An Experimental Study of Semi-Supervised EM algorithms in Audio Classification and Speaker Identification

Pedro J. Moreno, Shivani Agarwal[1]
Cambridge Research Laboratory
HP Laboratories Cambridge, MA
HPL-2003-258
December 8th , 2003*

semi-supervised, unlabeled, EM, classification

Most pattern recognition techniques assume the existence of large quantities of carefully labeled data for training classifiers. However, the generation of this labeled data is an expensive and time-consuming process. In applications like multimedia processing, vast amounts of data are generated daily, and labeling this data to refine classifiers becomes impossible. In the last years, a new body of techniques has emerged that explore how to take advantage of vast quantities of unlabeled data, i.e. data with no class assignment information. In this paper we study the applicability of these techniques to various audio classification tasks. We show very promising results that demonstrate a reduction in half of audio classification and speaker identification error rates.

* Internal Accession Date Only Approved for External Publication
[1] Ph.D student at the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801

1 Introduction

Traditional pattern recognition techniques assume the existence of a large set of training data where there is a known assignment of labels to each data point. Using this data, a classifier is trained to learn a mapping from data instances to corresponding labels, and the learned classifier is then applied to new instances; typically, the classifier is evaluated by measuring error rates on an independent test set.

This traditional picture has begun to change in the last years as researchers and practitioners have been confronted with the problem of generating good quality labeled training sets. Labeling data can be both costly and time-consuming, especially as human intervention is often required to carefully annotate data. For example, in the field of speech research, the Linguistic Data Consortium generates every year hundreds of hours of carefully transcribed audio databases. Similar efforts exist in other fields such as face recogntion and optical character recognition.

On the other hand, obtaining large amounts of data *without* labels is quite easy. Vast amounts of data are being generated in several domains such as multimedia (audio, video, images), genomics, data mining etc. Audio data can readily be collected from broadcasts, face images can be obtained from online cameras, and so on. Clearly, there is a need for learning techniques that can take advantage of such unlabeled data. In this paper, we study semi-supervised techniques for augmenting small sets of labeled data with large amounts of unlabeled data, and explore the applicability of these techniques to audio classification problems.

2 Previous Work

Learning techniques to explore the use of unlabeled data have a long history in some fields such as data mining and astronomical data analysis. Much of the research into such techniques has focused on the problem of clustering. In the area of classification, much of the previous work on semi-supervised learning has focused on text classification problems using variations of the EM algorithm and Naive Bayesian classifiers. Blum and Mitchell (Blum & Mitchell, 1998) introduced the co-training algorithm and applied it to the problem of web-page categorization. Nigam et al (Nigam et al., 2000) further explore the use of several EM algorithm variants again on the same dataset. Ghani (Ghani, 2001) explores several variants of the EM algorithm combined with Naive Bayes classifiers on similar problems.

Recently, techniques based on ensemble methods such as boosting have been modified to take advantage of unlabeled data too. Bennett et al (Bennett et al., 2002) have applied boosting based methods to several semi-supervised classification problems with promising results.

Our work is similar to the work of Ghani (Ghani, 2001), with two main differences. First, we focus on audio data where the distributions are continuous (as opposed to discrete in the case of text data). Second, we use Gaussian mixture models as our base classifiers, a more complex generative model than the one used by Ghani.

3 Learning Methods

In this section we describe the major approaches we have explored to take advantage of unlabeled audio data modeled with Gaussian mixtures.

3.1 Notation and Basic Classifier

For our particular problem we assume the existence of a partially labeled data set

$$\{(X_1, y_1), \ldots, (X_L, y_L), X_{L+1}, \ldots, X_{L+U}\}$$

where there are $L + U$ audio files, L labeled with their associated labels y and U unlabeled. Each audio file X is a sequence of N_X feature vectors, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X}\}$.
We define the probability density function for class c (out of a total of C classes) as a mixture of K Gaussians:

$$p(\mathbf{x}_i|Class = c) = \sum_{l=1}^{K} P(l)p(\mathbf{x}_i, \mu_{l,c}, \mathbf{\Sigma}_{l,c}) \tag{1}$$

