Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 1285–1294
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A new look at discriminative training for hidden Markov models

Xiaodong He *, Li Deng *

Microsoft Research, One Microsoft Way, Redmond, WA 98052, United States

Available online 12 April 2007

Abstract

Discriminative training for hidden Markov models (HMMs) has been a central theme in speech recognition research for many years. One most popular technique is minimum classification error (MCE) training, with the objective function closely related to the empirical error rate and with the optimization method based traditionally on gradient descent. In this paper, we provide a new look at the MCE technique in two ways. First, we develop a non-trivial framework in which the MCE objective function is re-formulated as a rational function for multiple sentence-level training tokens. Second, using this novel re-formulation, we develop a new optimization method for discriminatively estimating HMM parameters based on growth transformation or extended Baum–Welch algorithm. Technical details are given for the use of lattices as a rich representation of competing candidates for the MCE training.

Keywords: Hidden Markov model; Discriminative learning; Minimum classification error; Extended Baum–Welch algorithm; Growth transformation

1. Introduction

Hidden Markov models (HMMs) have been a well established framework for a variety of pattern recognition applications, including, most prominently, speech recognition applications (Rabiner and Juang, 1993; Bahl et al., 1987; Deng and O'Shaughnessy, 2003). One most attractive feature of the HMM framework is that its parameters can be learned automatically from the training data. In early days of HMMs, the parameters were learned by the maximum likelihood (ML) criterion based on the EM algorithm (e.g., Bahl et al., 1987; Rabiner and Juang, 1993). Improvement of parameter learning beyond ML has been pursued for many years (Brown, 1987; Chou, 2003; Deng et al., 2005a,b; Gopalakrishnan et al., 1991; He and Chou, 2003; Juang and Katagiri, 1992; Juang et al., 1997; Macherey et al., 2005; McDermott et al., 2007; Normandin, 1991; Povey and Woodland, 2002; Povey et al., 2003, 2004; Povey, 2004; Rathinavalu and Deng, 1998; Schluter et al., 2001), based on the concept of discrimination against classes, in contrast to maximizing likelihood of each individual class. The reason behind discriminative training is that complete knowledge of speech data distributions is lacking and training data is always limited. It is not until recently that discriminative training has shown uniform success in speech recognition over virtually all tasks, including especially large tasks (e.g., Woodland and Povey, 2000; Povey, 2004).

Among several types of discriminative training for HMMs, one prominent type is minimum classification error (MCE) training (Chou, 2003; Juang and Katagiri, 1992; Juang et al., 1997; He and Chou, 2003; Macherey et al., 2005; McDermott et al., 2007; Roux and McDermott, 2005; Rathinavalu and Deng, 1998). The essence of MCE is to define the objective function for optimization that is closely related to the empirical classification errors. This is more desirable than other types of discriminative training that are less closely related to the classification errors. The conventional MCE has been based on the sequential gradient-descent based technique, named generalized probabilistic descent (GPD), which optimizes the objective function as a highly complex function of the HMM parameters.

* Corresponding authors. Tel.: +1 425 706 4939 (X. He); +1 425 706 2719 (L. Deng); fax: +1 425 936 7329.
E-mail addresses: xiaohe@microsoft.com (X. He), deng@microsoft.com (L. Deng).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.11.022

Another significant advance in discriminative training is the development and application of a special type of optimization technique, called growth transformation (GT) or extended Baum–Welch (EBW) algorithm when it is used for HMM parameter estimation. GT is an iterative optimization scheme where if the parameter set Λ is subject to a transformation $\Lambda = T(\Lambda')$, then the objective function "grows" in its value $O(\Lambda) > O(\Lambda')$ unless $\Lambda = \Lambda'$. In (Gopalakrishnan et al., 1991), GT/EBW was developed for rational functions such as the mutual information as the optimization criterion. Maximization of mutual information (MMI) as a form of discriminative criterion for the discrete HMM was described in (Gopalakrishnan et al., 1991). This has been extended to the continuous-density HMM in (Normandin, 1991; Gunawardana and Byrne, 2001). The significance of GT/EBW lies in its effectiveness and closed-form parameter updating for large-scale optimization problems with difficult objective functions. Compared with the gradient based techniques which often require special and delicate care for tuning the parameter-dependent learning rate, GT/EBW mitigates such requirements and with the closed-form updating formula it is generally faster in reaching algorithm convergence.

