Enhanced Word Clustering for Hierarchical Text Classification

Inderjit S. Dhillon
inderjit@cs.utexas.edu

Subramanyam Mallela
manyam@cs.utexas.edu

Rahul Kumar
rahul@cs.utexas.edu

Department of Computer Sciences,
University of Texas, Austin, TX 78712.

March 1, 2002

Abstract

In this paper we propose a new information-theoretic divisive algorithm for word clustering applied to text classification. In previous work, such "distributional clustering" of features has been found to achieve significant improvements over feature selection in terms of classification accuracy, especially at lower number of features [2, 29]. However the existing clustering techniques are agglomerative in nature resulting in (i) sub-optimal word clusters and (ii) high computational cost. In order to explicitly capture the optimality of word clusters in an information theoretic framework, we first derive a global criterion for feature clustering. We then present a fast, divisive algorithm that monotonically decreases this objective function value, thus converging to a local minimum. We show that our algorithm minimizes the "within-cluster Jensen-Shannon divergence" while simultaneously maximizing the "between-cluster Jensen-Shannon divergence". In comparison to the previously proposed agglomerative strategies our divisive algorithm achieves higher classification accuracy especially at lower number of features. We further show that feature clustering is an effective technique for building smaller class models in hierarchical classification. We present detailed experimental results on the 20 News groups data set and a 3-level hierarchy of HTML documents collected from Dmoz Open Directory.

1 Introduction

Given a set of document vectors $\{d_1, d_2, \ldots, d_n\}$ and their associated class labels $c(d_i) \in \{c_1, c_2, \ldots, c_l\}$, text classification is the problem of estimating the true class label of a new document d. There exist a wide variety of algorithms for text classification, ranging from the simple but effective Naive Bayes algorithm to the more computationally demanding Support Vector Machines [24, 30, 31].

A common, and often overwhelming, characteristic of text data is its extremely high dimensionality. Typically the document vectors are formed using a vector-space or bag-of-words model [26]. Even a moderately sized document collection can lead to a dimensionality in thousands, for example, one of our test data sets contains 5,000 web pages from www.dmoz.org and has a dimensionality (vocabulary size) of 14,538. This high dimensionality can be a severe obstacle for classification algorithms based on Support Vector Machines, Linear Discriminant Analysis, k-nearest neighbor etc. The problem is compounded when the documents are arranged in a hierarchy of classes since a full-feature classifier needs to be applied at each node of the hierarchy.

A way to reduce dimensionality is by the distributional clustering of words/features [25, 2, 29]. Each word cluster can be treated as a single feature and thus, dimensionality can be drastically reduced. As shown by [2, 29], such feature clustering is more effective than feature selection [32], especially at lower number of features. Also, feature clustering appears to preserve classification accuracy as compared to a full-feature classifier. Indeed in some cases of small training sets and noisy features, word clustering can actually increase

accuracy in classification. However, the algorithms given in both [2] and [29] are agglomerative in nature thus yielding sub-optimal word clusters at a high computational cost.

In this paper, we first derive a global criterion that captures the optimality of word clustering in an information-theoretic framework. This leads to an objective function for clustering that is based on the generalized Jensen-Shannon divergence [20] among an arbitrary number of probability distributions. In order to find the best word clustering, i.e., the clustering that minimizes this objective function, we present a new divisive algorithm for clustering words. This algorithm is reminiscent of the k-means algorithm but uses Kullback Leibler divergences [18] instead of squared Euclidean distances. We prove that our divisive algorithm *monotonically* decreases the objective function value, thus converging to a local minimum. We also show that our algorithm minimizes "within-cluster divergence" and simultaneously maximizes "between-cluster divergence". Thus we find word clusters that are markedly better than the agglomerative algorithms of [2, 29]. The increased quality of our word clusters translates to higher classification accuracies, especially at small feature sizes and small training sets. We provide empirical evidence of all the above claims using a Naive Bayes classifier on the (a) CMU 20 newsgroup data set, and (b) an HTML data set comprising 5,000 web pages arranged in a 3-level hierarchy from the Open directory project (www.dmoz.org).