where $\mathbf{x}_i$ is a feature vector of dimension D, and $p(\mathbf{x}_i, \mu_{l,c}, \mathbf{\Sigma}_{l,c})$ is a multivariate Gaussian distribution with mean $\mu_{l,c}$ and covariance matrix $\mathbf{\Sigma}_{l,c}$. In this paper, we assume a diagonal covariance matrix.
We further assume that each audio class has a particular known *a priori* probability $P(Class = c)$. If $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X}\}$ denotes a single audio file composed of N_X vectors, and assuming that each vector in the sequence is independent and identically distributed, then the likelihood that the whole sequence has been generated by class c is $P(X|Class = c) = \prod_{i=1}^{N_X} p(\mathbf{x}_i|Class = c)$. To decide which class is most likely we can apply Bayes rule and obtain

$$P(Class = c|X) = \frac{P(X|c)P(c)}{\sum_{r=1}^{C} P(X|r)P(r)} \tag{2}$$

The Gaussian mixture models (GMMs) are trained using the well known EM algorithm (Dempster et al., 1977). The EM algorithm is used for maximum likelihood estimation in the presence of hidden or unobserved variables. The algorithm starts with an initial guess for the model parameters to be estimated, and then iterates over two steps: the E-step (expectation step) in which a probability distribution over the values of the hidden variables is computed assuming the current model parameter estimates, and the M-step (maximization step) in which maximum likelihood parameters are estimated using the expected values of the hidden variables based on the distributions computed in the E-step. The goal of the algorithm is to maximize the likelihood of the labeled files being produced by the learned GMMs. Effectively, the algorithm learns the parameters $\theta = \{P(l), \mu_{l,c}, \mathbf{\Sigma}_{l,c}\}$ for all the GMMs. In order to estimate the GMM parameters, EM needs to know the labels y_i for each audio file X_i.

3.2 Semi-Supervised Learning with Iterative EM

EM can be applied to unlabeled data by considering the unknown labels as the hidden variables. The E-step then computes the *a posteriori* probabilities of these unknown labels given the observed data and current model parameter estimates, allowing for the labels to be estimated,

and the M-step calculates the maximum likelihood model parameters based on these posterior distributions.
Initial GMMs are trained using the labeled data, and these are used to assign maximum *a posteriori* labels to the unlabeled data. Using these assigned labels on the unlabeled data, together with the known labels on the initial labeled data, the models are re-trained to obtain new parameter estimates. This process is then repeated until some suitable stopping criterion is reached.
Effectively we can think of this procedure as a two level EM algorithm in which given GMMs for each class we estimate labels for each audio file, and given these estimated labels we estimate GMM parameters $\theta = \{P(l), \mu_{l,c}, \mathbf{\Sigma}_{l,c}\}$ (priors, means, variances) for each class. Table 1 outlines the procedure.

- Begin with L labeled and U unlabeled audio files
- Given labeled pool of files, train C GMMs
- **For** $iter = 0, \ldots, N_{iters}$
  - **For** each unlabeled file X_i
    - Compute $P(X_i|Class = c)$, $c = 1, \ldots, C$
    - Assign label $y_i = \text{argmax}_c P(X_i|c)$ to file X_i
  - Using original L labeled files and labels assigned to the U unlabeled files, retrain the C GMMs
- Return final GMMs

Table 1: Iterative EM algorithm for audio classification.

3.3 Semi-Supervised Learning with Incremental EM

The iterative EM algorithm assigns labels to *all* the unlabeled audio files on each iteration, regardless of the confidence assigned to each audio file. An incremental version of the algorithm selects on each iteration the unlabeled audio files that are classified with highest confidence (i.e. high posterior probability) by the current model, and assigns labels only to these. The unlabeled data points that receive labels are then added to the labeled set, and this new, augmented labeled set is then used to re-estimate model parameters; this process is repeated until all unlabeled points have been labeled. Table 2 outlines the algorithm.

4 Experimental Setup

We conducted experiments with the above algorithms on two different audio classification tasks: gender identification and speaker identification. The first is a binary classification problem involving two classes, while the second task is a more complex problem involving a large number of classes (we used a database containing audio from 50 different speakers).

