MMIHMM: Maximum Mutual Information Hidden Markov Models

Nuria Oliver
Microsoft Research
nuria@microsoft.com

Ashutosh Garg
Univ. Illinois, Urbana-Champaign
ashutosh@ifp.uiuc.edu

January 2002

Technical Report
MSR-TR-2002-13

This paper proposes a new family of Hidden Markov Models named Maximum Mutual Information Hidden Markov Models (MMIHMMs). MMIHMMs have the same graphical structure as HMMs. However, the cost function being optimized is not the joint likelihood of the observations and the hidden states. It consists of the weighted linear combination of the mutual information between the hidden states and the observations and the likelihood of the observations and the states. We present both theoretical and practical motivations for having such a cost function. Next, we derive the parameter estimation (learning) equations for both the discrete and continuous observation cases. Finally we illustrate the superiority of our approach in different classification tasks by comparing the classification performance of our proposed Maximum Mutual Information HMMs (MMIHMMs) with standard Maximum Likelihood HMMs (HMMs), in the case of synthetic and real, discrete and continuous, supervised and unsupervised data. We believe that MMIHMMs are a powerful tool to solve many of the problems associated with HMMs when used for classification and/or clustering.

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

1 Introduction

It has been claimed [16] that a fundamental problem in formalizing our intuitive ideas about information is to provide a quantitative notion of 'meaningful' or 'relevant' information. These issues were missing in the original formulation of information theory, where the attention was focused on the problem of transmitting information rather than evaluating its value to the recipient. Information theory has therefore traditionally been seen as a theory of communication. However, in recent years there has been growing interest of applying information theoretic principles in machine learning and statistics. It has been argued that information theory provides a natural quantitative approach to the question of 'relevant' information.

There are many situations when we would like to compress or summarize dynamic time data (for example speech or video). One possible approach to solving that problem is having an additional 'hidden' variable that determines what is relevant. In the case of speech, for example, it could be the transcription of the signal, if we are interested in the speech recognition problem, or it might be the speaker's identity if speaker identification is our goal. The formal underlying structure of such problems would consist of extracting the information from one variable that is relevant for the prediction of another variable.

In this paper we formalize the idea of using information theory in the framework of Hidden Markov Models (HMMs). In the case of HMMs, we will enforce the hidden state variables to capture relevant information about the observations. At the same time, we would like our models to explain the generative process of the data as accurately as possible. Therefore, we propose a cost function that combines both the information theoretic (MI) and the maximum likelihood (ML) criteria.

The paper is organized as follows: First, in Section 2 we review the most relevant previous work. Next, Section 3 motivates and describes the proposed cost function to be optimized. The learning algorithm that estimates the parameters of the model (in the discrete and continuous, supervised and unsupervised cases) while optimizing such function is presented in Section 4. Experimental results are presented in Section 6. Finally we summarize our work and discuss future directions of research in Section 7.

2 Previous Work

In this work we introduce a new formulation for Hidden Markov Models. Numerous variations of the standard formulation of Hidden Markov Models have been proposed in the past, such as Parametrized-HMM (PHMM) [17], Entropic-HMM [4], Variable-length HMM (VHMM) [7], Coupled-HMM (CHMM) [5, 13], Input-Output-HMM (IOHMM) [2], Factorial-HMM [10] and Hidden-Markov Decision Trees (HMDT) [12], to name a few. Each of these models attempts to solve some of the deficiencies of standard HMMs given the particular problem or set of problems at hand. Given that most of them aim at modeling the data and learning the parameters using ML, in many cases their main differences lie in the conditional independence assumptions made while modeling the data, i.e.

in their graphical structure. Conversely, the graphical structure of the model presented in this paper remains the same as that of a standard HMM, but the optimization function is different. Even though we develop here the learning equations for HMMs, the framework that we present could easily be extended to any graphical model.

Tishby's et al. work on the Information Bottleneck [16] method and its extensions has been one of the sources of inspiration for our work. The Information Bottleneck method is an unsupervised non-parametric data organization technique. Given a joint distribution $P(A, B)$, the method constructs, using information theoretic principles, a new variable T that extracts partitions, or clusters, over the values of A that are informative about B. In particular, consider two random variables X and Q with their assumed joint distribution $P(X, Q)$, where X is the variable that we are trying to compress with respect to the 'relevant' variable Q. They propose the introduction a soft partitioning of X through an auxiliary variable T, and the probabilistic mapping $P(T|X)$, such that the mutual information $I(T;X)$ is minimized (maximum compression) while the probabilistic mapping $P(T|X)$, the relevant information $I(T;Q)$ is maximized.