We now give a brief outline of the paper. In Section 2, we discuss related work and contrast it with our work. In Section 3 we briefly review some useful concepts from information theory such as Kullback-Leibler(KL) divergence and Jensen-Shannon(JS) divergence, while in Section 4 we review Naive Bayes and show how to interpret it in terms of KL-divergence. Section 5 poses the question of finding optimal word clusters in terms of preserving mutual information between two random variables. Section 5.1 gives the algorithm that directly minimizes the resulting objective function which is based on KL-divergences, and presents some pleasing results about the algorithm, such as convergence and simultaneous maximization of "between-cluster JS-divergence". In Section 6 we present experimental results that show the superiority of our word clustering, and the resulting increase in classification accuracy. Finally, we present our conclusions in Section 7.

A word about notation: upper-case letters such as X, Y, C, W will denote random variables, while script upper-case letters such as $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{C}$, $\mathcal{W}$ denote sets. Individual set elements will often be denoted by lower-case letters such as x, w or x_i, w_t. Probability distributions will be denoted by p, q, p_1, p_2, etc. when the random variable is obvious or by $p(X)$, $p(C|w_t)$ to make the random variable explicit.

2 Related Work

Text classification has been extensively studied, especially since the emergence of the internet. Most algorithms are based on the bag-of-words model for text [26]. A simple but effective algorithm is the Naive Bayes method [24]. For text classification, different variants of Naive Bayes have been used, but McCallum and Nigam [21] showed that the variant based on the multinomial model leads to better results. For hierarchical text data, such as the topic hierarchies of Yahoo! (www.yahoo.com) and the Open Directory Project (www.dmoz.org), hierarchical classification has been studied in [17, 4, 10]. For some more details, see Section 4.1.

To counter high-dimensionality, various methods of feature selection have been proposed in [32, 17, 4]. Distributional clustering of words was first proposed by Pereira, Tishby and Lee in [25] where they used "soft" distributional clustering to cluster nouns according to their conditional verb distributions. Note that since our main goal is to reduce the number of features *and* the model size, we are only interested in "hard clustering" where each word can be represented by its (unique) word cluster. For text classification, Baker and McCallum used such hard clustering in [2], while more recently, Slonim and Tishby have used the so-called Information Bottleneck method for clustering words in [29]. Both these works use an identical agglomerative clustering strategy that makes a greedy move at every agglomeration. Both [2, 29] showed that the feature size can be aggressively reduced by such clustering without any noticeable loss in classification accuracy using Naive Bayes. Similar results have been reported for Support Vector Machines [3].

Two other dimensionality/feature reduction schemes are used in latent semantic indexing (LSI) [6] and its probabilistic version [16]. Typically these methods have been applied in the *unsupervised* setting and as shown in [2], LSI results in lower classification accuracies than feature clustering.

We now list the main contributions of this paper and contrast them with earlier work. As our first

contribution, we derive a global criterion that explicitly captures the optimality of word clusters in an information theoretic framework. This leads to an objective function in terms of the generalized Jensen-Shannon divergence among an arbitrary number of probability distributions. As our second contribution, we present a divisive algorithm that uses Kullback-Leibler divergence as the distance measure, and explicitly minimizes the global objective function. This is in contrast to [29] who considered the merging of *just two* word clusters at every step and derived a local criterion based on the Jensen-Shannon divergence of *two* probability distributions. Their agglomerative algorithm, which is similar to Baker and McCallum's algorithm [2], greedily optimizes this merging criterion. Thus, their resulting algorithm can yield sub-optimal clusters and is computationally expensive (the algorithm in [29] is $O(m^3 l)$ in complexity where m is the total number of words and l is the number of classes). In contrast our divisive algorithm is $O(mkl)$ where k is the number of word clusters required (typically $k \ll m$). Note that our hard clustering leads to a model size of $O(k)$, whereas soft clustering methods such as probabilistic LSI [16] lead to a model size of $O(wk)$. Finally, we show that our enhanced word clustering leads to higher classification accuracy, especially when the training set is small and in hierarchical classification of HTML data.

3 Some Information Theory Concepts

In this section, we quickly review some concepts from information theory which will be used heavily in this paper. For more details see the authoritative treatment in the book by Cover & Thomas [5].