- Begin with L labeled and U unlabeled audio files
- Given labeled pool of files, train C GMMs
- **While** there are unlabeled files
  - **For** each unlabeled file X_i
    - Compute $P(X_i|Class = c)$, $c = 1, \ldots, C$
  - **For** each class $c = 1, \ldots, C$
    - Let $S = \{X_i : P(X_i|c) > P(X_i|c') \; \forall c'\}$
    - Sort the files X_i in S according to $P(X_i|c)$
    - Select top N files in S; assign them label c, add them to the pool of labeled files; remove them from pool of unlabeled files
  - Using new set of labeled files, re-train the C GMMs
- Return final GMMs

Table 2: Incremental EM algorithm for audio classification.

4.1 Audio Databases

We used a spoken audio database distributed by the Linguistic Data Consortium for our experiments; in particular we used the HUB4 1996 and 1997 datasets. The audio is from broadcast sources, sampled at 16kHz. We had a total of about 30 hours of audio for the gender identification task, and 25 hours for the speaker identification task. To extract features we used a standard mel cepstrum representation popular in the speech recognition community. Each utterance is broken up into frames of 25.6ms each with a frame rate of 100 frames/sec. A Hamming window is applied to each frame and then 256 power spectrum coefficients are computed. The spectrum is then warped according to the Mel scale, its logarithm computed and a final discrete cosine transform applied, resulting in 13 mel-cepstrum coefficients. The first and second time derivatives are computed and appended to the feature vector, resulting in a 39-dimensional vector extracted every 10ms.

We constructed training and testing sets as follows. For each task, we randomized the order of audio files in the database, and then split the database into a small set of labeled data for training initial models, a large set of unlabeled data, and a fairly large set of testing data. For the gender identification task, we had a total of 17,000 audio files, which were split into 2,000 labeled files for use in training initial models, 10,000 unlabeled files for experimentation with unlabeled data algorithms, and 5,000 test files. The speaker identification task had a similar split of a total of 13,217 files into 2,000 labeled files for training initial models, 8,000 unlabeled files, and 3,217 test files.

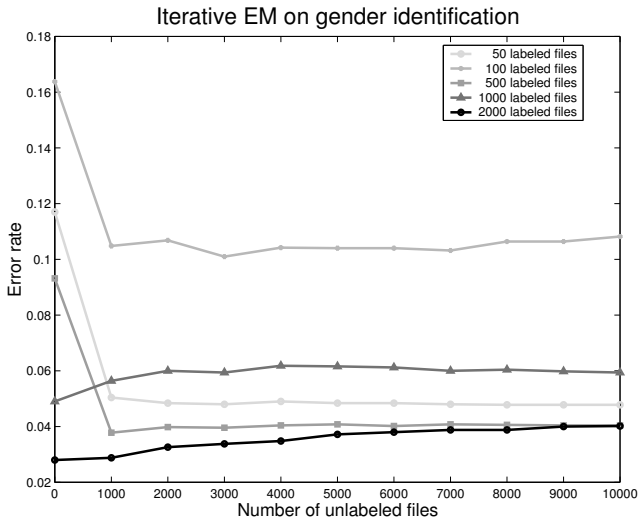

Figure 1: Results of combining labeled and unlabeled data using iterative EM (with 32 Gaussians per class) on gender identification. The error rates shown were measured on a test set of 5,000 files.

4.2 Experiments

The experiments consisted of training initial models using different amounts of labeled data, and then adding different amounts of unlabeled data using one of the two algorithms described in section 3. Initial GMMs were trained using EM on a small amount of labeled data, and the effect of adding increasing amounts of unlabeled data with the different algorithms was then studied.

5 Experimental Results

This section describes the results of our experiments with the different algorithms on each of the two classification tasks.

5.1 Results on Gender Identification

Figures 1 and 2 shows the results of using the iterative and incremental versions of EM on the gender identification task.
For all experiments on the gender identification database, we used 32 Gaussians per GMM, each with a diagonal covariance. (Experiments using 8 and 64 Gaussians per GMM showed similar trends in performance.) The results shown for iterative EM correspond to a single iteration of the EM algorithm. For the incremental EM experiments, upto 50 unlabeled files per class (i.e. a total of 100 files) were addded to the labeled set on each iteration. The incremental EM algorithm was run to completion, *i.e.* until all unlabeled files were added to the labeled set.

