An Experimental Study of EM-Based Algorithms for Semi-Supervised Learning in Audio Classification

Pedro J. Moreno PEDRO.MORENO@HP.COM
Hewlett Packard, Cambridge Research Laboratory, Cambridge, MA 02142, USA

Shivani Agarwal SAGARWAL@CS.UIUC.EDU
Department of Computer Science, University of Illinois, Urbana, IL 61801, USA

Abstract

We explore the applicability of EM-based algorithms for semi-supervised learning to problems in audio classification. Each audio class is modeled with a Gaussian mixture model, the parameters of which are themselves estimated through EM; this leads to a two-stage algorithm that makes use of EM at both stages. We find that adding unlabeled data through such algorithms to small amounts of labeled data can reduce audio classification error rates by more than half.

1. Introduction

With the ever-growing volume of multimedia content available through digital libraries, electronic databases and the World Wide Web, automatic methods for organizing multimedia data are becoming increasingly important. Since organizing such data often involves classifying multimedia documents into different categories as a key step, methods for automatic classification of audio files, images, video clips etc. are currently of tremendous interest. Machine learning techniques have played an important role in developing such methods, whereby a classifier is automatically trained from a set of sample training data. However, current methods are limited by the need to generate large sets of good quality labeled training data for each new classification problem; for example, in the field of speech research, the Linguistic Data Consortium generates every year hundreds of hours of carefully transcribed audio databases.

In this paper, we explore the possibility of making use of *unlabeled* data to train multimedia classifiers in a semi-supervised setting. As in many other domains, obtaining large amounts of multimedia data *without* labels is often easy: audio data can be readily collected from broadcasts, face images can be obtained from online cameras, and so on. Methods for exploiting such unlabeled data can therefore lead to large savings in both the time and cost required for training classifiers.

In the last few years, a number of algorithms have been developed to combine labeled and unlabeled data for training classifiers in a semi-supervised setting. Blum & Mitchell (1998) introduced the co-training algorithm, which relies on partitioning the feature space into two independent but redundant feature sets. Two classifiers are then trained, one on each feature set, using the labeled data, and each is used to incrementally label remaining examples for the other. Nigam & Ghani (2000) present a detailed discussion on the applicability of co-training, together with comparisons with various other algorithms on text classification problems. Ghani (2001) combines co-training with error-correcting output codes, showing promising results for semi-supervised learning in multi-class problems. Another interesting approach that has been proposed involves using kernel expansions across labeled as well as unlabeled examples (Szummer & Jaakkola, 2001). Recently, ensemble methods such as boosting have also been modified to take advantage of unlabeled data (Bennett et al., 2002).

Here we explore the applicability of semi-supervised learning methods that are based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to problems in audio classification. The use of EM for semi-supervised learning has been proposed in (Miller & Uyar, 1997). More recently, Nigam et al (2000) have studied its application to text classification problems, in which each text class is modeled with a multinomial distribution, corresponding to a naive Bayes clas-

Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, Washington DC, 2003.

sifier. They also consider an extension in which each class is modeled with a mixture of multinomials. In our work, each audio class is modeled with a Gaussian mixture model (GMM). The parameters of each GMM are themselves estimated through EM, giving rise to a two-stage algorithm that makes use of EM at both stages. We apply the algorithm using different variants of EM to both binary and multi-class audio classification problems; for both types of problems, we find that using unlabeled data to augment small amounts of labeled data can reduce audio classification error rates by more than half.

The learning methods we use are discussed in Section 2. Section 3 describes the audio data sets used in our experiments and presents our experimental results. We discuss our results and possible future directions in Section 4, with concluding remarks in Section 5.

2. Learning Methods

We first describe the use of GMMs for supervised learning in audio classification in Section 2.1, and then discuss their extension to semi-supervised learning through different variants of the EM algorithm in Section 2.2.