Let X be a discrete random variable that takes on values from the set $\mathcal{X}$ with probability distribution $p(x)$. The (Shannon) entropy of X [28] is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

The relative entropy or Kullback-Leibler(KL) divergence [18] between two probability distributions $p_1(x)$ and $p_2(x)$ is defined as

$$KL(p_1, p_2) = \sum_{x \in \mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)}.$$

KL-divergence is a measure of the "distance" between two probability distributions; however it is not a true metric since it is not symmetric and does not obey the triangle inequality [5, p.18]. KL-divergence is always non-negative but can be unbounded; in particular when $p_1(x) \neq 0$ and $p_2(x) = 0$, $KL(p_1, p_2) = \infty$. In contrast, the Jensen-Shannon divergence between p_1 and p_2 defined by

$$\begin{aligned} JS_\pi(p_1, p_2) &= \pi_1 KL(p_1, \pi_1 p_1 + \pi_2 p_2) + \pi_2 KL(p_2, \pi_1 p_1 + \pi_2 p_2) \\ &= H(\pi_1 p_1 + \pi_2 p_2) - \pi_1 H(p_1) - \pi_2 H(p_2), \end{aligned}$$

where $\pi_1 + \pi_2 = 1$, $\pi_i \geq 0$, is clearly a symmetric measure and is bounded [20]. The Jensen-Shannon divergence can be generalized to measure the distance between any finite number of probability distributions as:

$$JS_\pi(\{p_i : 1 \leq i \leq n\}) = H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i), \tag{1}$$

which is symmetric in the p_i's ($\sum_i \pi_i = 1, \pi_i \geq 0$).

Let Y be another random variable with probability distribution $p(y)$. The mutual information between X and Y, $I(X;Y)$, is defined as the KL-divergence between the joint probability distribution $p(x, y)$ and the product distribution $p(x)p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p_(x)p(y)} \\ &= KL(p(x, y), p(x)p(y)). \end{aligned} \tag{2}$$

Intuitively, mutual information is a measure of the amount of information that one random variable contains about the other. The higher its value the less is the uncertainty of one random variable due to knowledge about the other. Formally, it can be shown that $I(X;Y)$ is the reduction in entropy of one variable knowing the other: $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ [5].

4 Naive Bayes Classifier

Let $\mathcal{C} = \{c_1, c_2, \ldots, c_l\}$ be the set of l classes, and let $\mathcal{W} = \{w_1, \ldots, w_m\}$ be the set of words/features contained in these classes. Given a new document d, the probability that d belongs to class c_i is given by Bayes rule,

$$p(c_i|d) = \frac{p(d|c_i)p(c_i)}{p(d)}.$$

Assuming a generative multinomial model [21] and further assuming class-conditional independence of words yields the Naive Bayes classifier, which computes the most probable class for d as

$$c^*(d) = \text{argmax}_{c_i} p(c_i|d) = p(c_i) \prod_{t=1}^{m} p(w_t|c_i)^{n(w_t,d)}, \tag{3}$$

where $n(w_t, d)$ is the number of occurrences of word w_t in document d, and the quantities $p(w_t|c_i)$ are usually maximum likelihood estimates with a Laplace prior:

$$p(w_t|c_i) = \frac{1 + \sum_{d_j \in c_i} n(w_t, d_j)}{m + \sum_{t=1}^{m} \sum_{d_j \in c_i} n(w_t, d_j)}. \tag{4}$$

The class priors $p(c_i)$ are estimated by the maximum likelihood estimate

$$p(c_i) = \frac{|c_i|}{\sum_j |c_j|}.$$

We now manipulate the Naive Bayes rule in order to interpret it in an information theoretic framework. Rewrite formula (3) by taking logarithms and dividing by the length of the document $|d|$ to get

$$c^*(d) = \text{argmax}_{c_i} \log p(c_i) + \sum_{t=1}^{m} p(w_t|d) \log p(w_t|c_i), \tag{5}$$

where the document d may be viewed as a probability distribution over words: $p(w_t|d) = n(w_t, d)/|d|$. Adding the entropy of $p(W|d)$, i.e., $-\sum_{t=1}^{m} p(w_t|d) \log p(w_t|d)$ to (5), and negating, we get

$$\begin{aligned} c^*(d) &= \text{argmin}_{c_i} \sum_{t=1}^{m} p(w_t|d) \log \frac{p(w_t|d)}{p(w_t|c_i)} - \log p(c_i) \\ &= \text{argmin}_{c_i} KL(p(W|d), p(W|c_i)) - \log p(c_i), \end{aligned} \tag{6}$$

where $KL(p, q)$ denotes the KL-divergence between p and q as defined in Section 3. Note that here we have used W to denote the random variable that ranges over the set of all words $\mathcal{W}$. Thus, assuming equal class priors, we see that Naive Bayes may be interpreted as finding the class which has minimum KL-divergence from the given document. As we shall see again later, KL-divergence seems to appear "naturally" in our setting.

