Infinite Markov-Switching Maximum Entropy Discrimination Machines

Sotirios P. Chatzis SOTERI0S@ICLOUD.COM
Department of Electrical Engineering, Computer Engineering, and Informatics
Cyprus University of Technology

Abstract

In this paper, we present a method that combines the merits of Bayesian nonparametrics, specifically stick-breaking priors, and large-margin kernel machines in the context of sequential data classification. The proposed model employs a set of (theoretically) infinite interdependent large-margin classifiers as model components, that robustly capture local nonlinearity of complex data. The employed large-margin classifiers are connected in the context of a Markov-switching construction that allows for capturing complex temporal dynamics in the modeled datasets. Appropriate stick-breaking priors are imposed over the component switching mechanism of our model to allow for data-driven determination of the optimal number of component large-margin classifiers, under a standard nonparametric Bayesian inference scheme. Efficient model training is performed under the maximum entropy discrimination (MED) framework, which integrates the large-margin principle with Bayesian posterior inference. We evaluate our method using several real-world datasets, and compare it to state-of-the-art alternatives.

1. Introduction

In this work, we focus on the problem of classifying data with temporal interdependencies by application of large-margin techniques. In the last years, several researchers, inspired from the literature of support vector machines (SVMs), have proposed large-margin methods capable of classifying sequential data under the large-margin paradigm. For example, in Sha & Saul (2007), a large-margin generative model is proposed for sequential data classification; in Altun et al. (2004), an extension of SVMs suitable for structured output prediction is proposed, and is further applied to sequential data classification. The power and popularity of such large-margin approaches stems in part from the fact that their inference and training reduce to convex optimization problems, thus not suffering from the possibility of getting stuck to spurious local optima, which is often the case with alternative approaches. However, learning only a single large-margin model may often be less than sufficient to capture the underlying patterns (e.g., temporal clusters) in modeled data with rich and complex dynamics.

To address this issue, recently, a mixture-of-experts (Collobert et al., 2002; Fu et al., 2010; Zhu et al., 2011) model was proposed that uses a set of SVM classifiers, each one trained to perform modeling in a coherent subregion of the observations space. As such, each one of these classifiers, and, hence, the derived model as a whole, can capture much more subtle underlying patterns than a single SVM expert. However, a drawback of this approach is its complete lack of an explicit mechanism for capturing temporal dynamics in sequential data, encapsulated in the context of an appropriate component switching mechanism.

Inspired by these advances, in this paper we propose a Markov-switching mixture of large-margin classifiers for sequential data classification. A first basic concept underlying our approach is that, in data with temporal dynamics, one large-margin classifier is not enough for capturing rich underlying temporal structures; therefore, use of a set of local experts is needed. A second key-concept of our approach that differentiates it from existing approaches consists in the introduction of an appropriate mechanism describing how subsequent observations may belong to different subregions in the considered observations space. Indeed, simply considering that the data are generated from these subregions as draws from independent distribu-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

tions is not expected to allow for effective modeling. Rather, one would expect that such subregions could be interpreted as temporal states or subpatterns in the modeled data; therefore, transition from one state to another should be described by an appropriate model of temporal dependencies. To account for these facts, in this work we employ a latent first-order Markov chain to capture the temporal dynamics of the allocation of successive observations to the postulated model component large-margin classifiers.

A challenge in the field of Markov-switching models consists in the data-driven determination of the number of their latent states (model components) required to represent the modeled data (model order). The most common data-driven methodologies for model order selection are based on the popular Bayesian information criterion (BIC) or other related model size selection criteria (McLachlan & Peel, 2000). However, such model selection methods require training of multiple models (to select from), a procedure which can be applied only up to a limited extent, due to its computational demands. In addition, they are also well-known for their overfitting proneness, hence often leading to models much larger than necessary (McLachlan & Peel, 2000).

Nonparametric Bayesian modeling techniques, especially Dirichlet process (DP) prior-based models, have become very popular in statistics over the last few years, for performing nonparametric density estimation (Walker et al., 1999; Neal, 2000; Muller & Quintana, 2004). Briefly, a realization of a DP prior-based model can be seen as an infinite mixture of distributions with given parametric shape (e.g., Gaussian, HMM, etc.). Indeed, although theoretically a DP prior gives rise to an infinite number of parameters for the model, it turns out that inference for the model is possible, since only the parameters of a finite number of model components need to be represented explicitly (Neal, 2000; Antoniak, 1974).

Motivated by these results, formulation of our model is based on the introduction of appropriate nonparametric priors over the employed large-margin component classifiers of our model. Specifically, we utilize appropriate stick-breaking priors under a truncated nonparametric Bayesian inference scheme (Sethuraman, 1994). This way, our model combines the advantages of Bayesian nonparametrics to allow for automatic, data-driven determination of the appropriate number of model components (states), and large-margin classifiers to capture local nonlinearity in the context of a convex optimization scheme, not suffering from getting trapped into spurious local optima.

