Optimal feature weighting for the Continuous HMM

Oualid Missaoui and Hichem Frigui
CECS dept, University of Louisville

Abstract

We propose new Continuous Hidden Markov Model (CHMM) structure that integrates feature weighting component. We assume that each feature vector could include different subsets of features that come from different sources of information or different feature extractors. We modify the probability density function that characterizes the standard CHMM to include state and component dependent feature relevance weights. To learn the optimal feature weights from the training data, we modify the maximum likelihood based Baum-Welch algorithm and we derive the necessary conditions. The proposed approach is validated using synthetic and real data sets. The results are shown to outperform the standard CHMM.

1. Introduction

Hidden Markov Models (HMMs) have emerged as a powerful paradigm for modeling stochastic processes and pattern sequences. Originally, HMMs have been applied to the domain of speech recognition, and became the dominating technology [13]. In recent years, they have attracted growing interest in automatic target detection and classification [14], computational molecular biology [2], bioinformatics [11], mine detection [5], handwritten character/word recognition [12], and other computer vision applications [3]. HMMs are categorized into discrete and continuous models. An HMM is called continuous if the observation probability density functions are continuous and discrete if the observation probability density functions are discrete.

For most real applications, the problem of selecting or weighting the best subset of features constitutes an important part of the design of a good learning algorithm. Several features may be needed to capture the large intra- and inter-class variations. However, not all features are equally relevant and irrelevant features could degrade the generalization performance of the learning algorithms significantly. This problem has been researched extensively for the case of static data sets. In particular, several methods have been proposed for feature selection and weighting [1, 10]. In feature selection, the task's dimensionality is reduced by completely eliminating irrelevant features. This amounts to assigning binary relevance weights to the features (1 for relevant, and 0 for irrelevant). Feature weighting is an extension of the selection process where the features are assigned continuous weights which can be regarded as degrees of relevance. Because it provides a richer feature relevance representation, continuous weighting tends to outperform feature selection from an accuracy point of view in tasks where some features are useful but less important than others.

The feature selection or weighting problem has not received much attention for sequential data. This is despite the fact that irrelevant features could be present at multiple observations of the sequence, and thus, could have a more significant negative impact. Only few methods have been proposed to address this issue to discriminate between the audio and visual streams in speech recognition using continuous HMM [7, 16]. In these methods, the feature space is partitioned into different subspaces, and different probability density functions (pdf) are learned for the different spaces. The relevance weights for each subspace could be fixed a priori by an expert [4], or learned via Minimum Classification Error (MCE) approach or Generalized Probabilistic Descent (GPD) [16]. In [7], the authors have adapted the Baum-Welch algorithm [9] to learn the feature relevance weights. However, to derive the maximum likelihood equations, the model was restricted to include only one Gaussian component per state. This extension was not possible without the use of new constraints.

In this paper, we introduce aa novel structure for the continuous HMM that integrates feature weighting. We generalize the Baum-Welch algorithm of the feature relevance weights. We show that the proposed approach outperforms the baseline CHMM.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

2 Baseline Continuous HMM

An HMM is a model of a doubly stochastic process that produces a sequence of random observation vectors at discrete times according to an underlying Markov chain. At each observation time, the Markov chain may be in one of N_s states $s_1, ..., s_{N_s}$ and, given that the chain is in a certain state, there are probabilities of moving to other states. These probabilities are called the transition probabilities. An HMM is characterized by three sets of probability density functions, the transition probabilities (**A**), the state probability density functions (**B**), and the initial probabilities (π). Let T be the length of the observation sequence (i.e., number of time steps), let $O = [o_1, \ldots, o_T]$ be the observation sequence,where each observation vector v_i is characterized by p features (i.e. $v_i \in \mathbb{R}^p$), and let $Q = [q_1, \ldots, q_T]$ be the state sequence. The compact notation

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi) \tag{1}$$

is generally used to indicate the complete parameter set of the HMM model. In (1), A=$[a_{ij}]$ is the state transition probability matrix, where a_{ij}=$Pr(q_t$=$j|q_{t-1}$=$i)$ for i, j=1,. . ., N_s; π=$[\pi_i]$, where π_i=$Pr(q_1 = s_i)$ are the initial state probabilities; and $\mathbf{B} = \{b_i(o_t), i$=1,. . ., $N_s\}$, where $b_i(o_t)$=$Pr(o_t|q_t = i)$ is the set of observation probability distribution in state i.

