LARGE MARGIN HMMS FOR SPEECH RECOGNITION

Xinwei Li, Hui Jiang, Chaojun Liu

Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA
Email: {xwli, hj, cliu}@cs.yorku.ca

ABSTRACT

In this paper, motivated by large margin classifiers in machine learning, we propose a novel method to estimate continuous density hidden Markov model (CDHMM) in speech recognition according to the principle of maximizing the minimum multi-class separation margin. The approach is named as large margin HMM. Firstly, we will show that this type of large margin HMM estimation problem can be formulated as a standard constrained minimax optimization problem. Secondly, we propose an *iterative localized optimization* approach to perform the above minimax optimization for one model at each time to guarantee that the optimal value of the objective function always exists in the course of model parameter optimization. Then, we show that during each step the optimization can be solved by the GPD (generalized probabilistic descent) algorithm if we approximate the objective function by a differentiable function, such as summation of exponential functions. The large margin HMM-based classifiers are evaluated in a speaker-independent E-set speech recognition task by using the OGI ISOLET database. Experimental results show that the large margin HMMs can achieve significant word error rate (WER) reduction over the conventional HMM training methods, such as maximum likelihood estimation (MLE) and minimum classification error (MCE) training.

1. INTRODUCTION

Recently, the large margin methods have attracted a lot of research attentions in the field of machine learning. The fact that it is the margin in classification rather than the raw training error that matters has become a key tool in recent year when dealing with classifiers. The concept of large margin has been identified as a unifying principle for analyzing many different approaches in pattern classification, including Boosting, Mathematical Programming, Neural Networks and Support Vector Machine[10]. On the other hand, hidden Markov models (HMMs) have been successfully applied to many pattern classification tasks, ranging from automatic speech recognition to text document processing, and etc. In most cases, HMMs are usually estimated from training data based on maximum likelihood estimation (MLE) [5] or discriminative training (DT). The representative discriminative training approaches for HMMs include maximum mutual information estimation (MMIE) [9, 11] and minimum classification error (MCE) training [7, 6, 3]. However, most discriminative training approaches suffer the problem of poor generalization, as demonstrated in many speech recognition tasks[11, 3]. It is a very interesting topic how to build large margin HMM-based classifiers to improve generalization capability of HMMs in many practical classification problems. In [1, 2], the authors proposed the so-called Hidden Markov Support Vector machines (HMSVM) for label sequence learning problem. In HMSVM, discrete HMMs (DHMMs) are estimated based on the large margin principle. As shown in [1, 2], estimation of DHMMs for large margin turns out to be a quadratic programming problem under some constraints. The problem can be solved by many standard optimization software tools.

In this paper, motivated by some recent advances in machine learning about large margin classifiers, we propose to estimate HMMs discriminatively based on a new criterion, such as maximum separation margin, as in other large margin classifiers. Based on the theoretical results in machine learning, a large margin classifier implies a good generalization power and generally yields much lower generalization errors in new test data as shown in support vector machine and boosting method. As we know, Gaussian mixture continuous density HMM (CDHMM) is the most popular model for speech signals in speech recognition. In this paper, we will study how to estimate CDHMMs based on the above large margin principle for speech recognition. We propose a GPD-based localized optimization method to estimate the large margin HMMs iteratively. The large margin HMM-based classifiers are evaluated in a speaker-independent E-set speech recognition task by using the OGI ISOLET database. Experimental results show that the large margin HMMs can achieve significant word error rate (WER) reduction over the conventional HMM training methods, such as maximum likelihood estimation (MLE) and minimum classification error (MCE) training.

The remainder of this paper is organized as follows. First, in section 2 we will introduce the large margin training criterion. Next, in section 3 we will give our solution, namely *iterative localized optimization* for estimating large margin CDHMM parameters using the GPD algorithm. Experimental results will be presented in section 4. Finally a summary will be given in section 5.

2. LARGE MARGIN HMM

In ASR, given any speech utterance X, a speech recognizer will choose the word $\hat{W}$[1] as output based on the MAP decision rule as follows:

$$\begin{aligned}\hat{W} &= \arg\max_{W} \; p(W|X) = \arg\max_{W} \; p(W)\cdot p(X|W) \quad (1)\\ &= \arg\max_{W} p(W)\cdot p(X|\lambda_W) = \arg\max_{W} \mathcal{F}(X|\lambda_W)\end{aligned}$$

where λ_W denotes the HMM representing the word W and $\mathcal{F}(X|\lambda_W) = p(W)\cdot p(X|\lambda_W)$ is called discriminant function. In

[1]Depending on the problem of interest, a word W may be any linguistic unit, e.g., a phoneme, a syllable, a word, a phrase, a sentence, etc..

this work, we are only interested in HMM λ_W and assume $p(W)$ is fixed.

