Submitted (3/98) to the *15th International Conference on Machine Learning (ICML-98).*

Employing EM in Pool-Based Active Learning for Text Classification

Andrew McCallum[‡†]
mccallum@jprc.com
[‡]Just Research
4616 Henry Street
Pittsburgh, PA 15213

Kamal Nigam[†]
knigam@cs.cmu.edu
[†]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This paper shows how a text classifier's need for labeled training data can be reduced by a combination of active learning and Expectation Maximization (EM) on a pool of unlabeled data. Query-by-Committee is used to actively select documents for labeling, then EM with a naive Bayes model further improves classification accuracy by concurrently estimating probabilistic labels for the remaining unlabeled documents and using them to improve the model. We also present a metric for better measuring disagreement among committee members; it accounts for the strength of their disagreement and for the distribution of the documents. Experimental results show that our method of combining EM and active learning requires only half as many labeled training examples to achieve the same accuracy as either EM or active learning alone.

Keywords:
text classification
active learning
unsupervised learning
information retrieval

1 Introduction

In many settings for learning text classifiers, obtaining labeled training examples is expensive, while obtaining large quantities of unlabeled examples is very cheap. For example, consider the task of learning which web pages a user finds interesting; the user may not have the patience to hand-label a thousand training pages as interesting or not, yet multitudes of unlabeled pages are readily available on the Internet.

This paper presents the integration of *Active Learning* and *Expectation Maximization* for efficient learning of naive Bayes text classifiers. The resulting method works well with limited amounts of training data by taking advantage of a large pool of unlabeled data.

In previous work [Nigam *et al.*, 1998] we show that clustering of unlabeled and labeled documents via Expectation Maximization (EM) reduces text classification error by one-third over traditional supervised learning on several real-world data sets. In our setting for EM, a classifier is trained with whatever limited labeled examples are available, and this classifier is used to probabilistically fill in the "missing labels" on the unlabeled data. Then a new classifier is trained on the combination of the labeled and probabilistically "labeled" data, and the process iterates. Thus, EM uses the unlabeled data to better model the underlying distribution of the data—finding the classifier parameters that locally maximize the probability of both the labeled and unlabeled data. This method demonstrates one way of learning well with only a few labeled documents.

Active learning approaches this same problem in a different way. Unlike our setting for EM, the active learner can request the true class label for unlabeled examples it selects. However, each request is considered an expensive operation and the point is to perform well with as few queries as possible. A specific subsetting appropriate for text learning is *selective sampling*, where a learner must choose these examples from a stream or fixed pool, instead of creating unlabeled synthetic examples. In this paper, we consider *pool-based sampling*. Active learning aims to select the most informative examples for labeling from the pool. Informative examples are those that, if their class label were known, would reduce classification error and variance over the distribution of examples. Some methods measure this expected classification variance reduction directly [Cohn *et al.*, 1996]. Another method, *Query-by-Committee* (QBC), is easier to apply when closed-form calculation of variance would be prohibitively complex (although traditional approaches do not explicitly model the distribution of examples) [Freund *et al.*, 1997]. QBC measures the variance indirectly, by examining the disagreement among class labels assigned by a set of classifier variants—the variants being sampled from the probability distribution of classifiers resulting from the labeled training examples.

This paper presents results showing that the *combination* of EM and QBC learns with fewer labeled examples than either individually. In experimental results on a real-world text data set, EM applied to the the labeled examples that have been chosen by active learning and the remaining unlabeled examples requires only half as many labeled examples to achieve the same accuracy as either active learning or EM alone.

We advocate a paradigm we call *pool-leveraged sampling* that uses estimated statistics from the pool of data to improve active learning document selection. Within this context, we describe a selection metric for QBC that explicitly models the distribution of data. Experimental results show that the application of this paradigm improves performance.

We also discuss a stronger method of pool-leveraged sampling that interleaves QBC and EM such that the "unsupervised structure" of the pool of unlabeled examples actually informs the selection of active learning queries. By having the QBC committee members each perform EM, we hope to (1) avoid selecting examples whose labels can be reliably filled in by EM, and (2) encourage the selection of examples that will help EM find a local maximum with higher classification accuracy.

