REVIEW ON A MODIFIED BAUM-WELCH ALGORITHM FOR HIDDEN MARKOV MODELS WITH MULTIPLE OBSERVATION SPACES

Naveen Parihar

Mississippi State University
Mississippi State, MS 39762 USA
email: parihar@isip.mstate.edu

ABSTRACT

Usually, for estimating the parameters of a hidden Markov model (HMM), the Baum-Welch algorithm relies on a high dimensional common feature set. The papers proposes an algorithm based on the Baum-Welch algorithm for estimating parameters of a hidden Markov model(HMM) that relies on a low dimensional specific set of features for each state. Each feature set is chosen specifically for that state to be a sufficient statistic for the discrimination of the given state from a common “white-noise” state. The parameter set of each state must include the common-state as a special case. A simulated data example is provided showing that for a given training data set, the performance (State Classification Error probability) is superior over the conventional hidden Markov model. The paper has a very sound theoretical aspect which is very well supported by the simulated example. In this paper, we will go over all the theoretical aspects of the algorithm as well as the simulation example provided[1].

1. INTRODUCTION

We know the conventional hidden Markov model can model a process. Let there be N number of states from S_1 to S_N. If the observed sequences of data is denoted by $x[t]$, for time steps t=1, 2,...,T, then the following parameters completely describe the hidden Markov model. (1) the initial state prior probabilities μ_j, (2) the state transition matrix $A = a_{ij}$, (3) the observation densities at each state $B_j(x)$, where both i and j vary from 1 to N. A very nice way of estimating $\{\mu_j, A, B_j(x)\}$, for a model with N states is to train it iteratively using the Baum-Welch algorithm[2][3].

Usually the $x[t]$ is in high dimensions and because of computational and memory constrains, the high dimensional raw data is reduced to a low dimensional feature set $z[t] = T[x[t]]$. This process of reduction in dimensions of raw data is the feature extraction. Now, the new hidden markov Model is defined by the same (1) prior probabilities μ_j, (2) the state transition matrix $A = a_{ij}$ but a different low-dimensional observation densities $B_j(z)$. Speech processing commonly employs this technique where the observations are set of cepstral coefficients. Since the dimensions of $z[t]$ are low, different PDF estimation method such as Gaussian Mixtures are employed for estimating the state observational probabilities. It has been observed that the it is very difficult to estimate the PDF’s non parametrically above five dimensions and it is impossible above 20 [6] unless the features are exceptionally well-behaved (are close to independent or multivariate Gaussian). It is common for high-dimensional PDF estimators to give very good results as classifiers in many applications because of the fact that in high-dimensional space the data is inherently separable and any PDF estimator may do as good as another[4]. Only 5 to 10 features cannot contain all the information that is needed for speech recognition. So, the dimensionality reduction has been a major field of research. Some of the various approaches are feature selection (either use a smaller and insufficient features set or use more features and suffer PDF estimation errors), projection pursuits and subspace analysis. All these methods involve assumptions and approximations that do not hold in general. Feature selection assumes that most of the information for discriminating all data classes is contained in a few set of features. Projection-based methods assume that the information is linearly separable. The method proposed in the paper is completely general based on the concept of “sufficient statistic” and “class specific classifier”[5]. For a given observation sequence $x[t]$, for time steps t=1, 2,...,T, all the parameters for the hidden Markov model are estimated using the classical Baum-Welch algorithm except for the state likelihood function $b_j(x)$. Consider a common-state S_0 for which the raw data is a pure iid Gaussian noise. Thus, we have a likelihood ratio

$$\frac{b_j(x)}{b_0(x)} = \frac{b_j(z_j)}{b_0(z_j)}, 1 \le j \le N \tag{1}$$

The left hand side of (1) is the likelihood function of a conventional HMM scaled by a factor $b_0(x)$.The right hand side of (1) gives the testing criteria for the sufficiency of z_j for state S_j vs. state S_0. The dimensions at each state j are sufficient to discriminate it from the common state S_0. Clearly dimensions of each z_j ‘s are lower than z. For this sufficiency of the features, we require to know the PDF of x, that is what statistic best distinguishes the State S_j from the common-state S_0. This prior information is not completely known in the many real-world applications and hence the sufficiency of features can never be established theoretically. This is the same problem while selecting the features and the sufficiency is approximated. In the class-specific specifier[4}, each sub-set of states represent each class and has a different set of statistic for the class of data that it represents. The densities of each $b_j(z_j)$ at each state can be estimated using fewer number of training samples than the conventional HMM since the densities for less number of features is to be estimated at each state. Hence, we now conclude that the method proposed in the paper is theoretically very sound that can be applied to the applications like time-series (Speech) Analysis and Image recognition[5].