For a speech utterance X_i, assuming its true word identity as W_i^T, following [1, 2], the multi-class separation margin for X_i^T is similarly defined as:

$$d(X_i) = \mathcal{F}(X_i|\lambda_{W_i^T}) - \max_{W_j \in \Omega\ W_j \neq W_i^T} \mathcal{F}(X_i|\lambda_{W_j}) \quad (2)$$

$$= \min_{W_j \in \Omega\ W_j \neq W_i^T} \Big[\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j})\Big] \quad (3)$$

where Ω denotes the set of all possible words.

Obviously, if $d(X_i) \leq 0$, X_i will be incorrectly recognized by the current HMM set, denoted as Λ; if $d(X_i) > 0$, X_i will be correctly recognized by the models Λ.

Given a set of training data $\mathcal{D} = \{X_1, X_2, \cdots, X_N\}$, we usually know the true word identities for all utterances in $\mathcal{D}$, denoted as $\mathcal{L} = \{W_1^T, W_2^T, \cdots, W_N^T\}$. Thus, we can calculate the separation margin (or margin for short hereafter) for every utterance in $\mathcal{D}$ based on the definition in eq.(2) or (3). If we want to estimate the HMM parameters Λ, one desirable estimation criterion is to minimize the total number of utterances in whole training set which have negative margin as in the standard MCE estimation [6]. Furthermore, motivated by the large margin principle in machine learning, even for those utterances which all have positive margin, we may still want to maximize the minimum margin among them towards an HMM-based large margin classifier. Based on the machine learning theory, a large margin classifier usually leads to much lower generalization error rate in a new testing set and shows a more robust and better generalization capability. In this work, we will study how to estimate HMMs for speech recognition based on the above-mentioned principle of maximizing minimum multi-class separation margin.

First of all, from all utterances in $\mathcal{D}$, we need to identify a subset of utterances, $\mathcal{S}$, as:

$$\mathcal{S} = \{X_i \mid X_i \in \mathcal{D} \text{ and } 0 \leq d(X_i) \leq \gamma\} \quad (4)$$

where $\gamma > 0$ is a pre-set positive number. Analogically, we call $\mathcal{S}$ as *support vector set* and each utterance in $\mathcal{S}$ is called a support token which has relatively small positive margin among all utterances in training set $\mathcal{D}$. In other words, all utterances in $\mathcal{S}$ are relatively close to the classification boundary even though all of them locate in the right decision regions. To achieve a better generalization power, it is desirable to adjust decision boundaries, which are implicitly determined by all models, through optimizing HMM parameters Λ to make all support tokens as far from the decision boundaries as possible, which will result in a robust classifier with better generalization capability. This idea leads to estimating the HMM models Λ based on the criterion of maximizing the minimum margin of all support tokens, which is named as large margin estimation (LME) of HMM.

$$\tilde{\Lambda} = \arg\max_{\Lambda} \min_{X_i \in \mathcal{S}} d(X_i) \quad (5)$$

where the above maximization and minimization are performed subject to the constraints that $d(X_i) > 0$ for all $X_i \in \mathcal{S}$. The HMM models, $\tilde{\Lambda}$, estimated in this way, are called large margin HMMs.

Considering eq.(3), large margin HMMs can be equivalently estimated as follows:

$$\tilde{\Lambda} = \arg\max_{\Lambda} \max_{X_i \in \mathcal{S}} \min_{W_j \in \Omega\ j \neq i} \Big[\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j})\Big] \quad (6)$$

subject to

$$\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j}) > 0 \quad (7)$$

for all $X_i \in \mathcal{S}$ and $W_j \in \Omega\ j \neq i$.

Finally, the above optimization can be converted into a standard minimax optimization problem as:

$$\tilde{\Lambda} = \arg\min_{\Lambda} \max_{X_i \in \mathcal{S}} \max_{W_j \in \Omega\ j \neq i} \Big[\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i^T})\Big] \quad (8)$$

where the minimax optimization is subject to the following constraint:

$$\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i^T}) < 0 \quad (9)$$

for all $X_i \in \mathcal{S}$ and $W_j \in \Omega\ W_j \neq W_i^T$.