To perform inference, we employ the maximum entropy discrimination (MED) framework (Jaakkola et al., 1999), which integrates the large-margin principle with Bayesian posterior inference in an elegant and computationally efficient fashion, allowing to leverage existing high-performance techniques for DP and SVM models. We dub our approach the infinite Markov-switching maximum entropy discrimination machine (iM^{2}EDM) for sequential data classification.

The remainder of this work is organized as follows: In Section 2, we briefly review the theoretical background of our method, namely DP priors and the MED framework. In Section 3, we introduce the iM^{2}EDM approach, and derive its training and inference algorithms. In Section 4, we perform experimental evaluations, considering several applications dealing with semantic classification in real-world video sequences. In the final section of this paper, we summarize and discuss our results.

2. Theoretical Background

2.1. Dirichlet process models

DP models were first introduced in Ferguson (1973). A DP is characterized by a base distribution G_0 and a positive scalar α, usually referred to as the innovation parameter, and is denoted as $\mathrm{DP}(G_0, \alpha)$. Essentially, a DP is a distribution placed over a distribution. Let us suppose we randomly draw a sample distribution G from a DP, and, subsequently, we independently draw N random variables $\{\Theta_n^*\}_{n=1}^N$ from G:

$$G|\{G_0, \alpha\} \sim \mathrm{DP}(G_0, \alpha) \tag{1}$$

$$\Theta_n^*|G \sim G, \quad n = 1, \dots N \tag{2}$$

Integrating out G, the joint distribution of the variables $\{\Theta_n^*\}_{n=1}^N$ can be shown to exhibit a clustering effect. Specifically, given the first $N-1$ samples of G, $\{\Theta_n^*\}_{n=1}^{N-1}$, it can be shown that a new sample Θ_N^* is either (a) drawn from the base distribution G_0 with probability $\frac{\alpha}{\alpha+N-1}$, or (b) is selected from the existing draws, according to a multinomial allocation, with probabilities proportional to the number of the previous draws with the same allocation (Blackwell & MacQueen, 1973).

A characterization of the (unconditional) distribution of the random distribution G drawn from a Dirichlet process $\mathrm{DP}(G_0, \alpha)$ is provided by the *stick-breaking construction* of Sethuraman (1994). Consider two infinite collections of independent random variables $\boldsymbol{v} = (v_c)_{c=1}^{\infty}$, $\{\Theta_c\}_{c=1}^{\infty}$, where the v_c are drawn from

the Beta distribution $\text{Beta}(1,\alpha)$, and the Θ_c are independently drawn from the base distribution G_0. The stick-breaking representation of G is then given by (Sethuraman, 1994)

$$G = \sum_{c=1}^{\infty} \pi_c(\boldsymbol{v}) \delta_{\Theta_c} \tag{3}$$

where

$$\pi_c(\boldsymbol{v}) = v_c \prod_{j=1}^{c-1} (1 - v_j) \quad \in [0,1] \tag{4}$$

$$v_c|\alpha \sim \text{Beta}(1,\alpha) \tag{5}$$

and

$$\sum_{c=1}^{\infty} \pi_c(\boldsymbol{v}) = 1 \tag{6}$$

Note that, typically, due to the significant effect of the innovation parameter α on the internal data allocation mechanism of the DP, an appropriate prior is also imposed over α in the context of model inference. Usually, a conjugate Gamma prior is imposed, s.t.

$$p(\alpha) = \mathcal{G}(\alpha|\eta_1,\eta_2) \tag{7}$$

2.2. Maximum entropy discrimination

Let us denote as $\boldsymbol{x} \in \mathbb{R}^d$ an observation vector from a modeled input space, and as y the classification label assigned to it, taking values from the finite set $\{1,\ldots,L\}$. Let us also consider a large-margin classifier for this problem, and denote as $F(y,\boldsymbol{x};\boldsymbol{w})$ its discriminant function with parameter vector $\boldsymbol{w}$. In conventional large-margin classifiers, such as SVMs, a point-estimate is obtained for the parameter vector $\boldsymbol{w}$ by resolving a (typically convex) constrained optimization problem. A major drawback of such point-estimates is their lack of a direct probabilistic interpretation. As a consequence, such approaches prove to underperform, and be vulnerable to input noise, since point-estimates cannot account for the uncertainty in the modeled datasets.

MED is a method that allows for resolving these issues of conventional large-margin classifiers, by learning a distribution $q(\boldsymbol{w})$ obtained as the solution of the following entropic regularized risk minimization problem (Jaakkola et al., 1999)

$$\min_{q(\boldsymbol{w})} \text{KL}\left(q(\boldsymbol{w})||p(\boldsymbol{w})\right)) + \gamma\, \mathcal{R}(q(\boldsymbol{w})) \tag{8}$$

where $p(\boldsymbol{w})$ is a prior imposed over the parameter vector $\boldsymbol{w}$, γ is a positive regularization constant, and $\text{KL}\left(q||p\right))$ stands for the Kullback-Leibler divergence between $q(\boldsymbol{w})$ and $p(\boldsymbol{w})$. $\mathcal{R}(q(\boldsymbol{w}))$ is the hinge-loss function; it encodes the large-margin principle underlying the considered classifier, and is defined as

$$\mathcal{R}(q(\boldsymbol{w})) \triangleq \sum_n \max_y \Big(\delta_n(y) + \langle F(y,\boldsymbol{x}_n;\boldsymbol{w}) - F(y_n,\boldsymbol{x}_n;\boldsymbol{w})\rangle_{q(\boldsymbol{w})}\Big) \tag{9}$$

where $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ is the set of available training examples (input/output pairs), $\delta_n(y)$ is usually defined as a binary function equal to one if the computed output y and the true (training) label y_n are different, and $\langle\cdot\rangle_{q(\cdot)}$ is the expectation of a quantity with respect to the posterior distribution $q(\cdot)$.

