelling all the words in the document with the same label. If we insert the document in class C_i, then the new probabilities of the system are given by

$$P'_k = \begin{cases} \frac{K_k + L(\mathbf{x})}{K + L(\mathbf{x})} & \text{if } k = i, \\ \frac{K_k}{K + L(\mathbf{x})} & \text{otherwise} \end{cases} \qquad (5)$$

and

$$p'_{kj} = \begin{cases} \frac{k_{kj} + x_j}{K_k + L(\mathbf{x})} & \text{if } k = i, \\ \frac{k_{kj}}{K_k} & \text{if } k \neq i. \end{cases} \qquad (6)$$

Since the assignment of the new document $\mathbf{x}$ is to the class that gives rise to the largest decrease in entropy, the discriminant function $g_i()$ for class C_i may be written as

$$g_i(\mathbf{x}) = -(E'_i(W,C) - E(W,C)), \qquad (7)$$

where $E'_i(W,C)$ denotes the entropy of the system after $\mathbf{x}$ has been inserted in C_i. Since $E(W,C)$ is independent of C_i, the discriminant becomes

$$g_i(\mathbf{x}) = -E'_i(C) - \sum_{k=1}^{c} P'_k E'_i(W_k) \qquad (8)$$

where $E'_i(C) = -\sum_{k=1}^{c} P'_k \log P'_k$ and $E'_i(W_k) = -\sum_{j=1}^{d} p'_{kj}\log p'_{kj}$. Since $p\log p = 0$ when $p \to 0$, words that do not occur in the system will not contribute to the entropy. The computational complexity

of the proposed minimum system entropy (MSE) classification algorithm is $\mathcal{O}(cd)$, where c is the number of classes and d is the size of the vocabulary. Note that for any given i, $E_i'(W_k)$ needs to be computed only for $k = i$, since for the remaining values of k, $E_i'(W_k)$ remains equal to $E(W_k)$. The algorithm can be quite slow when c or d is very large. In the following section, we derive an approximation to the discriminant function that is computationally more efficient. The approximate discriminant function also provides some insights into the minimum system entropy criterion.

3. Approximate Algorithm for Classification

3.1. Approximate Computation of Change in System Entropy

In this section, we derive a second order approximation to the change in entropy, ΔE, when a new document is inserted into the system. Since entropy function is not well behaved at zero (i.e., the derivative is not finite), a Taylor's series expansion with respect to the original p_{ij}'s can lead to problems. Therefore, we use the new probabilities p'_{ij} as the basis for expansion. (It is to be noted that when a new document is inserted into the system, some of the p_{ij}'s, which were zero before the insertion, can assume small non-zero values.) To accommodate the constraint on the probabilities, without loss of generality, in the following we treat P'_c or p'_{id} as the dependent variable. In other words, $P'_c = 1 - \sum_{i=1}^{c-1} P'_i$ and $p'_{id} = 1 - \sum_{j=1}^{c-1} p'_{ij}$. This implies that

$$dP'_c = -\sum_{i=1}^{c-1} dP'_i \quad \text{and} \quad dp'_{id} = -\sum_{j=1}^{d-1} dp'_{ij}. \tag{9}$$

$E_i'(W, C)$, the entropy when the new document is inserted into class C_i, is defined by

$$\begin{aligned} E_i'(W,C) &= -\sum_{k=1}^{c} P'_k \log P'_k - \sum_{k=1}^{c} P'_k \sum_{j=1}^{d} p'_{kj} \log p'_{kj} \\ &= E_i'(C) + \sum_{k=1}^{c} P'_k E_i'(W_k), \end{aligned} \tag{10}$$

where P'_k and p'_{kj} are defined as in (5) and (6). The change, $\Delta E_i'(W, C)$, in entropy can be written as

$$\Delta E_i'(W,C) = \Delta E_1(C) + \Delta E_2(C) + \sum_{k=1}^{c} P'_k \Delta E_i'(W_k), \tag{11}$$

where

$$\Delta E_1(C) = \sum_{k=1}^{c} \frac{\partial E_i'(C)}{\partial P'_k} \Delta P'_k + \frac{1}{2} \sum_{k=1}^{c} \sum_{j=1}^{c} \frac{\partial^2 E_i'(C)}{\partial P'_k \partial P'_j} \Delta P'_k \Delta P'_j \tag{12}$$