2.1. Audio Classification with GMMs

GMMs are popular as a generative model in many domains involving continuous data due to their flexibility and analytical tractability. Here we briefly describe their application to audio classification.

In a supervised audio classification problem, we are given a finite set of labeled training examples:

$$S^l = \{(X_1, y_1), \ldots, (X_L, y_L)\},$$

where X_i are audio files and y_i are corresponding labels in some set of class labels $Y = \{c_1, c_2, \ldots, c_M\}$. Each audio file X is represented as a sequence of feature vectors in some Euclidean space $\mathbb{R}^d$:

$$X = < \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X} >, \quad \mathbf{x}_n \in \mathbb{R}^d.$$

Note that the number of feature vectors, N_X, can be different for different audio files X. Each example (X_i, y_i) in the training set S^l is assumed to be drawn independently from some fixed (but unknown) underlying distribution. Using the training set, we are required to predict, for a new audio file X, the class j with highest posterior probability, $P(c_j|X)$, under this distribution (under the assumption that new examples are drawn from the same distribution).

Generative classifiers generally perform this prediction by modeling the class-conditional distributions $p(X|c_j)$, and then using Bayes' rule to compute the posteriors:

$$\begin{aligned} P(c_j|X) &= \frac{p(X|c_j)P(c_j)}{p(X)} \\ &= \frac{p(X|c_j)P(c_j)}{\sum_{j'=1}^{M} p(X|c_{j'})P(c_{j'})}. \end{aligned} \tag{1}$$

The class priors $P(c_j)$ are either taken to be uniform, or estimated by counting the number of times each label appears in the training data.

In using GMMs for audio classification, each of the class-conditional densities of feature vectors $\mathbf{x} \in \mathbb{R}^d$ is modeled as a mixture of Gaussians:

$$p(\mathbf{x}|c_j) = \sum_{k=1}^{K_j} \lambda_{jk} \; p(\mathbf{x}; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}), \tag{2}$$

where K_j is the number of mixture components in the model for class j, λ_{jk} are mixing weights, and $p(\mathbf{x}; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ is a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_{jk}$ and covariance matrix $\boldsymbol{\Sigma}_{jk}$. In this paper, we use the same (fixed) number of Gaussian components to model each class, i.e. $K_j = K \; \forall j$, and assume diagonal covariance matrices. A common (albeit incorrect) assumption that is made is that the feature vectors in the sequence representing an audio file are independent. The probability of a complete file $X = < \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X} >$ being generated from class j is then given by

$$\begin{aligned} p(X|c_j) &= \prod_{n=1}^{N_X} p(\mathbf{x}_n|c_j) \\ &= \prod_{n=1}^{N_X} \left(\sum_{k=1}^{K} \lambda_{jk} \; p(\mathbf{x}_n; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \right). \end{aligned} \tag{3}$$

As described above, this can be used with Bayes' rule (Eq. 1) to compute the posterior probabilities $P(c_j|X)$, in order to arrive at the maximum *a posteriori* classification for X:

$$j* = \arg\max_j P(c_j|X). \tag{4}$$

The parameters $\boldsymbol{\theta}_j = \{\lambda_{jk}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\}$ of the GMMs are estimated from the training data using the EM algorithm. 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 expectation step (E-step) in which expected values of the hidden variables are computed assuming the current model parameter estimates, and the maximization step (M-step) in which maximum likelihood parameters are estimated using the expected values of the hidden variables computed in the E-step. On convergence, the algorithm produces both (local) maximum likelihood estimates of the model parameters and expected values of the hidden variables. The application of EM to learning the parameters of a GMM is detailed in several texts, e.g. (Mitchell, 1997). Briefly, given a set of data points drawn from a GMM whose parameters are to be estimated, information about which component of the GMM generated each data point is missing. This information is encoded in the form of hidden variables, and EM is then used to find GMM parameters that maximize the likelihood of generating the observed data, together with expected values for the hidden variables. In our case, the GMM parameters $\boldsymbol{\theta}_j$ for each class j are estimated using the audio files in S^l that belong to class j.