By (5), we can clearly see that Naive Bayes is a linear classifier. Despite its crude assumption about the class-conditional independence of words, Naive Bayes has been found to yield surprisingly good classification performance, especially on text data. Plausible reasons for the success of Naive Bayes have been explored in [8, 12].

4.1 Hierarchical Naive Bayes

Hierarchical classification utilizes the hierarchical topic structure such as Yahoo! to decompose the classification task into a set of simpler problems, one at each node in the hierarchy. We can simply extend the Naive Bayes classifier to achieve hierarchical classification by constructing a classifier at each internal node of the tree with training data as the documents in its children. The tree is assumed to be "**is-a**" hierarchy, i.e., the training instances are inherited by the parents. Then classification is just a greedy descent down the tree until the leaf node is reached. This way of classification has been shown to be equivalent to the standard non-hierarchical classification over a flat set of leaf classes if maximum likelihood estimates of *all* features are used [23]. However, hierarchical classification along with feature selection has been shown to achieve better classification results than a flat classifier [17]. This is because each classifier can now utilize a different subset of features that are most relevant to the classification sub-task at hand. Furthermore the classifier now requires only a small number of features to classify since it needs to distinguish between a fewer number of classes. In this paper we propose a new divisive scheme for feature clustering to aggressively reduce the number of features associated with each node classifier in the hierarchy. We present detailed experiments with Dmoz Science hierarchy in Section 6.

5 Distributional Word Clustering

Let C be a discrete random variable that takes on values from the set of classes $\mathcal{C} = \{c_1, \ldots, c_l\}$, and let W be the random variable that ranges over the set of words $\mathcal{W} = \{w_1, \ldots, w_m\}$. The joint distribution $p(C, W)$ can be estimated from the training set. Now suppose we cluster words into the k clusters $\mathcal{W}_1, \ldots, \mathcal{W}_k$. Since our application is to reduce the number of features, we only look at "hard" clustering where each word belongs to exactly one word cluster, i.e,

$$\mathcal{W} = \cup_{i=1}^{k} \mathcal{W}_i, \quad \text{and} \quad \mathcal{W}_i \cap \mathcal{W}_j = \phi,\ i \neq j.$$

Let the random variable W^C range over the word clusters. In order to judge the quality of the word clusters we now introduce an information-theoretic measure.

The information about C captured by W can be measured by the mutual information $I(C;W)$. Ideally, we would like word clusters that *exactly* preserve the mutual information; however clustering always lowers the mutual information. Thus we would like to find a clustering that minimizes the decrease in the mutual information $I(C;W) - I(C;W^C)$. The following theorem states that this change in mutual information can be expressed in terms of the generalized Jensen-Shannon divergence of each word cluster.

Theorem 1 *The change in mutual information due to word clustering is given by*

$$I(C;W) - I(C;W^C) = \sum_{j=1}^{k} \pi(\mathcal{W}_j) JS_{\pi'}(\{p(C|w_t) : w_t \in \mathcal{W}_j\}) \tag{7}$$

where $\pi_t = p(w_t)$, $\pi(\mathcal{W}_j) = \sum_{w_t \in \mathcal{W}_j} \pi_t$, $\pi'_t = \pi_t / \pi(\mathcal{W}_j)$ *and JS denotes the generalized Jensen-Shannon divergence as defined in (1).*

Proof. By the definition of mutual information (see (2)), and using $p(c_i, w_t) = \pi_t p(c_i|w_t)$ we get

$$\begin{aligned} I(C;W) &= \sum_i \sum_t \pi_t p(c_i|w_t) \log \frac{p(c_i|w_t)}{p(c_i)} \\ \text{and} \quad I(C;W^C) &= \sum_i \sum_j \pi(\mathcal{W}_j) p(c_i|\mathcal{W}_j) \log \frac{p(c_i|\mathcal{W}_j)}{p(c_i)}. \end{aligned}$$

Since we are interested in hard clustering,

$$\begin{aligned} \pi(\mathcal{W}_j) &= \sum_{w_t \in \mathcal{W}_j} \pi_t \\ \text{and} \quad p(c_i|\mathcal{W}_j) &= \sum_{w_t \in \mathcal{W}_j} \frac{\pi_t}{\pi(\mathcal{W}_j)} p(c_i|w_t), \end{aligned}$$

thus implying that for all clusters $\mathcal{W}_j$,

$$\pi(\mathcal{W}_j)p(c_i|\mathcal{W}_j) = \sum_{w_t \in \mathcal{W}_j} \pi_t p(c_i|w_t), \tag{8}$$

$$p(C|\mathcal{W}_j) = \sum_{w_t \in \mathcal{W}_j} \frac{\pi_t}{\pi(\mathcal{W}_j)} p(C|w_t). \tag{9}$$