$$\begin{aligned} \Delta E_2(C) &= \sum_{k=1}^{c} \frac{\partial}{\partial P'_k} \left(\sum_{l=1}^{c} P'_l E_i'(W_l) \right) \Delta P'_k \\ &\quad + \frac{1}{2} \sum_{k=1}^{c} \sum_{j=1}^{c} \frac{\partial^2 \left(\sum_{l=1}^{c} P'_l E_i'(W_l) \right)}{\partial P'_k \partial P'_j} \Delta P'_k \Delta P'_j \end{aligned} \tag{13}$$

and

$$\begin{aligned} \Delta E_i'(W_k) &= \sum_{j=1}^{d} \frac{\partial E_i'(W_k)}{\partial p'_{kj}} \Delta p'_{kj} \\ &\quad + \frac{1}{2} \sum_{l=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 E_i'(W_k)}{\partial p'_{kl} \partial p'_{kj}} \Delta p'_{kl} \Delta p'_{kj} \end{aligned} \tag{14}$$

However,

$$\begin{aligned} \frac{\partial E_i'(C)}{\partial P'_k} &= -(1 + \log P'_k) \ , \\ \frac{\partial^2 E_i'(C)}{\partial P'_k \partial P'_j} &= \begin{cases} -\frac{1}{P'_k} & \text{if } k = j, \\ 0 & \text{otherwise} \end{cases} , \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\partial \left(\sum_{l=1}^{c} P'_l E_i'(W_l) \right)}{\partial P'_k} &= E_i'(W_k) \ , \\ \frac{\partial^2 \left(\sum_{l=1}^{c} P'_l E_i'(W_l) \right)}{\partial P'_k \partial P'_j} &= 0 \ , \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{\partial E_i'(W_k)}{\partial p'_{kj}} &= 1 + \log p'_{kj} \quad \text{and} \\ \frac{\partial^2 E_i'(W_k)}{\partial p'_{kl} \partial p'_{kj}} &= \begin{cases} -\frac{1}{p'_{ik}} & \text{if } l = j, \\ 0 & \text{otherwise.} \end{cases} . \end{aligned} \tag{17}$$

By substituting (15), (16) and (17) in (12), (13) and (14) respectively, and by using (9), we obtain

$$\begin{aligned} \Delta E_1(C) &= -\sum_{k=1}^{c} \log P'_k \Delta P'_k - \frac{1}{2} \sum_{k=1}^{c} \frac{1}{P'_k} \Delta P'^2_k, \\ \Delta E_2(C) &= \sum_{k=1}^{c} E_i'(W_k) \Delta P'_k \quad \text{and} \\ \Delta E_i'(W_k) &= -\sum_{j=1}^{d} \log p'_{kj} \Delta p'_{kj} - \frac{1}{2} \sum_{j=1}^{d} \frac{1}{p'_{kj}} (\Delta p'_{kj})^2. \end{aligned} \tag{18}$$

Let $L(\mathbf{x})$ denote the length of the test document $\mathbf{x}$ to be classified. Let $K' = K + L(\mathbf{x})$, $K'_i = K_i + L(\mathbf{x})$, and $k'_{ij} = k_{ij} + x_j$ denote the number of words in the system, the number of words in class C_i, and the number of w_j in class C_i respectively, after $\mathbf{x}$ has been

inserted in class C_i. If the new document $\mathbf{x}$ is assigned to class C_i, then we have

$$\Delta P'_l = \begin{cases} -\frac{L(\mathbf{x})}{K}(1 - P'_l) & \text{if } l = i, \\ \frac{P'_l L(\mathbf{x})}{K} & \text{otherwise.} \end{cases} \quad (19)$$

and

$$\Delta p'_{lj} = \begin{cases} -\frac{1}{K_l}(x_j - p'_{lj} L(\mathbf{x})) & \text{if } l = i, \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

