CROSS-VALIDATION EM TRAINING FOR ROBUST PARAMETER ESTIMATION

*T. Shinozaki**

Kyoto University, Kyoto, Japan
staka@u.washington.edu

M. Ostendorf

Department of Electrical Engineering
University of Washington, Washington, U.S.A
mo@ee.washington.edu

ABSTRACT

A new maximum likelihood training algorithm is proposed that compensates for weaknesses of the EM algorithm by using cross-validation likelihood in the expectation step to avoid overtraining. By using a set of sufficient statistics associated with a partitioning of the training data, as in parallel EM, the algorithm has the same order of computational requirements as the original EM algorithm. Analyses using a GMM with artificial data show the proposed algorithm is more robust for overtraining than the conventional EM algorithm. Large vocabulary recognition experiments on Mandarin broadcast news data show that the method makes better use of more parameters and gives lower recognition error rates than EM training.

Index Terms— EM training, overtraining, cross-validation

1. INTRODUCTION

A general problem in model estimation is overfitting to the training data. When maximum likelihood (ML) estimation is used, training set likelihood increases with the number of model parameters. However, the difference in likelihood of the training data and new data also grows. In other words, the model loses the ability to generalize. This problem can be more severe with unstable (high variance) classifiers, i.e. those that can lead to very different results for different randomly selected training sets. A variety of methods have been proposed for addressing this problem, including information-theoretic criteria, such as the Bayesian information criterion (BIC) and minimum description length (MDL) that trade-off likelihood gain and a penalty function related to the number of free parameters, and data-driven techniques such as cross-validation (CV) and bootstrapping [1].

When a model includes hidden variables, such as in a hidden Markov model (HMM), it is difficult to solve the ML problem directly, and the iterative expectation-maximization (EM) algorithm is used [2]. A complication with the EM algorithm is that there is in general no guarantee of reaching a global optimum, and local optima can at times be problematic, particularly with small training sets. For models that incorporate Gaussian mixture distributions, there is a further complication: the EM algorithm can be unstable. For example, a two-mixture Gaussian distribution gives arbitrarily large likelihood for training data if one of the Gaussians covers only one data point with very small variance and the other Gaussian spans the rest of the data points. Obviously, such a model is not desirable. A simple heuristic solution is to limit the likelihood by flooring variances of Gaussian components with a threshold [3], but the threshold needs to be tuned empirically since a large threshold increases class confusability and a small threshold may not eliminate the problem.

While information-theoretic model selection techniques have proved useful in the model selection problem for HMMs [4, 5], they do not address the potential problem of instability. CV has been used within decision tree design for improving question selection and determining size [6], which in turn determines model complexity via distribution tying. Here, we take this idea further, proposing a variation of the EM algorithm that incorporates CV within the iterative procedure, referred to as CV-EM. A key difference is that, rather than use CV for model selection, it is used in CV-EM to reduce the bias of the sufficient statistics and thereby improve the parameter estimates. Using parallel training techniques, the CV-EM algorithm is similar in complexity to the standard EM algorithm.

To analyze the basic behavior of the algorithm, the proposed method is first applied to train GMMs using artificial data and the model performance is evaluated by calculating likelihood for new data. In the analysis, it is shown that the models trained by CV-EM give similar or higher likelihood than EM. CV-EM is more robust for overtraining than EM and provides a means to automatically decide the optimal number of iterations.

CV-EM is then applied to large vocabulary recognition experiments on Mandarin broadcast news. The speech recognition experiments show that CV-EM provides more stable performance as a function of model size and it improves overall system performance.

The algorithm is introduced in the next section, followed by the analyses using GMMs with artificial samples and the experiments using HMMs in Mandarin broadcast news speech recognition. Alternative applications and open questions are raised in the discussion.