Our work is also related to the recently popular debate of conditional versus joint density estimation [**?**]. The 'conditional' approach (i.e. the maximization of the conditional likelihood of the variables of interest instead of the full likelihood) is closely related to the use of discriminative approaches in learning theory. Jebara nicely summarizes in [11] the advantages and disadvantages associated with joint and conditional density estimation. Standard HMMs perform joint density estimation of the hidden state and observation random variables. However, in situations where the resources are limited (complexity, data, structures), the system has to handle very high dimensional spaces or when the goal is to classify or cluster with the learned models, a conditional approach is probably superior to the full joint density approach. One can think of these two methods (conditional vs joint) as two extremes with our work providing a tradeoff between the two. Sections 3 and 5 analyze the properties of our approach and relate it to the purely probabilistic model more formally.

Finally we would like to point out how our work is different to the Maximum Mutual Information Estimation (MMIE) approach that is so popular in the speech recognition community. In particular, Bahl et al. [1] introduced the concept of Maximum Mutual Information Estimation (MMIE) for estimating the parameters of an HMM in the context of speech recognition, where typically a different HMM is learned for each possible class (e.g. one HMM for each word in the vocabulary). New waveforms are classified by computing their likelihood based on each of the models. The model with the highest likelihood is selected as the winner. However, in our approach, we learn a single HMM whose hidden states correspond to different classes. The algorithm in [1] attempts to maximize the mutual information between the choice of the HMM and the observation sequence to improve the discrimination across different models. In contrast, our algorithm aims at maximizing the mutual information between the observations and the hidden states, so as to minimize the classification error when the hidden states are used as the classification output.

3 Mutual Information, Bayes Optimal Error, Entropy and Conditional Probability

In the 'generative approach' to machine learning, the goal is to learn a probability distribution that defines the process that generated the data. Such an approach is particularly good in modeling the general form of the data and can give some useful insights into the nature of the original problem. Recently, there has been an increasing focus on connecting the performance of these generative models to their classification accuracy when they are used for classification tasks. In particular, Garg and Roth develop an extensive analysis in [9] of the relationship between the Bayes optimal error of a classification task using a probability distribution and the entropy between the random variables of interest. Consider the family of probability distributions over two random variables (X, Q) denoted by $P(X, Q)$. The classification task is to predict Q after observing X. The relationship between the conditional entropy $H(X|Q)$ and the Bayes optimal error, ϵ is given by

$$\frac{1}{2}H_b(2\epsilon) \leq H(X|Q) \leq H_b(\epsilon) + \log\frac{M}{2}. \tag{1}$$

with $H_b(p) = -(1-p)\log(1-p) - p\log p$.

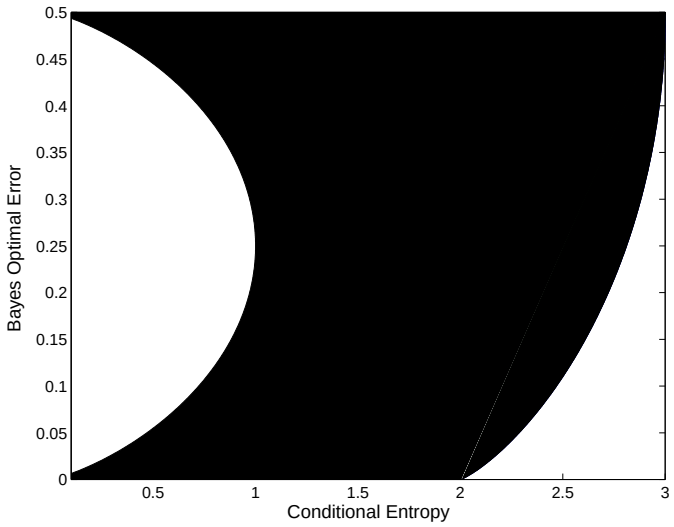


Figure 1: Bayes optimal error versus conditional entropy

Figure 1 illustrates this relationship between the conditional entropy and the Bayes optimal error. In Figure 1 the only realizable –and at the same time observable– distributions are those within the black region. One can conclude from Figure 1 that, if the data is generated according to a distribution that has high conditional entropy, the Bayes optimal error of any classifier for this data will be high. Even though this relationship is between the *true model* and

the *Bayes optimal error*, it also applies to a model that has been estimated from data, –assuming a consistent estimator has been used, such as Maximum Likelihood, and the model structure is the true one. As a result, when the learned distribution has high conditional entropy, it might not necessarily do well on classification. Therefore, if the final goal is classification, the graph in Figure 1 suggests that low entropy models should be preferred over high entropy ones. The cost function proposed in Eqn 2 favors low conditional entropy models to high entropy ones.

