KERNEL BASED HIDDEN MARKOV MODEL WITH APPLICATIONS TO EEG SIGNAL CLASSIFICATION

Wenjie Xu[1,2], Jiankang Wu[1]
[1]Institute for Infocomm Research
21 Heng Mui Keng Terrace
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

Cuntai Guan
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
{ctguan}@i2r.a-star.edu.sg

ABSTRACT

To enhance the performance of hidden Markov models for EEG signal classification, we present here a new model referred to as kernel based hidden Markov model (KHMM). Due to the embedded HMM structure, this model is capable of capturing well the temporal change of a time-series signal. Furthermore, KHMM has better discrimination and generalization capability inherited from kernel methods. We evaluate the kernel based hidden Markov model by applying it to EEG signal classification when motor imagery is performed, yielding positive experimental results.

KEY WORDS

Kernel method, Hidden Markov Model, EEG classification

1 Introduction

The hidden Markov model (HMM) is a statistical model that has been widely applied to many scientific and engineering areas [1, 2]. It can well model temporal or sequential structures of signals by combining the observation and hidden state in an elegant manner. Therefore, it is particularly suitable for modeling temporal signals, such as speech and bio-signals.

On the other hand, support vector machine (SVM) is a maximum margin classifier with solid background in statistical learning theory. In principle, SVM constructs a hyperplane in the kernel space so as to maximize the margin of separation between positive and negative examples [3]. Thereby, it has a good generalization performance while retaining the advantage of discriminative approaches.

Considering the advantage of HMM and SVM, a few recent endeavors resort to combining HMMs and SVM in a unified framework which can capture well the temporal information as well as preserve the superior classification and generalization capability. An impressing work in this field was given by Altun et al. [4] who proposed a so-called hidden Markov support vector machine (HM-SVM) for sequences labeling. It aims at identifying the states of individual observation by effectively learn nonlinear discriminant function while retaining the ability to capture correlations in structured examples. More recently, Taskar et al. [5] proposed an effective optimization algorithm based on quadratic program such that it has an polynomial-size formulation, as opposed to HM-SVM which requires an exponential number of constraints.

HM-SVM and similar methods can identify well the label sequences of signals belonged to a single class. However, these methods may not be effective in classifying continuous signals. In our case, we are given a number of trials of EEG data. Each trial of EEG corresponds to a particular mental activity (for example, imagination of left or right hand movement). We know that each trial does contain a segment of signal which is elicited by one of the mental activities. But we have no information about when this segment starts and how long it lasts. To classify these trials, a straightforward method is to convert these continuous signals into the sequences of discrete symbols by HM-SVM, and then classify the sequences. Unfortunately, this method may bring distortions and therefore degrade the performance of classification. Therefore, it will be advantageous to model and classify the continuous signals directly.

In this paper we propose a novel model to address the signal classification problem, by combining hidden Markov model and maximum margin principle in a unified kernel-based framework. We refer to the model as the kernel-based hidden Markov model (KHMM). The training is formulated as finding the maximum margin between the true model and the best runner-up while minimizing the classification error simultaneously. Unlike previous methods, it may not compulsively know the hidden states of individual observation in advance. Besides, it has been applied to motor imagery classification tasks, yielding positive experimental results.

The organization of the paper is as follows. Section 2 presents the KHMM framework by introducing a loss function. Section 3 presents how to train the model using the maximum margin optimization, followed by 4. In that section we propose a training algorithm. Section 5 is devoted to evaluate the kernel based hidden Markov model by a motor imagery classification task. Finally, we conclude our paper in the last section.

2 Loss function and kernels

To build hidden Markov models for multi-class classification problem, we need to to learn the models $\{\lambda_k\}$ from the corresponding training data. In the classification phase, recognition is performed according the following criterion,

$$h_\lambda(\mathbf{O}) = \arg\max_k P(\mathbf{q}|\mathbf{O}, \lambda_k) \tag{1}$$

where $\mathbf{q}$ is a state sequence related to the observation sequence $\mathbf{O}$ such that the class conditional probability is maximum.

