Clustering using the Information Bottleneck Method with Annealing
Lisa Miller
Machine Learning, ICS 663
Prof. S. Still
Spring 2010

1 Introduction

The purpose of this project is to examine the performance of an existing clustering method, Information Bottleneck with Annealing (IBANN), when applied to the task of document clustering. The information bottleneck method (IB) \[9\] is known as one of the best methods for clustering multi-dimensional data \[5\] and this variation of IB uses annealing in order to eliminate a preprocessing step normally required of unsupervised clustering algorithms, it is able to find the number of clusters present in the data \[8\]. This method has previously been shown to produce good results with lower dimensional data, but it has not been used on a very high-dimensional task such as document clustering. It is intended in this project to adjust the existing algorithm to deal with high-dimensional data in order to compare the results with other methods commonly used for unsupervised document clustering, such as k-means, and with some commonly used supervised methods of classification, such as Support Vector Machines.

In order to properly compare our results with previous unsupervised work, and to compare with supervised methods, it is necessary to run the algorithm on a commonly used, labeled dataset. Additionally, due to the high-dimensional nature of document clustering, preprocessing of the documents to reduce dimensionality is commonly done before evaluating performance. Therefore, prior to any work being done on the clustering algorithm itself, a dataset needed to be chosen and a preprocessing method developed in a way such that our results are comparable with previous work.

2 Clustering Method

The method of clustering that this project intends to implement is the one introduced by Still and Bialek in *How Many Clusters? An Information-Theoretic Perspective* \[8\]. The algorithm used in this method is based on a bootstrapping EM-type method which gradually forces the data into the maximal number of resolvable clusters as the temperature value is lowered. An outline of this algorithm is seen in Algorithm 1

3 Data Set Selection

In examining the literature on document classification, it was determined that there are four commonly used labeled data collections: 20 Newsgroups \[1\], Reuters-21578 \[4\], OHSUMED \[2\], and RCV1 \[3\]. Of the three, 20 newsgroups is probably the most commonly used but the individual
Algorithm 1 IBANN algorithm

{starting with all data in one cluster}

while Maximum possible clusters are not filled do
    split the clusters and perturb $p(y|c)$ values slightly
    lower temperature
    {enter EM}
    while Convergence criteria not met do
        {E-step}
        \[
p(c|x) = \frac{p(c) * e^{-\frac{1}{T}D_{KL}[p(y|x)||p(y|c)]}}{\sum_c p(c) * e^{-\frac{1}{T}D_{KL}[p(y|x)||p(y|c)]}}
        \]
        {M-step}
        \[
p(x|c) = \frac{p(c|x) * p(x)}{\sum_x p(c|x) * p(x)}
        \]
        \[
p(y|c) = \sum_x p(y|x) * p(x|c)
        \]
        Calculate $p_{\text{new}}(c)$
        Convergence criteria = $\sum_c D_{KL}[p_{\text{new}}(c)||p_{\text{prev}}(c)]$
    end while
    merge clusters
    end while

if Solution not deterministic or $p(c)$ not converged then
    {continue annealing}
    split the clusters and perturb $p(y|c)$ values slightly
    lower temperature
    {enter EM}
    while Convergence criteria not met do
        {E-step}
        \[
p(c|x) = \frac{p(c) * e^{-\frac{1}{T}D_{KL}[p(y|x)||p(y|c)]}}{\sum_c p(c) * e^{-\frac{1}{T}D_{KL}[p(y|x)||p(y|c)]}}
        \]
        {M-step}
        \[
p(x|c) = \frac{p(c|x) * p(x)}{\sum_x p(c|x) * p(x)}
        \]
        \[
p(y|c) = \sum_x p(y|x) * p(x|c)
        \]
        Calculate $p_{\text{new}}(c)$
        Convergence criteria = $\sum_c D_{KL}[p_{\text{new}}(c)||p_{\text{prev}}(c)]$
    end while
    merge clusters
    Check if solution deterministic
    Calculate $p_{\text{new}}(c) - p_{\text{old}}(c)$
end if
documents are very short, OHSUMED contains abstracts of medical journal articles which are in general the longest documents of the three but it is newer and appears less commonly in the literature at this time, RCV1 is a new Reuters collection that is just coming into popular usage. It was decided to start evaluation with the Reuters-21578 collection due to the longer article length and its long history of usage in document classification, particularly it was used in Joachims’ seminal paper on Support Vector Machines, Text categorization with Support Vector Machines: Learning with many relevant features [6].