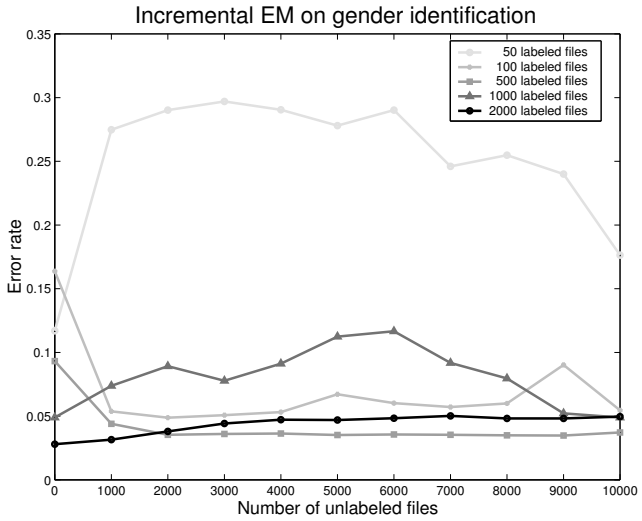

Figure 2: Results of combining labeled and unlabeled data using incremental EM (with 32 Gaussians per class) on gender identification. The error rates shown were measured on a test set of 5,000 files.

As seen in the figures, both forms of EM are successful in improving classification performance with unlabeled data. When labeled data is plenty and the initial models are therefore already good, unlabeled data tends to degrade performance (refer to the plots for 1,000 and 2,000 labeled files); this is in tandem with previous observations (e.g., (Nigam et al., 2000)). However, when only a small amount of labeled data is available, and the initial models are therefore relatively poor, unlabeled data is seen to give a significant improvement in performance. The iterative EM algorithm is especially effective, and can deal well even with initial models trained with very few labeled examples, a case on which the incremental version of the algorithm seems to fail (refer to the plots for 50 labeled examples; even with this small labeled set, the iterative EM algorithm reduces the error rate by more than half, from 11.7% to 5.04%, with the addition of just 1,000 unlabeled examples - and further to 4.78% with the addition of 10,000 unlabeled examples).

5.2 Results on Speaker Identification

Figures 3 and 4 show the results of using the different algorithms on the speaker identification task. The results for all experiments on the speaker identification database are with 64 Gaussians per GMM, each with a diagonal covariance (Experiments using 8 and 32 Gaussians per GMM showed similar trends in performance).
As in the gender identification experiments, the results shown for iterative EM correspond to a single iteration of the EM algorithm. For the incremental EM experiments, up to 2 unlabeled files per class (*i.e.* a total of 100 files) were addded to the labeled set on each iteration. As before, incremental EM was run to completion, *i.e.* until all unlabeled files were added to the labeled set.

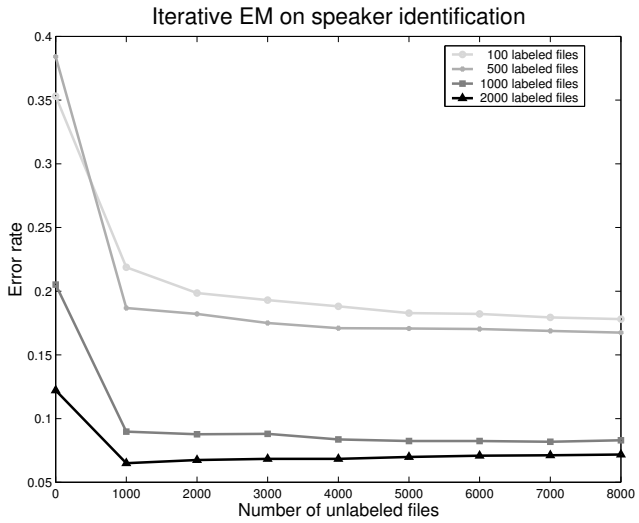

Figure 3: Results of combining labeled and unlabeled data using iterative EM (with 64 Gaussians per class) on speaker identification. The error rates shown were measured on a test set of 3,217 files.