A Hidden Markov Model (HMM) is a probability distribution over a set of random variables, some of which are referred to as the hidden states (as they are normally not observed and they are discrete) and others are referred to as the observations (continuous or discrete). Traditionally, the parameters of Hidden Markov Models are estimated by maximizing the joint likelihood of the hidden states Q and the observations X, $P(X,Q)$. Conventional Maximum Likelihood (ML) techniques would be optimal in the case of very large datasets (so that the estimate of the parameters is correct) if the true distribution of the data was in fact an HMM. However none of the previous conditions is normally true in practice. The HMM assumption might be in many occasions highly unrealistic and the available data for training is normally very limited, leading to important problems associated with the ML criterion (such as overfitting). Moreover, ML estimated models are often used for clustering or classification. In these cases, the evaluation function is different to the optimization function, which suggests the need of an optimization function that correctly models the problem at hand. The cost function defined in Eqn 2 is designed to tackle some of these problems associated to ML estimation.

When formulating our optimization functional, we exploit the relationship between the conditional entropy of the data and the Bayes optimal error previously described. In the case of Hidden Markov Models (HMMs), the X variable corresponds to the observations and the Q variable to the hidden states. We would like to maximize the joint probability distribution $P(Q,X)$ while forcing the Q variable to contain maximum information about the X variable (i.e. to maximize their mutual information or minimize the conditional entropy). In consequence, we propose to maximize both the joint likelihood and the mutual information between the hidden variables and the observations. This leads to the following cost function

$$F = (1-\alpha)I(Q,X) + \alpha \log P(X_{obs}, Q_{obs}) \tag{2}$$

where $\alpha \in [0,1]$, provides a way of deciding the appropriate weighting between the Maximum Likelihood (ML) (when $\alpha = 1$) and Maximum Mutual Information (MMI) (when $\alpha = 0$) criteria, and $I(Q,X)$ refers to the mutual information between the states and the observations. However, very often one does not observe the state sequence[1]. In such a scenario, the cost function reduces to

$$F = (1-\alpha)I(Q,X) + \alpha \log P(X_{obs}) \tag{3}$$

In fact, mutual information is closely related to conditional likelihood. Learning the parameters is equivalent to learning the conditional dependencies be-

[1] We will refer to this case as the unsupervised case while referring to the former as the supervised case.

tween the variables (edges in the graphical model). This relationship is made explicit in the following theorem by Bilmes et al.:

Theorem 1 Mutual Information and Likelihood [3] *Given three random variables X, Q^a and Q^b, where $I(X, Q^a) > I(X, Q^b)$, the conditional likelihood of X given Q^a is higher than that of X given Q^b, for a sample size large enough.*

The theorem also holds true for conditional mutual information, such as $I(X, Z|Q)$, or for a particular value of q, $I(X, Z|Q = q)$

Therefore, given a graphical model in general (and an HMM in particular) whose parameters have been learned by maximizing the joint likelihood $P(X, Q)$, if we were to add some edges according to mutual information the resulting dynamic graphical model would yield higher conditional likelihood score than before the modification [3]. In an HMM we are maximizing the joint likelihood of the hidden states and the observations, $P(X, Q)$. At the same time, it would be desirable to make sure that the states Q are good predictors of the observations X. According to theorem 1, maximizing the mutual information between states and observations increases the conditional likelihood of the observations given the states $P(X|Q)$. This justifies, to some extent, why the cost function defined in Eqn 2 combines the two desirable properties of maximizing the conditional and joint likelihood of the states and the observations.

4 MMIHMMs

We develop in this section the learning algorithms for discrete and continuous, supervised and unsupervised MMIHMMs. For the sake of clarity and simplicity, we will start with the supervised case, where the 'hidden' states are actually observed in the training data.

Consider a Hidden Markov Model with $\mathbf{Q}$ as the states and $\mathbf{X}$ as the observations. Let F denote the cost function to maximize,

$$F = (1 - \alpha)I(Q, X) + \alpha \log P(X_{obs}, Q_{obs}) \tag{4}$$

The mutual information term $I(Q, X)$ can be expressed as $I(Q, X) = H(X) - H(X/Q)$, where $H(\cdot)$ refers to the entropy. Since $H(X)$ is independent of the choice of the model and is characteristic of the generative process, we can reduce our cost function to

$$\begin{aligned} F &= -(1 - \alpha)H(X/Q) + \alpha \log P(X_{obs}, Q_{obs}) \\ &= (1 - \alpha)F_1 + \alpha F_2 \end{aligned}$$

