A maximum margin discriminative learning algorithm for temporal signals

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

Abstract

We propose a new maximum margin discriminative learning algorithm here for classification of temporal signals. It is superior to conventional HMM in the sense that it does not need prior knowledge of the data distribution. It learns the classifier by using a nonlinear discriminative procedure based on a maximum margin criterion, providing a strong generalization mechanism. This maximum margin discriminative learning method is presented together with a two-step learning algorithm. We evaluate the kernel based hidden Markov model by applying it to some simulation and real experiments. The preliminary results have shown significant improvement in classification accuracy.

1 Introduction

Maximum likelihood (ML) estimation, as the most popular learning method for hidden Markov model [9], may not lead to an optimal performance. This is due in part to the mismatch between the chosen distribution form and the actual signal distribution that is typically not available. To address this issue, a few recent endeavors resort to discriminative training approaches, such as maximum mutual information (MMI) estimation [2] and minimum classification error (MCE) estimation [6]. These approaches have their roots in maximum a posteriori (MAP) decision theory. Different from ML, here the learning is applied to all categories in the training phase. In the case of inadequate sparse training samples, they can usually demonstrate significant performance over the traditional ML approach. However, the performance of these learning methods still largely depends on consistency to actual data distribution.

We expect a nonparametric method that can be used with arbitrary distributions and without the assumption that forms of the underlying densities are known. Support vector machine (SVM), for example, is a nonparametric classification method with solid background in statistical learning theory [11]. In principle, SVM constructs a hyperplane in the kernel space so as to maximize the margin of separation between positive and negative examples, which guarantees strong generalization compared with the traditional discriminative approaches used to train HMM models. However, SVM suffers from an apparent lack of considering the underlying process of signal generation so that it may fail to classify temporal signals.

Motivated by this dilemma, we propose a new nonparametric learning for classification of temporal signal in this paper. It incorporates kernel-based discriminative learning approaches into hidden Markov model, having no need of prior knowledge of signal distribution. The learning is formulated as finding the maximum margin of separation between the category of the sample and the best runner-up in the kernel space. By contrast, previous margin-based approaches [1, 10] try to maximum the summation of margins between the true states of all the observations and the best runner-up(s). The formulation is by imposing the explicit constraint to the cost function so that the inferred state sequence from the designed model is the most possible state sequence. By minimizing an auxiliary cost function which is associated to the inferred state sequence, we present a two-step learning algorithm that alternatively estimates the parameters of the designed model and the most possible state sequences until convergence. Besides, our algorithm has been applied to the synthetic and real data of motor imagery classification tasks, yielding positive experimental results.

2 Classification of temporal signal

Multiclass classification is to learn a function $h : \mathcal{X} \mapsto \mathcal{Y}$ that maps an instance x of $\mathcal{X}$ into an element y of $\mathcal{Y}$. In general $\mathcal{Y}$ is a countable set and has $\mathcal{Y} = \{1, \cdots, K\}$. In this paper, we consider the problem of the signal classification where a signal $\mathbf{x}$ is a sequence from the set $\mathcal{X} = \{\mathcal{X}_1 \times \cdots \times \mathcal{X}_T\}$. In a motor imagery signal classification task [8], for example, the goal is to determine from the EEG signal, a time sequence signal for several seconds, which action the user is imagining.

A popular family of classification function h for the problem of the signal classification is statistically based. To achieve the minimum classification error, the optimal classifier, according to the classical Bayes decision theory, is the one that employs the decision rule of Eq. (1), which is called the *maximum a posteriori* (MAP) decision.

$$h(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} P(y|\mathbf{x}). \tag{1}$$

There are some methods to compute the posterior probabilities of Eq. (1). Here we approximate those by only considering the most likely state sequence, that is[1]

$$P(y|\mathbf{x}) \approx \max_{\mathbf{q} \in \mathcal{Q}|_y} P(\mathbf{q}|\mathbf{x}) \tag{2}$$

where $\mathcal{Q}|_y$ is the subset of state sequences $\mathbf{q}$ which belong to model y.

The theorem of random fields [7] provides a way to approximate the conditional probability $P(\mathbf{q}|\mathbf{x})$ directly

$$P(\mathbf{q}|\mathbf{x}) \propto \exp\left[\sum_{C \in \mathcal{C}} V_C(\mathbf{q}|_C, \mathbf{x})\right]$$

where $\mathcal{C}$ is the set of cliques for a graph, $\mathbf{q}|_C$ the set of components of $\mathbf{q}$ associated with the clique C, and V_C is called a potential.