Plugging in (12) - (20) into (11), we obtain

$$\begin{aligned} \Delta E_i(W, C) &= \frac{L(\mathbf{x})}{K}\left(\log P'_i + E'_i(C) - E'_i(W_i)\right) \\ &+ \frac{L(\mathbf{x})}{K}\sum_{l=1}^{c} P'_l E'_i(W_l) - \frac{L^2(\mathbf{x})}{2K'^2}\left(\frac{1 - P'_i}{P'_i}\right) \\ &+ \frac{P'_i}{K_i}\sum_{j=1}^{d}\left(\log p'_{ij} + E'_i(W_i)\right) x_j \\ &- \frac{P'_i}{2K_i^2}\left(\sum_{j=1}^{d}\frac{x_j^2}{p'_{ij}} - L^2(\mathbf{x})\right) \end{aligned} \quad (21)$$

Ignoring the terms that are common to all classes C_i, and assuming that $P'_i = K'_i/K' \approx K_i/K$ we can obtain the following discriminant function $g^s_i()$ for class C_i.

$$\begin{aligned} g^s_i(\mathbf{x}) &= \frac{1}{K}\sum_{j=1}^{d} x_j \log p'_{ij} + \frac{1}{K} L(\mathbf{x}) \log P'_i \\ &- \frac{1}{2K'K_i}\left(\sum_{j=1}^{d}\frac{x_j^2}{p_{ij}}\right) - \frac{L(\mathbf{x})^2}{2K'^2}. \end{aligned} \quad (22)$$

If we include only the terms that arise out of the first-order approximation, we obtain

$$g^f_i(\mathbf{x}) = \frac{1}{K}\sum_{j=1}^{d} x_j \log p'_{ij} + \frac{1}{K} L(\mathbf{x}) \log P'_i, \quad (23)$$

which is equivalent to

$$g^f_i(\mathbf{x}) = P_i'^{L(\mathbf{x})} \prod_{j=1}^{d} p'^{\,x_j}_{ij}. \quad (24)$$

It is to be noted that in (23) and (24) the summation or product over j needs to be carried out only for j such that $x_j > 0$. Therefore the corresponding p'_{ij} are guaranteed to be non-zero, and (22), (23), and (24) are always well defined. This would not be true if we used $E(C, W)$ as the reference point to do the second-order expansion. In general, for a given document, the number of non-zero entries in $\mathbf{x}$ are much smaller than d. Therefore, computation of the approximate discriminant $g^s_i(\mathbf{x})$ or $g^f_i(\mathbf{x})$, which has a computational complexity of $\mathcal{O}(cL(\mathbf{x}))$, is much more efficient than that of the exact one. In practice, we have noticed that the speedup can be over an order of magnitude.

3.2. Comparison with Naive Bayes

In the case of the multinomial model, the discriminant function for class C_i may be written as

$$g_i(\mathbf{x}) = P(L(\mathbf{x}))L(\mathbf{x})!P_i \prod_{j=1}^{d} \frac{p_{ij}^{x_j}}{x_j!}. \quad (25)$$

Since the terms involving factorials are common to all classes, we see that (24) and (25) are similar, except that the prior probability is raised to the power $L(\mathbf{x})$ in the case of the first-order MSE classifier. However, if the prior probabilities of the classes are equal, then the first order MSE classifier becomes identical to the multinomial case. Moreover, the connection between the multinomial model and cross entropy between the probability distributions of the class and the test vector is well known (Baker & McCallum, 1998). In fact, in (23), the term $\sum_{j=1}^{d} x_j \log p'_{ij}$ is essentially the cross-entropy term, since x_j is a scaled estimate of the probability of word w_j in the new document $\mathbf{x}$.

In the case of the multivariate Bernoulli model, each element in $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ is either 1 or 0 depending on whether word w_j occurs or does not occur in document D_i. The discriminant function for class C_i may be written as

$$g_i(\mathbf{x}) = P_i \prod_{j=1}^{d} p_{ij}^{x_j}(1 - p_{ij})^{1 - x_j}. \quad (26)$$