To learn the models, a good approximation to the observation probability $P(\mathbf{O}|\lambda)$ has to be found. The most popular representation, for continuous signals or observations, is a finite mixture of Gaussian densities. In this paper, we represent directly the conditional probability $P(\mathbf{q}|\mathbf{O}, \lambda)$ in another way using the theorem of random fields [6]

$$P(\mathbf{q}|\mathbf{O}, \lambda_k) \propto \prod_{(i,j)\in E} \psi_{ij}(\mathbf{O}, q_i, q_j) \tag{2}$$

where ψ_{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 as

$$P(\mathbf{q}|\mathbf{O}, \lambda_k) = \prod_t \exp[\mathbf{w}^k_{q_t} \cdot \mathbf{f}(\mathbf{O}, q_t, q_{t-1})] \tag{3}$$

where $\mathbf{w}^k_{q_t}$ is the weight modeling the correlation between observation and state for class k at state q_t, and the dependency between state q_{t-1} and stateq_t.

Let us denote the basis function for the observation model in (3) as $\mathbf{f}(\mathbf{O}, q_t, q_{t-1}) = \mathbf{f}_t(\mathbf{O})$. A possible basis function is

$$\mathbf{f}_t(\mathbf{O}) = \rho(q_t, q_{t-1})\Phi(o_t, o_{t-1}) \tag{4}$$

where $\rho(q_t, q_{t-1})$ is an indicator function for the state transaction and $\Phi(o_t, o_{t-1})$ could be a kernel feature of the observation vector o_t and o_{t-1}. According to statistical learning theory [7], the inner-products of kernel features can be replaced by a kernel function $\mathbf{K}(\cdot,\cdot)$ that satisfies Mercer's conditions.

By combining (1) with (3), we obtain the following objective function in logarithm form

$$h_\lambda(\mathbf{O}) = \arg\max_k \sum_t \mathbf{w}^k_{q_t} \cdot \mathbf{f}_t(\mathbf{O}) \tag{5}$$

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 loss function can be estimated using the following piecewise linear bound [8]

$$\mathcal{I} \le \frac{1}{m}\sum_{i=1}^{m}\left[\max_k \left\{\sum_t \mathbf{w}^k_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) + 1 - \delta_{y_i,k}\right\} - \sum_t \mathbf{w}^{y_i}_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i)\right] \tag{6}$$

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

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

$$\forall i \qquad \max_k \left\{\sum_t \mathbf{w}^k_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) + 1 - \delta_{y_i,k}\right\} - \sum_t \mathbf{w}^{y_i}_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) = 0 \tag{7}$$

Equivalently, (7) could be rewrite as

$$\forall i,k \quad \sum_t \mathbf{w}^{y_i}_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) + \delta_{y_i,k} - \sum_t \mathbf{w}^k_{q_i} \cdot \mathbf{f}_t(\mathbf{O}_i) \ge 1 \tag{8}$$

However, in practice the samples may not be necessarily linearly separable. In this case, we add slack variables $\xi_i \ge 0$ and modify (8) to have

$$\forall i,k \quad \sum_t \mathbf{w}^{y_i}_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) + \delta_{y_i,k} - \sum_t \mathbf{w}^k_{q_i} \cdot \mathbf{f}_t(\mathbf{O}_i) \ge 1 - \xi_i \tag{9}$$

3 Maximum margin Optimization

In the maximum margin classification, one has to either restrict the norm of $\mathbf{w}$, or fix the functional margin [3]. Here we employ the latter strategy and thus have the following Lagrangian optimization

$$J(\mathbf{w}, \xi, \eta) = \frac{1}{2}\beta\sum_k\sum_{q_t} ||\mathbf{w}^k_{q_t}||^2_2 + \sum_i \xi_i + \sum_{i,k} \eta_{i,k}\left[\sum_t \mathbf{w}^k_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) - \sum_t \mathbf{w}^{y_i}_{q_t} \cdot \mathbf{f}_t(\mathbf{O}_i) - \delta_{y_i,k} + 1 - \xi_i\right] \tag{10}$$

where $\eta_{i,k}$ are a dual set of variables for the constraints (9).

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

$$\frac{\partial J}{\partial \xi_i} = 1 - \sum_k \eta_{i,k} = 0 \qquad \Longrightarrow \qquad \sum_k \eta_{i,k} = 1 \tag{11}$$

$$\frac{\partial J}{\partial \mathbf{w}^k_{q_t}} = \beta\mathbf{w}^k_{q_t} + \sum_i \eta_{i,k} \sum_{q_r=q_t} \mathbf{f}_r(\mathbf{O}_i) - \sum_{i,y_i=k}\sum_k \eta_{i,k} \sum_{q_r=q_t} \mathbf{f}_r(\mathbf{O}_i) = 0 \tag{12}$$