Mutual information is naturally in the form of a rational function and MMI is obviously suited to GT/EBW optimization. However, as a discriminative criterion, it is only indirectly related to classification errors. On the other hand, MCE as a discriminative criterion is closely related to classification errors, but it is not naturally in the form of a rational function when there are multiple utterance tokens in the training data. Hence, it has been a tradition to use the gradient-descent techniques (GPD) for optimizing the MCE criterion (Chou, 2003; Juang et al., 1997; McDermott et al., 2007; Rathinavalu and Deng, 1998). In this paper, we break this long-held tradition and take a fresh look at the MCE. This new analysis and formulation of the MCE covers two main issues. First we re-examine the MCE criterion. Second the results of the re-examination permit the use of the new GT/EBW optimization technique for optimizing the MCE criterion with respect to the HMM parameters.

The organization of this paper is as follows. In Section 2, an overview of the traditional MCE is provided. Then, in Section 3, we reformulate the MCE criterion (with multiple training tokens) into a rational functional form. We provide a rigorous proof by induction for the correctness of the rational functional form. Given this non-trivial reformulation, in Section 4, we present in detail a novel GT/EBW based optimization technique for estimating the parameters of the Gaussian HMMs. In Section 5, the lattice-based MCE training is described, and a summary is given in Section 6.

2. Overview of minimum classification error (MCE) training

We denote by Λ the parameter set of the generative model expressed in terms of a joint statistical distribution

$$p_\Lambda(X, S) = p_\Lambda(X|S)P(S), \tag{1}$$

on the observation training data sequence X and on the corresponding label sequence S, where we assume the parameters in the "language model" $P(S)$ are not subject to optimization. We use $r = 1, \ldots, R$ as the index for "token" (e.g., a single sentence or utterance) in the training data, and each token consists of a "string" of an observation data sequence: $X_r = x_{r,1}, \ldots, x_{r,Tr}$, with the corresponding label (e.g., word) sequence: $S_r = w_{r,1}, \ldots, w_{r,Nr}$. That is, S_r denotes correct label sequence for token r. Further, we use s_r to denote all possible label sequences for the rth token, including the correct label sequence S_r and all other incorrect label sequences.

MCE learning was originally introduced for multiple-category classification problems where the smoothed error rate is minimized for isolated "tokens" (Juang and Katagiri, 1992). It was later generalized to minimize the smoothed "sentence token" or string-level error rate (Juang et al., 1997; Chou, 2003), which is known as "embedded MCE".

The MCE objective function is defined first based on a set of class discriminant functions and a special type of loss function. Then the model is estimated to minimize the expected loss that is closely related to the recognition error rate of the classifier.

In embedded MCE training, a set of discriminant functions is first defined based on the correct string S_r and the N most confusable competing strings, $s_{r,1}, \ldots, s_{r,N}$. Define the top N best competing strings as

$$s_{r,1} = \arg\max_{s_r : s_r \neq S_r} \{\log p_\Lambda(X_r, s_r)\},$$
$$s_{r,i} = \arg\max_{s_r : s_r \neq S_r, s_r \neq s_{r,1}, \ldots, s_{r,i-1}} \{\log p_\Lambda(X_r, s_r)\} \quad i = 2, \ldots, N.$$

Then, the discriminant functions for the correct string and the N competing strings take the form of

$$g_{s_r}(X_r; \Lambda) = \log p_\Lambda(X_r, s_r), \quad s_r \in \{S_r, s_{r,1}, \ldots, s_{r,N}\}.$$

And the decision rule for the recognizer or classifier is the one that for the observation data sequence, X_r,

$$C(X_r) = s_r^* \quad \text{if } s_r^* = \arg\max_{s_r} g_{s_r}(X_r; \Lambda).$$

Next, a misclassification measure in MCE is defined. For the general N-best MCE training, the following misclassification measure has been widely used (Juang et al., 1997):

$$d_r(X_r, \Lambda) = -\log p_\Lambda(X_r, S_r) + \log \left\{ \frac{1}{N} \sum_{s_r, s_r \neq S_r} \exp\left[\eta \log p_\Lambda(X_r, s_r)\right] \right\}^{\frac{1}{\eta}}. \tag{2}$$

This misclassification measure function emulates the decision rule, i.e., $d_r(X_r, \Lambda) \geqslant 0$ implies misclassification and $d_r(X_r, \Lambda) < 0$ implies a correct classification. The second term in (2) is a soft-max function, which counts the scores of all N competitive candidates. It can be looked