2.2. Incorporating Unlabeled Data with EM

In many practical applications, the number L of labeled training examples is insufficient to obtain accurate parameter estimates. We discuss how EM-based procedures can be used to take advantage of unlabeled data to improve parameter estimates in such cases.

The scenario we consider now is when the labeled training set S^l can be augmented with a set of unlabeled examples S^u, so that the complete set of training data, $S = S^l \cup S^u$, is given by

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

In this case, the unknown labels $y_{L+1}, \ldots, y_{L+U}$ corresponding to the unlabeled audio files $X_{L+1}, \ldots, X_{L+U}$ constitute missing information. This missing information can be encoded in the form of hidden variables; for each unlabeled example X_i, we can define M hidden variables z_{ij}, $j = 1, \ldots, M$, such that

$$z_{ij} = \begin{cases} 1 & \text{if } y_i = c_j \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

The EM algorithm can then be used to find GMM parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j\}$ that maximize the *combined* likelihood of the labeled and unlabeled data in the presence of these hidden variables, $\mathbf{z} = \{z_{ij}\}$. Applying EM in its basic form consists of starting with an initial parameter estimate $\hat{\boldsymbol{\theta}}^{(0)}$ (which can be obtained by training the GMMs using the labeled data only, as described in the preceding section), and then iterating over the following two steps:

- [E-step] Set $\hat{\mathbf{z}}^{(t+1)} = E[\mathbf{z}|S; \hat{\boldsymbol{\theta}}^{(t)}]$.
- [M-step] Set $\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} P(S, \hat{\mathbf{z}}^{(t+1)}|\boldsymbol{\theta})$.

```
• Set $t = 0$.
• [Initial M-step] Initialize $\hat{\boldsymbol{\theta}}^{(0)} = \arg\max_{\boldsymbol{\theta}} P(S^l|\boldsymbol{\theta})$.
• Repeat till convergence:
    • [E-step] Set $\hat{\mathbf{z}}^{(t+1)} = E[\mathbf{z}|S; \hat{\boldsymbol{\theta}}^{(t)}]$.
    • For $i = L+1, \ldots, L+U$ do:
        • Set $j^* = \arg\max_j \hat{z}_{ij}^{(t+1)}$.
        • Set $\hat{z}_{ij}^{\text{hard}(t+1)} = \begin{cases} 1 & \text{if } j = j^* \\ 0 & \text{otherwise} \end{cases}$ , $j = 1, \ldots, M$.
    • [M-step] Set $\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} P(S, \hat{\mathbf{z}}^{\text{hard}(t+1)}|\boldsymbol{\theta})$.
    • Set t = t+1.
• Output $\hat{\boldsymbol{\theta}}^{(t)}$.
```

Table 1. Iterative EM-based algorithm (Section 2.2.1) for estimating maximum likelihood values for parameters $\boldsymbol{\theta}$ given training data $S = S^l \cup S^u$, where S^l contains labeled examples and S^u contains unlabeled examples (see text).

As discussed in (Nigam et al., 2000), since $E[z_{ij}|S; \boldsymbol{\theta}] = P(y_i = c_j|S; \boldsymbol{\theta}) = P(c_j|X_i; \boldsymbol{\theta}_j)$, the E-step effectively assigns probabilistic labels to the unlabeled examples, according to their posterior probabilites (which can be computed using Eqs. 1 and 3). Consequently, if the expected values of z_{ij} are used as such, the M-step requires computing maximum likelihood parameters with "fractional" examples in each class. Instead, we use the expected values of z_{ij} (i.e. posteriors) to estimate a "hard" labeling for the unlabeled examples at each step. Different strategies for assigning these hard labels give rise to two different algorithms for estimating the parameters; the first is iterative in nature, the second incremental.

