KERNEL BASED HIDDEN MARKOV MODEL WITH APPLICATIONS TO EEG SIGNAL CLASSIFICATION

Wenjie Xu\(^1\), Jiankang Wu\(^1\)
\(^1\)Institute for Infocomm Research
Singapore, 119613
\{wenjie, jiankang\}@i2r.a-star.edu.sg

Zhiyong Huang\(^2\)
\(^2\)School of Computing
National University of Singapore
Singapore, 117543
huangzy@comp.nus.edu.sg

ABSTRACT
Learning the dependencies between neighboring signals/data is an important issue in multi-class pattern classification. For nonlinear dependencies, this paper presents a new model referred to as kernel based hidden Markov model. Using kernel feature space as a potential nonlinear manifold, the model integrates Markov model with the maximum margin method. The incorporation of these two kinds of techniques can ensure the generalization property of KHMM, simultaneously retaining the rich structure knowledge of sequential data. We evaluate the kernel based hidden Markov model by applying it to motor imagery classification tasks, yielding positive experimental results.

KEYWORDS
Kernel method, Hidden Markov Model, Motor Imagery

1 Introduction
Hidden Markov model (HMM), as one of the statistical models, is well understood and widely applied to many scientific and engineering areas [1, 2, 3]. This model characterizes the statistical properties of the real-world signals by parsimoniously choosing the model such that the conditional probability of the observation sequence is locally maximized

$$\arg \max_{\lambda} P(O|\lambda) = (A, B, \pi) \quad (1)$$

Accordingly, a set of such models can be chosen for each class, and then classifying a given pattern can be achieved by comparing the individual joint probability to paired observations and hidden states sequence $P(q, O|\lambda_i)$.

In addition to its probabilistic nature, the reason that hidden Markov model is suitable for sequential signal, such as speech and bio-signal, is that HMM well models the temporal or sequential structure by combining the observation and hidden state. However, it is known that HMM based classifier at least has a major limitation [4]. As state before, the individual model for each class is obtained by the maximum likelihood (ML) principle. Thereby, this does not explicitly take into account the decision boundaries between the different classes. In some cases in which there are only a small training set this limitation would cause over-fitting problems and degenerate the performance.

It is known that kernel method has been well established as an efficient way to nonlinear analysis in recent years. In particular, support vector machine (SVM), as one of the kernel based methods, has demonstrated impressive empirical success on a broad range of pattern recognition tasks. By maximizing the margin of separation between positive and negative examples, SVM constructs a hyperplane in the kernel space [5]. As such, it can provide a strong generalization guarantee for pattern recognition problems, while it does not need to incorporate prior knowledge.

Recently, there are some attempts to combine HMM with SVM in a unify framework. Altun et al. [6] propose a so called hidden Markov support vector machine (HM-SVM) for label sequences learning. Their method can effectively learn nonlinear discriminant function by employing kernel functions, while retaining the ability to capture correlations in structured examples. Furthermore, Taskar et al. [7] propose an effective optimization algorithm based on quadratic program such that it has a polynomial-size formulation, as opposed to HM-SVM which requires an exponential number of constraints.

The goal of this paper is to propose an efficient approach, namely kernel based hidden Markov model (KHMM), to classify the sequence data instead of sequences label itself. Unlike the previous methods [6, 7] which maximize the difference between the true sequences labels and the best runner-up, our method is to introduce the margin maximization between the true model and the best runner-up and thus minimizes the classification error. Another important difference between our approach and theirs is that we employ distinct weight parameters for individual state, as it characterizes the state-observation dependence and state-state transaction. Additionally, the states in our proposed approach could have the “hidden” property. The true states sequences related to the observations are not necessary to be known in the training phase. Our method would predict the states by applying the Viterbi algorithm.

The organization of the paper is as follows. Section 2 constructs the KHMM framework by introducing a lose function. Section 3 presents how to train the model using
the maximum margin optimization, and section 4 follows a training algorithm. Section 5 evaluates the kernel based hidden Markov model by applying to motor imagery classification tasks. Finally, we conclude our paper in the last section.

2 Lose function and kernels

In the HMM based multi-class problem, the task is to learn the models \( \{\lambda_k\} \) for every class from the corresponding training data. The classification function \( h: \mathcal{O} \rightarrow \mathcal{Y} \) is the following form

\[
h_{\lambda}(\mathcal{O}) = \arg \max_k P(\mathcal{O} | \lambda_k)
\]  

(2)

where \( \mathcal{O} \) is a state sequence related to the observation sequence \( \mathcal{O} \) such that the class conditional probability is maximum.