2 Probabilistic Framework and Naive Bayes

This section presents a probabilistic framework for text classification without EM or active learning. The next two sections add EM and active learning by building on this framework. We approach the task of text classification from a Bayesian learning perspective. With four commonly used assumptions [Domingos and Pazzani, 1997; Joachims, 1997] about the nature of the generative parametric model, we use training data to calculate Bayes optimal estimates of the model parameters. Then, armed with these estimates, we can classify new test documents by using Bayes rule to turn the generative model around and calculate the probability that a class would have generated the test document in question.

We first assume that text documents are generated by a mixture model, parameterized by θ. Secondly, we assume that the mixture model contains $j = \{1, \ldots, |\mathcal{C}|\}$ mixture components, where each component corresponds to a single class of documents. This dictates that a document, d_i, is created by (1) selecting a class according to the class priors, $\mathrm{P}(c_j|\theta)$, then (2) having the corresponding mixture component generate a document according to its own parameters, with distribution $\mathrm{P}(d_i|c_j;\theta)$. The probability of generating document d_i independent of its class is thus a sum of total probability over all mixture components:

$$\mathrm{P}(d_i|\theta) = \sum_{j=1}^{|\mathcal{C}|} \mathrm{P}(c_j|\theta)\mathrm{P}(d_i|c_j;\theta). \tag{1}$$

A document is comprised of an ordered sequence of word events, drawn from a vocabulary V. We assume that the lengths of documents are evenly distributed, and are independent of class. Our final assumption is the naive Bayes assumption: that the probability of each word event in a document is independent of the word's context and position in the document. Thus, each document d_i is drawn from a multinomial distribution of words with as many independent trials as the number of words in d_i. We write $w_{d_{ik}}$ for the word in position k of document d_i, where the subscript of w indicates an index into the vocabulary. Then, the probability of a document given its class is:

$$\mathrm{P}(d_i|c_j;\theta) = \mathrm{P}(|d_i|) \prod_{k=1}^{|d_i|} \mathrm{P}(w_{d_{ik}}|c_j;\theta). \tag{2}$$

The mixture model parameterization θ is composed of disjoint sets of parameters, θ_j, for each class c_j. These parameters define the multinomial distribution for each class. For each class, θ_j is composed of probabilities for each word, such that $\theta_{jt} = \mathrm{P}(w_t|c_j;\theta)$, where $\theta_{jt} > 0$ and $\sum_t \theta_{jt} = 1$. The only other parameters in the model are the class prior probabilities, written $\theta_{0j} = \mathrm{P}(c_j|\theta)$.

Given a set of labeled training documents, $\mathcal{D}$, we can calculate Bayes optimal estimates for the parameters of the model that generated the documents. These are maximum likelihood estimates, straightforward counting of events, supplemented by simple 'smoothing' that primes each word's count with a count of one to avoid probabilities of zero [Vapnik, 1982]. We define $N(w_t, d_i)$ to be the count of the number of times word w_t occurs in document d_i, and define $\mathrm{P}(c_j|d_i) = \{0, 1\}$ as given by the document's class label. Then, the estimate of the probability of word w_t in class c_j is

$$\hat{\theta}_{jt} = \mathrm{P}(w_t|c_j;\hat{\theta}) = \frac{1 + \sum_{i=1}^{|\mathcal{D}|} N(w_t, d_i)\mathrm{P}(c_j|d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|\mathcal{D}|} N(w_s, d_i)\mathrm{P}(c_j|d_i)}. \tag{3}$$

The class prior parameters are set by the maximum likelihood estimate

$$\hat{\theta}_{0j} = \mathrm{P}(c_j|\hat{\theta}) = \frac{\sum_{i=1}^{|\mathcal{D}|} \mathrm{P}(c_j|d_i)}{|\mathcal{D}|}. \tag{4}$$

Given estimates of these parameters calculated from the training documents, it is possible to calculate probabilistic labels for a new, unlabeled test document—the probability that each class component generated it. We formulate this by first applying Bayes rule, and then substituting for $\mathrm{P}(d_i|c_j;\hat{\theta})$ and $\mathrm{P}(d_i|\hat{\theta})$ using Equations 1 and 2.