2.2.1. Iterative EM-Based Algorithm

One method for assigning hard labels to the unlabeled examples is to find, for each unlabeled example X_i, the variable z_{ij} with the highest expected value, and assign X_i to the corresponding class j. This leads to the algorithm outlined in Table 1. The algorithm constitutes an iterative procedure: at each iteration, the E-step re-labels every unlabeled example with the maximum *a posteriori* class predicted by the current model parameters, and the M-step then re-estimates the model parameters assuming the current labeling of the data. Each M-step in the algorithm involves an inner use of EM to estimate maximum likelihood GMM parameters, as described in Section 2.1.

2.2.2. Incremental EM-Based Algorithm

The iterative algorithm above assigns hard labels to *all* the unlabeled examples at each iteration, irrespective

```
• Set $t = 0$.
• [Initial M-step] Initialize $\hat{\boldsymbol{\theta}}^{(0)} = \arg\max_{\boldsymbol{\theta}} P(S^l|\boldsymbol{\theta})$.
• While $S^u \neq \emptyset$ do:
  • [E-step] Set $\hat{\mathbf{z}}^{(t+1)} = E[\mathbf{z}|S;\hat{\boldsymbol{\theta}}^{(t)}]$.
  • For $j = 1, \ldots, M$ do:
    • Set $S^u_j = \{X_i \in S^u : \hat{z}^{(t+1)}_{ij} > \hat{z}^{(t+1)}_{ij'} \ \forall \ j' \neq j\}$.
    • Set $i^* = \arg\max_{\{i:X_i \in S^u_j\}} \hat{z}^{(t+1)}_{ij}$.
    • Set $S^l = S^l \cup \{(X_{i^*}, c_j)\}$.
    • Set $S^u = S^u \setminus \{X_{i^*}\}$.
  • [M-step] Set $\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} P(S^l|\boldsymbol{\theta})$.
  • Set t = t+1.
• Output $\hat{\boldsymbol{\theta}}^{(t)}$.
```

Table 2. Incremental EM-based algorithm (Section 2.2.2) for estimating maximum likelihood values for parameters $\boldsymbol{\theta}$ given training data $S = S^l \cup S^u$, where S^l contains labeled examples and S^u contains unlabeled examples (see text).

of the confidence in the assignments. An alternative approach is to assign hard labels only to the examples for which the maximum *a posteriori* classification according to the current model can be made with high confidence (i.e. the predicted class has high posterior probability). This leads to the algorithm outlined in Table 2. The algorithm proceeds in an incremental fashion, adding the most confidently classified example in each class to the labeled set on each iteration. In practice, due to time considerations we allow the algorithm to label on each iteration the n most confidently classified examples in each class, for some appropriate number n. Again, each M-step involves an inner use of EM for GMM parameter estimation.

The incremental algorithm above is equivalent to the self-training algorithm of (Nigam & Ghani, 2000).

3. Experiments

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 (in our case, the task was to distinguish between 50 different speakers).

3.1. Data Sets

We used a spoken audio database distributed by the Linguistic Data Consortium[1]. In particular, we used the HUB4 1996 and 1997 data sets. The audio is from broadcast sources, sampled at 16kHz.

For the gender identification task, we had a total of 17,000 audio files, corresponding to a total of about 30 hours of spoken audio. To construct training and test sets, we randomized the order of audio files in the data set, and then split the data into a small set of 2,000 labeled files for use in training initial models, an unlabeled training set of 10,000 files, and a test set of 5,000 files.

For the speaker identification task, we had a total of 13,217 audio files, corresponding to a total of about 25 hours of spoken audio. As for the gender identitfication task, we randomized the order of audio files in the data set, and then split the data into a small set of labeled data for training initial models, a large set of unlabeled training data, and a set of test data. In this case we had a labeled training set of 2,000 audio files, an unlabeled training set of 8,000 files, and a test set of 3,217 files.