To learn the models, a good approximation to the observation probability \( P(\mathcal{O}) \) has to be found. The most general representation of the pdf, for continuous signals or observations, is a finite mixture of Gaussian densities. In this paper, we represent directly the conditional probability \( P(\mathcal{O} | \lambda_k) \) in another way using the theorem of random fields [8]

\[
P(\mathcal{O} | \lambda_k) \propto P(q, \lambda_k) \propto \prod_{(i,j) \in E} \psi_{ij}(\mathcal{O}, q_i, q_j)
\]  

(3)

where \( \psi_{ij} \) are the network potentials and \( E \) is the set of dependencies between states. For simplicity, here we assume state-state interaction is the first order Markov chain. Therefore, the conditional probability can be derived

\[
P(\mathcal{O} | \lambda_k) = \prod_{t} \exp[w_{q_t} \cdot f(\mathcal{O}, q_t, q_{t-1})]
\]  

(4)

where \( w_{q_t} \) is the weight modeling the correlation between observation and state for class \( k \), state \( q_t \). We think this kind of correlation is state dependence.

Let us denote the basis function for the observation model in (4) as \( f(\mathcal{O}, q_t, q_{t-1}) = f_i(\mathcal{O}) \). In particular, a possible basis function is derived

\[
f_i(\mathcal{O}) = \rho(q_t, q_{t-1}) \Phi(o_t)
\]  

(5)

where \( \rho(q_t, q_{t-1}) \) could be an indicator function for the state transaction and \( \Phi(o_t) \) could be a kernel feature of the observation vector \( o_t \). According to statistical learning theory [9], the inner-products of kernel features can be replaced by a kernel function \( K(\cdot, \cdot) \) that satisfies Mercer’s conditions.

By combining (2) with (4), we obtain the classification function as the following logarithm form

\[
h_{\lambda}(\mathcal{O}) = \arg \max_k \sum_{t} w_{q_t}^k \cdot f_i(\mathcal{O})
\]  

(6)

Unlike the basic HMM, where the maximum likelihood (ML) criterion estimates the model parameters such that the class conditional probability of the training data is maximized, our proposed method is to choose the models carefully such that the corresponding classification error is minimized.

By taking into account the misclassification and margin simultaneously, the lose function can be estimated using the following piecewise linear bound [10]

\[
\mathcal{I} \leq \frac{1}{m} \sum_{i=1}^{m} \left[ \max_k \left\{ \sum_{t} w_{q_t}^k \cdot f_i(\mathcal{O}_i) + 1 - \delta_{y_i,k} \right\} - \sum_{t} w_{y_t}^i \cdot f_i(\mathcal{O}_i) \right]
\]  

(7)

where \( \delta_{p,q} \) is equal 1 if \( p = q \) and 0 otherwise, and \( y_i \) is true class for the i-th example.

To minimize the classification error, the lose function \( \mathcal{I} \) has to be minimum by optimizing a set of \( w \). When a sample set \( S \) is linearly separable, the above loss function value is equal to zero when the following constraints are satisfied for all the examples in \( S \)

\[
\forall i \quad \max_k \left\{ \sum_{t} w_{q_t}^k \cdot f_i(\mathcal{O}_i) + 1 - \delta_{y_i,k} \right\} - \sum_{t} w_{y_t}^i \cdot f_i(\mathcal{O}_i) = 0
\]  

(8)

Thereby, the set \( \{w\} \) that satisfies (8) would also satisfy the constraints,

\[
\forall i, k \quad \sum_{t} w_{q_t}^k \cdot f_i(\mathcal{O}_i) + \delta_{y_i,k} - \sum_{t} w_{q_t}^i \cdot f_i(\mathcal{O}_i) \geq 1
\]  

(9)

However, the samples might not be linearly separable. In this kind of more general cases, we therefore add slack variables \( \xi_i \geq 0 \) and modify (9) to be

\[
\forall i, k \quad \sum_{t} w_{q_t}^k \cdot f_i(\mathcal{O}_i) + \delta_{y_i,k} - \sum_{t} w_{q_t}^i \cdot f_i(\mathcal{O}_i) \geq 1 - \xi_i
\]  