2. THE ALGORITHM

The modified Baum-Welch algorithm to estimate the parameters of Class-Specific HMM using the low-dimensional features is provided here[1].

2.1. The Class-Specific HMM

2.1.1. The class-specific forward procedure

1. Initialization:

$$\alpha_1(j) = u_j \frac{b_j(z[1])}{b_0(z_j[1])}, (1 \le j \le N) \tag{2}$$

2. Induction:

$for(1 \le t \le T-1, 1 \le j \le N)$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] \frac{b_j(z_j)[t+1]}{b_0(z_j)[t+1]} \tag{3}$$

3. Termination:

$$\frac{p(x([1], \dots, x[T])|\Lambda)}{p(x([1], \dots, x[T])|H_0)} = \sum_{i=1}^{N} \alpha_T(i) \tag{4}$$

where H_0 is the condition that state S_0 is true at every t.

2.1.2. The class-specific backward procedure

1. Initialization:

$$\beta_T(i) = 1 \tag{5}$$

2. Induction:

$for(t= T-1, \dots, 1, 1 \le i \le N)$

$$\beta_t(j) = \sum_{i=1}^{N} a_{ji} \frac{b_i(z_i)[t+1]}{b_0(z_i)[t+1]} \beta_{t+1}(i) \tag{6}$$

2.1.3. HMM Reestimation formulas

Define $\gamma_t(j)$ as $p(\theta_t = j[x[1], \dots, x[T]])$,

We have,

$$\gamma_t(j) = (\alpha_t(j)\beta_t(j)) / \left(\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)\right) \tag{7}$$

Let

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}\frac{b_j(z_j)[t+1]}{b_0(z_j)[t+1]}\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{m=1}^{N}\alpha_t(i)a_{im}\frac{b_m(z_m)[t+1]}{b_0(z_m)[t+1]}\beta_{t+1}(m)} \quad (8)$$

The updated state priors are $\hat{\mu}_i = \gamma_1(i)$. The updated state transition matrix is

$$\hat{a}_{ij} = \left(\sum_{t=1}^{T-1}\xi_t(i,j)\right) \Big/ \left(\sum_{t=1}^{T-1}\gamma_t(i)\right) \quad (9)$$

2.1.4. Gaussian mixture reestimation formulas

We assume the following Gaussian mixture representation for each $b_j(z_j)$:

$$b_j(z_j) = \sum_{k=1}^{M} c_{jk}N(z_j, \mu_{jk}, U_{jk}), (1 \le j \le N) \quad (10)$$

where

$$N(z_j, \mu, U) \cong (2\pi)^{-P_j/2}|U|^{-1/2}e^{\left\{-\frac{1}{2}(s-\mu)'U^{-1}(s-\mu)\right\}}$$

and P_j is the dimension of z_j. We may let M ne independent of j as long as M is sufficiently large. Let

$$\gamma_t(j,m) = \gamma_t(j)\left[\frac{c_{jm}N(z_j[t], \mu_{jm}, U_{jm})}{b_j(z_j[t])}\right] \quad (11)$$

$$\hat{c}_{jm} = \left(\sum_{t=1}^{T}\gamma_t(j,m)\right) \Big/ \left(\sum_{t=1}^{T}\sum_{l=1}^{M}\gamma_t(j,l)\right) \quad (12)$$

$$\hat{\mu}_{jm} = \left(\sum_{t=1}^{T}\gamma_t(j,m)z_j[t]\right) \Big/ \left(\sum_{t=1}^{T}\gamma_t(j,m)\right) \quad (13)$$

$$\hat{U}_{jm} = \frac{\sum_{t=1}^{T}\gamma_t(j,m)(z_j[t]-\mu_{jm})(z_j[t]-\mu_{jm})'}{\sum_{t=1}^{T}\gamma_t(j,m)} \quad (14)$$

2.2. Relationship to conventional algorithm

All the estimated parameters are same as those estimated by the conventional Baum-Welch algorithm except for the Gaussian mixtures parameters $b_j(z_j)$.

