A New Feature Selection Score for Multinomial Naive Bayes Text Classification Based on KL-Divergence

Karl-Michael Schneider
Department of General Linguistics
University of Passau
94032 Passau, Germany
schneide@phil.uni-passau.de

Abstract

We define a new feature selection score for text classification based on the KL-divergence between the distribution of words in training documents and their classes. The score favors words that have a similar distribution in documents of the same class but different distributions in documents of different classes. Experiments on two standard data sets indicate that the new method outperforms mutual information, especially for smaller categories.

1 Introduction

Text classification is the assignment of predefined categories to text documents. Text classification has many applications in natural language processing tasks such as E-mail filtering, prediction of user preferences and organization of web content.

The Naive Bayes classifier is a popular machine learning technique for text classification because it performs well in many domains, despite its simplicity (Domingos and Pazzani, 1997). Naive Bayes assumes a stochastic model of document generation. Using Bayes' rule, the model is inverted in order to predict the most likely class for a new document.

We assume that documents are generated according to a multinomial event model (McCallum and Nigam, 1998). Thus a document is represented as a vector $d_i = (x_{i1} \dots x_{i|V|})$ of word counts where V is the vocabulary and each $x_{it} \in \{0, 1, 2, \dots\}$ indicates how often w_t occurs in d_i. Given model parameters $p(w_t|c_j)$ and class prior probabilities $p(c_j)$ and assuming independence of the words, the most likely class for a document d_i is computed as

$$\begin{aligned} c^*(d_i) &= \operatorname*{argmax}_j p(c_j)p(d|c_j) \\ &= \operatorname*{argmax}_j p(c_j) \prod_{t=1}^{|V|} p(w_t|c_j)^{n(w_t,d_i)} \end{aligned} \tag{1}$$

where $n(w_t, d_i)$ is the number of occurrences of w_t in d_i. $p(w_t|c_j)$ and $p(c_j)$ are estimated from training documents with known classes, using maximum likelihood estimation with a Laplacean prior:

$$p(w_t|c_j) = \frac{1 + \sum_{d_i \in c_j} n(w_t, d_i)}{|V| + \sum_{t=1}^{|V|} \sum_{d_i \in c_j} n(w_t, d_i)} \tag{2}$$

$$p(c_j) = \frac{|c_j|}{\sum_{j'=1}^{|C|} |c_{j'}|} \tag{3}$$

It is common practice to use only a subset of the words in the training documents for classification to avoid overfitting and make classification more efficient. This is usually done by assigning each word a score $f(w_t)$ that measures its usefulness for classification and selecting the N highest scored words. One of the best performing scoring functions for feature selection in text classification is mutual information (Yang and Pedersen, 1997). The mutual information between two random variables, $MI(X;Y)$, measures the amount of information that the value of one variable gives about the value of the other (Cover and Thomas, 1991).

Note that in the multinomial model, the word variable W takes on values from the vocabulary V. In order to use mutual information with a multinomial model, one defines new random variables $W_t \in \{0, 1\}$ with $p(W_t = 1) = p(W = w_t)$ (McCallum and Nigam, 1998; Rennie, 2001). Then the mutual information between a word w_t and the class variable C is

$$MI(W_t; C) = \sum_{j=1}^{|C|} \sum_{x=0,1} p(x, c_j) \log \frac{p(x, c_j)}{p(x)p(c_j)} \tag{4}$$

where $p(x, c_j)$ and $p(x)$ are short for $p(W_t = x, c_j)$ and $p(W_t = x)$. $p(x, c_j)$, $p(x)$ and $p(c_j)$ are estimated from the training documents by counting how often w_t occurs in each class.