For the continuous HMM, $b_i(o_t)$'s are defined by a mixture of some parametric probability density functions. The most common parametric pdf used in continuous HMM is the mixture Gaussian density where

$$b_i(o_t) = \sum_{j=1}^{M_i} c_{ij} b_{ij}(o_t), i = 1, ..., N_s. \tag{2}$$

In (2), M_i is the number of components in state i, c_{ij} is the mixture coefficient for the j^{th} mixture component in state i, and satisfies the constraints $c_{ij} \geq 0$, and $\sum_{j=1}^{Mi} c_{ij} = 1$, for $i = 1, ..., N_s$, and $b_{ij}(o_t)$ is a D-dimensional multivariate Gaussian density with mean μ_{ij} and covariance matrix Σ_{ij}. Typically, Baum-Welch algorithm [13] is used to estimate the parameters **A** and **B** iteratively using:

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{3}$$

$$\overline{c}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{4}$$

$$\overline{\mu}_{ijn} = \frac{\sum_{t=1}^{T} \gamma_t(i,j) o_{tn}}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{5}$$

$$\overline{\sigma}^2_{ijn} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)(o_{tl} - \overline{\mu}_{ijn})^2}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{6}$$

In (3) , (4), (5) and (6),

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}, \tag{7}$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}, \tag{8}$$

and

$$\gamma_t(i,k) = \gamma_t(i) \left[\frac{c_{ij} b_{ij}(o_t)}{b_i(o_t)} \right] \tag{9}$$

The variables $\alpha_t j$ and $\beta_t(j)$ are computed using the Forward and Backward algorithms [13] respectively.

3 Continuous HMM with feature discrimination

For high dimensional observations, the feature space can be very sparse. In this case, Gaussian components cannot model the data effectively. To overcome this limitation, we assume that the p features have been partitioned into L subsets: $FS_1, FS_2, \ldots, FS_L$. Then, instead of learning Gaussians components in the high dimensional space, we learn Gaussian components in the lower dimensional subspaces. Moreover, a feature relevance weight will be learned for each Gaussian component. The motivation is that some Gaussian components will cover dense subspaces, while other components may cover sparse subspaces. Intuitively, the latter components are less reliable in modeling the data and should be assigned smaller weights.

Formally, let $\phi(o_t^{(k)}, \mu_{ij}^{(k)}, \Sigma_{ij}^{(k)})$ be the j^{th} component in state i using only feature subset k into account, and let $w_{ij}^{(k)}$ be the feature relevance weight of this component. To cover the entire feature space, we use a mixture of L components, i.e.,

$$\sum_{k=1}^{L} w_{ij}^{(k)} \phi(o_t^{(k)}, \mu_{ij}^{(k)}, \Sigma_{ij}^{(k)}), \tag{10}$$

To model each state by multiple components, we use

$$b_i(o_t) = \sum_{j=1}^{M} v_{ij} \sum_{k=1}^{L} w_{ij}^{(k)} \phi(o_t^{(k)}, \mu_{ij}^{(k)}, \Sigma_{ij}^{(k)}) \tag{11}$$

subject to:

$$\sum_{k=1}^{L}(w_{ij}^{(k)})^m = K \text{ and } \sum_{j=1}^{M} v_{ij} = 1 \qquad (12)$$

where K is a constant, usually set to one. The constant $m \in (1, \infty)$ controls the discrimination between the different subspaces. For lower values of m (close to 1), the discrimination is more emphasized, however for higher values of m, the different feature subsets tend to have equal contribution to the discrimination.

The parameter v_{ij} is similar to the mixing coefficient in the standard HMM. In this case, it is a weight assigned to the mixture of L component that cover the entire feature space and not to a single component.