$$\begin{aligned} \mathrm{P}(c_j|d_i;\hat{\theta}) &= \frac{\mathrm{P}(c_j|\hat{\theta})\mathrm{P}(d_i|c_j;\hat{\theta})}{\mathrm{P}(d_i|\hat{\theta})} \\ &= \frac{\mathrm{P}(c_j|\hat{\theta})\prod_{k=1}^{|d_i|}\mathrm{P}(w_{d_{ik}}|c_j;\hat{\theta})}{\sum_{r=1}^{|C|}\mathrm{P}(c_r|\hat{\theta})\prod_{k=1}^{|d_i|}\mathrm{P}(w_{d_{ik}}|c_r;\hat{\theta})} \end{aligned} \tag{5}$$

To classify a document, simply select the class with the largest $\mathrm{P}(c_j|d_i;\hat{\theta})$ as the predicted class for a document. Despite the fact that the mixture model and word independence assumptions are strongly violated with real-world data, naive Bayes performs classification very well. Domingos and Pazzani discuss why the violation of the word independence assumption does little damage to classification accuracy [Domingos and Pazzani, 1997].

3 Using EM to Incorporate Unlabeled Data

When naive Bayes is given just a small set of labeled training data, classification accuracy will suffer because variance in the parameter estimates of the generative model will be high. However, by augmenting this small set with a large set of unlabeled data and combining the two pools with EM, we can improve our parameter estimates. EM concurrently generates probabilistic labels for the unlabeled documents, and a more probable model with smaller parameter variance that predicts the same probabilistic labels. This section describes how to use EM to combine these pools for better parameter estimation within the probabilistic framework of the previous section.

EM is a class of iterative algorithms for maximum likelihood estimation in problems with incomplete data [Dempster *et al.*, 1977]. Given a model of data generation, and data with some missing values, EM will converge to a set of generative parameters that locally maximizes the likelihood of both the labeled and unlabeled data. By treating the class labels of the unlabeled data as missing values, and running EM on the entire data set, the resulting parameter estimates give higher classification accuracy for new documents when the pool of labeled examples is small [Nigam *et al.*, 1998]. This is a special case of a more general missing values formulation [Ghahramani and Jordan, 1994].

In practice, EM is an iterative two-step process. The E-step calculates probabilistic class labels, $\mathrm{P}(c_j|d_i)$, for every unlabeled document using a current estimate of θ and Equation 5. The M-step calculates a new maximum likelihood estimate for θ using all the labeled data, both original and probabilistically labeled, by Equations 3 and 4. We initialize the process with parameter estimates using just the labeled training data, and iterate until $\hat{\theta}$ reaches a fixed point. The resulting $\hat{\theta}$ has smaller variance in its parameter estimates because it used more documents in forming the estimates. See [Nigam *et al.*, 1998] for more details.

4 Active Learning with EM

Instead of estimating class labels for unlabeled documents as EM does, active learning requests the true class labels for selected unlabeled documents. Optimally, a learner selects labels for documents that minimize classification error over the document distribution. With statistical learners, this is equivalent to minimizing classification variance. This section first presents the mechanics of applying Query-by-Committee to active learning, then explains our method for selecting documents to be labeled, and finally describes the integration of active learning with EM.

Query-by-Committee measures expected reductions in classification variance indirectly by creating a set of classifier variants (committee members), and measuring the disagreement among committee members for each potential query. Following theoretical and empirical work on QBC [Freund *et al.*, 1997; Dagan and Engelson, 1995], our committee members are created by sampling classifiers according to the distribution of classifier parameters specified by the training data. Parameters θ_{jt} are sampled from Normal distributions with mean given by Equation 3, and variance $\mathrm{P}(w_t|c_j)(1-\mathrm{P}(w_t|c_j))/n_j$, where n_j is the total number of word occurrences in the training data for class c_j; that is, $n_j = \sum_{d_i}\sum_{w\in V} N(w,d_i)\mathrm{P}(c_j|d_i)$. We sample k times, to create k committee members. Individual committee members are denoted by m.

Next, some number of documents, L, are selected for class label requests.[1] Dagan and Engelson do this by sampling from the unlabeled documents to produce a stream; then for each document in the stream, they measure the classification disagreement among committee members using *vote entropy*, and heuristically convert this to a probability of selecting the document for labeling. Vote entropy is the entropy of the class label distribution resulting from having each committee member deterministically "vote" for its winning class. The heuristic must be tuned.