3. ITERATIVE LOCALIZED OPTIMIZATION OF LARGE MARGIN CDHMM

The constraints in eq.(9) can not guarantee the existence of the minimax point. As an illustration of this, let's assume a simple case with only two classes $m1$ and $m2$ and there is a support token X close to the decision boundary. If we pull $m1$ and $m2$ together at the same time, we can keep the boundary unchanged but increase the margin defined in eq.(3) as much as we want. As models move toward X, the absolute values of both $\mathcal{F}(X|m1)$ and $\mathcal{F}(X|m2)$ increase, so does the margin as well, although the relative position of X related to the boundary actually does not change at all.

More constraints must be introduced in the above minimax optimization procedure to make sure that the optimal point does exist. In this work, we adopt a localized optimization strategy to add more constraints in the above minimax optimization, which is named as the *Iterative Localized Optimization* method. In this method, instead of optimizing parameters of all models at the same time, we will only adjust one selected model in each step, and then the process iterates to update another model until the minimum margin is maximized. There are different approaches to solve this problem. Please refer to a companion paper [8], where we reformulate the large margin estimation to maximize the so-called *relative margin*, which is bounded by definition. In that case, any regular optimization approach can be used for updating all model parameters jointly.

Algorithm 1 Iterative Localized Optimization

repeat

1. Identify the support set $\mathcal{S}$ based on the current model set $\Lambda^{(n)}$.
2. Choose the support token, say X_k, from $\mathcal{S}$ which currently gives the minimum margin; Choose the true model of X_k, say $\lambda_k^{(n)}$ for optimization in this iteration.
3. Minimizing the margin by ONLY updating the model λ_k: $\lambda_k^{(n)} \Rightarrow \lambda_k^{(n+1)}$.
4. $n = n + 1$.

until some convergence conditions are met

In the above iterative localized optimization method, in each iteration, only one model, to say λ_k, is updated based on the minimax optimization given in eq.(8) so that we only need to consider those functions which are relevant to the currently selected model. The minimax optimization can be re-formulated as:

$$\lambda_k^{(n+1)} = \arg\min_{\lambda_k} \max_{X_i \in \mathcal{S}\ i\neq j\ j=k\ or\ i=k} \left[\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i^T})\right] \tag{10}$$

subject to the constraints in eq.(9).

This localized minimax optimization can be numerically solved by using some optimization software tools. Given a large number of parameters in CDHMMs, it usually is too slow to use a general-purpose minimax tool to solve this optimization problem.

In this work, we alternatively adopt a GPD-based algorithm [7] to solve the minimax problem in eq.(10) in an approximate way.

First of all, based on eq.(10), we construct a differentiable objective function as follows:

$$\begin{aligned} Q(\lambda_k) &= \frac{1}{\eta}\log\Big\{ \sum_{X_i\in\mathcal{S}\ j\neq i\ i=k\ or\ j=k} \exp[\eta\cdot\mathcal{F}(X_i|\lambda_{W_j}) \\ &\quad -\eta\cdot\mathcal{F}(X_i|\lambda_{W_i})]\Big\} \end{aligned} \tag{11}$$

where $\eta > 1$ is a constant . As $\eta \to \infty$, $Q(\lambda_k)$ will approach the maximization in eq.(10). Then, the GPD algorithm can be used to update the model parameters, λ_k, in order to minimize the above approximate objective function, $Q(\lambda_k)$.

Assume each speech unit, e.g., a word W, is modeled by an N-state CDHMM with parameter vector $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, $A = \{a_{ij}|1 \le i, j \le N\}$ is transition matrix, and θ is parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\cdots,K}$ for each state i, where K denotes number of Gaussian mixtures in each state. The state observation p.d.f. is assumed to be a mixture of multivariate Gaussian distribution. In many cases, we prefer to use multivariate Gaussian distribution with diagonal precision matrix. Given any speech utterance $X_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{iR}\}$, $\mathcal{F}(X_i|\lambda_{W_j})$ can be calculated as:

$$\begin{aligned} \mathcal{F}(X_i|\lambda_{W_j}) &= \log(p(X_i|\lambda_{W_j})p(W_j)) \\ &\approx \log p(W_j) + \log\pi_{s_1^*} + \sum_{t=2}^{R}\log a_{s_{t-1}^* s_t^*} + \prod_{t=1}^{R}\log\omega_{s_1^* l_1^*} \\ &\quad + \frac{1}{2}\sum_{t=1}^{R}\sum_{d=1}^{D}\left[\log r_{s_t^* l_t^* d} - r_{s_t^* l_t^* d}(\mathbf{x}_{itd} - m_{s_t^* l_t^* d})^2\right] \end{aligned} \tag{12}$$