3.2. Feature Extraction

To extract features from the audio files, we used a standard Mel-cepstrum representation popular in the speech recognition community (Rabiner & Juang, 1993). Each audio file is broken up into overlapping frames of 25.6 milliseconds each with a frame rate of 100 frames per second. 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 (known as the delta and delta-delta features) are computed and appended to the feature vector, resulting in a 39-dimensional vector extracted every 10 milliseconds. (Thus according to the notation of Section 2, $d = 39$ in our experiments.)

3.3. Experimental Results

The experiments consisted of training initial models using different amounts of labeled data, and then studying the effect of adding increasing amounts of unlabeled data with the two algorithms described in Section 2.2. The following sections discuss the results on each task.

3.3.1. Results on Gender Identification

Figures 1 and 2 show the results of using the iterative and incremental EM-based algorithms on the gender identification task. For all experiments reported on

[1] `http://www.ldc.upenn.edu/`

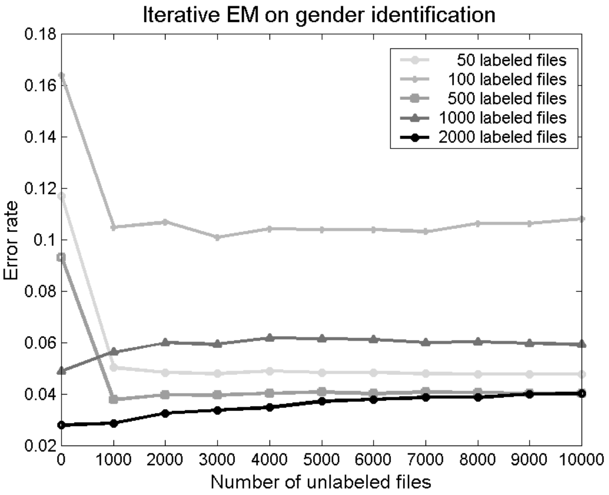

Figure 1. Results of combining labeled and unlabeled data using the iterative EM-based algorithm of Section 2.2.1 on the gender identification task. The error rates shown were measured on a test set of 5,000 files.

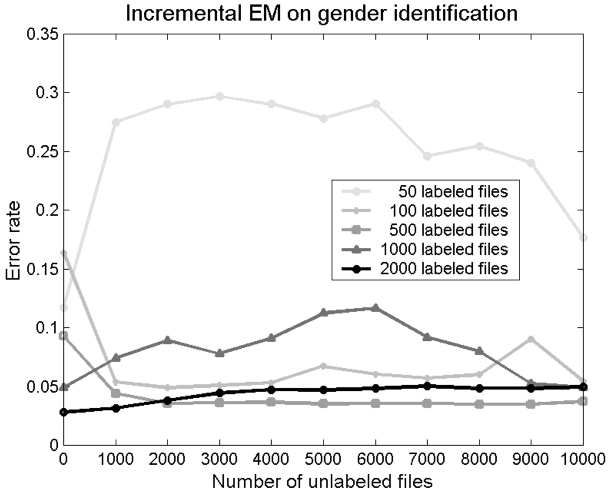

Figure 2. Results of combining labeled and unlabeled data using the incremental EM-based algorithm of Section 2.2.2 on the gender identification task. The error rates shown were measured on a test set of 5,000 files.

this task, each class was modeled with a mixture of 32 Gaussians, each with a diagonal covariance matrix. The results shown for iterative EM correspond to a single iteration of the algorithm. For the incremental EM experiments, upto 50 unlabeled files per class (i.e. a total of 100 files) were added 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.

As seen in the figures, both algorithms are capable of improving classification accuracy with unlabeled data, although the iterative algorithm appears to give more consistent results. When labeled data is plenty and the initial parameter estimates are therefore already accurate, adding unlabeled data with either algorithm 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 parameter estimates are therefore relatively poor, unlabeled data is seen to give a significant improvement in performance. The iterative algorithm is especially effective, and can deal well even with initial models trained with very few labeled examples, a case on which the incremental algorithm seems to fail (refer to the plots for 50 labeled examples; even with this small labeled set, the iterative EM-based 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).