After obtaining labels for some documents, the active learning process repeats, creating new committee members as above after incorporating the newly labeled documents and requesting labels for batches of L documents. After active learning is completed, a final, single classifier is created without variance perturbations, and is used for testing new documents. Our approach to QBC differs from Dagan and Engelson's in three ways:

(1) We select documents from the entire pool of unlabeled documents. Rather than tuning a heuristic for picking documents from a stream, we choose the documents from the entire pool with the highest disagreement.

(2) Rather than using vote entropy, we measure committee disagreement for each document using *Kullback-Leibler divergence to the mean* [Pereira *et al.*, 1993]. All the unlabeled documents are probabilistically classified by each committee member, resulting in class distributions, $\mathrm{P}_m(C|d_i)$, from each committee member m for each document d_i. Unlike vote entropy, which compares only the committee members' top ranked class, KL divergence can measure the strength of the certainty of disagreement of committee members by calculating differences in the committee members' class distributions, $\mathrm{P}_m(C|d_i)$. KL divergence to the mean is an average of the KL divergence of each distribution to the mean of all the distributions:

$$\frac{1}{k}\sum_{m=1}^{k} D\left(\mathrm{P}_m(C|d_i)||\mathrm{P}_{avg}(C|d_i)\right). \tag{6}$$

[1]Selecting more than one document at a time is only a computational convenience.

where $\mathrm{P}_{avg}(C|d_i)$ is the class distribution mean over all committee members: $\mathrm{P}_{avg}(C|d_i) = (\sum_m \mathrm{P}_m(C|d_i))/k$.

KL divergence, $D(\cdot||\cdot)$, is an information theoretic measure of the inefficiency resulting from sending messages sampled from the first distribution using a code that is optimal for the second. The KL divergence between class distributions $\mathrm{P}_1(C)$ and $\mathrm{P}_2(C)$ is:

$$D(\mathrm{P}_1(C)||\mathrm{P}_2(C)) = -\sum_{j=1}^{|\mathcal{C}|} \mathrm{P}_1(c_j) \log\left(\frac{\mathrm{P}_2(c_j)}{\mathrm{P}_1(c_j)}\right). \tag{7}$$

(3) In addition to preferring documents with differing committee classifications, we incorporate into our disagreement metric a preference for documents that will reduce the classification variance of many other documents (as prescribed by theory [Cohn *et al.*, 1996]). The stream approach approximates this implicitly in that the stream is produced by sampling the underlying distribution. However, we accomplish this more accurately, especially when labeling only a small number of documents, by modeling the document density explicitly. We approximate the density estimation by measuring a document's distance to its class centroid, using a "similarity-based" metric proposed for word co-occurrence probabilities [Dagan *et al.*, 1994]. This is an appropriate approximation because given the assumption that data is generated according to a mixture model, density is highest near the centroids of the mixture components. Distance, Z, between document d_i and class c_j is defined as

$$Z(d_i, c_j) = e^{-D(\mathrm{P}(W|d_i)||\mathrm{P}(W|c_j))}, \tag{8}$$

where W is a random variable over all words in the vocabulary, $\mathrm{P}(W|d_i)$ is the maximum likelihood estimate of words sampled uniformly from document d_i; $(\mathrm{P}(w_t|d_i) = N(w_t, d_i)/|d_i|)$, and $\mathrm{P}(W|c_j)$ for individual words is given in Equation 3. In future work we will explore density estimation that is independent of class.

Thus, the overall importance of selecting an unlabeled document for labeling is the committee classification disagreement, weighted by its distance to the centroid. That is:

$$-\frac{1}{k}\sum_{m=1}^{k}\sum_{j=1}^{|\mathcal{C}|} Z(d_i, c_j)\mathrm{P}_m(c_j) \log\left(\frac{\mathrm{P}_{avg}(c_j)}{\mathrm{P}_m(c_j)}\right). \tag{9}$$