In the following we will use the standard HMM notation for the transition a_{ij} and observation b_{ij} probabilities,

$$a_{ij} = P(q_t = i, q_{t+1} = j), \quad b_{ij} = P(x_t = j|q_t = i) \tag{5}$$

Expanding each of the terms F_1 and F_2 separately we obtain,

$$F_1 = -H(X|Q) = \sum_X \sum_Q P(X, Q) \log \prod_{t=1}^{T} P(x_t|q_t)$$

$$
\begin{aligned}
&= \sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{i=1}^{N} P(x_t = j|q_t = i)P(q_t = i)\log P(x_t = j|q_t = i) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{i=1}^{N} P(q_t = i)b_{ij}\log b_{ij} \\
F_2 &= \log \pi_{q_1^o} + \sum_{t=2}^{T} \log a_{q_{t-1}^o, q_t^o} + \sum_{t=1}^{T} \log b_{q_t^o, x_t^o}
\end{aligned}
$$

Combining F_1 and F_2 and adding the appropriate Lagrange multipliers to ensure that the a_{ij} and b_{ij} coefficients sum to 1, we obtain:

$$
\begin{aligned}
F_L &= (1-\alpha)\sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{i=1}^{N} P(q_t = i)b_{ij}\log b_{ij} \\
&\quad +\alpha \log \pi_{q_1^o} + \alpha \sum_{t=2}^{T} \log a_{q_{t-1}^o, q_t^o} + \alpha \sum_{t=1}^{T} \log b_{q_t^o, x_t^o} \\
&\quad +\beta_i \left(\sum_j a_{ij} - 1 \right) + \gamma_i \left(\sum_j b_{ij} - 1 \right)
\end{aligned} \tag{6}
$$

Note that in the case of continuous observation HMMs, we can no longer use the concept of entropy as previously defined. As a result, we will be using the counterpart differential entropy. Because of this important distinction, we will carry out the analysis for discrete and continuous observation HMMs separately.

4.1 Discrete MMIHMMs

To obtain the parameters that maximize the cost function, we take the derivative of F_L from Eqn 6 and will equate it to zero. First solving for b_{ij}, we obtain

$$
\frac{\partial F_L}{\partial b_{ij}} = (1-\alpha)(1+\log b_{ij}) \left(\sum_{t=1}^{T} P(q_t = i) \right) + \frac{N_{ij}^b \alpha}{b_{ij}} + \gamma_i = 0 \tag{7}
$$

where N_{ij}^b is the number of times one observes state j when the hidden state is i. Eqn 7 can be expressed as

$$
\log b_{ij} + \frac{W_{ij}}{b_{ij}} + g_i + 1 = 0 \tag{8}
$$

where

$$
\begin{aligned}
W_{ij} &= \frac{N_{ij}^b \alpha}{(1-\alpha)\left(\sum_{t=1}^{T} P(q_t = i)\right)} \\
g_i &= \frac{\gamma_i}{(1-\alpha)\left(\sum_{t=1}^{T} P(q_t = i)\right)}
\end{aligned}
$$

The solution of Eqn 8 is given by

$$b_{ij} = -\frac{W_{ij}}{\mathsf{LambertW}(-\mathsf{W_{ij}e^{1+g_i}})} \tag{9}$$

where $\mathsf{LambertW(x)} = \mathsf{y}$ is the solution of the equation $ye^y = x$.

Now we are going to solve for a_{ij}. Let's first look at the derivative of F_1 with respect to a_{lm}.

$$\frac{\partial F_1}{\partial a_{lm}} = \sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{i=1}^{N} b_{ij}\log b_{ij}\frac{\partial P(q_t=i)}{\partial a_{lm}} \tag{10}$$

To solve the above equation, we need to compute $\frac{\partial P(q_t=i)}{\partial a_{lm}}$. This can be computed using the following iteration

$$\frac{\partial P(q_t=i)}{\partial a_{lm}} = \begin{cases} \sum_j \frac{\partial P(q_{t-1}=j)}{\partial a_{lm}} a_{ji} & \text{if } m \neq i, \\ \sum_j \frac{\partial P(q_{t-1}=j)}{\partial a_{lm}} a_{ji} + P(q_{t-1}=l) & \text{if } m = i \end{cases} \tag{11}$$

with the initial conditions

$$\frac{\partial P(q_2=i)}{\partial a_{lm}} = \begin{cases} 0 & \text{if } m \neq i, \\ \pi_l & \text{if } m = i \end{cases} \tag{12}$$

Taking the derivative of F_L, with respect to a_{lm}, we obtain,

$$\begin{aligned} \frac{\partial F}{\partial a_{lm}} &= (1-\alpha)\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{M} b_{ik}\log b_{ik}\frac{\partial P(x_t=i)}{\partial a_{lm}} \\ &\quad +\alpha\frac{N_{lm}}{a_{lm}} + \beta_l \end{aligned}$$

where N_{lm} is the count of the number of occurrences of $q_{t-1}=l, q_t=m$ in the data set. The update equation for a_{lm} is obtained by equating this quantity to zero and solving for a_{lm}

$$a_{lm} = \frac{\alpha N_{lm}}{(1-\alpha)\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{M} b_{ik}\log b_{ik}\frac{\partial P(x_t=i)}{\partial a_{lm}} + \beta_l} \tag{13}$$

where β_l is chosen so that $\sum_m a_{lm} = 1, \forall l$.