The Reuters set is freely distributed and the documents are stored in a standardized, SGML marked-up form. There have been several different methods of determining which documents in the set to use for classification, all of which are easily recreateable using the SGML markup tags. It was chosen to go with the same version as that used by Joachims and many others, the Modified Apte Split. This gives a training set of 9,603 documents and a test set of 3,299 documents with 90 financial market-related categories each having at least one training and one test document.

4 Data Preprocessing

The IBANN method requires that the input data be in the form of normalized histograms representing $P(\text{word}|\text{doc})$. The word dimensionality quickly becomes very large with the 12,902 documents in the combined training and test sets yielding 32,012 unique words.

Several dimensionality reduction techniques were investigated including the stemming of words and removal of stop words used in Joachims work [6] but it was decided that the most appropriate route to take was the information theoretic one used by Slonim and Tishby [7]. They calculated each word’s contribution to the mutual information between words and documents

$$i(y) \equiv p(y) \sum_{x \in X} p(x|y) \log \frac{p(x|y)}{p(x)}$$

and took the top 2000 contributing words.

We decided to graph the output of the cumulative information contribution for all the words

$$I(S) \equiv \sum_{y \in S} i(y)$$

versus the word count $|S|$ to see if some obvious cutoff point would be present.

Figure 1 shows the graph of $I(S)$ vs. $|S|$ for the entire dataset. As can be seen no real obvious cutoff point appeared. We decided to try setting the initial cutoff (red line in the figure) to 90% of the information, or 7,941 words for the entire test/train dataset. This cuts the word dimensionality by about 75% but is still much, much higher than was previously attempted by Still and Bialek [8] or Slonim and Tishby [7].

5 Code Adjustments

In order to accommodate the high dimensionality, the code for the IBANN algorithm needed to be modified. The large matrices it uses $p(y|x)$ and $p(x,y)$ have been converted to hash tables. This
works excellently due to the sparseness of the data. Additionally, all other matrices in the code, \( p(y|c), p(c|x) \), and the one holding \( D_{KL} [p(y|x) \parallel p(y|c)] \) all had to be declared so that they were not being held on the memory stack. Just declaring them caused the program to blow up the stack and throw a segmentation fault. It may be necessary to eventually convert all of these matrices to hash tables in order to speed up execution.

6 Current State of Project and Future Work

At this point, the program is runnable and seems to be working correctly though slowly. Now we must undertake adjusting the parameters to get the best performance. The program has eight major variables to adjust: \( \beta \)-the inverse temperature, \( \alpha \)-the annealing rate, \( N \)-the maximum number of clusters, the convergence criteria threshold for the EM part, the size of the perturbation to the split cluster elements, the allowable tolerance for merging clusters, the deviation allowed when determining if the solution is deterministic, and the allowable error threshold for determining if \( p(c) \) has converged while annealing.

Future work on the project will be adjusting these parameters to get the best performance and possibly replacing more matrices with hash tables to speed up execution. Beyond that, post-processing of the output must be implemented so that the results can be compared with other methods. Also, the number of words used in the algorithm should be varied and the results compared to see if it is feasible to get equivalent results keeping less words and less of the mutual information between words and documents. Additionally, more research needs to be done into
which other methods would be most appropriate to compare our results with. Then, some of these other methods will need to be implemented (hopefully from available toolkits) for direct comparison on our dataset. Finally, it would also be interesting to also obtain results using some of the other datasets, especially 20 Newsgroups.

References