Note that the probability distribution $p(C|\mathcal{W}_j)$ is the (weighted) mean distribution of the constituent distributions $p(C|w_t)$. Thus,

$$I(C;W) - I(C;W^C) = \sum_i \sum_t \pi_t p(c_i|w_t) \log p(c_i|w_t) - \sum_i \sum_j \pi(\mathcal{W}_j) p(c_i|\mathcal{W}_j) \log p(c_i|\mathcal{W}_j), \tag{10}$$

with the extra $\log(p(c_i))$ terms cancelling due to (8). The first term in (10), after rearranging the summation, may be written as

$$\begin{aligned} & \sum_j \sum_{w_t \in \mathcal{W}_j} \pi_t \left(\sum_i p(c_i|w_t) \log p(c_i|w_t) \right) \\ = & -\sum_j \sum_{w_t \in \mathcal{W}_j} \pi_t H(p(C|w_t)) \\ = & -\sum_j \pi(\mathcal{W}_j) \sum_{w_t \in \mathcal{W}_j} \frac{\pi_t}{\pi(\mathcal{W}_j)} H(p(C|w_t)). \end{aligned} \tag{11}$$

Similarly, the second term in (10) may be written as

$$\begin{aligned} & \sum_j \pi(\mathcal{W}_j) \left(\sum_i p(c_i|\mathcal{W}_j) \log p(c_i|\mathcal{W}_j) \right) \\ = & -\sum_j \pi(\mathcal{W}_j) H(p(C|\mathcal{W}_j)) \\ = & -\sum_j \pi(\mathcal{W}_j) H\left(\sum_{w_t \in \mathcal{W}_j} \frac{\pi_t}{\pi(\mathcal{W}_j)} p(C|w_t) \right) \end{aligned} \tag{12}$$

where (12) is obtained by substituting the value of $p(C|\mathcal{W}_j)$ from (9). Substituting (11) and (12) in (10) and using the definition of Jensen-Shannon divergence from (1) gives us the desired result. □

The above theorem gives us a global measure of the goodness of wood clusters. The informal interpretation of Theorem 1 is as follows:

1. The quality of word cluster $\mathcal{W}_j$ is measured by the Jensen-Shannon divergence between the individual word distributions $p(C|w_t)$ (weighted by the word priors, $\pi_t = p(w_t)$). The smaller the Jensen-Shannon divergence the more "compact" is the word cluster, i.e., smaller is the increase in entropy due to clustering (see (1)).

2. The overall goodness of the word clustering is measured by the sum of the qualities of individual word clusters (weighted by the cluster priors $\pi(\mathcal{W}_j) = p(\mathcal{W}_j)$).

Given the global criterion of Theorem 1, we would now like to find an algorithm that searches for the optimal word clustering that minimizes this criterion. We now rewrite this criterion in a way that will suggest a "natural" algorithm.

Lemma 1 *The generalized Jensen-Shannon divergence of a finite set of probability distributions can be expressed as the (weighted) sum of Kullback-Leibler divergences to the (weighted) mean, i.e.,*

$$JS_\pi(\{p_i : 1 \leq i \leq n\}) = \sum_{i=1}^{n} \pi_i KL(p_i, m) \tag{13}$$

where $\pi_i \geq 0, \sum_i \pi_i = 1$ *and* m *is the (weighted) mean probability distribution,* $m = \sum_i \pi_i p_i$.

Proof. Use the definition of entropy to expand the expression for JS-divergence given in (1). The result follows by appropriately grouping terms and using the definition of KL-divergence. □

5.1 The Algorithm

By Theorem 1 and Lemma 1, the decrease in mutual information due to word clustering may be written as

$$\sum_{j=1}^{k} \pi(\mathcal{W}_j) \sum_{w_t \in \mathcal{W}_j} \frac{\pi_t}{\pi(\mathcal{W}_j)} KL(p(C|w_t), p(C|\mathcal{W}_j)).$$

As a result the quality of word clustering can be measured by the objective function

$$\begin{aligned} Q(\{\mathcal{W}_j\}_{j=1}^{k}) &= I(C;W) - I(C;W^C) \\ &= \sum_{j=1}^{k} \sum_{w_t \in \mathcal{W}_j} \pi_t KL(p(C|w_t), p(C|\mathcal{W}_j)). \end{aligned} \tag{14}$$