4.2 Continuous MMIHMM

For the sake of clarity and without loss of generality, we will restrict our attention to the case when the $P(x|q)$ is a single Gaussian. Under this assumption, the HMM is characterized by the following parameters

$$\begin{aligned} &P(q_t=j|q_{t-1}=i) = a_{ij} \\ &P(x_t|q_t=i) = \frac{1}{\sqrt{2\pi|\Sigma_i|}}\exp(-\frac{1}{2}(x_t-\mu_i)^T\Sigma_i^{-1}(x_t-\mu_i)) \end{aligned}$$

where Σ_i is the covariance matrix when the hidden state is i and $|\Sigma_i|$ is the determinant of the covariance matrix. Now, for the cost function given in Eqn 2 F_1 and F_2 can be written as

$$\begin{aligned}
F_1 &= -H(X|Q) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N}\int P(q_t=i)\log P(x_t|q_t=i)dP(x_t|q_t=i) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N} P(q_t=i)\int(-\frac{1}{2}\log(2\pi|\Sigma_i|) \\
&\quad -\frac{1}{2}(x_t-\mu_i)^T\Sigma_i^{-1}(x_t-\mu_i))dP(x_t|q_t=i)y \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N} P(q_t=i)(-\frac{1}{2}\log(2\pi|\Sigma_i|)-\frac{1}{2}) \\
F_2 &= \log P(Q_{obs}, X_{obs}) \\
&= \sum_{t=1}^{T}\log P(x_t|q_t) + \log\pi_{q_1^\circ} + \sum_{t=2}^{T}\log a_{q_{t-1}^\circ, q_t^\circ}
\end{aligned}$$

Following the same steps as for the discrete case, we again form the Lagrange F_L, take its derivative with respect to each of the unknown parameters and obtain the corresponding update equations. First the means of the Gaussians

$$\mu_i = \frac{\sum_{t=1,q_t=i}^{T} x_t}{N_i} \tag{14}$$

where N_i is the number of times $q_t = i$ in the observed data. Note that this is the standard update equation for the mean of a Gaussian, and it is the same as for ML estimation in HMMs. This is because the conditional entropy is independent of the mean.

Next, the update equation for a_{lm} is same as in Eqn 13 except for replacing $\sum_k b_{ik}\log b_{ik}$ by $-\frac{1}{2}\log(2\pi|\Sigma_i|)-\frac{1}{2}$. Finally, the update equation for Σ_i is

$$\begin{aligned}
\Sigma_i &= \frac{\alpha\sum_{t=1,q_t=i}^{T}(x_t-\mu_i)(x_t-\mu_i)^T}{N_i\alpha+(1-\alpha)\sum_{t=1}^{T}P(q_t=i)} \\
&= \frac{\sum_{t=1,q_t=i}^{T}(x_t-\mu_i)(x_t-\mu_i)^T}{N_i+\frac{(1-\alpha)}{\alpha}\sum_{t=1}^{T}P(q_t=i)}
\end{aligned} \tag{15}$$

It is interesting to note that the update equation for Σ_i in Eqn 15 is very similar to the one obtained when using ML estimation, except for the term in the denominator $\frac{(1-\alpha)}{\alpha}\sum_{t=1}^{T}P(q_t=i)$, which can be thought of as a regularization term. Because of this positive term, the covariance Σ_i is smaller than what it would have been otherwise. This corresponds to lower conditional entropy, as desired.

4.3 Unsupervised Case

The above analysis can easily be extended to the unsupervised case, i.e. when only X_{obs} is given and Q_{obs} is not available. In this case, we use the cost function given in Eqn 3. The update equations for the parameters are very similar to the ones obtained in the supervised case. The only difference is that now we replace N_{ij} in Eqn 7 by $\sum_{t=1,x_t=j}^{T} P(q_t = i|X_{obs})$, N_{lm} is replaced in Eqn 13 by $\sum_{t=2}^{T} P(q_{t-1=l}, q_t = m|X_{obs})$, and N_i is replaced in Eqn 15 by $\sum_{t=1}^{T} P(q_t = i|X_{obs})$. These quantities can be easily computed using the Baum-Welch algorithm by means of the forward and backward variables.