Using (11), We generalize the Baum-Welch learning algorithm to learn the parameter of the Gaussian component and their relevance weights. It can be shown that maximizing the likelihood of the training data results in the following update equations:

$$w_{ij}^{(k)} = \left[K \frac{\sum_{t=1}^{T} \gamma_t(i,j,k)}{\sum_{t=1}^{T} \gamma_t(i,j)} \right]^{\frac{1}{m}} \qquad (13)$$

$$v_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \qquad (14)$$

$$\mu_{ijd}^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(i,j,k) o_{td}^{(k)}}{\sum_{t=1}^{T} \gamma_t(i,j,k)} \qquad (15)$$

$$\sigma_{ijd}^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(i,j,k)(o_{td}^{(k)} - \mu_{ijd}^{(k)})^2}{\sum_{t=1}^{T} \gamma_t(i,j,k)} \qquad (16)$$

where

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N_s} \alpha_t(j)\beta_t(j)} \qquad (17)$$

$$\gamma_t(i,j) = \gamma_t(i) \frac{v_{ij} b_{ij}(o_t)}{b_i(o_t)} \qquad (18)$$

$$\gamma_t(i,j,k) = \gamma_t(i) \frac{v_{ij} w_{ij}^{(k)} \phi(o_t^{(k)}, \mu_{ij}^{(k)}, \Sigma_{ij}^{(k)})}{b_j(o_t)} \qquad (19)$$

$$b_{ij}(o_t) = \sum_{k=1}^{L} w_{ij}^{(k)} \phi(o_t^{(k)}, \mu_{ij}^{(k)}, \Sigma_{ij}^{(k)}) \qquad (20)$$

4 Experiments

4.1 Synthetic data set

To validate the proposed CHMM, we generate a synthetic data set as used in [15, 8]. In particular, we generate a multivariate time series data set that includes 2 classes. Each class has 3 states that generate data according to a normal distribution. For class 1,

$$(\mu_1^1, \mu_2^1, \mu_3^1) = ([10\ 2\ 5\ 1], [5\ 6\ 2\ 3], [2\ 10\ 1\ 5])$$
$$(\Sigma_1^1, \Sigma_2^1, \Sigma_3^1) = ([1\ 1\ 1\ 1], [0.5\ 0.5\ 1\ 1], [1\ 1\ 0.5\ 0.5])$$

For class 2, $\mu_1^2 = \mu_1^1$, $\mu_2^2 = [1\ 2\ 2\ 3]$,$\mu_3^1 = \mu_3^2$, and $\Sigma_i^2 = \Sigma_i^1, \forall\, i = 1, 2, 3$.

From each multivariate Gaussian we first generate a static data set with N=150 points. Then, we build multivariate sequences of length T=15. For each class, we assume that all the sequences start by an item from s_1. We also assume that for both classes the transition between these states is governed by the transition matrix $A = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 1 \end{bmatrix}$.

For this example, we partition the 4-D feature space into 2 subspaces. The first one includes the first 2 dimensions. Training the CHMM using this data results in symbols with various feature relevance weights. In general, higher weights are assigned to feature subset with low variance.

Table 1 displays the classification results using the baseline CHMM and the proposed CHMM that include feature weighting. As it can be seen, learning symbol and state dependent feature relevance weights improves the classification rate significantly.

4.2 Landmine data

The proposed CHMM is applied to the problem of buried landmine detection using Ground Penetrating Radar (GPR). Landmines (and other buried objects) appear in time-domain GPR as shapes that are similar to hyperbolas corrupted by noise. Thus, we use features that are based on the degree to which edges occur in the diagonal and anti-diagonal directions. These features are extracted using Gabor filters [6] at multiple scales. To fit this application into the HMM context, we take the traversal path as the time variable and use a sequence of T=15 observations to produce a confidence that a mine is present at various positions. We refer the reader to [6, 5] for more details about this application.

The training data set for this application includes 275 mine signatures and 885 false alarm signatures. For testing, we use another set that includes 272 mine signature and 886 false alarm signature. We extract Gabor features at 2 scales and 4 orientations. Features extracted at the same scale are treated as one subset, and for each symbol a relevance weight is learned for each scale.

Figure 1 displays the ROC curve that shows the performance of different CHMMs. We compare the new feature weighting based CHMM to the baseline CHMM using all the features as one subset. As it can be seen, the feature weighting approach outperforms the baseline CHMM. This is because different weights are learned for the different Gaussian components. In particular, some components capture information at lower scale, while others capture information at a higher scale.