Note that in this case, P_i is computed as n_i/n, i.e., the fraction of documents labelled as C_i. Comparing with (24), we see that minimum entropy classification does not directly incorporate the negative evidence provided by the non-occurrence of words. However, for the special case when there is only one binary feature, $E(W_i)$ may be written as

$$E(W_i) = -p_{i0} \log p_{i0} - p_{i1} \log p_{i1}, \quad (27)$$

where p_{i0} is the probability that the feature is '0' in class C_i and p_{i1} is the probability that the feature is '1'. Moreover, if the new document to be inserted has a feature value of x, we may write

$$\Delta P'_l = \begin{cases} \frac{K_l}{K} - \frac{K_l + 1}{K + 1} = \frac{P'_l - 1}{K} & \text{if } l = i, \\ \frac{K_l}{K} - \frac{K_l}{K + 1} = \frac{P'_l}{K} & \text{otherwise.} \end{cases} \quad (28)$$

and

$$\begin{cases} \Delta p'_{i0} = \frac{k_{i0}}{K_i} - \frac{K_i + 1 - x}{K_i + 1} & = \frac{p'_{i0} + x - 1}{K_i} \\ \Delta p'_{i1} = \frac{k_{i1}}{K_i} - \frac{ki1 + x}{K_i + 1} & = \frac{p'_{i1} - x}{K_i} \\ \Delta p'_{l0} = \Delta p'_{l1} = 0 & \text{when } l \neq i \end{cases} \quad (29)$$

Hence, we obtain

$$\Delta E_i(C,W) = \frac{L(x)}{K}\left(\log P_i' + E_i'(C) + \sum_{l=1}^{c} P_l' E_i'(W_l)\right) + \frac{1}{K}\left((1-x)\log p_{i0}' + x \log p_{i1}'\right) \quad (30)$$

Again, ignoring terms that are common to all classes C_i, we obtain the discriminant function

$$g_i(\mathbf{x}) = P_i p_{i1}^{x} p_{i0}^{(1-x)} = P_i p_{i1}^{x}(1-p_{i1})^{(1-x)} \quad (31)$$

Thus, the minimum system entropy criterion becomes identical to the multivariate Bernoulli criterion.

4. Experimental Results

In this section, we provide empirical evidence for the good performance of the minimum system entropy (MSE) method. The experiments are based on comparisons with Naive Bayes and Support Vector Machines on several well-known data sets such as **WebKB**, **20 Newsgroups**, and **Syskill Webert** data. In addition, to demonstrate the scalability of MSE, we have also included results on the Internet directory, **dmoz** [1].

4.1. Data Sets

In this section, we describe the data sets and how we used them. We did not use any kind of feature selection (stopword removal, pruning of vocabulary based on information gain, de-tagging, removal of terms that occur only once, etc.) or manipulation (stemming, etc.) for any of the data sets. This is different from the way some of the data was used in the previous works by Nigam et al. (1998) and Joachims (1998), where the authors chose to perform different pre-processing steps on different data sets based on whichever procedure provided the best results.

The **WebKB** data set [2] (Craven et al., 1997) contains Web pages gathered from the computer science departments of several US universities. As done by Nigam et al. (1999), we have chosen the four most populous and popular categories: *student, faculty, course* and *project*, altogether consisting of 4,199 documents. We removed the server headers before tokenizing the documents. We found that this data set has 80,897 unique words.

The **Newsgroups** [2] data set contains 19,997 articles evenly divided among twenty UseNet discussion groups (Joachims, 1997). There are five categories that discuss issues related to computer science, three related to religion and a couple related to sports such as baseball and hockey. While tokenizing, we discarded the subject and newsgroup category along with the complete UseNet header. Total number of unique words for this data set was 131,686.

The **dmoz** data set was created from the RDF distribution of the open directory project from Mozilla. The RDF distribution currently has 3,466,271 sites in 450,767 categories. The RDF distribution has sites and their short description. We crafted two data sets from the distribution, one consisting of 2000 randomly picked sites belonging to the 18 top categories ('Arts', 'Business', etc.), which we shall call **DmozTop** and another consisting of all the sites in the four most populous categories under 'Arts' (i.e., 'Music', 'Movies', 'Literature', and 'Television'), which we shall call **4DmozArts**. We considered only the description of the sites and not their original contents for classification. **DmozTop** has a total of 18 classes with 2000 documents each and the vocabulary size is 55,842. **4DmozArts** 85,530 documents with a vocabulary size of 61,350, and is divided into 4 classes.