By incorporating (11), we rewrite (12) and thus have

$$\mathbf{w}_{q_t}^{k} = \beta^{-1}\left[\sum_i (\delta_{y_i,k} - \eta_{i,k}) \sum_{q_r = q_t} \mathbf{f}_r(\mathbf{O}_i)\right] \tag{13}$$

To postulate the dual problem for our primal problem, we first expand (10), as follows:

$$\begin{aligned} J(\mathbf{w}, \xi, \eta) = & \sum_i \xi_i \left(1 - \sum_k \eta_{i,k}\right) \\ & + \sum_{i,k} \eta_{i,k} \left[\sum_t \mathbf{w}_{q_t}^{k} \cdot \mathbf{f}_t(\mathbf{O}_i)\right] \\ & - \sum_{i,k} \eta_{i,k} \left[\sum_t \mathbf{w}_{q_t}^{y_i} \cdot \mathbf{f}_t(\mathbf{O}_i)\right] \\ & + \sum_{i,k} \eta_{i,k}(1 - \delta_{y_i,k}) + \frac{1}{2}\beta \sum_k \sum_{q_t} ||\mathbf{w}_{q_t}^{k}||_2^2 \end{aligned} \tag{14}$$

The first term on the right-hand side of (14) is zero by virtue of the optimality condition of (11). Furthermore, from (12) we have

$$\begin{aligned} & \sum_{i,k} \eta_{i,k} \left[\sum_t \mathbf{w}_{q_t}^{k} \cdot \mathbf{f}_t(\mathbf{O_i})\right] \\ & = \beta^{-1} \sum_{i,j} \mathbf{f}_i \cdot \mathbf{f}_j \sum_k \eta_{i,k}(\delta_{y_j,k} - \eta_{j,k}) \end{aligned} \tag{15}$$

where $\mathbf{f}_i \cdot \mathbf{f}_j = \sum_t \sum_{q_r = q_t} \mathbf{f}_t(\mathbf{O}_i) \cdot \mathbf{f}_r(\mathbf{O}_j)$. Similarly, the third and fifth terms of (14) can be expressed as follows, following forms, respectively.

$$\begin{aligned} & \sum_{i,k} \eta_{i,k} \left[\sum_t \mathbf{w}_{q_t}^{y_i} \mathbf{f}_t(\mathbf{O_i})\right] \\ & = \beta^{-1} \sum_{i,j} \mathbf{f}_i \cdot \mathbf{f}_j \sum_k \delta_{y_i,k}(\delta_{y_j,k} - \eta_{j,k}) \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{1}{2}\beta \sum_k \sum_{q_t} ||\mathbf{w}_{q_t}^{k}||_2^2 = & \frac{1}{2}\beta^{-1} \sum_{i,j} \mathbf{f}_i \cdot \mathbf{f}_j \\ & \sum_k (\delta_{y_i,k} - \eta_{i,k})(\delta_{y_j,k} - \eta_{j,k}) \end{aligned} \tag{17}$$

Accordingly, by setting the objective function $J(\mathbf{w}, \xi, \eta) = \mathcal{Q}(\eta)$, we may reformulate (14) as

$$\begin{aligned} \mathcal{Q}(\eta) = & -\frac{1}{2}\beta^{-1} \sum_{i,j} \mathbf{f}_i \cdot \mathbf{f}_j \sum_k (\delta_{y_i,k} - \eta_{i,k})(\delta_{y_j,k} - \eta_{j,k}) \\ & + \sum_{i,k} \eta_{i,k}(1 - \delta_{y_i,k}) \end{aligned} \tag{18}$$

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

4 A Training Algorithm

In this section, we introduce our model learning algorithm. To find the optimum model parameters in the framework, the "hidden" states sequence needs to be estimated first. This inference can be done by performing the Viterbi algorithm, as the state-state transaction cost $\alpha_t(i)$ is solve recursively.

$$\alpha_1(i) = \mathbf{w}_i^k \cdot \Phi(o_1) \tag{19}$$

$$\alpha_{t+1}(i) = \alpha_t + \mathbf{w}_i^k \cdot \mathbf{f}(o_{t+1}, i, q_t) \tag{20}$$

When the state sequence is estimated, new model parameters will be found by performing maximum margin optimization.

Algorithm 4.1 (learning algorithm)

1. *initialize the models using the k-means clustering.*
2. *search the best states sequences by using Viterbi algorithm*
3. *find the optimum models parameters given $(\mathcal{O}, \mathcal{Y}, \mathcal{Q})$ based on the maximum margin optimization.*
4. *classify the evaluation set based on the above optimum models.*
5. *stop when classification performance decreases or predefined steps have reached.*
6. *goto 2.*

5 Experiments

We evaluate our approach on the classification of EEG signal for motor imagery, to distinguish left and right hand movement imagination [9]. The experiments were performed by one male subject (38 years old).