5 Discussion

5.1 Convexity

From the law of large numbers, it is known that, in the limit (i.e. as the number of samples approaches infinity), the likelihood of the data tends to the negative of the entropy, $P(X) \approx -H(X)$.

Therefore, in the limit, the negative of our cost function for the supervised case can be expressed as

$$\begin{aligned} -F &= (1-\alpha)H(X|Q) + \alpha H(X,Q) \\ &= H(X|Q) + \alpha H(Q) \end{aligned} \tag{16}$$

Note that $H(X|Q)$ is a strictly concave function of $P(X|Q)$, and $H(X|Q)$ is a linear function of $P(Q)$. Consequently, in the limit, the cost function from Eqn 16 is strictly convex (its negative is concave) with respect to the distributions of interest.

In the unsupervised case and in the limit again, our cost function can be expressed as

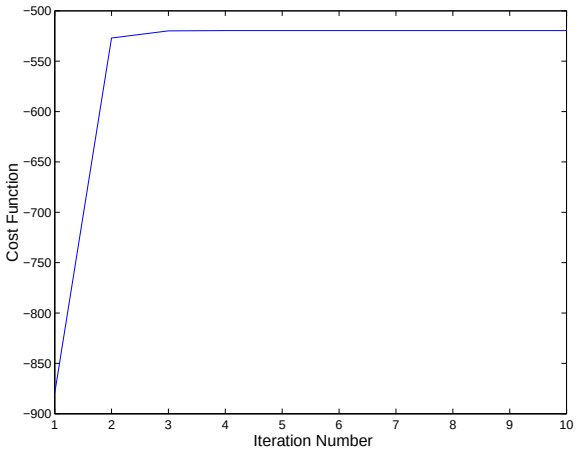
$$\begin{aligned} F &= -(1-\alpha)H(X|Q) - \alpha H(X) \\ &= -H(X) + (1-\alpha)(H(X) - H(X|Q)) \\ &= -H(X) + (1-\alpha)I(X,Q) \approx P(X) + (1-\alpha)I(X,Q) \end{aligned}$$

The unsupervised case thus reduces to the original case with α replaced by $1-\alpha$. Maximizing F is, in the limit, the same as maximizing the likelihood of the data and the mutual information between the hidden and the observed states, as expected.

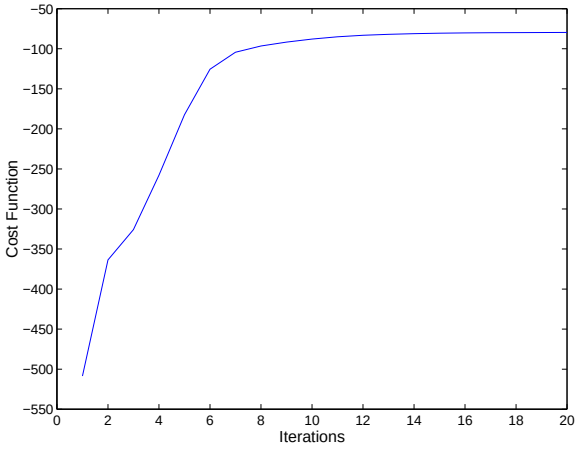
5.2 Convergence

We analyze next the convergence of the MMIHMM learning algorithm in the supervised and unsupervised cases. In the supervised case, HMMs are directly learned without any iteration. However, in the case of MMIHMM we do not have a closed form solution for the parameters b_{ij} and a_{ij}. Moreover these parameters are inter-dependent (i.e. in order to compute bij, we need to compute $P(q_t = i)$ which requires the knowledge of a_{ij}). Therefore an iterative solution is needed.

Fortunately, the convergence of the iterative algorithm is extremely fast, as it is illustrated in Figure 2. This figure shows the cost function with respect to the iterations for a particular case of the speaker detection problem (a) (see section 6), and for synthetically generated data in an unsupervised situation (b). From Figure 2 it can be seen that the algorithm typically converges after only 2-3 iterations.



(a)



(b)

Figure 2: Value of the cost function with respect to the iteration number in (a) the speaker detection experiment; (b) a continuous unsupervised case with synthetic data.

6 Experimental Results

In this section we describe the set of experiments that we have carried out to obtain quantitative measures of the performance of MMIHMMs when compared to HMMs in various classification tasks. We have conducted experiments with synthetic and real, discrete and continuous, supervised and unsupervised data.