2. CROSS-VALIDATION EM TRAINING

For distributions in the exponential family, the EM algorithm iterates between two steps:

- *E-step:* given the observations O and the current model parameters $\theta^{(p)}$, compute the expected sufficient statistics $t^{(p)} = E[t|O, \theta^{(p)}]$ associated with the hidden variables.
- *M-step:* Use these statistics to update the model parameters $\theta^{(p+1)}$ in ML estimation, as if they were based on the observed variables.

(Examples of sufficient statistics t are the first and second order sample averages for a Gaussian, which are weighted by the state occupancy posteriors to get $t^{(p)}$ in an HMM.) Because the E-step and the M-step use the same training data, the algorithm is susceptible to reinforcing bad choices, such as assigning too few samples to a

*Most of the work was carried out when the author was at Department of Electrical Engineering, University of Washington, Seattle, Washington, U.S.A and International Computer Science Institute, Berkeley, California, USA

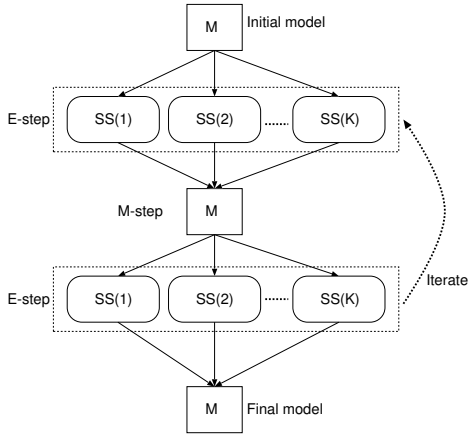

Fig. 1. Parallel EM training. SS(i) denotes the sufficient statistics for the i-th data subset.

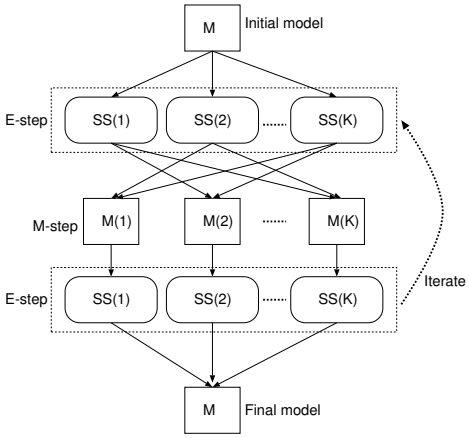

Fig. 2. CV-EM training. M(i) denotes the i-th CV model estimated without using the i-th data subset.

Gaussian mixture component with high likelihood or a poor state alignment in HMM modeling. The key idea behind the proposed CV-EM method is to compute the sufficient statistics of subsets of data using a model that was estimated from an independent set of data. Since there is no overlap in the data used for the E-step and the M-step, we reduce the potential for overfitting.

2.1. Algorithm

The procedure of CV-EM training is similar to that for parallel EM training [7]. In parallel EM training (Figure 1), the training data is partitioned into K subsets, sufficient statistics are independently calculated for each subset in the E-step, and then accumulated. The model parameters are updated in the M-step. In CV-EM, the partition is leveraged to efficiently design multiple models, as shown in Figure 2, so that the M-step and E-step procedures use different data. Specifically, the first E-step is identical to the parallel EM algorithm, and K sufficient statistics files are computed for the partitions. Then, instead of making a single model by accumulating all the sufficient statistics, K cross-validation models are generated by excluding the sufficient statistics from one subset. Each cross-validation model is used in the E-step to estimate the new sufficient statistics for the data subset that has been excluded from the parameter estimation of that model. The E-step and the M-step are repeated as in conventional EM training, and the final model is obtained by merging all the sufficient statistics.

The use of K-fold CV in the EM updates has the potential problems of a pessimistic bias and higher variance associated with using a smaller effective training set for each model. However, the problems of data fragmentation are minimal, because the model is estimated from a large fraction $(K-1)/K$ of the data. In addition, the increase in expected error can be offset by a reduction in error associated with avoiding an optimistic (overfitting) bias and problematic local optima. Probability distributions that are highly specialized to a particular data point cannot earn large likelihood, since the data point used in the M-step does not appear in the E-step. Thus, the inclusion of K-fold CV within each iteration of EM makes the algorithm as a whole more robust to the problem of local optima, though there is still no guarantee of convergence to a global optimum.