In our experimental paradigm, the subject was instructed to fixate on a computer screen about 180cm in front of him. Each trial was 6 seconds long, starting with a blank screen which indicated a pause. At 2nd second, the blank screen was replaced by a prompting arrow stimulus, pointing either to the left or to the right lasting for 4 seconds. Following the direction of the arrow, the subject performed motor imagery accordingly. The complete experiment consisted of five runs, each run consisted of 20 trials. The number of left and right hand imaginations are balanced.

EEG signals were recorded using the Neuroscan SynAmp2 system, sampled at 250 Hz. 28 channels of EEG around the C3 and C4 region related to the sensorimotor cortex were then chosen from the 64 scalp electrodes. EEG signals between 100 ms before stimuli and 4000 ms after stimuli were extracted for later processing. The extracted signal is filtered using the Infinite Impulse Response (IIR) band-pass filter with the frequency bandwidth of 8-36Hz.

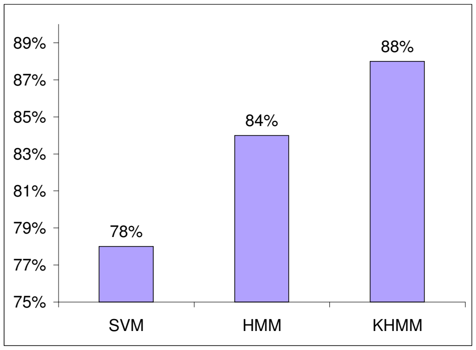

Figure 1. Motor imagery classification accuracy.

All data were divided into 20 folds of 95 training and 5 test samples each. Before classification, the time sequences are first divided into segments of 200ms length for feature extraction. There are 100ms overlap between neighboring segments for HMM and KHMM, while we use is 900 ms with 50ms overlap in the case of SVM. For the purpose of comparison, common spatial patterns (CSP) features are employed in all classification methods. For more details about the preprocessing and feature extraction please refer to [10]. 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 [3]. The classification result, shown in Fig 1, are averages over these 20 folds. We compare our proposed algorithm with other two classification approaches, SVM and HMM. In this dataset, our proposed approach gives the highest classification accuracy of 88%, compared to the SVM (78%) and HMM (84%). The low classification accuracy of SVM may be due to the fact that it does not explicitly take the temporal dynamic of the signals into account.

6 Conclusion

We presented here a kernel based hidden Markov model for classifying multi-class sequential data. The model is capable of both exploring the temporal dynamic of the signals and maximizing the margins between classes in an efficient way, by taking advantage of the rich language of Markov model and the kernel techniques. Our results on motor imagery classification have shown that HMM can exploit the nature of sequential signals and significantly outperform other non-structural methods on the EEG signals, which carry a lot of information on physiological changes. On the other hand, the proposed approach learns the models by using a nonlinear discriminative procedure based on a maximum margin criterion, providing a strong generalization mechanism. The good experimental results attest to the excellent performance of the proposed model for EEG signal classification and brain computer interfaces.

In this paper, the experiment was performed on a 100-trials data set for evaluating the performance of our approach. We anticipate more experimental works on EEG signal classification. In addition, a comparative study on the date set of different sizes will be performed to evaluate the generalization performance of the algorithm. We will also extend the approach to other interesting classification problems in brain computer interfaces.

7 Acknowledgements

We would like to thank Mr. Ranganatha Sitaram for the data collection. We also thank Mr. Haihong zhang and the anonymous reviewers for their valuable comments.

References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of The IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[2] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller, "Hidden markov models for online classification of single trial eeg data," *Pattern Recognition Letters*, vol. 22, pp. 1299–1309, 2001.

[3] S. Haykin, *Neural Networks. A Comprehensive Foundation*, Prentice Hall, New Jersey, USA, 1999.

[4] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov support vector machines," in *Proc. ICML*, 2003.

[5] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004, MIT Press.

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001.

[7] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[8] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[9] G. Pfurtscheller and C Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, pp. 1123–1134, July 2001.

[10] W. Xu, C. Guan, E. S. Chng, S. Ranganatha, M. Thulasidas, and J. Wu, "High accuracy classification of eeg signal," in *17-th ICPR*, 2004, vol. 2, pp. 391–394.