Table 1. Classification Rates

Classifier	Classification rate
Baseline CHMM	85.75%
CHMM with feature discrimination	93.075%

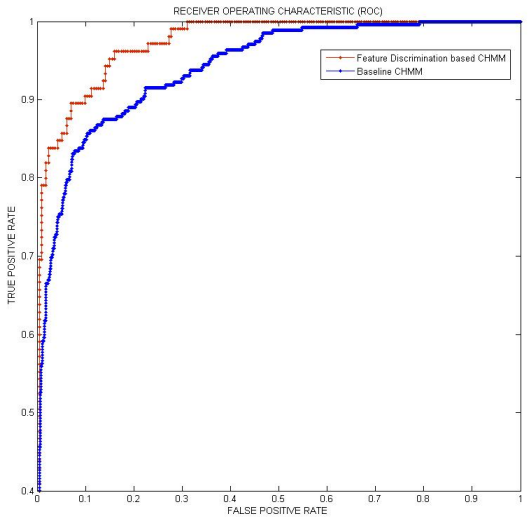

Figure 1. CHMM's ROCs comparison.

5. Conclusion

We have proposed a general method that allows maximum likelihood optimization of subset feature relevance weights for continuous HMM. In particular, the probability density function of the CHMM is approximated by a linear combination of weighted mixture of Gaussians. Hence, this pdf form facilitates the relevance weights optimization within the Baum-Welch learning scheme along with the rest of the CHMM parameters. Preliminary results on a synthetic data set and a library of GPR signatures show that the introduction of the feature weighting component allows the CHMM to perform better than considering all the features as one single subset. The feature relevance weights could also be learned using discriminative or corrective training instead of maximum likelihood based training to improve the discriminative capability of the CHMM. We are currently investigating this alternative.

ACKNOWLEDGMENTS

This work was supported in part by an Office of Naval Research award number $\mathrm{N00014-05-10788}$.

References

[1] H. Almuallim and T. Dietterich. Learning with many irrelevant features,. *9th National Conf. on Arti. Intell., pp. 547-552*, 1991.

[2] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. McClure. Hidden markov models of biological primary sequence information. *In Nat. Acad. Science, pp. 91(3):1059-1063, USA,*, 1994.

[3] H. Bunke and T. Caelli. Hidden markov models: Applications in computer vision. *Kulwer Academic,*, 2001.

[4] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia, Vol. 2, No. 3, September.*

[5] H. Frigui, P. Gader, and D. Ho. Real time landmine detection with ground penetrating radar using discriminative and adaptive hidden markov models. *EURASIP Journal on Applied Signal Processing, Vol. 2005 , Issue 1 , pp: 1867 - 1885*, January 2005.

[6] H. Frigui, O. Missaoui, and P. Gader. Landmine detection using discrete hidden markov models with gabor features. *Proc. SPIE Vol. 6553 , Apr. 27*, 2007.

[7] J. Hernando. Maximum likelihood weighting of dynamic speech features for cdhmm speech recognition. *Acoustics, Speech, and Signal Processing,ICASSP.,IEEE, on pp. 1267-1270 vol.2.*

[8] M. Kadous. Temporal classification: Extending the classification paradigm to multivariate time series. *PhD thesis, School of Computer Science and Engineering, University of New South Wales*, 2002.

[9] S. Kapadia. Discriminative training of hidden markov models. *PhD thesis, University of Cambridge.*, March 18, 1998.

[10] K. Kira and L. Rendell. The feature selection problem: traditional methods and a new algorithm. *10th National Conf. on Arti. Intell., pp. 129-134*, 1992.

[11] T. Koski. *Hidden Markov Models for Bioinformatics.* Kluwer Academic Publishers, Netherlands, 2001.

[12] M. Mohamed and P. Gader. Generalized hidden markov models part 2: Applications to handwritten word recognition. *IEEE Trans. Fuzzy Systems, 8:186-194,*, 2000.

[13] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. 1989.

[14] P. Runkle, P. Bharadwaj, and L. Carin. Hidden markov model multi-aspect target classification. 1999.

[15] N. Saito. Local feature extraction and its application using a library of bases. *PhD thesis, Yale University*, December 1994.

[16] A. Torre, A. Peinado, A. Rubio, and C. Segura, J. adn Benitez. Discriminative feature weighting for hmm-based continuous speech recognizers. *Speech Communication 38, on pp. 267-286.*