Here we only consider a simple case, where we only re-estimate mean vectors of CDHMMs based on the large margin principle while keeping all other CDHMM parameters constant during the large margin estimation. For any utterance X_i in the support token set $\mathcal{S}$, we can re-write $\mathcal{F}(X_i|\lambda_i)$ and $\mathcal{F}(X_i|\lambda_j)$ according to eq.(12) as follows:

$$\mathcal{F}(X_i|\lambda_i) \approx C' - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D} r_{s_t' l_t' d}(x_{itd} - m_{s_t' l_t' d})^2 \tag{13}$$

$$\mathcal{F}(X_i|\lambda_j) \approx C^{"} - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D} r_{s_t^{"} l_t^{"} d}(x_{itd} - m_{s_t^{"} l_t^{"} d})^2 \tag{14}$$

where C' and $C^{"}$ are two constants independent from mean vectors. In this case, the discriminant functions $\mathcal{F}(X_i|\lambda_i)$ and $\mathcal{F}(X_i|\lambda_j)$ can be represented as a summation of some quadratic functions related to mean values of CDHMMs. Then we can represent the decision margin $\mathcal{F}(X_i|\lambda_i) - \mathcal{F}(X_i|\lambda_j)$ as:

$$\begin{aligned} \mathcal{F}(X_i|\lambda_i) - \mathcal{F}(X_i|\lambda_j) &\approx C - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}\Big[r_{s_t' l_t' d}(x_{itd} - m_{s_t' l_t' d})^2 \\ &\quad - r_{s_t^{"} l_t^{"} d}(x_{itd} - m_{s_t^{"} l_t^{"} d})^2\Big] \end{aligned} \tag{15}$$

where $C = C' - C^{"}$.

From eqs.(11) and (15), it is straightforward to calculate the gradient of the objective function, $Q(\lambda_k)$, with respect to each mean vector in the model λ_k.

At last, we can use the GPD algorithm to adjust λ_k to minimize the objective function $Q(\lambda_k)$ as follows:

$$\mu_{sql}^{(n+1)} = \mu_{sql}^{(n)} - \epsilon\frac{\partial Q(\lambda_k)}{\partial\mu_{sql}}\bigg|_{\lambda_k=\lambda_k^{(n)}} \tag{16}$$

$$\tag{17}$$

where $\mu_{sql}^{(n+1)}$ denotes the l-th dimension of Gaussian mean vector for the q-th mixture component of state s of HMM model λ_k at $(n+1)$-th iteration.

4. EXPERIMENTAL RESULTS

The LME algorithm described above is tested on the English E-set vocabulary of OGI ISOLET database, consisting of {B, C, D, E, G, P, T, V, Z}. ISOLET is a database of letters of the English alphabet spoken in isolation. The database consists of 7800 spoken letters, two productions of each letter by 150 speakers, 75 male and 75 female. The recordings were done under quiet, laboratory conditions with a noise-canceling microphone. The data were sampled at 16 kHz with 16-bit quantization. ISOLET is divided into five parts named ISOLET 1-5. In our experiment, only the first production of each letter in ISOLET 1-4 is used as training data (1080 utterances). All data in ISOLET 5 is used as testing data (540 utterances). The feature vector is of 39 dimensions, which include 12-d static MFCC, log-energy, delta and acceleration coefficients. An HMM recognizer with 16-state whole-word based models is trained based on different training criterion. Here CDHMMs with1-mixture per state and 2-mixture per state are experimented. Only means will be updated in the experiment. We always use the best MCE models as the initial models in the large margin estimation.

Tables 1 gives a performance comparison of the best results obtained by different training criteria.

Table 1. Results (word accuracy %) of different training criteria on E-set test data.

	1-mixture	2-mixture
ML	85.56	90.56
MCE	91.48	94.07
LME	92.78	95.19

It is clearly demonstrated that LME achieves the best results. For example, the LME-trained models with 2-mixture per state

achieve the word accuracy of 95.19%, which indicates 18.9% errors reduction over the corresponding MCE-trained models, which get 94.07% in accuracy.