2.2. Implementation Details

In general, K-fold CV requires making K models, which will increase the overall computational cost by a factor of K if these models are separately trained from scratch. However, that increase can be mostly avoided by using sufficient statistics. For training sets that are large compared to K, the overall computational cost of CV-EM is mostly determined by computing the K sets of sufficient statistics in the E-step, which has the same computational complexity as that of EM. The K cross-validation models are obtained by subtracting the corresponding sufficient statistics from the overall sum without repeating their accumulation. The storage requirement is mostly determined by the collection of sufficient statistics files and is linear in K.

Since there are multiple models trained with hidden variables, a potential problem is that the models could learn different interpretations of these variables that would be lost in the combination of the sufficient statistics from different models. To illustrate with a more concrete example, if Gaussian mixture models are trained independently, merging the statistics estimated from different models would not make sense, because there is no correspondence between a specific mixture component in one model vs. another. This problem is avoided by initializing the algorithm with a single model and by having a sufficiently large K so that the different models have a large amount of data in common, so it is less likely that they diverge drastically in interpretation of hidden variables. Note that for K-fold CV-EM training, any combination of two cross-validation models share $(K-2)/(K-1)$ of their training data, or 95% for $K = 21$.

3. EXPERIMENTS

3.1. Simulated Data

Analyses were performed using training and test data sampled from 4-dimensional 8-mixture Gaussian distributions whose component diagonal Gaussians and weights were randomly defined. GMMs with 8 mixture components were trained by first initializing their parameters using the global mean and variance and then applying the CV-EM or the EM algorithm. The models were trained with different training set sizes and different numbers of CV folds. The performance of the models were evaluated by likelihood calculated for the

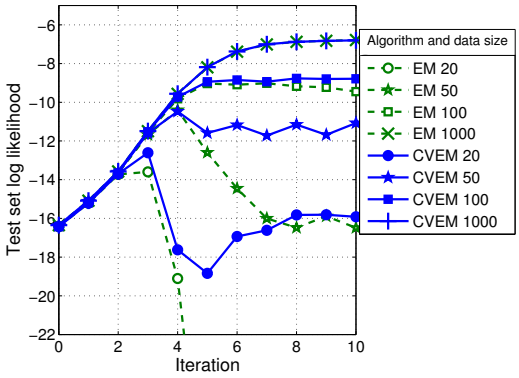

Fig. 3. Test set likelihood of GMMs trained by EM and CV-EM with varying training set sizes.

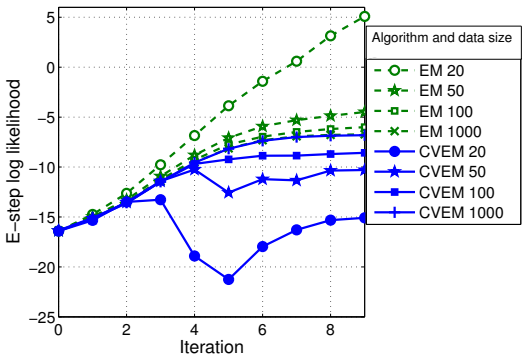

Fig. 4. Training set likelihood obtained in the E-step.

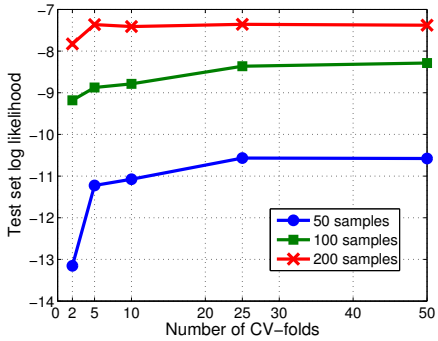

Fig. 5. Model performance as a function of number of CV folds (K).

test sets with 1000 samples. To eliminate the randomness of the results, the experiments were repeated for 100 times for each training condition using data sampled from the different random population distributions and their likelihood was averaged.

