Pattern Recognition Letters 30 (2009) 1301–1309

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A stochastic version of Expectation Maximization algorithm for better estimation of Hidden Markov Model

Shamsul Huda [a,*], John Yearwood [a], Roberto Togneri [b]

[a] Center for Informatics and Applied Optimization, School of Information Technology and Mathematical Science (ITMS), University of Ballarat, VIC 3353, Victoria, Australia
[b] Center for Intelligent Information Processing, School of Electrical, Electronic and Computer Engineering, University of Western Australia, WA, Australia

ARTICLE INFO

Article history:
Received 28 November 2008
Received in revised form 18 May 2009
Available online 23 June 2009

Communicated by R.C. Guido

Keywords:
Hidden Markov Model
Expectation Maximization
Speech recognition
Constraint-based Evolutionary Algorithm
Stochastic EM

ABSTRACT

This paper attempts to overcome the local convergence problem of the Expectation Maximization (EM) based training of the Hidden Markov Model (HMM) in speech recognition. We propose a hybrid algorithm, Simulated Annealing Stochastic version of EM (SASEM), combining Simulated Annealing with EM that reformulates the HMM estimation process using a stochastic step between the EM steps and the SA. The stochastic processes of SASEM inside EM can prevent EM from converging to a local maximum and find improved estimation for HMM using the global convergence properties of SA. Experiments on the TIMIT speech corpus show that SASEM obtains higher recognition accuracies than the EM.

1. Introduction

The Hidden Markov Model (HMM) is the most successful statistical modeling technique widely used for signal classification, Automatic Speech Recognition (ASR) (Rabiner, 1989; Levinson et al., 1983) and time series classification. This is because HMM has a powerful ability to model the temporal nature of signals statistically as well as has the ability to represent arbitrarily complex probability density functions of the underlying systems. In a Bayesian classification scenario (for signal classification or recognition), the HMM provides the posterior probability of the signal given its class label/phoneme label (where a phoneme is a basic theoretical unit of speech sound for speech signal) in signal classification. Therefore, success of the recognition/classification of a signal depends heavily on how precisely the estimated HMM can represent the underlying phoneme/signal classes in the training data.

The standard method of estimating the parameters of HMM is the Expectation Maximization (EM) (Rabiner, 1989; Levinson et al., 1983; Dempster et al., 1977) algorithm. The EM (Rabiner, 1989; Levinson et al., 1983; Dempster et al., 1977) algorithm is attractive and used for estimation of the HMM as well as for estimation of many other probabilistic models such as the Finite Mixture Models (FMM) McLachlan and Basford, 1988 and the Gaussian Mixture Models (GMM) McLachlan and Basford, 1988 because EM is computationally efficient and can approximate the underlying distribution from the set of observed data which has missing or hidden components (Bilmes, 1998). Unfortunately, the estimation of HMM parameters computed by the EM approach is not always the best (Rabiner, 1989; Levinson et al., 1983). The reason is that the EM algorithm is strongly dependent on the selection of the initial values of model parameters and increases the values for likelihood function at each iteration which guarantees to produce a local rather than a global maximum of the likelihood function (Rabiner, 1989; Levinson et al., 1983; Wu, 1983). This gives a non-optimized estimation of the parameters of HMM and consequently lowers the recognition accuracy in ASR systems.

To circumvent the local convergence problem of EM, recently several investigators have applied hybrid algorithms using Evolutionary Algorithm (EA) in combination with EM for optimal estimation of a Gaussian Mixture Model (GMM) in a non-linear classification problem and unsupervised clustering. These hybrid algorithms in (Martinez and Vitria, 2000; Pernkopf et al., 2005; Martinez and Vitria, 2001; Majdi-Nasab et al., 2006) ignore the constraints of the GMM and assume equal mixture weights which may fail in many practical situations where the mixture weights of individual mixtures of GMM are not the same. Therefore, these algorithms (Martinez and Vitria, 2000; Pernkopf et al., 2005; Martinez and Vitria, 2001; Majdi-Nasab et al., 2006) cannot be

* Corresponding author. Tel.: +61 0431271690; fax: +61 3 5327 9289.
E-mail addresses: shuda9203@yahoo.com, shuda@ballarat.edu.au (S. Huda), j.yearwood@ballarat.edu.au (J. Yearwood), roberto@ee.uwa.edu.au (R. Togneri).

0167-8655/$ - see front matter
doi:10.1016/j.patrec.2009.06.006

applied on the constraint based models (like HMM). HMM (with continuous density, CDHMM) combines several GMMs into a single model that contains a large number of parameters and mixture constraints aggregated from the GMMs. The HMM also has transition and observation probability constraints for each state. These constraints must be satisfied while estimating the HMM parameters. Therefore, the algorithms proposed in (Martinez and Vitria, 2000; Pernkopf et al., 2005; Martinez and Vitria, 2001; Majdi-Nasab et al., 2006) cannot be applied on the HMM. In (Paul, 1985), a Simulated Annealing (SA) technique for training of Discrete Hidden Markov Model (DHMM) was proposed using a simple random-perturbation neighborhood-operator of SA. All DHMM parameters have values in the range of (0–1). DHMM does not need to satisfy the HMM state based constraints such as the mixture constraint of CDHMM. However the approach in (Paul, 1985) will not work for training of CDHMM. This is because CDHMM parameters can have values with a very high range (negative or positive). In a constraint based search space with high range for domain-values such as CDHMM, the simple random-perturbation neighborhood-operator of the SA technique in Paul (1985) is not feasible. In (Andrieu and Doucet, 2000), a Simulated Annealing algorithm based on data augmentation and stochastic simulation of hidden Markov chain has been proposed for HMM training. However, the algorithm in Andrieu and Doucet (2000) was not tested for accuracy-levels on any practical data set/rel-life systems (e.g. ASR systems).