1. Synthetic Discrete Supervised Data:

We generated 10 different datasets of randomly sampled synthetic discrete data with 4 hidden states and 5 observation values. We used 100 samples for training and 100 for testing. The training was supervised for both HMMs and MMIHMMs. MMIHMMs had an average improvement over the 10 datasets of **12**%, when compared to HMMs of exactly the same structure. The optimal α variables ranged from 0.05 to 0.95, depending on the dataset.The best accuracy of HMMs and MMIHMMs for each of the 10 datasets is depicted in Figure 3, together with the optimal α for each of the datasets. A summary of the accuracy of HMMs and MMIHMMs is shown in Table 1.

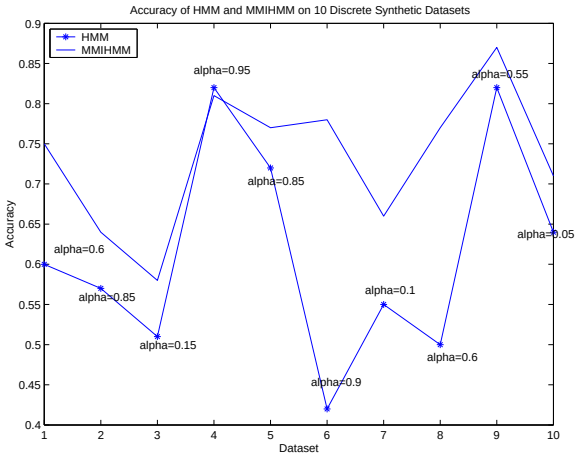


Figure 3: Accuracies and optimal value of α for MMIHMM and HMM (star-line) on 10 different datasets of synthetic discrete data.

2. Speaker Detection:

An estimate of the person's state is important for the reliable functioning of any interface that relies on speech communication. In particular, detecting when users are speaking is a central component of open mike speech-based user interfaces, specially given their need to handle multiple people in noisy environments. We carried out some experiments in a speaker detection task. The speaker detection dataset was the same that appears in [8]. It consisted of five sequences of one user playing blackjack in a simulated casino setup using CRL's Smart Kiosk [14]. The sequences were of varying duration from 2000 to 3000 samples, with a total of 12500 frames. The original feature space had 32 dimensions that resulted from quantizing five binary features (skin color presence, face texture presence, mouth motion presence, audio silence presence

and contextual information). Only the 14 most significant dimensions were selected out of the original 32-dimensional space.

The learning task in this case was supervised for both HMMs and MMIHMMs. Three were the variables of interest: the presence/absence of a speaker, the presence/absence of a person facing frontally, and the existence/absence of an audio signal or not. The goal was to identify the correct state out of four possible states: (1) no speaker, no frontal, no audio; (2) no speaker, no frontal and audio; (3) no speaker, frontal and no audio; (4) speaker, frontal and audio. Figure 4 illustrates the classification error for HMMs (dotted line) and Mhttp://www.meany.org/MIHMMs (solid line) with α varying from 0.05 to 0.95 in .1 increments. Note how in this case MMIHMMs outperformed HMMs for all the values of α. The accuracies of HMMs and MMIHMMs are summarized in table 1. The accuracy reported in [8] using a bi-modal (audio and video) DBN was of about 80%.

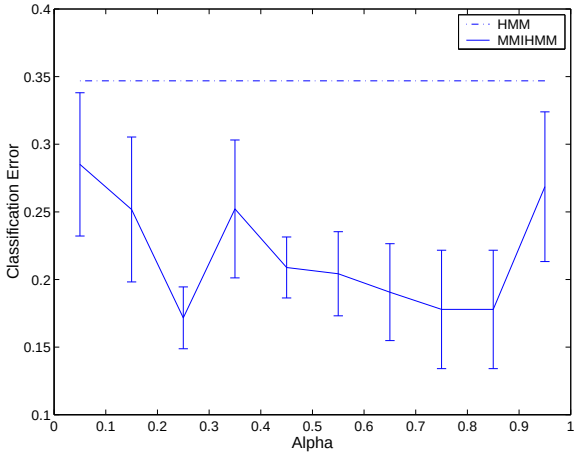


Figure 4: Error bars for the Speaker Detection data in MMIHMMs and HMMs

3. Protein Data:

Gene identification and gene discovery in new genomic sequences is certainly an important computational question addressed by bioinformatics scientists. In this example, we tested both HMMs and MMIHMMs in the analysis of the Adh region in Drosophila. More specifically, part of an annotated drosophila sequence was used to conduct the experiments and to obtain the measure for the algorithms' performance (7000 data points on training and 2000 on testing). MMIHMMs were superior tohttp://www.meany.org/ HMMs for different values of alpha. The best results were obtained for a value of alpha of 0.5 as Table 1 reflects.