(10)

3 Maximum margin Optimization

In the general maximum margin classification, one has to either restrict the norm of \( w \), or fix the functional margin [5]. Here we employ the latter strategy and thus have the following Lagrangian optimization

\[
J(w, \xi, \eta) = \frac{1}{2} \sum_{k} \sum_{q_t} ||w_{q_t}^k||^2 + \sum_{i} \xi_i + \sum_{i,k} \eta_{i,k} \left[ \sum_{t} w_{q_t}^k \cdot f_i(\mathcal{O}_i) - \sum_{t} w_{y_t}^i \cdot f_i(\mathcal{O}_i) - \delta_{y_i,k} + 1 - \xi_i \right]
\]  

(11)
It is known that the saddle point of the Lagrangian would be the minimum for the primal variables {w, ξ} and the maximum for the dual variables η. To find the minimum over the primal variables, we require the following two conditions

$$\frac{\partial J}{\partial \xi_i} = 1 - \sum_k \eta_{i,k} = 0 \quad \Rightarrow \quad \sum_k \eta_{i,k} = 1 \tag{12}$$

$$\frac{\partial J}{\partial w_{q_t}^k} = \beta w_{q_t}^k + \sum_i \eta_{i,k} \sum_{q_{t-1}=q_t} f_i(O_t)$$

$$- \sum_{i,j,q_{t-1}=q_t} \sum_k \eta_{i,k} \sum_{q_{t-1}} f_{i,j}(O_t) = 0 \tag{13}$$

By incorporating (12), we rewrite (13) and thus have

$$w_{q_t}^k = \beta^{-1} \left[ \sum_i (\delta_{y_{i,k}} - \eta_{i,k}) \sum_{q_{t-1}=q_t} f_i(O_t) \right] \tag{14}$$

To postulate the dual problem for our primal problem, we first expand (11), term by term, as follows:

$$J(w, \xi, \eta) = \sum_i \xi_i \left( 1 - \sum_k \eta_{i,k} \right)$$

$$+ \sum_{i,k} \eta_{i,k} \left[ \sum_t w_{q_t}^k \cdot f_t(O_t) \right]$$

$$- \sum_{i,k} \eta_{i,k} \left[ \sum_t w_{q_t}^{y_i} \cdot f_t(O_t) \right]$$

$$+ \sum_{i,k} \eta_{i,k}(1 - \delta_{y_{i,k}}) + \frac{1}{2} \beta \sum_{k} \sum_{q_t} ||w_{q_t}^k||_2^2 \tag{15}$$

The first term on the right-hand side of (15) is zero by virtue of the optimality condition of (12). Furthermore, from (13) we have

$$\sum_{i,k} \eta_{i,k} \left[ \sum_t w_{q_t}^k \cdot f_t(O_t) \right]$$

$$= \sum_{i,k} \eta_{i,k} \sum_t \beta^{-1}$$

$$\left[ \sum_j (\delta_{y_{j,k}} - \eta_{j,k}) \sum_{q_{t-1}=q_t} f_j(O_j) \right] \cdot f_t(O_t)$$

$$= \beta^{-1} \sum_{i,j} \left[ \sum_t \sum_{q_{t-1}=q_t} f_i(O_i) \cdot f_j(O_j) \right]$$

$$\sum_k \eta_{i,k}(\delta_{y_{j,k}} - \eta_{j,k})$$

$$= \beta^{-1} \sum_{i,j} \sum_k \eta_{i,k}(\delta_{y_{j,k}} - \eta_{j,k}) \tag{16}$$

where $f_i \cdot f_j = \sum_t \sum_{q_{t-1}=q_t} f_i(O_i) \cdot f_j(O_j)$. Similarly, the third and fifth terms of (15) have the following forms, respectively.

$$\sum_{i,k} \eta_{i,k} \left[ \sum_t w_{q_t}^{y_i} \cdot f_t(O_t) \right]$$

$$= \beta^{-1} \sum_{i,j} f_i \cdot f_j \sum_k \delta_{y_{j,k}} - \eta_{j,k} \tag{17}$$

$$\frac{1}{2} \beta \sum_{k} \sum_{q_t} \|w_{q_t}^k\|^2 = \frac{1}{2} \beta^{-1} \sum_{i,j} f_i \cdot f_j$$

$$\sum_k (\delta_{y_{i,k}} - \eta_{i,k})(\delta_{y_{j,k}} - \eta_{j,k}) \tag{18}$$