It is interesting to note that adding just a small amount of unlabeled data seems to give almost all of the advantage that is gained from adding larger numbers of unlabeled examples. One possible explanation for this is that adding increasing amounts of unlabeled data leads to parameter estimates that depend very little on the labeled data for which reliable class information is actually known, and therefore may not be optimal in terms of improving classification accuracy. De-weighting the contribution of the unlabeled examples may be one way to overcome this phenomenon; this is to be investigated in future work.

3.3.2. Results on Speaker Identification

Figures 3 and 4 show the results of using the two algorithms on the 50-class speaker identification task. The results for all experiments reported on this task are with mixtures of 64 Gaussians per class, each with a diagonal covariance matrix. As in the gender identification experiments, the results shown for iterative EM correspond to a single iteration of the algorithm. For the incremental EM experiments, up to 2 unlabeled files per class (i.e. a total of 100 files) were added 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.

As in the gender identification case, both EM-based algorithms are able to take advantage of unlabeled data to improve performance, although again the iterative algorithm is especially effective. Even with only 100

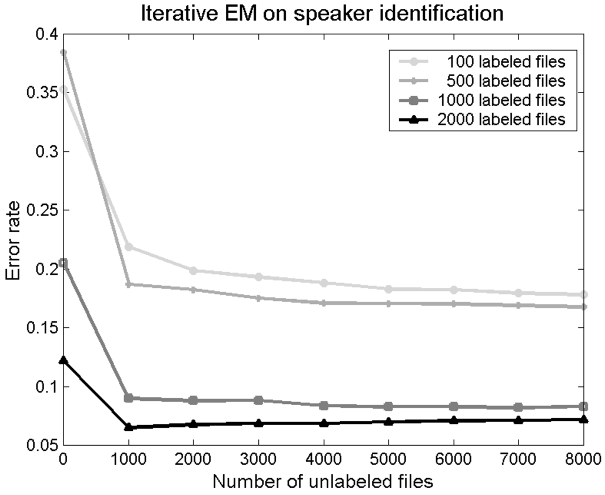

Figure 3. Results of combining labeled and unlabeled data using the iterative EM-based algorithm of Section 2.2.1 on the speaker identification task. The error rates shown were measured on a test set of 3,217 files.

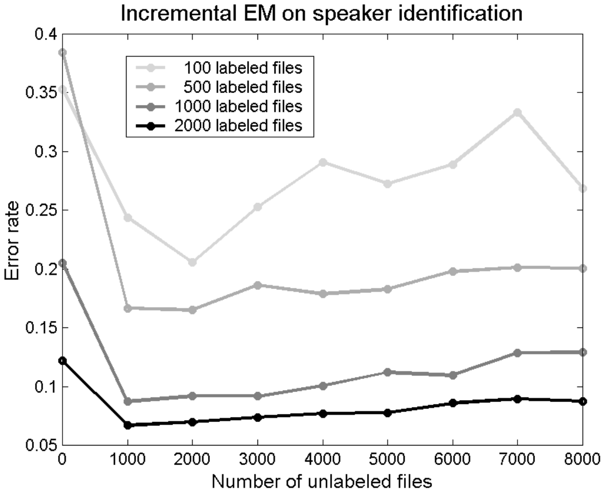

Figure 4. Results of combining labeled and unlabeled data using the incremental EM-based algorithm of Section 2.2.2 on the speaker identification task. The error rates shown were measured on a test set of 3,217 files.

labeled files (i.e. only 2 labeled files per class), the iterative algorithm reduces the error rate from 35.31% to 21.88% with the addition of just 1,000 unlabeled files, and further to 17.81% with the addition of 8,000 unlabeled files. In this case the error rate is reduced even when the initial training set has a larger number of labeled examples (refer to the plots for 1,000 and 2,000 unlabeled files). This is because for this problem, even 2,000 labeled files mean only 40 files per class, and the initial parameter estimates are therefore not as good as in the binary gender identification problem.