- Loop while adding documents:
  - Build an initial estimate of $\hat{\theta}$ from the labeled documents only. (Equations 3 and 4)
  - Loop k times, once for each committee member:
    - Create committee member, m, by sampling from the $\hat{\theta}_j$ distributions with mean and variance indicated by training data. (Page 7)
    - *Starting with the sampled classifier apply EM with the unlabeled data. Loop while classifier parameters change:*
      - *Use the current classifier to probabilistically label the unlabeled documents (Equation 5).*
      - *Recalculate the classifier parameters $\hat{\theta}_{0j}$ and $\hat{\theta}_j$ given the probabilistically assigned labels (Equations 3 and 4).*
    - Use the current classifier to probabilistically label all unlabeled documents (Equation 5)
  - Calculate the disagreement of all unlabeled documents (Equation 9), and request class labels for the L documents that are most important.
- Build the classifier (Equations 3 and 4).
- *Starting with this classifier, apply EM as above.*

Table 1: Our active learning algorithm. Traditional Query-by-Committee omits the EM steps, indicated by italics, and uses a different scoring metric.

This *density-weighted KL* metric will tend to select a document that each committee member thinks is strongly prototypical for a different class. In our experience, vote entropy in pool-based sampling for document classification tends to select outliers—documents that have high committee disagreement simply because there are short or unusual.

Combining Active Learning and EM

Active learning can be straightforwardly combined with EM by running EM to convergence after actively selecting all the training data that will be labeled. This can be understood as using active learning to select a better starting point for EM hill climbing, instead of randomly selecting documents to label for the starting point.

A more interesting approach to *pool-leveraged sampling* is to interleave EM with active learning, so that EM not only builds on the results of active learning, but active learning is also informed by EM. To do this we run EM to convergence on each committee member before performing the disagreement calculations. The intended effect is (1) to avoid requesting labels for examples whose label can be reliably filled in by EM, and (2) to encourage the selection of examples that will help EM find a local maximum with higher classification accuracy. With more accurate committee members, QBC should pick more informative documents to label. The complete active learning algorithm, both with and without EM, is summarized in Table 1.

Unlike previous work in which queries must be generated [Cohn, 1994], and previous work in which the unlabeled data is available only as a stream [Dagan and Engelson, 1995; Freund *et al.*, 1997], our assumption about the availability of a pool of unlabeled data makes the leverage possible. This pool is often present for many real-world tasks in which efficient use of labels is important, especially in text learning.

5 Related Work

Our approach to QBC active learning without EM follows that of Dagan and Engelson [Dagan and Engelson, 1995]. They use stream-based sampling and vote entropy; instead, we use pool-based sampling and density-weighted KL. Several other studies have investigated active learning for text categorization. Lewis and Gale examine uncertainty sampling and relevance sampling [Lewis and Gale, 1994; Lewis, 1995]. These techniques select queries based on only a single classifier instead of a committee, and thus cannot approximate classification variance reduction. Liere and Tadepalli use committees of Winnow learners for active text learning [Liere and Tadepalli, 1997]. They select documents for which two randomly selected committee members disagree on the class label. Both text studies learn binary classifiers for classes with low frequency, as in our Reuters results, and use only titles instead of full documents.

In previous work, we show that EM with unlabeled data reduces text classification error by one-third [Nigam *et al.*, 1998]. Two other studies have used EM to combine labeled and unlabeled data without active learning for classification, but on non-text tasks [Miller and Uyar, 1997; Shahshahani and Landgrebe, 1994]. Ghahramani and Jordan use EM with mixture models to fill in missing values [Ghahramani and Jordan, 1994].

6 Experimental Results

This section provides evidence that using a combination of active learning and EM does better than using either individually. The results are based on data sets from UseNet and newswires.[2]

The Newsgroups data set, collected by Ken Lang, contains about 20,000 articles evenly divided among 20 UseNet discussion groups [Joachims, 1997]. We use five confusable `comp.*` classes as our data set. When tokenizing this data, we skip the UseNet headers (thereby discarding the subject line); tokens are formed from contiguous alphabetic characters. Best performance was obtained with no feature selection, no stemming, and by normalizing word counts by document length. The resulting vocabulary, after removing words that occur only once, has 22958 words. On each trial, 20% of the documents are randomly selected for placement in the test set.