Accordingly, setting the objective function $J(w, \xi, \eta) = Q(\eta)$, we may reformulate (15) as

$$Q(\eta) = -\frac{1}{2} \beta^{-1} \sum_{i,j} f_i \cdot f_j \sum_k (\delta_{y_{i,k}} - \eta_{i,k})(\delta_{y_{j,k}} - \eta_{j,k})$$

$$+ \sum_{i,k} \eta_{i,k}(1 - \delta_{y_{i,k}}) \tag{19}$$

Now the objective function $Q$ is only the function of $\eta$ and concave in this variable. Therefore, the maximum value of $Q(\eta)$ is unique and could be found by using standard quadratic programming (QP) techniques.

### 4 A Training Algorithm

The goal of this section is to propose a model learning algorithm. To find the optimum model parameters in our framework, the “hidden” states sequence has to be predicted in advance. This inference can be done by performing the Viterbi algorithm, as the state-state transaction cost $\alpha_t(i)$ is solve inductively.

$$\alpha_1(i) = \mathbf{w}_i \cdot \Phi(a_1) \tag{20}$$

$$\alpha_{t+1}(i) = \alpha_t + w_i^k \cdot f(o_{t+1}, i, q_t) \tag{21}$$

After that, a new model parameters will be found related to the states sequences of all the examples by performing maximum margin optimization.

**Algorithm 4.1 (learning algorithm)**

1. predict the states sequences $Q$ for all examples using basic HMM

2. repeat

3. find the optimum model parameters related to $(O, Y, Q)$ based on the maximum margin optimization

4. search the runner-up states sequences by using Viterbi algorithm

5. until no more errors or predefined steps done
5 Experiments

We evaluate our approach on the motor task, distinguishing left and right hand movement imagination [11]. The experiments were performed on one male subject, between 35 and 40 years of age.

In our experimental paradigm, the subject was instructed to fixate on a computer monitor about 180cm in front of him. Each trial was 5 seconds long, starting with a blank screen which indicated a pause. At 2 s, the blank screen was replaced by a stimulus arrow, pointing either to the left or to the right for 4 seconds. Depending on the direction of the arrow, the subject was instructed to carry out a motor imagery. The experimental session consisted of five experimental runs of 20 randomized trials.

The EEG signals were recorded using the Neuroscan SynAmp2 system and were sampled at 250 Hz. 28 channels of EEG data were then chosen from the 64 scalp electrodes, around the Cz, CPz, Fz and FCz region related to the sensorimotor cortex. These sampled epochs, filtered at 8-36 Hz using the Infinite Impulse Response (IIR) band-pass filter, were then extracted in the range of 100 ms before stimulus and 4000 ms after that.

The data set is divided into 20 folds of 95 training and 5 test samples. The classification results, shown in Fig 1, are averages over these 20 folds. We compare our proposed algorithm with other two state-of-art classification approaches, SVM and HMM. Before classification, the time sequences are first divided to segments for extracting features. We perform the 200 ms long windows overlapping 100 ms each other on the data, except those used to extract features for the SVM. In this case 900 ms long 850 overlapped windows are employed, as the SVM ignores structure knowledges of sequential data. For Comparison, common spatial patterns (CSP) features are employed in our all classification tasks. More details about the preprocessing and feature extraction please refer to [12]. Additionally, both HMM and KHMM consist of 3 states for capturing the structure of EEG data. The kernel function used in SVM and KHMM is the RBF kernel [5]. Fig 1 shows the classification results. In this dataset, our proposed approach give the highest classification accuracy of 88%, compared with the SVM (60%) and HMM (84%).

6 Conclusion

This paper has presented a kernel based hidden Markov model for classifying the multi-class sequential data. The model is capable of exploring both the structure and the nonlinear dependency among the data in an efficient way, by taking advantage of the rich language of Markov model and the kernel techniques. Our results on motor imagery tasks show that these kind of structure models exploit the nature of sequential signals and significantly outperform other non-structure methods on the EEG signals with physiological changes. On the other hand, our approach learn the models by using a nonlinear discriminative procedure based on a maximum margin criterion, providing a strong generalization guarantee. Overall, the positive results attest to the excellent performance of the proposed model for EEG signal classification and BMI system.

References