Figure 3 shows the test set likelihood of the GMMs trained by the EM and the CV-EM algorithms with varying training set sizes as a function of the number of the iterations. The zeroth iteration means the likelihood is evaluated using the initial model. The number of CV folds K for the CV-EM training was 10. For CV-EM training, general models that integrate all the K sufficient statistics were generated at each iteration along with the CV models and used for the evaluation. When the same initial model is used, CV-EM gives the same general model as EM for the fist iteration since the first E-step gives the same results. The difference of the two algorithms appears after the second iteration.

As can be seen in the figure, CV-EM training always gives about the same or higher likelihood than EM training. The most prominent advantage of CV-EM is the stability for the large training iterations. Although it is guaranteed for EM that it always increases the training set likelihood at each iteration, it does not hold for the test set likelihood. Especially, when the training data is small, the algorithm tends to make the model specialized for particular training samples and the generality of the model is lost as the iteration proceeds. As a result, the test set likelihood first increases but then begins to decrease after reaching an optimal point. As shown in the figure, CV-EM training is not completely free from overtraining, but it is much more stable than EM training. Because a large number of iterations can be safely specified, the stability is useful when the optimal number of iterations is not known and when composite models (such as HMMs with GMM observation distributions) have components that may benefit from different numbers of iterations.

Figure 3 also shows that, as the training set size increases, the difference of the likelihood after EM vs. CV-EM training becomes small. In other words, the problem of overtraining is minor when a large number of training samples is used relative to the number of model parameters.

Figure 4 shows the training set likelihood obtained in the E-step of the EM and the CV-EM training. Because the E-steps are parallelized, the likelihood is obtained as a weighted average of their output. The zeroth iteration means the first E-step that uses the initial model. Unlike the EM training, the CV-EM likelihood is not monotonic with the number of iterations. While the EM likelihood increases as the training set size become small, CV-EM has smaller likelihood for the smaller training set. In other words, the E-step likelihood of the CV-EM training has a similar trend as the test set likelihood. This is because the CV-EM likelihood is calculated for each data subset using a model that is estimated without using that subset and thus is more reliable. The likelihood by EM and CV-EM training is similar when large training set is used, because the model parameters and the likelihood are estimated properly regardless of the method. This property of the CV-EM likelihood makes it possible for CV-EM to automatically decide the optimal number of iterations. This possibility is investigated in [8].

Figure 5 shows the model performance as a function of the number of CV folds K. The GMMs were trained with different numbers of CV folds for the same training set. The CV-EM training was iterated for 10 times. When small K is used, the effective training set size becomes small and the model performance degrades especially when small training set is used.

3.2. Speech Recognition

Speech recognition experiments were conducted using training data consisting of Mandarin Hub4 and the broadcast news programs from the TDT4 corpus. The total amount of training data was 97 hours. The development and test set were the dev04 and eval04 of the Mandarin RT-04 and included half an hour and one hour of broadcast news programs, respectively. The experiments build on the Decipher recognition system for Mandarin broadcast news [9], using a slightly simplified version with a trigram language model and maximum likelihood acoustic model training. The dictionary included 49k words. Acoustic models were word internal, tied-state triphones

Table 1. CER for development set and evaluation set

Model		Dev04 set		Eval04 set	
# states	# mixes	EM	CV-EM	EM	CV-EM
2500	32	8.8	9.2	18.9	18.9
5000	32	9.3	9.1	18.9	18.8
1800	128	9.1	8.8	18.8	18.4
3000	128	9.4	8.9	18.7	18.6
6000	128	9.9	9.6	19.7	18.6

with untied mixture weights. A small variance floor was used (10^{-20}), which was the standard configuration for this system. Vocal tract length normalization and speaker adaptation were applied. Evaluation was based on character error rate (CER). The baseline models were trained using the parallel EM algorithm. CV-EM training used a number of CV folds of $K = 21$. The training data was randomized before the partitioning for the CV-EM training.[1] The same number of the iterations (five) was used for both methods using the same Gaussian mixture HMM as the input.