Again, we observe a steep improvement in performance with the addition of a small amount of unlabeled data, but relatively little further improvement on adding larger numbers of unlabeled examples. This requires further investigation.

4. Discussion

It is clear from our experiments that unlabeled data, in conjunction with EM-like algorithms, has the potential to improve audio classification accuracy when labeled audio data is limited in quantity. The experiments also suggest that an iterative EM-based algorithm is better suited to this task than an incremental one, although more extensive experimentation with parameters such as the number of iterations and size of increment would be required to confirm this.

The results of our experiments lead naturally to a number of interesting questions. From a practical viewpoint, an important open question is how to determine the extent to which unlabeled data will help for a given set of labeled training data; as we have seen in our experiments, this can vary significantly with the sizes of the training sets involved. It would be useful to develop a theoretical understanding of this issue; for example, it may be possible to develop PAC-type bounds for semi-supervised learning that quantify the expected gain or loss in accuracy for given sizes of labeled and unlabeled training sets.

It is also important to understand why the benefit gained from unlabeled data seems to wither out as more unlabeled data is added. One possible reason for this lies in the formulation of the objective function that is maximized; in the EM-based methods used in this paper, we seek to maximize the combined likelihood of the labeled data and the unlabeled data. Consequently, as more and more unlabeled files are added, the contribution of the labeled data becomes insignificant over time and the algorithms do not offer any guarantees in terms of reducing the error rate; in effect, the algorithms can end up finding a set of GMM parameters that maximize the likelihood of the data with no reliable label information present. Using an objective function that directly aims to minimize some measure of the error rate may therefore prove to be more effective. Techniques such as kernel expansions (Szummer & Jaakkola, 2001) that use as an objective function the likelihood of only the labeled data, given some distance measure across labeled and unlabeled data, are also worth exploring.

It may also be possible to make use of ideas that have been developed in the speech understanding community. For example, adaptation techniques such as maximum likelihood linear regression (Leggetter & Woodland, 1995), which constrain the manner in which model parameters can evolve and therefore limit the deviation from initial parameter estimates, can prove to be useful in the context of semi-supervised learning.

5. Conclusion

In this paper we have studied the use of different variants of the EM algorithm for semi-supervised learning in audio classification. We applied the algorithms to both binary and multi-class classification problems; our experiments suggest that for both types of problems, audio classification 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. As discussed in the previous section, there are several possible avenues for further improvement and investigation. Many modeling issues also remain to be explored; for example, as more data is labeled the structure of the GMMs can potentially be re-adjusted, adding more component Gaussians to the mixtures. Nevertheless, our initial experiments indicate that semi-supervised learning can be profitably applied to audio data. Furthermore, it is likely that our results with GMMs can be extended to other multimedia domains, such as images and video, that also involve primarily continuous-valued attributes.

Acknowledgments

The authors would like to thank Nuno Vasconcelos for helpful discussions, and the anonymous referees for their constructive suggestions. This research was conducted while Shivani Agarwal was an intern at the HP Cambridge Research Laboratory.

References

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

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory.*

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *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*, *9*, 171–185.

Miller, D. J., & Uyar, H. S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in Neural Information Processing Systems 9* (pp. 571–577).

Mitchell, T. (1997). *Machine learning.* McGraw Hill.

Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. *Proceedings of the International Conference on Information and Knowledge Management.*

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.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition.* Prentice Hall.

Szummer, M., & Jaakkola, T. (2001). Kernel expansions with unlabeled examples. *Advances in Neural Information Processing Systems 13.*