Note that it is natural that the KL-divergence emerges as the distance measure in the above objective function, since mutual information is just the KL-divergence between the joint distribution and the product distribution (see Section 3). Writing the objective function in the above manner suggests an iterative algorithm that repeatedly (i) re-partitions the distributions $p(C|w_t)$ by their closeness in KL-divergence to the cluster distributions $p(C|\mathcal{W}_j)$, and (ii) subsequently given the new word clusters, re-computes these cluster distributions using (9). Figure 1 describes the algorithm in detail. Note that this divisive algorithm bears some resemblance to the k-means or Lloyd-Max algorithm, which usually uses squared Euclidean distances [11, 9, 15].

Note that our initialization strategy is crucial to our algorithm, see step 1 in Figure 1 (also see [7, Section 5.1]). This strategy guarantees absolute continuity of each $p(C|w_t)$ with at least one cluster distribution $p(C|\mathcal{W}_j)$, i.e., guarantees that at least one KL-divergence is finite. This is because our initialization strategy ensures that every word w_t is part of some cluster $\mathcal{W}_j$. Thus by the formula for $p(C|\mathcal{W}_j)$ in step 2, it cannot happen that $p(c_i|w_t) \neq 0$, and $p(c_i|\mathcal{W}_j) = 0$. Note that we can still get some infinite KL-divergence values but these do not lead to any difficulty (indeed in an implementation we can handle such "infinity problems" without an extra "if" condition due to the handling of "infinity" in the IEEE floating point standard [14, 1]).

We now discuss the computational complexity of our algorithm. Step 3 of each iteration requires the KL-divergence to be computed for every pair, $p(C|w_t)$ & $p(C|\mathcal{W}_j)$. This is the most computationally demanding task and costs a total of $O(mkl)$ operations. Generally, we have found that the algorithm converges in $10-15$ iterations independent of the size of the data set. Thus the total complexity is $O(mkl)$, which grows linearly with m (note that $k \ll m$). In contrast, the agglomerative algorithm of [29] costs $O(m^3 l)$ operations.

The algorithm in Figure 1 has certain pleasing properties. As we will prove in Theorem 3, our algorithm decreases the objective function value at every step and thus is guaranteed to converge to a local minimum in a finite number of steps (note that finding the global minimum is NP-complete [13]). Also, by the equivalence of (7) and (14) we see that our algorithm minimizes the "within-cluster" Jensen-Shannon divergence. It turns out that (see Theorem 4) we can show that our algorithm *simultaneously maximizes* the "between-cluster" Jensen-Shannon divergence. Thus the different word clusters produced by our algorithm are "maximally" far apart.

We now give formal statements of our results with proofs.

Algorithm Divisive_Clustering($\mathcal{P}$,Π,l,k,$\mathcal{W}$)

Input: $\mathcal{P}$ is the set of distributions, $\{p(C|w_t) : 1 \leq t \leq m\}$,
Π is the set of all word priors, $\{\pi_t = p(w_t) : 1 \leq t \leq m\}$
l is the number of document classes,
k is the number of desired clusters.

Output: $\mathcal{W}$ is the set of word clusters $\{\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_k\}$.

1. Initialization: for every word w_t, assign w_t to $\mathcal{W}_j$ such that $p(c_j|w_t) = \max_i p(c_i|w_t)$. This gives l initial word clusters; if $k \geq l$ split each cluster into approximately k/l clusters, otherwise merge the l clusters to get k word clusters.

2. For each cluster $\mathcal{W}_j$, compute

$$\begin{aligned} \pi(\mathcal{W}_j) &= \sum_{w_t \in \mathcal{W}_j} \pi_t \\ p(C|\mathcal{W}_j) &= \sum_{w_t \in \mathcal{W}_j} \frac{\pi_t}{\pi(\mathcal{W}_j)} p(C|w_t). \end{aligned}$$

3. Re-compute all clusters: For each word w_t, find its new cluster index as

$$\text{argmin}_i \;\; KL(p(C|w_t), p(C|\mathcal{W}_i)),$$

resolving ties arbitrarily. Thus compute the new word clusters $\mathcal{W}_j$, $1 \leq j \leq k$, as

$$\begin{aligned} \mathcal{W}_j = \{w_t : KL(p(C|w_t), p(C|\mathcal{W}_j)) \leq \\ KL(p(C|w_t), p(C|\mathcal{W}_i)),\ 1 \leq i \leq k\}. \end{aligned}$$