4. Real-time Emotion Data:

Finally we carried out an emotion recognition task using the emotion data described in [6]. The data had been obtained from a database of five people that had been instructed to display facial expressions corresponding to the following six types of emotions: anger, disgust, fear, happiness, sadness and surprise. The

Table 1: Classification accuracies for HMMs and MMIHMMs on different datasets

DATASET	HMM	MIHMM
SYNTDISC	55%	**66**%($\alpha_{\mathbf{optimal}} = .\mathbf{1}$)
SPEAKERID	64%	**88**%($\alpha_{\mathbf{optimal}} = .\mathbf{75}$)
GENE	68%	**84**%($\alpha_{\mathbf{optimal}} = .\mathbf{5}$)
EMOTION	67%	**74**%($\alpha_{\mathbf{optimal}} = .\mathbf{8}$)

data collection method is described in detail in [15]. We used the same video database as the one used in [15]. It consisted of six sequences of each facial expression for each of the five subjects. In the experiments reported here, we used unsupervised training of continuous HMMs and MMIHMMs. The accuracy results of both types of models are displayed in Table 1.

7 Summary and Future Work

We have presented a new framework for estimating the parameters of Hidden Markov Models. We have motivated, proposed and justified a new cost function that linearly combines the mutual information and the likelihood of the hidden states and the observations in an HMM. We have derived the parameter estimation equations in the discrete and continuous, supervised and unsupervised cases. Finally we have shown the superiority of our approach in a classification task when compared to standard HMMs in different synthetic and real datasets.

Future lines of research include automatically estimating the optimal α, extending the approach to other graphical models with different structures, and better understanding the connection between MMIHMMs and other information theoretic and discriminative approaches. We are also exploring how to apply our framework to a number of applications and real-life problems.

8 Acknowledgements

We would like to thank you CRL for sharing the protein and speaker detection data, and Prof. T. Huang and I. Cohen for sharing the emotion data with us.

References

[1] Bahl, Brown, de Souza, and Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proc. ICASSP*, pages 999–999, 1986.

[2] Yoshua Bengio and Paolo Frasconi. An input output HMM architecture. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 427–434. The MIT Press, 1995.

[3] J. Bilmes. Dynamic bayesian multinets. In *Proc. of the 16th conf. on Uncertainty in Artificial Intelligence. Morgan Kaufmann, 2000.*, 2000.

[4] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[5] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden markov models for complex action recognition. In *Proc. of CVPR97*, pages 994–999, 1996.

[6] I. Cohen, A. Garg, and T. S. Huang. Emotion recognition using multilevel hmms. In *Workshop on Affective Computing in NIPS'00*, 2000.

[7] A. Galata, N. Johnson, and D. Hogg. Learning variable length markov models of behaviour. *International Journal on Computer Vision, IJCV-2001*, pages 398–413, 2001.

[8] A. Garg, V. Pavlovic, J. Rehg, and T. S. Huang. Audio-visual speaker detection using dynamic bayesian networks. In *Proceed. of the Intl. Conference on Face and Gesture (FG'00)*, 2000.

[9] A. Garg and D. Roth. Understanding probabilistic classifiers. In *Proceed. of the 12th European Conference on Machine Learning*, 2001.

[10] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden Markov models. In David S. Touretzky, Michael C. Mozer, and M.E. Hasselmo, editors, *NIPS*, volume 8, Cambridge, MA, 1996. MITP.

[11] Tony Jebara. Action-reaction learning: Analysis and synthesis of human behaviour. Master's thesis, MIT Media Lab, 1998.

[12] Michael I. Jordan, Zoubin Ghahramani, and Lawrence K. Saul. Hidden Markov decision trees. In David S. Touretzky, Michael C. Mozer, and M.E. Hasselmo, editors, *NIPS*, volume 8, Cambridge, MA, 1996. MITP.

[13] N. Oliver. *Towards Perceptual Intelligence: Statistical Modeling of Human Individual and Interactive Behaviors.* PhD thesis, Massachusetts Institute of Technology, MIT, 2000.

[14] J.M. Rehg, M. Loughlin, and K. Waters. Vision for a smart kiosk. In *CVPR 97*, pages 690–696, 1997.

[15] L. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multimodal information. In *Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97)*, pages pp. 397–401, 1997.

[16] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *In Proc. of the 37-th Allerton Conference on Communication and Computation*, 1999.

[17] A. Wilson and A. Bobick. Recognition and interpretation of parametric gesture. In *Proc. of International Conference on Computer Vision (ICCV'98)*, pages 329–336, 1998.