For simplicity, here we assume state-state interaction is the first order Markov chain. In this special case, the potential V_C only models the interactions of each consecutive state pair and state-observation pair. Therefore, the conditional probability can be derived as

$$P(\mathbf{q}|\mathbf{x}) \propto \exp\left[\mathbf{w} \cdot \sum_t \boldsymbol{\varphi}(q_t, q_{t-1}, \mathbf{x}_t)\right]. \tag{3}$$

The basis functions are the features associated to the relationship of the observable signal and states sequence. In this paper, we define the features as follows:

$$\boldsymbol{\varphi}(q_t, q_{t-1}, \mathbf{x}) = \rho(q_{t-1}, q_t)\boldsymbol{\phi}(q_t, \mathbf{x}_t)$$

where $\rho(q_{t-1}, q_t)$ is an indicator function for the the state transaction and $\boldsymbol{\phi}(q_t, \mathbf{x}_t)$ represents the kernel features of the observation $\mathbf{x}_t$ given the state q_t.

By substituting Eqs. (2) and (3) into Eq. (1), the classification function employed for the signal classification has the logarithm form

$$h(\mathbf{x}) = \arg\max_k \left\{\max_{\mathbf{q}} \left[\mathbf{w}_k \cdot \sum_t \boldsymbol{\varphi}(q_{t-1}, q_t, \mathbf{x}_t)\right]\right\}. \tag{4}$$

3 Maximum Margin Discriminative Learning

Maximum Margin discriminative learning can be used to find the optimal decision surface, increasing the "confidence" of the classification. The basic idea is to construct the decision surface in such a way that the margin between the true class and the best runner-up is maximized [4]. Given the training sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the margin r has the upper bound as follows:

$$r \leq \min_i \left\{ \max_{\mathbf{q} \in \mathcal{Q}|_{y_i}} [\mathbf{w}_{y_i} \cdot \boldsymbol{\varphi}(\mathbf{q}, \mathbf{x}_i)] - \max_{\substack{k \neq y_i \\ \mathbf{q} \in \mathcal{Q}|_k}} [\mathbf{w}_k \cdot \boldsymbol{\varphi}(\mathbf{q}, \mathbf{x}_i)] \right\} \tag{5}$$

where we denote $\boldsymbol{\varphi}(\mathbf{q}, \mathbf{x}) = \sum_t \boldsymbol{\varphi}(q_{t-1}, q_t, \mathbf{x}_t)$ for simplicity.

Unfortunately, it may be difficult to maximize the margin of separation directly. Similar to the support vector machine, this optimization problem is equivalent to minimizing the Euclidean norm of the weight vector $\mathbf{w}$ while keeping the margin $r = 1$. Furthermore, we can also extend the constrained problem to the linearly non-separable case by introducing a new set of nonnegative slack variables $\{\xi_i\}_{i=1}^N$. Therefore, the constrained optimization problem that we have to solve may now be stated as:

Given the training sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, find the optimum values of the weight vector $\mathbf{w}$ such that they satisfy the constraints

$$\forall i, k, \mathbf{q} \qquad \mathbf{w}_{y_i} \cdot \boldsymbol{\varphi}(\hat{\mathbf{q}}_i, \mathbf{x}_i) + \delta_{k,y_i} - \mathbf{w}_k \cdot \boldsymbol{\varphi}(\mathbf{q}, \mathbf{x}_i) \geq 1 - \xi_i$$

and the weight vector $\mathbf{w}$ minimizes the cost function:

$$J(\mathbf{w}) = \frac{1}{2} \sum_k \|\mathbf{w_k}\|_2^2 + C \sum_i \xi_i$$

4 Two-step learning algorithm

Because the underlying stochastic process is not usually observable and thus the optimal state sequence has to be estimated, the constrained optimization problem given in section 3 can not be solved directly using standard quadratic programming (QP) techniques. In this section, we present a two-step learning algorithm for solving the constrained optimization problem. It can be seen that this two-step algorithm is similar to the mathematics of standard Expectation-Maximization (EM) technique [5], although our optimization problem is not directly related to probability estimation.