The 'ModApte' train/test split of the Reuters 21578 Distribution 1.0 data set consists of 12902 Reuters newswire articles in 135 overlapping topic categories. Following several other studies [Joachims, 1998; Liere and Tadepalli, 1997] we build binary classifiers for each of the 10 most populous classes. We ignore words on a stoplist, but do not use stemming. The resulting vocabulary has 19371 words. Results are reported on the complete test set as precision-recall breakeven points, a standard information retrieval measure for binary classification.

For both data sets, an initial classifier was trained with one random document per class. Active learning proceeds as described in Table 1. Each active learning iteration selects five documents (L=5) for labeling from the original pool of 9601 unlabeled training examples for Reuters, and ten documents (L=10) out of the pool of about 4000 for Newsgroups. Experiments were run for 100 active learning iterations. For QBC we use a committee size of three (k=3); initial experiments show that committee size has little effect. All EM runs perform seven EM iterations; we never found classification accuracy to improve beyond the seventh iteration. All the results in this paper are averages of ten runs per condition.

The left-hand graph in Figure 1 shows a comparison of different selection metrics for QBC without EM on the Newsgroups data set. The best performer is our density-weighted KL metric, requiring, for example, 215 labelings to

[2]These data sets are available on the Internet. See http://www.cs.cmu.edu/~textlearning and http://www.research.att.com/~lewis.

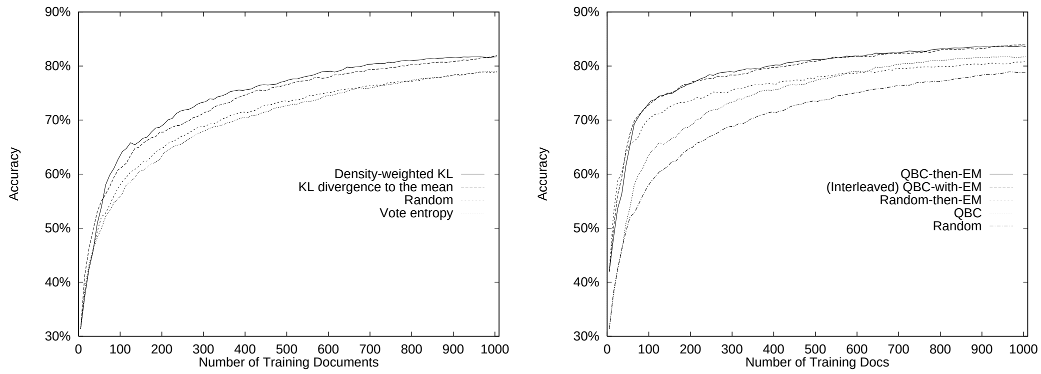

Figure 1: Active learning on the Newsgroups data. On the left, a comparison of disagreement metrics for QBC shows that by considering density, density-weighted KL does better than other metrics. On the right, combinations of QBC and EM outperform stand-alone QBC or EM. All have faster learning rates than random example selection. In these cases, QBC uses the density-weighted KL selection metric. Note that for resolution, the vertical axes do not range from 0 to 100.

achieve 75% accuracy. The next best is KL divergence to the mean, without weighting, requiring 275 labelings for 75% accuracy. The baseline of comparison is random selection, which is equivalent to traditional supervised learning, and it needs 345 labelings. Vote entropy performs slightly worse than the baseline, needing 385 labelings. As expected, density-weighted KL captures some density information and is thus statistically significantly the best performer ($p < 0.05$). Vote entropy performs worst because it does not attempt to capture either density information, or the strengths of disagreement. Note that previous uses of vote entropy have done so in a stream-based sampling setting, which implicitly captures some density information.

Now we consider the addition of EM to the learning scheme. Three variants are applied to the Newsgroups data set. Our EM baseline post-processes random selection with runs of EM (Random-then-EM). The most straightforward method of combining EM and active learning is to run EM after active learning completes (QBC-then-EM). We also interleave EM and active learning, by running EM on each committee member (QBC-with-EM). This also includes a post-processing run of EM. In QBC, disagreement is measured according to density-weighted KL, as the previous experiment indicated was appropriate. Random selection (Random) and QBC without EM (QBC) are repeated from the previous experiment.