4. Stop if the change in objective function value given by (14) is "small" (say 10^{-3}); Else go to step 2.

Figure 1: Divisive Algorithm for word clustering based on KL-divergences

Lemma 2 *Given probability distributions $p_1, \ldots, p_n$, the distribution that is closest (on average) in KL-divergence is the mean probability distribution m, i.e., given any probability distribution q,*

$$\sum_i \pi_i KL(p_i, q) \geq \sum_i \pi_i KL(p_i, m), \tag{15}$$

where $\pi_i \geq 0$, $\sum_i \pi_i = 1$ and $m = \sum_i \pi_i p_i$.

Proof. Use the definition of KL-divergence to expand the left-hand side(LHS) of (15) to get

$$\sum_i \pi_i \sum_x p_i(x) \left(\log p_i(x) - \log q(x)\right).$$

Similarly the RHS of (15) equals

$$\sum_i \pi_i \sum_x p_i(x) \left(\log p_i(x) - \log m(x)\right).$$

Subtracting the RHS from LHS leads to

$$\begin{aligned} \sum_i \pi_i \sum_x p_i(x) \left(\log m(x) - \log q(x)\right) = \\ \sum_x m(x) \log \frac{m(x)}{q(x)} = KL(m, q). \end{aligned}$$

The result follows since the KL-divergence is always non-negative [5, Theorem 2.6.3]. □

Theorem 2 *The Algorithm in Figure 1 monotonically decreases the value of the objective function given in (14).*

Proof. Let $\mathcal{W}_1^{(i)}, \ldots, \mathcal{W}_k^{(i)}$ be the word clusters at iteration i, and let $p(C|\mathcal{W}_1^{(i)}), \ldots, p(C|\mathcal{W}_k^{(i)})$ be the corresponding cluster distributions. Then

$$\begin{aligned} Q(\{\mathcal{W}_j^{(i)}\}_{j=1}^k) &= \sum_{j=1}^k \sum_{w_t \in \mathcal{W}_j^{(i)}} \pi_t KL(p(C|w_t), p(C|\mathcal{W}_j^{(i)})) \\ &\geq \sum_{j=1}^k \sum_{w_t \in \mathcal{W}_j^{(i+1)}} \pi_t KL(p(C|w_t), p(C|\mathcal{W}_j^{(i)})) \\ &\geq \sum_{j=1}^k \sum_{w_t \in \mathcal{W}_j^{(i+1)}} \pi_t KL(p(C|w_t), p(C|\mathcal{W}_j^{(i+1)})) \\ &= Q(\{\mathcal{W}_j^{(i+1)}\}_{j=1}^k) \end{aligned}$$

where the first inequality is due to step 3 of the algorithm, and the second inequality follows from step 2 and Lemma 2. Note that if equality holds, i.e., if the objective function value is equal at consecutive iterations, then step 4 terminates the algorithm. □

Theorem 3 *The Algorithm in Figure 1 always converges to a local minimum in a finite number of iterations.*

Proof. The result follows since the algorithm monotonically decreases the objective function value, which is bounded from below (by zero). For more details, see [27]. □

We now show that the total Jensen-Shannon divergence (which is constant for a given set of probability distributions) can be written as the sum of two terms, one of which is the objective function (14) that our algorithm minimizes.

Theorem 4 *Let $p_1, \ldots, p_n$ be a set of probability distributions and let $\pi_1, \ldots, \pi_n$ be corresponding scalars such that $\pi_i \geq 0$, $\sum_i \pi_i = 1$. Suppose $p_1, \ldots, p_n$ are clustered into k clusters $\mathcal{P}_1, \ldots, \mathcal{P}_k$, and let m_j be the (weighted) mean distribution of $\mathcal{P}_j$, i.e.,*

$$m_j = \sum_{w_t \in \mathcal{P}_j} \frac{\pi_t}{\pi(\mathcal{P}_j)} p_t, \quad \pi(\mathcal{P}_j) = \sum_{p_t \in \mathcal{P}_j} \pi_t. \tag{16}$$

Then the total JS-divergence between $p_1, \ldots, p_n$ can be expressed as the sum of "within-cluster JS-divergence" and "between-cluster JS-divergence", i.e.,

$$\begin{aligned} JS_\pi(\{p_i : 1 \leq i \leq n\}) &= \sum_{j=1}^k \pi(\mathcal{P}_j) JS_{\pi'}(p_t : p_t \in \mathcal{P}_j) \\ &\quad + JS_{\pi''}(\{m_i : 1 \leq i \leq k\}), \end{aligned}$$

where $\pi'_t = \pi_t / \pi(\mathcal{P}_j)$ and we use π'' as the subscript in the last term to denote $\pi''_j = \pi(\mathcal{P}_j)$.