The EM algorithm is an iterative optimization technique to solve the parameters estimation problem while we are

[1] in Eq. (2), $p(\mathbf{q}|\mathbf{x}) = p(y, \mathbf{q}|\mathbf{x})$ because we can certainly identify every state sequence $\mathbf{q}$ as the unique category.

not given some "hidden" nuisance variables. In particular, an auxiliary function which averages over the values of the hidden variables given the parameters at the previous iteration is defined. By minimizing this auxiliary function, we will always carry out an improvement over the previous estimated parameters, unless finding the optimal values of parameters. In our case, the hidden variables are the most possible state sequences $\hat{\mathbf{q}}_i$. Instead of considering the expected values over the distribution on these unobservable state sequences, we just consider the sequences of states that minimize the cost, given the previous values of the parameters:

$$Q(\mathbf{w},\bar{\mathbf{w}}) \overset{\text{def}}{=} \frac{1}{2}\sum_{k}\|\mathbf{w}_k\|_2^2 + C\sum_{i}\xi_i + \sum_{i,k,\mathbf{q}}\eta_{i,k,\mathbf{q}}Q_1(i,k,\mathbf{q}). \tag{6}$$

The Q_1 has the following form:

$$Q_1(i,k,\mathbf{q}) = \mathbf{w}_k\cdot\boldsymbol{\varphi}(\mathbf{q},\mathbf{x}_i) - \mathbf{w}_{y_i}\cdot\boldsymbol{\varphi}(\hat{\mathbf{q}}_i(\bar{\mathbf{w}}_{y_i}),\mathbf{x}_i) - \delta_{y_i,k} + 1 - \xi_i \tag{7}$$

where $\hat{\mathbf{q}}_i(\bar{\mathbf{w}}_{y_i})$ is the most possible state sequence of the sample $\mathbf{x}_i$ given the previous value of weight $\bar{\mathbf{w}}_{y_i}$.

The next step is to find a new set of weights $\mathbf{w}$ which minimizes $Q(\mathbf{w},\bar{\mathbf{w}})$ where $\bar{\mathbf{w}}$ is the previous set of weights. Accordingly, we may solve this optimization subproblem using Karush-Kuhn-Tucker (KKT) theorem [3]. Due to the limit of space, here we omit the technical details of the derivation. We obtain the optimization subproblem in the dual formulation as follows:

$$\begin{aligned}
&\max_{\alpha}\left\{\begin{aligned}&-\frac{1}{2}\sum_{k}\sum_{i,\mathbf{q}}\sum_{j,\mathbf{q}'}\alpha_{i,k,\mathbf{q}}\alpha_{j,k,\mathbf{q}'}K(\mathbf{q},\mathbf{x}_i,\mathbf{q}',\mathbf{x}_j)\\ &+\sum_{i,k,\mathbf{q}}\alpha_{i,k,\mathbf{q}}\delta_{y_i,k}\end{aligned}\right\}\\
&s.t.\quad \sum_{k,\mathbf{q}}\alpha_{i,k,\mathbf{q}} = 0, \forall i;\quad \alpha_{i,k,\mathbf{q}} \le C\delta_{y_i,k}\delta_{\hat{\mathbf{q}}_i,\mathbf{q}}, \forall i,k,\mathbf{q};
\end{aligned} \tag{8}$$

where $K(\mathbf{q},\mathbf{x}_i,\mathbf{q}',\mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{q},\mathbf{x}_i)\cdot\boldsymbol{\varphi}(\mathbf{q}',\mathbf{x}_j)$. Having determined the optimum Lagrange multipliers, denoted by $\alpha_{i,k,\mathbf{q}}$, we may compute the optimum weights $\mathbf{w}$, yielding:

$$\mathbf{w}_k = \sum_{i}\sum_{\mathbf{q}\in\mathcal{Q}|_k}\alpha_{i,k,\mathbf{q}}\boldsymbol{\varphi}(\mathbf{q},\mathbf{x}_i) \tag{9}$$

where $\mathcal{Q}|_k$ is the subset of state sequences $\mathbf{q}$ which belong to model k.

The algorithm consists of steps of repeatedly replacing $\bar{\mathbf{w}}$ by $\mathbf{w}$ using update Eq. (9) until convergence. Theorem 4.1 guarantees that such an approach will converge in a finite number of iterations to a solution so that the cost function $J(\mathbf{w})$ reaches the minimal point. Due to the limit of space, here we omit the proof.