The right-hand graph of Figure 1 includes results combining EM and active learning on the Newsgroups data set. As expected, Random selection and straight QBC give the slowest learning rates, 595 and 365 labelings to reach 75% accuracy respectively. Random-then-EM improves upon both, especially before random performance starts to plateau; it needs 255 labelings to reach 75%. These results are consistent with earlier results in this domain [Nigam *et al.*, 1998]. QBC-then-EM is impressive, needing only 165 labelings. Interleaved QBC-with-EM marginally improves over QBC-then-EM at this accuracy, needing 155 documents—less than 30% of the training data as random, less than 45% of the labeled examples as QBC alone, and less than 65% of the labeled examples as EM alone. QBC-with-EM and QBC-then-EM are statistically significantly better than the other methods for accuracies above 66% ($p < 0.05$). At 77% accuracy QBC-with-EM requires less than 50% of the labeled examples as either Random-then-EM or QBC.

These results indicate that the combination of EM and active learning provides a large benefit. However, QBC interleaved with EM does not perform better than QBC followed by EM—not what we were expecting. We hypothesize that this discrepancy is caused by the extreme 1/0 classification probabilities assigned by naive Bayes. Due to its independence assumption, naive Bayes typically assigns probability very near one to the winning class, and probabilities near zero to the other classes. While having only small amounts of training data tends to curb this effect, EM brings the effect back because it incorporates a lot of unlabeled training data. Thus, outlier documents that should not be selected are given higher disagreement weights in QBC-with-EM than with QBC-then-EM. In future work, we will try improved score-to-probability mappings that we believe will cause QBC-with-EM to select documents that fall close to different centroids in different runs of EM—exactly those documents that, if labeled, should help EM form correct clusters.

In comparison to previous active learning studies in text classification domains [Lewis and Gale, 1994; Liere and Tadepalli, 1997], the magnitude of our classification accuracy increase is relatively modest. Both of these previous studies consider binary classifiers in which the positive class has very small priors. With a very infrequent positive class, random selection should perform extremely poorly; nearly all documents selected for labeling will be from the negative class. In tasks where the class priors are more even, random selection will perform much better, making the improvement of active learning less dramatic. With an eye towards testing this hypothesis, we perform a subset of our previous experiments on the Reuters data set, which has these characteristics