The **Syskill and Webert** [3] Web page rating data set contains the HTML source Web pages plus a user's ratings of these Web pages. The Web pages are on four separate subjects (Bands - recording artists; Goats; Sheep; and BioMedical). Each Web page has been rated as either 'hot', 'medium' or 'cold'. The classification task is to guess the rating of unseen Web pages. As in the previous works by Pazzani and Billsus (1997) We have considered only the 'hot' and 'cold' classes, since the number of pages in the 'medium' class is very small. There are total of 225 documents with 12,971 unique words.

4.2. Experiments

We present results pertaining to three experiments. In the first experiment, we compare the classification accuracy of MSE against the multinomial Naive Bayes algorithm and SVM. In the second experiment, we study the influence of feature selection on the classification accuracy of MSE, Naive Bayes and SVM. In the third experiment, we compare the effect of the size of the training data on the performance of these methods. For the implementation of the naive Bayes and SVM classifiers, we used the *rainbow* package in the bow distribution (McCallum, 1996).

For the comparison experiment, we used the standard definition of classification accuracy, also used by Mc-

[1] http://dmoz.org

[2] http://www-2.cs.cmu.edu/~TextLearning/datasets.html

[3] http://kdd.ics.uci.edu

Callum and Nigam (1998). We randomly chose 60% of documents in the corpus for training and the rest for testing. We conducted a total of 20 such trials for each corpus. Our classifier and *rainbow* were both trained with the same set of training documents. Further, *rainbow* was configured to use Laplace smoothing for naive Bayes, linear kernel for SVM, and no feature selection. Table 1 presents the average classification accuracies over 20 random samples for various data sets. In the case of MSE, we performed the classification using three different methods. The brute force method, which uses (8), the first-order approximation method, which uses (24), and the second-order approximation method, which uses (22).

Table 1. Average classification accuracy(%) of Naive Bayes, SVM, and MSE

Data sets	Naive Bayes	SVM	Min. System Entropy		
			First Order	Second Order	Brute Force
WebKB	80.05	87.8	80.1	83.3	85.2
Newsgroup	77.65	79.78	76.9	80.48	81.1
4DmozArt	92.86	-	93.79	94.33	94.89
DmozTop	74.44	-	79.04	81.72	82.01
Syskill	71.80	78.95	79.54	80.35	80.80

From the table, we see that the Brute Force and Second order methods outperform Naive Bayes. The performance of First Order is comparable to that of Naive Bayes. The table also shows that Second Order is a good approximation of the MSE criterion. The results with respect to SVM are somewhat mixed. On the WebKB data set, SVM outperforms MSE. However, on the remaining data sets, MSE outperforms SVM. Owing to the large size of **dmoz**, we were unable to train SVM on it even on a dual-processor P III, 1 GHz, machine with 1 GB RAM. Therefore, the results could not be included for comparison.

In the second experiment, we studied the influence of feature selection by restricting the vocabulary to the top N features that have the most information gain (Yang & Pedersen, 1997) in the training sample, where N varied from 200 to the entire vocabulary of the training set. We again used the feature selection method incorporated in *rainbow* for this experiment. The training set was a random 60%-sample of the corpus. Figure 1 shows the plot of classification accuracy of the various algorithms for the **Newsgroups** data set when the top N features with the highest information gain criterion are selected. Figure 2 shows a similar plot for the **WebKB** data set. From the two plots, we see that the behaviour of MSE is very similar to that of Naive Bayes. The plots also indicate that the performance of SVM is more heavily influenced by feature selection, whereas MSE is relatively unaffected. SVM can outperform MSE at certain levels of feature selection.

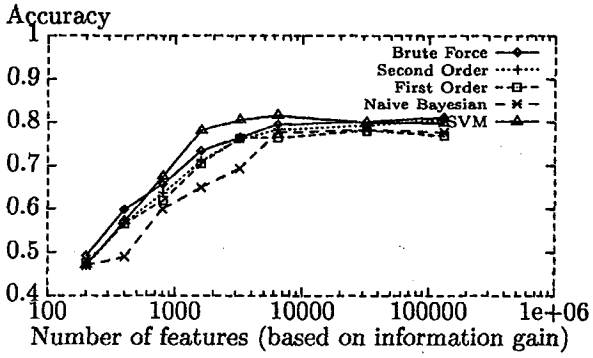

Figure 1. Classification accuracy on **NewsGroups** when top N features by information gain are selected