2 Naive Bayes and KL-Divergence

There is a strong connection between Naive Bayes and KL-divergence (Kullback-Leibler divergence, relative entropy). KL-divergence measures how

much one probability distribution is different from another (Cover and Thomas, 1991). It is defined (for discrete distributions) by

$$KL(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \quad (5)$$

By viewing a document as a probability distribution over words, Naive Bayes can be interpreted in an information-theoretic framework (Dhillon et al., 2002). Let $p(w_t|d) = n(w_t,d)/|d|$. Taking logarithms and dividing by the length of d, (1) can be rewritten as

$$\begin{aligned} & c^*(d) \\ &= \operatorname*{argmax}_j \log p(c_j) + \sum_{t=1}^{|V|} n(w_t,d) \log p(w_t|c_j) \\ &= \operatorname*{argmax}_j \frac{1}{|d|} \log p(c_j) + \sum_{t=1}^{|V|} p(w_t|d) \log p(w_t|c_j) \end{aligned} \quad (6)$$

Adding the entropy of $p(W|d)$, we get

$$\begin{aligned} & c^*(d) \\ &= \operatorname*{argmax}_j \frac{1}{|d|} \log p(c_j) - \sum_{t=1}^{|V|} p(w_t|d) \log \frac{p(w_t|d)}{p(w_t|c_j)} \\ &= \operatorname*{argmin}_j KL(p(W|d), p(W|c_j)) - \frac{1}{|d|} \log p(c_j) \end{aligned} \quad (7)$$

This means that Naive Bayes assigns to a document d the class which is "most similar" to d in terms of the distribution of words. Note also that the prior probabilities are usually dominated by document probabilities except for very short documents.

3 Feature Selection using KL-Divergence

We define a new scoring function for feature selection based on the following considerations. In the previous section we have seen that Naive Bayes assigns a document d the class c^* such that the "distance" between d and c^* is minimized. A classification error occurs when a test document is closer to some other class than to its true class, in terms of KL-divergence.

We seek to define a scoring function such that words whose distribution in the individual training documents of a class is much different from the distribution in the class (according to (2)) receive a lower score, while words with a similar distribution in all training documents of the same class receive a higher score. By removing words with a lower score from the vocabulary, the training documents of each class become more similar to each other, and therefore, also to the class, in terms of word distribution. This leads to more homogeneous classes. Assuming that the test documents and training documents come from the same distribution, the similarity between the test documents and their respective classes will be increased as well, thus resulting in higher classification accuracy.

We now make this more precise. Let $S = \{d_1, \ldots, d_{|S|}\}$ be the set of training documents, and denote the class of d_i with $c(d_i)$. The average KL-divergence for a word w_t between the training documents and their classes is given by

$$KL_t(S) = \frac{1}{|S|} \sum_{d_i \in S} KL(p(w_t|d_i), p(w_t|c(d_i))). \quad (8)$$

One problem with (8) is that in addition to the conditional probabilities $p(w_t|c_j)$ for each word and each class, the computation considers each individual document, thus resulting in a time requirement of $\mathcal{O}(|S|)$.[1] In order to avoid this additional complexity, instead of $KL_t(S)$ we use an approximation $\widetilde{KL}_t(S)$, which is based on the following two assumptions: (i) the number of occurrences of w_t is the same in all documents that contain w_t, (ii) all documents in the same class c_j have the same length. Let N_{jt} be the number of documents in c_j that contain w_t, and let

$$\tilde{p}_d(w_t|c_j) = p(w_t|c_j) \frac{|c_j|}{N_{jt}} \quad (9)$$

be the average probability of w_t in those documents in c_j that contain w_t (if w_t does not occur in c_j, set $\tilde{p}_d(w_t|c_j) = 0$). Then $KL_t(S)$ reduces to

$$\widetilde{KL}_t(S) = \frac{1}{|S|} \sum_{j=1}^{|C|} N_{jt} \tilde{p}_d(w_t|c_j) \log \frac{\tilde{p}_d(w_t|c_j)}{p(w_t|c_j)}. \quad (10)$$

Plugging in (9) and (3) and defining $q(w_t|c_j) = N_{jt}/|c_j|$, we get

$$\widetilde{KL}_t(S) = -\sum_{j=1}^{|C|} p(c_j) p(w_t|c_j) \log q(w_t|c_j). \quad (11)$$

Note that computing $\widetilde{KL}_t(S)$ only requires a statistics of the number of words and documents for each

[1] Note that $KL_t(S)$ cannot be computed simultaneously with $p(w_t|c_j)$ in one pass over the documents in (2): $KL_t(S)$ requires $p(w_t|c_j)$ when each document is considered, but computing the latter needs iterating over all documents itself.

class, not per document. Thus $\widetilde{KL}_t(S)$ can be computed in $\mathcal{O}(|C|)$. Typically, $|C|$ is much smaller than $|S|$.