3. EXPERIMENT

3.1. The Simulation

A synthetic six state $\{S_1, S_2, \ldots, S_6\}$ hidden markov model was created. The following state transition matrix and an equiprobable initial state priors were used.

$$A = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0.3 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0.7 & 0.3 \\ 0.1 & 0 & 0 & 0 & 0 & 0.9 \end{bmatrix}$$

For each segment t, a state was randomly choose according to the known initial state distribution and then an $N = 256$ sample time series $x[t] = \{x_1[t], x_2[t], \ldots, x_N[t]\}$ for this segment t was formed according to the state distribution (statistical model). Features were then calculated from each of these segments to be used as the data $z[t]$ for to train the HMM. The common-state S_0 was chosen to be a iid Gaussian noise of mean zero and unit variance.The six signals which are modelled by the six states of the HMM were chosen according to the following criterion (1) the signals are easy to produce and describe, (2) a sufficient statistic $\{z_1, z_2, \ldots, z_6\}$ was known exactly for each model that can discriminate it from the common-state, though this won't hold generally in the real-world, (3) these sufficient-statistic had a known density under state S_o, (4) all the signals and statistics were diverse and statistically dependent. the description of each of these six signal types, sufficient statistics and its distribution under the common state is given by the Table 1 in the paper under review[1].

The simulation was implemented using class-specific algorithm, and the two different approaches of conventional Baum-Welch algorithm. The two classical approaches were different in the way the

Gaussian Mixtures were implemented. One used the full or the general covariance matrices (“CL” approach) while the other used the diagonal covariance matrix that assumes all the features are statistically independent (“IA” approach). The IA approach is usually employed in speech recognition so as to estimate a fewer number of parameters. The paper refers these approaches as CS, CL and IA. The CS approach used 1 or 2 features per state but for CL and IA approach, all the features were combined to form a 7-dimensional feature set.

The parameters of the model were estimated using Baum-Welch algorithm for each with CS, CL and IA approach on 1, 2, 5, 10, 20, 40, 80, 160 and 320 records. Each of these record consisted of 99 data segments. Each segment from which features were calculated, consisted of N = 256 time samples. To evaluate the performance as a function of number of records the Viterbi algorithm was used for decoding a separate pool of data with 640 records. The state classification error probability was used as a measure of performance by dividing the total number of errors by the total number of segments (total number of time-steps or observations of HMM).

3.2. The results

During the parameter estimation using Baum-Welch for CL and IA approaches, caused the catastrophic errors during decoding because the initialization (μ and A) was not seeded properly. To overcome this problem, a separate pool of labelled data was used to train teach of the six states. This better estimate of PDF (μ and A) was used as the starting point to train the model both for the CL and IA approaches. This reduced the errors to a “lower bound”.

While the CL and IA encountered catastrophic errors while decoding, the CS approach proposed by the paper did not encounter such problem. This seems to be one of the advantages of this algorithm. Also, the Figure 1 of the paper [1] shows that the state classification error probability for CS approach is lower than both the CS and IA with less number of records (from 1 to more than 100) used for parameters estimation. As the number of records for training increase, both the CS and CI approach converge. The IA approach has the higher error rate because of the fact that it assumes independence among the features which is not true.

4. SUMMARY AND CONCLUSION

The theoretical approach of the class-specific implementation of the Baum-Welch was demonstrated very well in the paper. The claim that this new class-specific algorithm that employs low-dimensional feature set has a higher performance than the CL or IA approaches given a training data-set. this has been extremely well supported by the simulation example provided in the paper where for class-specific approach the feature dimensions were either one or two whereas for the CL and IA approached the feature dimensions were combined as seven. The reason for improvement in performance is correctly given in the paper as the low-dimension feature set that, the parameters for which can be estimated using less training data. The reduction in dimensions of the features comes from the prior knowledge about the sufficient feature set for a given state.

REFERENCES

[1] P. M. Baggenstoss, “A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces”, *IEEE Conference on Acoustics, Speech, Signal Processing*, vol. 2, pp. II717–II720, 2000.

[2] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 257-286, February1989.

[3] J. Picone, “Pattern Recognition Notes,” *ECE 8990, Pattern Recognition, Mississippi State University,* pp.11-18*, 2001.*

[4] P. M. Baggenstoss, “Class-specific features in classification,” *IEEE Transactions on Signal Processing,* vol. 47, no. 12, December, 1999.

[5] P. M. Baggenstoss and H. Niemann, “A theoretical optimal probabilistic classifier using class-specific features,” *IEEE Conference on Pattern Recognition*, vol. 2, pp. 763-768, 2000.

[6] D. W. Scott, *Multivariate Density Estimation.*

Wiley, 1992.