Theorem 4.1. *Suppose that $\mathbf{w}^{(p)}$ for $p = 0,1,2,\ldots$ is an instance of the two-step learning algorithm such that:*

1. *the sequence $J(\mathbf{w}^{(p)})$ is bounded, and*
2. *$Q(\mathbf{w}^{(p+1)},\mathbf{w}^{(p)}) - Q(\mathbf{w}^{(p)},\mathbf{w}^{(p)}) \le 0$ for all p.*

Then the sequence $\mathbf{w}^{(p)}$ converges to some $\mathbf{w}^*$ in the closure of Ω.

5 Experimental Results

Several signal classification experiments were conducted to study the characteristics of our proposed learning algorithm, and the difference in classification performances as compared to traditional learning method. We report two sets of experimental results here: one involves a set of synthetic data sequences with mixture distributions and the other pertains to the EEG signal related to the motor imagery, well known to the brain computer interfaces community.

In the first experiment, we individually generate two classes of the synthetic data set from two different first-order hidden Markov models. Each model is a left-right model and consists of three states. Every state is modeled as a Gaussian mixture with two components. To evaluate the performance of our method, these two models are slightly different so that there are very big overlap between them.

In order to evaluate our proposed learning method (NPL), we carry out 8 runs of 10-fold cross-validation on the data set containing 1000 samples with 15 time sequences. Each run performs the conventional maximum likelihood (ML) and our proposed learning with different size training set. Fig. 1 shows our experiments on HMM and NPL method. The results show that our proposed learning method outperforms traditional ML learning. Importantly, our proposed method give a quite flat classification accuracy curve after employing 60 training samples, while ML algorithm has a less stable curve. It shows that, compared to conventional ML method, our algorithm need less training samples and has a good generalization performance.

In the second experiment, We evaluate our approach on the classification of EEG signal for motor imagery, to distinguish left and right hand movement imagination [8]. The experiments were performed by a male subject (38 years old).

In our experimental paradigm, the subject was instructed to fixate on a computer screen about in 180 cm 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

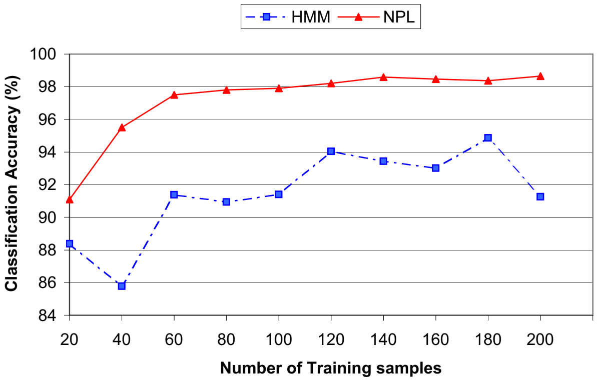

Figure 1. Average classification performance for HMM and our proposed method

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.

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 $900ms$ length with $250ms$ overlap for feature extraction. 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 [12]. Additionally, both HMM and our proposed method consist of 3 states for capturing the structure of EEG data. The kernel function used in SVM and NPL is the RBF kernel [11]. The classification results, shown in table 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 93%, 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.

	SVM	HMM	NPL
Classification accuracy (%)	78	84	93

Table 1. Average classification performance for SVM, HMM and our proposed method.

6 Conclusion

We presented here a non-parametric learning for classifying multi-class temporal signals. The model is capable of both exploiting the temporal dynamics of the signals and maximizing the margins between classes in an effective way, by taking advantage of the rich language of hidden Markov model and superior separability of the kernel techniques. The most important contribution here in this article is the proposed maximum margin discriminative learning method. It was presented with a two-step learning algorithm for constructing the classifier.

The experimental results on synthetic data and real motor imagery EEG signal classification have shown that our proposed algorithm can exploit the nature of sequential signals and significantly outperforms the non-structural methods, and the HMM based parametric methods.

References

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

[2] A. Ben-Yishai and D. Burshtein. A discriminative training algorithm for hidden Markov models. *IEEE Trans. On Speech and Audio Processing*, 12(3):204–217, May 2004.

[3] D. Bertsekas. *Nonlinear Programming*. Athenas Scientific, Belmont, MA, 1995.

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

[5] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[6] B. Juang, W. Chou, and C. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. On Speech and Audio Processing*, 5(3):257–265, May 1997.

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

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

[9] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of The IEEE*, 77:257–286, Feb. 1989.

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

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

[12] W. Xu, C. Guan, E. S. Chng, S. Ranganatha, M. Thulasidas, and J. Wu. High accuracy classification of EEG signal. In *17-th ICPR*, volume 2, pages 391–394, 2004.