Another important thing to note is the following. By removing words with an uneven distribution in the documents of the same class, not only the documents in the class, but also the classes themselves may become more similar, which reduces the ability to distinguish between different classes. Let $p(w_t)$ be the number of occurrences of w_t in all training documents, divided by the total number of words, $q(w_t) = \sum_{j=1}^{|C|} N_{jt}/|S|$ and define

$$\widetilde{K}_t(S) = -p(w_t) \log q(w_t). \tag{12}$$

$\widetilde{K}_t(S)$ can be interpreted as an approximation of the average divergence of the distribution of w_t in the individual training documents from the global distribution (averaged over all training documents in all classes). If w_t is independent of the class, then $\widetilde{K}_t(S) = \widetilde{KL}_t(S)$. The difference between the two is a measure of the increase in homogeneity of the training documents, in terms of the distribution of w_t, when the documents are clustered in their true classes. It is large if the distribution of w_t is similar in the training documents of the same class but dissimilar in documents of different classes. In analogy to mutual information, we define our new scoring function as the difference

$$KL(w_t) = \widetilde{K}_t(S) - \widetilde{KL}_t(S). \tag{13}$$

We also use a variant of *KL*, denoted *dKL*, where $p(w_t)$ is estimated according to (14):

$$p'(w_t) = \sum_{j=1}^{|C|} p(c_j)p(w_t|c_j) \tag{14}$$

and $p(w_t|c_j)$ is estimated as in (2).

4 Experiments

We compare *KL* and *dKL* to mutual information, using two standard data sets: 20 Newsgroups[2] and Reuters 21578.[3] In tokenizing the data, only words consisting of alphabetic characters are used after conversion to lower case. In addition, all numbers are mapped to a special token NUM. For 20 Newsgroups we remove the newsgroup headers and use a stoplist consisting of the 100 most frequent words of the British National Corpus.[4] We use the ModApte split of Reuters 21578 (Apté et al., 1994) and use only the 10 largest classes. The vocabulary size is 111868 words for 20 Newsgroups and 22430 words for Reuters.

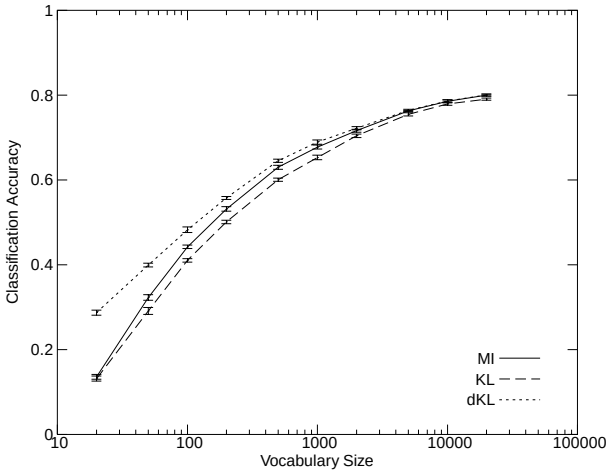

Figure 1: Classification accuracy for 20 Newsgroups. The curves have small error bars.

Experiments with 20 Newsgroups are performed with 5-fold cross-validation, using 80% of the data for training and 20% for testing. We build a single classifier for the 20 classes and vary the number of selected words from 20 to 20000. Figure 1 compares classification accuracy for the three scoring functions. *dKL* slightly outperforms mutual information, especially for smaller vocabulary sizes. The difference is statistically significant for 20 to 200 words at the 99% confidence level, and for 20 to 2000 words at the 95% confidence level, using a one-tailed paired t-test.

For the Reuters dataset we build a binary classifier for each of the ten topics and set the number of positively classified documents such that precision equals recall. Precision is the percentage of positive documents among all positively classified documents. Recall is the percentage of positive documents that are classified as positive.