In our earlier research work (Huda et al., 2009), a Constraint based Evolutionary Learning approach to EM (CEL-EM) is proposed for training of HMM where a constraint based Evolutionary Algorithm (EA) is hybridized with standard EM. Several constraint based versions of CEL-EM including a Preserving Feasibility of Solutions and Penalty Function (PFS-PF) based CEL-EM (similar to standard constraint handling technique of EA (Michalewicz and Schoenauer, 1996)) have also been proposed to maintain the constraints of HMM during the training of HMM. In this earlier work (Huda et al., 2009), it is seen that hybrids of a constraint based EA and EM in the CEL-EM (Huda et al., 2009) achieve a better estimation for HMM and higher recognition accuracies in ASR systems. However, the hybrids of EA and EM in the CEL-EM (Huda et al., 2009) require high computational time. The population-based approach CEL-EM obtains improved recognition accuracy by sacrificing computational cost.

To overcome the EM problem, Celeux and Diebolt (1985) proposed a Stochastic version of EM (SEM) Celeux and Diebolt (1985) and formulated SEM for the Finite Mixture Model (FMM). However SEM shows erratic behavior if there is not sufficient training data (Celeux and Diebolt, 1992). Celeux and Diebolt (1992) also proposed a modified version of SEM known as Stochastic Approximation version of EM (SAEM) (Celeux and Diebolt, 1992) which uses SEM at the beginning iterations of the estimation process and EM is used when termination is closer. However both SEM (Celeux and Diebolt, 1985) and SAEM (Celeux and Diebolt, 1992) were formulated and tested for Finite Mixture Models (FMM). An HMM has two stochastic processes which is different from FMM. In the HMM, one stochastic process combines several mixture models and the other process incorporates the state transition distribution. Therefore, the formulation of SEM (Celeux and Diebolt, 1985) and SAEM (Celeux and Diebolt, 1992) proposed by Celeux and Diebolt (1985) and Celeux and Diebolt (1992) cannot be applied for estimation of the parameters of HMM.

In this paper, we therefore propose a single candidate model-based hybrid algorithm, the Simulated Annealing Stochastic version of EM (SASEM) that hybridizes the Simulated Annealing (SA) and the EM. SASEM provides a stochastic reformulation for the HMM estimation process by incorporating a stochastic step between the EM steps where the random process inside the EM steps is controlled using the Simulated Annealing (SA) (Kirkpatrick et al., 1983) technique. Our training approach (SASEM) is novel in the following ways:

(1) SASEM is a single candidate model-based approach that hybridizes SA and EM to avoid the local maximum problem of EM.
(2) SASEM provides a stochastic reformulation of HMM estimation process and computes better estimation for the HMM by using the global convergence properties of SA.
(3) SASEM is more computationally efficient training method for HMM than other population-based hybrid approaches (Huda et al., 2009; Martinez and Vitria, 2000; Pernkopf et al., 2005; Martinez and Vitria, 2001; Majdi-Nasab et al., 2006). Because SASEM is a single candidate model-based approach.

In our proposed hybrid algorithm, SASEM, overcomes the local maximum problem of HMM training and provides better estimation for HMM by using the Simulated Annealing (SA) (Kirkpatrick et al., 1983) technique and introducing stochastic steps in the EM steps. SASEM reformulates the HMM estimation process by implementing a stochastic step between the EM steps. The stochastic reformulation of the HMM in the SASEM is used to generate the neighborhood points for the SA where SA provides better control over the acceptance of stochastic step and EM steps. However, in the SASEM, the random-perturbation of the stochastic step between the EM steps prevents EM converging to a local maximum point and finds a better maximum of the likelihood function as well as a better estimation for the HMM using the global convergence properties of SA.

SASEM has been tested using the TIMIT (Garofolo et al., 1993) speech corpus and compared to standard EM as well as our population-based hybrid approach CEL-EM (Huda et al., 2009). Results show that SASEM achieves higher recognition accuracies than the EM. In some cases, SASEM obtains higher recognition accuracies than our population-based hybrid approach, PFS-PF based version of CEL-EM (Huda et al., 2009). Moreover SASEM takes less computational time than our population-based approach CEL-EM (Huda et al., 2009).

The remainder of this paper is organized as follows. In the next section, a brief description of the HMM, its parameters and the constraints are discussed. Section 3 briefly describes the EM algorithm and its problems in estimating the HMM parameters. Section 4 explains the formulation of Stochastic EM (SEM) and Stochastic Approximation EM (SASEM) for the Finite Mixture Model (FMM) in brief. The proposed Simulated Annealing Stochastic version of EM (SASEM) algorithm for HMM training is described in Section 5. The experimental procedure and results are presented in Section 6. The significance of the results is analyzed in Section 7. Conclusions of this study are given in the last section.

2. Hidden Markov Model (HMM) for signal classification/recognition, its parameters and constraints

The HMM is a doubly stochastic process. One stochastic process comprises the distribution of observations at each state which is a multimodal Gaussian mixture for a Continuous Density HMM (CDHMM). The other stochastic process involves the transitions between the HMM states which are the transition probabilities, $A = a_{ij}$. HMM parameters can be represented as $\{\lambda = c_{jn}, \mu_{jn}, \Sigma_{jn}, a_{ij}\}$ where $i,j = 1,2,\ldots,K$ (K = total number of states), $n = 1,2\ldots,M$ (M = total number of mixtures) and (c_{jn} = mixture weight, μ_{jn} = mean vector, Σ_{jn} = co-variance for jth state, nth mixture). The probability for the tth observation at the jth state is considered as $b_j(O_t) = \sum_{n=1}^{M} c_{jn} b_j^n(O_t)$ where