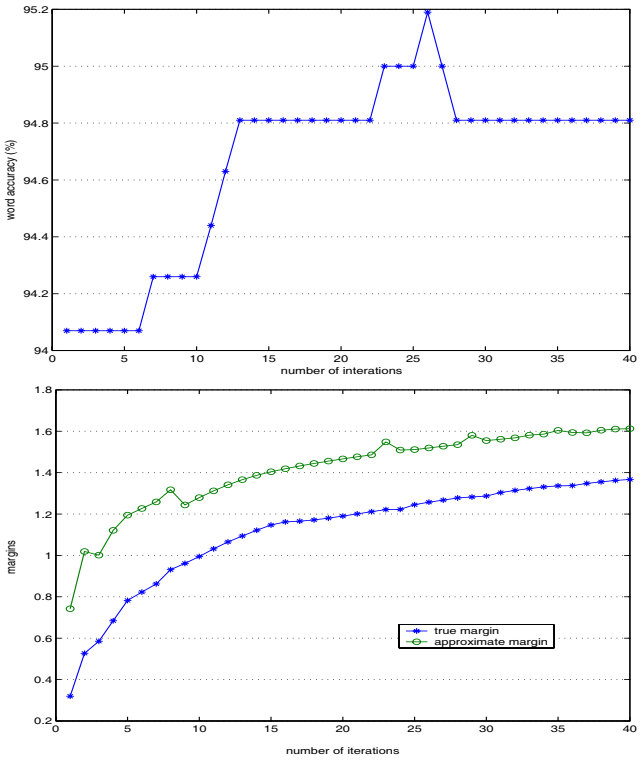

Fig. 1. Curves for LME training of a 2-mixture model on E-set task. Top figure is word accuracy of LME model on testing set The middle figure includes the curve of approximate margin which was being maximized during LME-training and the curve of corresponding true margins.

Figure 1 plots the recognition accuracy, approximate margin, $-Q(\lambda)$ (where $Q(\lambda)$ is given in eq.(11)), and true margin on the testing set , $d(X_i)$ as given in eq.(3) as a function of the number of iterations of the LME training procedure. We can see from the curves that with the number of iterations going up, the approximate margin keeps increasing, which is consistent with the goal of GPD optimization. Meanwhile the recognition accuracy on the testing set keeps increasing (or unchanged for a short period) before it reaches the best point. At 26 iterations, the LME model achieves 95.19% accuracy on the testing set, representing a 18.9% reduction in recognition error. Also we can see that the true margin keeps increasing accordingly as the the objective function increases. It can be proved that the objective function, i.e., the approximate margin, is an upper-bound of the true margin.

The current LME training algorithm only estimates models based on the support vector set, which consists of only correctly recognized tokens in the training set. In case there is any recognition error in the training set, a different algorithm, which is similar to MCE formulation, was proposed in [4] to handle the set of error tokens. However, in the work reported here, there is no recognition error in the training set, so we are not concerned with this issue.

5. SUMMARY

In this paper, we have proposed a new training method, large margin HMM (LME), for continuous HMM based speech recognition. The LME approach aims at improving the poor generalization capability of existing discriminative training algorithms. Motivated by large margin classifier in machine learning, the new training criterion is trying to maximize the minimum multi-class seperation margin. We investigated its performance on a speaker-independent isolated-word task. The LME method provides up to 18.9% reduction in error rate, compared to the popular MCE method. Further research and experiments on continuous speech recognition and sub-word based system are underway, which will be reported in the future.

6. REFERENCES

[1] Y. Altun and T. Hfomann, "Large margin methods for label sequence learning," *Proc. of Eurospeech 2003*, pp.993–996, Geneva, Switzerland, Sep. 2003.

[2] Y. Altun, I. Tsochantaridis and T. Hofmann, " Hidden Markov Support Vector Machines," *Proc. of the 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

[3] H. Jiang, O. Siohan, F. Soong and C.-H. Lee, "A dynamic in-search discriminative raining approach for large vocabulary speech recognition," Proc. of 2002 IEEE international Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002), pp.I-113-116, Orlando, Florida, May 2002.

[4] H. Jiang, "Discriminative Training for Large Margin HMMs", *Technical Report CS-2004-01, CSE Department, York University*, March 2004.
(http://www.cs.yorku.ca/techreports/2004/CS-2004-01.html)

[5] B.-H. Juang, S. E. Levinson and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. on Information Theory*, Vol. IT-32, No. 2, pp.307-309, 1986.

[6] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, pp.257-265, Vol.5, No.3, May 1997.

[7] S. Katagiri, B.-H. Juang and C.-H. Lee, "Pattern recognition using a generalized probabilistic descent method," *Proceedings of the IEEE*, Vol. 86, No. 11, pp.2345-2373, Nov. 1998.

[8] C.-J. Liu, H. Jiang, X.-W. Li, "Discriminative Training of CDHMMs for Maximum Relative Separation Margin", *submitted to ICASSP 2005*, Mar. 2005.

[9] Y. Normandin, R. Cardin and R. Demori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, Apr. 1994.

[10] A. J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans (ed.), *Advances in Large Margin Classifiers*, The MIT Press.

[11] P.C. Woodland and D. Povey, "Large Scale Discriminative Training of hidden Markov models for speech recognition," *Computer Speech & Language*, pp.25-47, Vol. 16, No. 1, January 2002.