In Figures 2 and 3 we report microaveraged and macroaveraged recall for each number of selected words. Microaveraged recall is the percentage of all positive documents (in all topics) that are classified as positive. Macroaveraged recall is the average of the recall values of the individual topics. Microaveraged recall gives equal weight to the documents and thus emphasizes larger topics, while macroaveraged recall gives equal weight to the topics and thus emphasizes smaller topics more than microav-

[2] http://www.ai.mit.edu/∼jrennie/20Newsgroups/

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[4] http://www.itri.brighton.ac.uk/∼Adam.Kilgarriff/bnc-readme.html

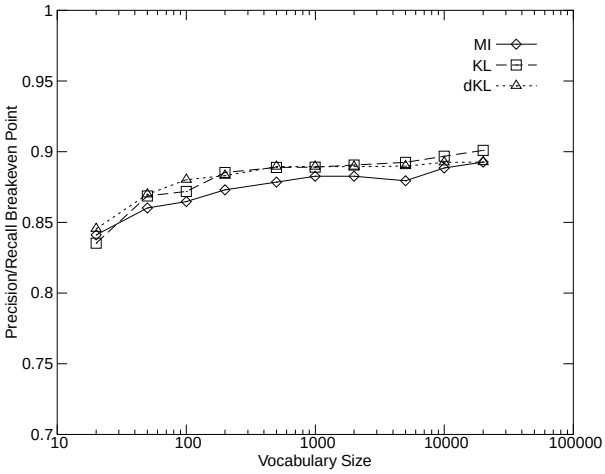

Figure 2: Microaveraged recall on Reuters at breakeven point.

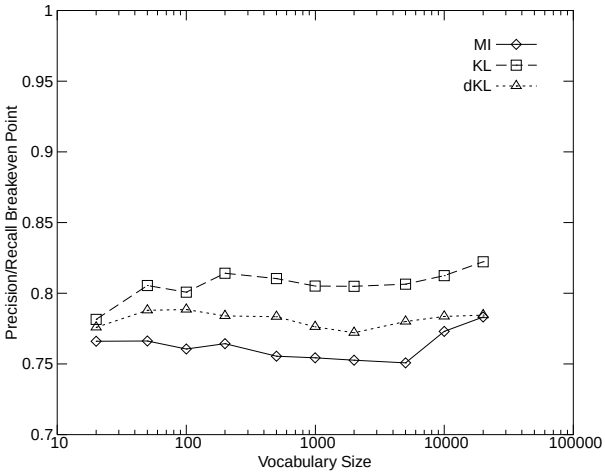

Figure 3: Macroaveraged recall on Reuters at breakeven point.

eraged recall.

Both *KL* and *dKL* achieve slightly higher values for microaveraged recall than mutual information, for most vocabulary sizes (Fig. 2). *KL* performs best at 20000 words with 90.1% microaveraged recall, compared to 89.3% for mutual information. The largest improvement is found for *dKL* at 100 words with 88.0%, compared to 86.5% for mutual information.

For smaller categories, the difference between the KL-divergence based scores and mutual information is larger, as indicated by the curves for macroaveraged recall (Fig. 3). *KL* yields the highest recall at 20000 words with 82.2%, an increase of 3.9% compared to mutual information with 78.3%, whereas *dKL* has its largest value at 100 words with 78.8%, compared to 76.1% for mutual information. We find the largest improvement at 5000 words with 5.6% for *KL* and 2.9% for *dKL*, compared to mutual information.

5 Conclusion

By interpreting Naive Bayes in an information theoretic framework, we derive a new scoring method for feature selection in text classification, based on the KL-divergence between training documents and their classes. Our experiments show that it outperforms mutual information, which was one of the best performing methods in previous studies (Yang and Pedersen, 1997). The KL-divergence based scores are especially effective for smaller categories, but additional experiments are certainly required.

In order to keep the computational cost low, we use an approximation instead of the exact KL-divergence. Assessing the error introduced by this approximation is a topic for future work.

References

Chidanand Apté, Fred Damerau, and Sholom M. Weiss. 1994. Towards language independent automated learning of text categorization models. In *Proc. 17th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, pages 23–30.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley, New York.

Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. 2002. Enhanced word clustering for hierarchical text classification. In *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 191–200.

Pedro Domingos and Michael Pazzani. 1997. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *Learning for Text Categorization: Papers from the AAAI Workshop*, pages 41–48. AAAI Press. Technical Report WS-98-05.

Jason D. M. Rennie. 2001. Improving multi-class text classification with Naive Bayes. Master's thesis, Massachusetts Institute of Technology.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proc. 14th International Conference on Machine Learning (ICML-97)*, pages 412–420.