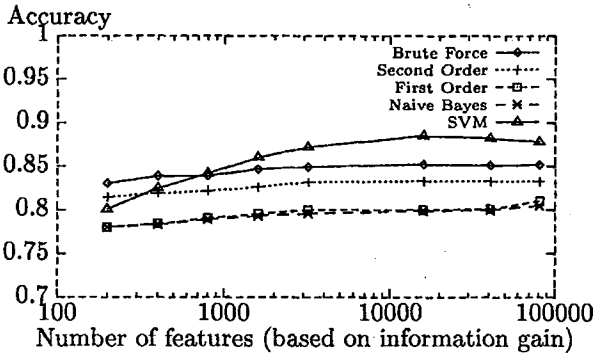

Figure 2. Classification accuracy on **WebKB** when top N features by information gain are selected

In the third experiment, we studied the effect of the size of the training data set on the classification accuracy. We varied the size of the training data from 10% of the corpus to 90%. Figure 3 shows the plot for the **Newsgroups** data set and Figure 4 shows a similar plot for the **WebKB** data set. From the figures, we see that the behavior of the MSE algorithm is similar to that of Naive Bayes. Moreover, since only a few manually-labelled documents are required to train the algorithm, it can be used for populating manually created taxonomies. The plots also show that SVM can suffer from overfitting when the percentage of training documents is high.

The scaled naive Bayes (SNB) algorithm (Nigam et al., 1999; Nigam et al., 2000) is another approach that has been shown to perform well on certain data sets. In SNB, the document vectors $\mathbf{x}_i$ are scaled in such a manner that all of them have the same length, i.e.,

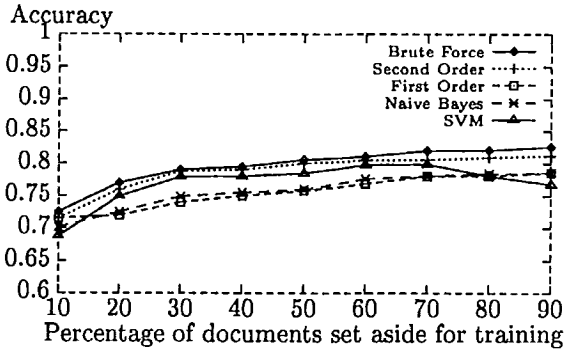

Figure 3. Effect of size of training data set on classification accuracy for **Newsgroups**

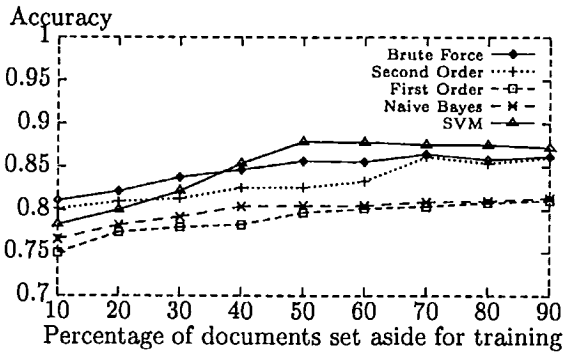

Figure 4. Effect of size of training data set on classification accuracy for the **WebKB** data set

$\sum_{j=1}^{d} x_{ij}$ is the same for all document vectors. It follows that when all classes have the same number of documents, the prior probabilities will all be equal in the case of SNB, regardless of the original lengths of the individual documents. We have noticed that when the class priors are very unequal, SNB does not perform as well. Our comparisons of MSE with SNB on data sets that have roughly equal priors (e.g. **Industry Sector** [3] and **NewsGroups**) indicates that MSE is better.

5. Summary and Conclusions

In this paper, we introduced a new approach to classification of text documents. The approach uses the entropy of the collection of training documents as a measure of the uncertainty associated with the joint distribution of words and labels of the documents. A new document is assigned the label that results in the maximum decrease in the system entropy.

The computational complexity for classification of a new document using a brute force version of the proposed method is $\mathcal{O}(cd)$, where c is the number of classes and d is the number of words in the vocabulary, i.e., the size of the dictionary. This can be slow, since the value of c and d can be fairly large in a given application. To overcome this drawback, we have derived an approximation that computes the expected change in the entropy when a document is inserted into class C_i. The computational complexity of the approximate expression is only $\mathcal{O}(cL(\mathbf{x}))$, where $L(\mathbf{x})$ is the length of the document to be classified. Thus, it is no worse than that of traditional naive Bayes algorithms.