In agreement with the gender identification case, both forms of EM are successful in taking advantage of unlabeled data to improve performance. Again the iterative EM algorithm is especially effective, giving fairly consistent results; even with only 100 labeled files (*i.e.* only 2 labeled files per class), it reduces the error rate from 35.31% to 21.88% with the addition of just 1,000 unlabeled files, and to 17.81% with the addition of 8,000 unlabeled files. This is remarkable given the multi-class nature of the problem. Another interesting observation is that in this case the error rate continues to be reduced even when the initial training set has a larger number of labeled examples; this is probably because for this problem even 2,000 labeled files mean only 40 files per class, and the initial models are therefore not as good as in the gender identification case.

6 Conclusions and Future Work

In this paper we have studied the use of different variants of the EM algorithm for audio classification in a semi-supervised setting. Although the EM algorithm is a general technique that can be used in the presence of any hidden variables, we find that it gives impressive performance when the labels are missing; indeed, our experiments suggest that error rates can be reduced by half using unlabeled data. These results are quite promising, especially given the high cost of human annotations involved in producing labeled training data.

We have only scratched the surface of what promises to be an important direction in audio and other multimedia organization tasks. Many modeling issues still remain to be explored; for example, as more data is labeled the structure of the GMMs could potentially be re-adjusted, adding more component Gaussians to the mixture. Forgetting factors have not been explored ei-

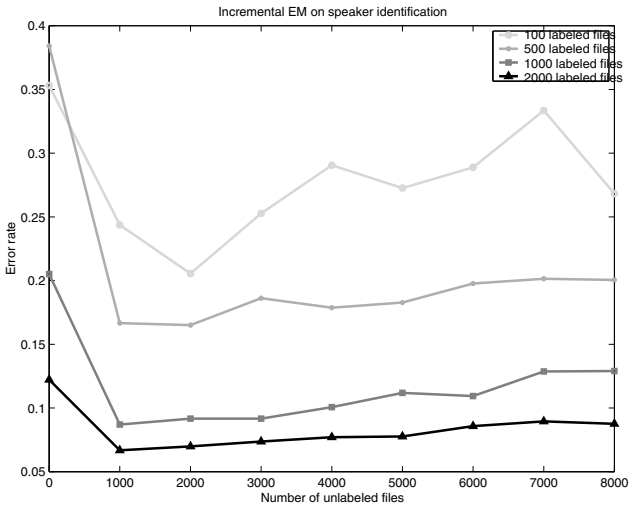

Figure 4: Results of combining labeled and unlabeled data using incremental EM (with 64 Gaussians per class) on speaker identification. The error rates shown were measured on a test set of 3,217 files.

ther; as more data is seen it is plausible that earlier data should be de-weighted in its contribution to re-estimating GMM parameters.

One of the limitations of the EM methods used in this research is the formulation of the cost function we are maximizing. In the work presented in this paper we seek to maximize the combined likelihood of the labeled data and the unlabeled data. Over time, as more and more unlabeled files are added, the contributions of the labeled data become insignificant and the EM algorithm does not offer much guarantees in terms of reducing the error rate. In effect, EM is simply finding a set of GMMs that maximize the likelihood of the data *with no reliable label information present*. Using a cost function that directly aims to minimize the error rate may therefore prove to be more effective. Techniques such as kernel expansions (Szummer & Jaakkola, 2001) that use as cost function the likelihood of the labeled data given the kernel distance across labeled and unlabeled data are also worth exploring.

The use of traditional EM adaptation techniques such MLLR (Leggetter & Woodland, 1995) using unlabeled data has not been explored. The idea of not learning the GMM parameters $\{P(l), \mu_{l,c}, \Sigma_{l,c}\}$, but rather constraining the ways in which they can evolve, is a relevant approach that should also be explored in the context of semi-supervised learning.

References

Bennett, K. P., Demiriz, A., & Maclin, R. (2002). Exploiting unlabeled data in ensemble methods. *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training.

COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data Using the EM Algorithm. *Journal of the Royal Society of Statistics, 39,* 1–38.

Ghani, R. (2001). Combining labeled and unlabeled data for text classification with a large number of categories. *Proceedings of the IEEE International Conference on Data Mining*.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*.

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

Szummer, M., & Jaakkola, T. (2001). Kernel expansions with unlabeled examples. *Advances in Neural Information Processing Systems*.