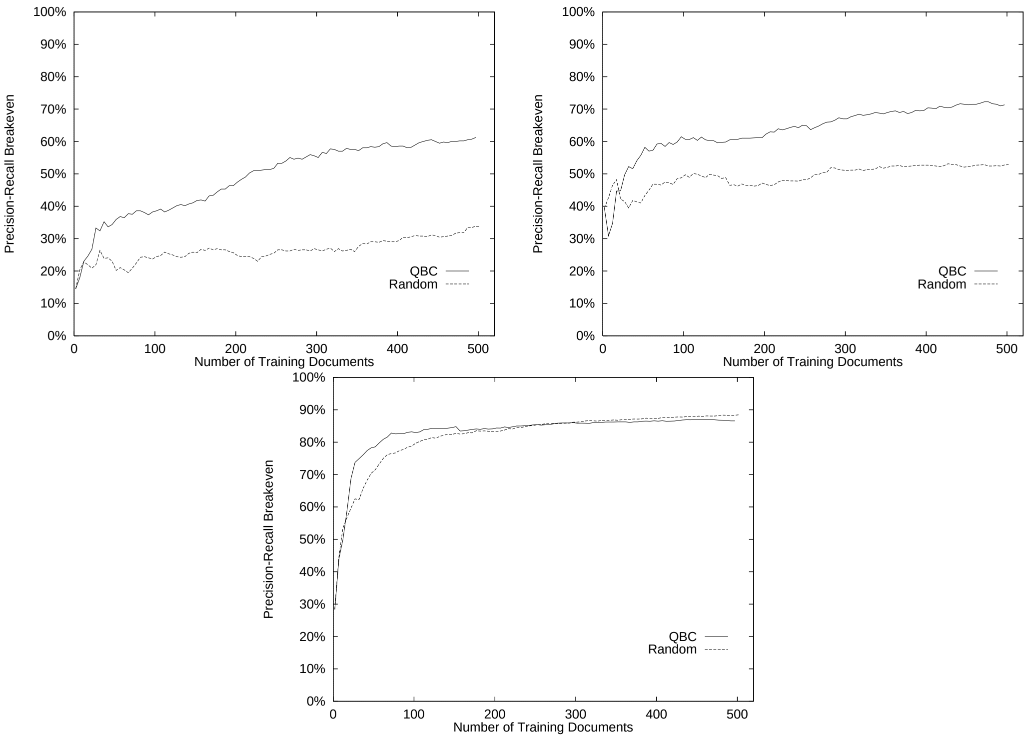

Figure 2: Active learning results on three categories of the Reuters data, corn, trade, and acq, respectively clockwise from the top left and in increasing order of frequency. Note that active learning with committees outperforms random selection and that the magnitude of improvement is larger for more infrequent classes.

of skewed priors. We compare Random against QBC with a density-weighted KL disagreement metric.

Figure 2 presents results for three of the ten binary classification tasks. The frequencies of the positive classes are 0.018, 0.038 and 0.184 respectively for the left (the corn category), right (trade) and bottom (acq) graphs. The results are representative of the spectrum of active learning results, and of the class frequencies. In all cases, active learning results in more accurate classifiers than Random. After 202 labelings, improvements of accuracy over random are from 25% to 47% for corn, 47% to 62% for trade, and 83% to 84% for acq. The distinct trend across all ten categories is that less frequently occurring positive classes have bigger improvements with active learning. Thus, we conclude that our earlier results are reasonable given that with even class priors, random provides a relatively strong baseline of performance.

7 Conclusions

This paper shows that the combination of active learning and EM for learning in the presence of unlabeled data provides a powerful improvement over either method alone. The paradigm of *pool-leveraged* selective sampling allows a collection of unlabeled examples to be used explicitly, without incurring the cost of requesting their labels. EM uses this pool of unlabeled examples to create more accurate classifiers. A new metric approximates the density of the pool, selecting documents that will better reduce future classification variance. Our results show that the combination of EM and active learning in pool-based sampling results in a factor of two reduction in labelings.

In ongoing work, we will explore methods for converting extreme naive Bayes "probabilities" to values that better match empirical probabilities of being correct. We plan a comparison of vote entropy and KL divergence selection metrics for stream-based sampling. We are also considering different density approximators that are not class-dependent or parametric.

Acknowledgments

We thank Doug Baker for help formatting the Reuters data set. This research was supported in part by the Darpa HPKB program under contract F30602-97-1-0215.

References

[Cohn *et al.*, 1996] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[Cohn, 1994] David Cohn. Neural network exploration using optimal experiment design. In *NIPS 6*, 1994.

[Dagan and Engelson, 1995] Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classifiers. In *ICML-95*, 1995.

[Dagan *et al.*, 1994] Ido Dagan, Fernando Pereira, and Lillian Lee. Similarity-based estimation of word cooccurrence probabilities. In *32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM. algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[Domingos and Pazzani, 1997] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning*, 29:103–130, 1997.

[Freund *et al.*, 1997] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

[Ghahramani and Jordan, 1994] Zoubin Ghahramani and Michael Jordan. Supervised learning from incomplete data via an EM approach. In *NIPS 6*, 1994.

[Joachims, 1997] Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *ICML-97*, 1997.

[Joachims, 1998] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *ECML-98*, 1998.

[Lewis and Gale, 1994] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of ACM SIGIR Conference*, 1994.

[Lewis, 1995] David D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. *SIGIR Forum*, 29(2):13–19, 1995.

[Liere and Tadepalli, 1997] Ray Liere and Prasad Tadepalli. Active learning with committees for text categorization. In *AAAI-97*, 1997.

[Miller and Uyar, 1997] David J. Miller and Hasan S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in Neural Information Processing Systems (NIPS 9)*, 1997.

[Nigam *et al.*, 1998] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Learning to classify text from labeled and unlabeled documents. In *Submitted to AAAI-98*, 1998. http://www.cs.cmu.edu/~mccallum.

[Pereira *et al.*, 1993] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993.

[Shahshahani and Landgrebe, 1994] B. Shahshahani and D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Trans. on Geoscience and Remote Sensing*, 32(5):1087–1095, Sept 1994.

[Vapnik, 1982] V. Vapnik. *Estimations of dependences based on statistical data.* Springer Publisher, 1982.