Table 1 shows the CER for the development and evaluation set of the speech recognition experiments. The system parameters were tuned for the EM-based models with the development set, and the same settings were used for the models trained by the proposed methods. CV-EM gives better results than EM when the models are relatively large and is more robust for data sparsity. By choosing optimal model sizes on the development set, CV-EM leads to 2.6% relative CER reduction on the evaluation set from 18.9% to 18.4%, which is statistically significant at the level of p=.021 using the NIST Matched Pairs Sentence-Segment Word Error (MAPSSWE) test.

4. DISCUSSION

In summary, we have explored the use of CV within the EM algorithm as a means of providing more robust parameter estimates. We experimentally analyzed the algorithm using GMMs with artificial data and showed it gives the similar or higher test set likelihood than EM and is stable for overtraining. In addition, it gives reliable likelihood in the E-step that can be used to predict the optimal number of iterations. We also evaluated the approach in training large HMMs with Gaussian mixture distributions and achieved small gains in error rate over standard EM with roughly the same computational requirements as the standard EM algorithm.

The same CV idea can be applied to other iterative training methods that can be parallelized using sufficient statistics such as discriminative training [10, 11]. Since discriminative training aims at eliminating the difference of likelihood-based optimization and error-rate-based evaluation, and CV training is intended to reduce the error of the parameter estimates, the two approaches are complementary. The CV-EM process is somewhat analogous to cross-adaptation [12], which uses statistics from different models representing different views of the same data. It would be interesting to assess whether similar gains could be achieved with a single model if the same CV-EM ideas are applied within the transform estimation process [13].

The results here represent an empirical advance, but the theoretical basis remains to be explored. While we have an intuitive argument for why the CV-EM strategy is useful within the EM framework, the proof of convergence for the EM algorithm no longer holds since CV-EM can decrease the likelihood of the training data. However, there are conditions under which iterative estimation algorithms may converge without guaranteed likelihood increases at each step [14], which may be relevant to this question.

[1] We also experimented with constraints in the CV partitioning that forced all utterances from the same speaker to be in a single subset to ensure independence of the subsets, but no additional improvement was obtained.

Acknowledgments

This work was supported by DARPA, contract No. HR0011-06-C-0023. Distribution is unlimited. The views herein are those of the authors and do not reflect the views of the funding agency.

5. REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, vol. Series B 39, no. 1, pp. 1–38, 1977.

[3] H. Melin, J. W. Koolwaaij, J. Lindberg, and F. Bimbot, "A comparative evaluation of variance flooring techniques in HMM-based speaker verification," in *Proc. ICSLP*, Sydney, 1998, pp. 2379–2382.

[4] S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with applications in speech recognition," in *Proc. ICASSP*, 1998, pp. 645–648.

[5] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EuroSpeech*, 1997, vol. 1, pp. 99–102.

[6] T. Shinozaki, "HMM state clustering based on efficient cross-validation," in *Proc. ICASSP*, Toulouse, 2006, vol. I, pp. 1157–1160.

[7] S. Young *et al.*, *The HTK Book*, Cambridge University Engineering Department, 2005.

[8] X. Anguera, T. Shinozaki, C. Wooters, and J. Hernando, "Model complexity selection and cross-validation EM training for robust speaker diarization," in *ICASSP*, Hawaii, 2007, Submitted.

[9] M. Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. ICSLP*, 2006, pp. 1233–1236.

[10] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986, pp. 49–52.

[11] D. Povey and P.C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, vol. I, pp. 105–108.

[12] H. Soltau, B. Kingsbury, L. Mangu, D. Povery, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. ICASSP*, 2005, vol. I, pp. 205–208.

[13] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. Eurospeech*, 1995, pp. 1155–1158.

[14] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, pp. 2049–2073, 2005.