Our analysis shows that the first order approximation of the proposed approach can be related to the naive Bayes techniques and the cross entropy methods under some conditions. The proposed approach also implicitly takes into account the length of the document to be classified. The experimental results indicate that MSE performs well even when no feature selection is used. Moreover, unlike SVM, it is easy to train even when the size and dimensionality of the data set is very large. In other words, it requires no feature selection, has no thresholds, parameters, or kernels to choose, while at the same time being computationally efficient both in terms and training and testing.

In our future work, we propose to extend the concept of system entropy to the more general case of a hierarchy of classes. In this case, the classification problem becomes one of finding the correct node in the hierarchy at which a given document should be inserted, i.e., the problem is to populate a given taxonomy (or ontology). It is to be noted that a document will be inserted at one of the leaf nodes only if it is sufficiently specific. Otherwise, it could be inserted at one of the intermediate nodes. Therefore, the entropy model needs to be able to facilitate such decisions.

References

Baker, L. D., & McCallum, A. K. (1998). Distributional clustering of words for text classification. *Proceedings of twenty first International Conference on Research and Development in Information Retrieval* (pp. 96–103). Melbourne, AU.

Cohen, W., & Singer, Y. (1996). Context-sensitive learning methods for text categorization. *Proceedings of nineteenth International Conference on Research and Development in Information Retrieval* (pp. 307–315). Zürich, CH.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (1997). Learning to extract symbolic knowledge from the World Wide Web. *Proceedings of the National Conference on Artificial Intelligence* (pp. 509–516).

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131–163.

Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *Proceedings of fourteenth International Conference on Machine Learning* (pp. 143–151). Nashville, US.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *Proceedings of 10th European Conference on Machine Learning* (pp. 137–142). Chemnitz, DE: Springer Verlag, Heidelberg, DE.

Kalt, T., & Croft, W. B. (1996). *A new probabilistic model for text classifiaction and retrieval* (Technical Report).

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. *Proceedings of fourteenth International Conference on Machine Learning* (pp. 170–178). Nashville, US.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of bayesian classifiers. *Proceedings of National Conference on Artificial Intelligence* (pp. 223–228).

Larkey, L. S., & Croft, W. B. (1996). Combining classifiers in text categorization. *Proceedings of nineteenth International Conference on Research and Development in Information Retrieval* (pp. 289–297). Zürich, CH.

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. *Proceedings of Tenth European Conference on Machine Learning* (pp. 137–142).

Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. *Proceedings of seventeenth International Conference on Research and Development in Information Retrieval* (pp. 331–339).

Li, H., & Yamanishi, K. (1997). Document classification using a finite mixture model. *In Proceedings of the thirty fifth Annual Meeting of the Association for Computaional Linguistics.*

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *Proceedings of Workshop on Learning for Text Categorizationi, American Association for Artificial Intelligence.*

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

McCallum, A. K., Rosenfeld, R., Mitchell, T. M., & Ng, A. Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. *Proceedings of fifteenth International Conference on Machine Learning* (pp. 359–367). Madison, US.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. *Proceedings of Workshop on Machine Learning for Information Filteringi, International Joint Conference on Artificial Intelligence*, 61–67.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*, 103–134.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (1998). Learning to classify text from labeled and unlabeled documents. *Proceedings of fifteenth Conference of the American Association for Artificial Intelligence* (pp. 792–799). Madison, US.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning, 27*, 313–331.

Robertson, S. E., & Spark-Jones, K. (1976). Relevance weighting of search terms. *The American Society for Information Science, 27*, 129–146.

S. T. Dumais, J. Platt, D. H., & Sahami, M. (1998). Inductive learning algorithms representations for text categorization. *Proceedings of seventh International Conference on Information and Knowledge Management.*

Slattery, S., & Craven, M. (1998). Combining statistical and relational methods for learning in hypertext domains. *Proceedings of eigth International Conference on Inductive Logic Programming* (pp. 38–52). Madison, US: Springer Verlag, Heidelberg, DE.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*, 69–90.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning* (pp. 412–420). Nashville, US.