Proof. By Lemma 1, the total JS-divergence may be written as

$$\begin{aligned} JS_\pi(\{p_i : 1 \leq i \leq n\}) &= \sum_{i=1}^n \pi_i KL(p_i, m) \\ &= \sum_{i=1}^n \sum_x \pi_i p_i(x) \log \frac{p_i(x)}{m(x)} \end{aligned} \tag{17}$$

where $m = \sum_i \pi_i p_i$. With m_j as in (16), and rewriting (17) in order of the clusters $\mathcal{P}_j$ we get

$$\sum_{j=1}^{k} \sum_{p_t \in \mathcal{P}_j} \sum_{x} \pi_t p_t(x) \left(\log \frac{p_t(x)}{m_j(x)} + \log \frac{m_j(x)}{m(x)} \right)$$

$$= \sum_{j=1}^{k} \pi(\mathcal{P}_j) \sum_{p_t \in \mathcal{P}_j} \frac{\pi_t}{\pi(\mathcal{P}_j)} KL(p_t, m_j) + \sum_{j=1}^{k} \pi(\mathcal{P}_j) KL(m_j, m)$$

$$= \sum_{j=1}^{k} \pi(\mathcal{P}_j) JS_{\pi'}(p_t : p_t \in \mathcal{P}_j) + JS_{\pi''}(\{m_i : 1 \leq i \leq k\}),$$

where $\pi''_j = \pi(\mathcal{P}_j)$, which proves the result. □

This concludes our formal treatment. We now see how to use the word clusters produced by divisive clustering in conjunction with the Naive Bayes classifier.

5.2 Naive Bayes with Word Clusters

The Naive Bayes method can be simply translated into using word clusters instead of words. This is done by estimating the new parameters $p(\mathcal{W}_s|c_i)$ for word clusters similar to the word parameters $p(w_t|c_i)$ in (4) as

$$p(\mathcal{W}_s|c_i) = \frac{\sum_{d_j \in c_i} n(\mathcal{W}_s, d_j)}{\sum_{s=1}^{k} \sum_{d_j \in c_i} n(\mathcal{W}_s, d_j)}$$

where $n(\mathcal{W}_s, d_j) = \sum_{w_t \in \mathcal{W}_s} n(w_t, d_j)$

Note that when estimates of quantities $p(w_t|c_i)$ are relatively poor, the corresponding word cluster parameters $p(\mathcal{W}_s|c_i)$ can provide more robust estimates resulting in higher classification scores.

Now the Naive Bayes rule (5) for classifying a test document d can be rewritten as

$$c^*(d) = \text{argmax}_{c_i} \log p(c_i) + \sum_{s=1}^{k} p(\mathcal{W}_s|d) \log p(\mathcal{W}_s|c_i),$$

where $p(\mathcal{W}_s|d) = n(\mathcal{W}_s|d)/|d|$.

6 Experimental Results

This section provides empirical evidence that our Divisive Clustering of Figure 1 outperforms other feature selection algorithms and agglomerative approaches. We compare our results with feature selection by Information Gain and Mutual Information [32], and feature clustering using the agglomerative algorithms in [2, 29]. We call the latter Agglomerative Clustering in this section for the purpose of comparison. We also show that Divisive Clustering achieves higher classification accuracy than the best performing feature selection method when the training data is *sparse* and show improvements over similar results reported in [29].

6.1 The Data Sets

The *20 Newsgroups* data set, collected by Ken Lang, contains about 20,000 articles evenly divided among 20 UseNet Discussion groups. This data set has been used for testing several text classification tasks[2, 29, 21]. Many of the news groups have similar topics (for example five groups talk about computers), and are quite confusable. In addition 4.5% of the documents are repeated, possibly due to cross posting across multiple news groups. During indexing we skipped headers, pruned words occurring in less than three documents, used a stop list, but did not use stemming. The resulting vocabulary had 35077 words.

We collected the *Dmoz* data from the Open Source Directory www.dmoz.org. The dmoz hierarchy contains about 3 million documents and 300,0000 classes. We chose the *Science* category at the top and crawled some