Finally, under the MED framework, the prediction rule of the derived large-margin classifier becomes

$$y^* = \arg\max_y \langle F(y,\boldsymbol{x};\boldsymbol{w})\rangle_{q(\boldsymbol{w})} \tag{10}$$

essentially utilizing the expectation of the employed discriminant function with respect to the parameters posterior $q(\boldsymbol{w})$ to obtain predictions in a way similar to conventional large-margin approaches.

The Bayesian-style formulation of MED renders it an elegant means of integrating the ideas of large-margin learning and Bayesian generative modeling, and includes SVM-type models as a special case. In addition, MED allows for incorporating latent variables in the derived models (Lewis et al., 2006), which comprise a key-tool in machine learning, as well as for performing structured output prediction (Zhu & Xing, 2009).

3. Proposed Approach

3.1. Model Formulation

As we have already discussed, using a single global MED/large-margin classifier to model the complex underlying patterns in sequential observations is rather unlikely to yield models with satisfactory recognition performance. In addition, existing mixture-of-expert approaches (e.g., Fu et al., 2010; Zhu et al., 2011) are not designed for handling sequential data with underlying temporal patterns and dynamics. Apparently, in such a setting, the temporally dependent observations cannot be accurately modeled as draws from independent distributions.

Under these considerations, we introduce the iM^{2}EDM method; it comprises a set of large-margin classifiers, each one fitted to capture complex structure in a subpart of the observations space (model state). The pattern under which successive observations are generated from different model states is captured by means of a latent first-order Markov chain that interconnects

the component large-margin classifiers of the model (states).

Let us consider an L-class classification problem, with class variables $y \in \{1, \ldots, L\}$, and M-dimensional input observations $\boldsymbol{x} \in \mathbb{R}^M$. Derivation of our model commences by introducing a latent Markov chain comprising infinite (latent) states, and considering that each modeled observation is associated with one latent model state. Let $s_t \in \{1, \ldots, \infty\}$ be a latent variable denoting the model state that generates the tth pair of input/output observations $\{\boldsymbol{x}_t, y_t\}$. To obtain an appropriate prior construction, we impose suitable stick-breaking priors over the latent state variables s_t, following the discussions of Section 2.1. Specifically, we impose stick-breaking priors over the latent state transitions in the Markov chain, of the form

$$p(s_t = j | s_{t-1} = i; \varpi_{ij}) = \varpi_{ij}(\boldsymbol{v}^{\varpi}), \; t > 1 \tag{11}$$

where the $\varpi_{ij}(\boldsymbol{v}^{\varpi})$ are the probabilities generated by a stick-breaking process with stick-variables $\boldsymbol{v}^{\varpi}$, such that $\boldsymbol{v}^{\varpi} = (v_{ij}^{\varpi})_{i,j=1}^{\infty}$

$$\varpi_{ij}(\boldsymbol{v}^{\varpi}) = v_{ij}^{\varpi} \prod_{k=1}^{j-1} (1 - v_{ik}^{\varpi}) \tag{12}$$

$$v_{ij}^{\varpi} \sim \text{Beta}(1, \alpha_i^{\varpi}) \tag{13}$$

$$\alpha_i^{\varpi} \sim \mathcal{G}(\eta_1, \eta_2) \tag{14}$$

and

$$\sum_{j=1}^{\infty} \varpi_{ij}(\boldsymbol{v}^{\varpi}) = 1, \; \forall i \tag{15}$$

Similar, we impose a stick-breaking prior for the initial state prior probabilities π_i of the latent Markov chain, such that $\boldsymbol{v}^{\pi} = (v_i^{\pi})_{i=1}^{\infty}$

$$p(s_1 = i | \pi_i) = \pi_i(\boldsymbol{v}^{\pi}) \tag{16}$$

$$\pi_i(\boldsymbol{v}^{\pi}) = v_i^{\pi} \prod_{k=1}^{i-1} (1 - v_k^{\pi}) \tag{17}$$

$$v_i^{\pi} \sim \text{Beta}(1, \alpha^{\pi}) \tag{18}$$

$$\alpha^{\pi} \sim \mathcal{G}(\varepsilon_1, \varepsilon_2) \tag{19}$$

and

$$\sum_{i=1}^{\infty} \pi_i(\boldsymbol{v}^{\pi}) = 1 \tag{20}$$

Subsequently, on the basis of this construction, and conditional on the latent Markov chain states generating each observation, we employ a set of conditional discriminant functions for our model of the form

$$F(y_t, \boldsymbol{x}_t | s_t = c; W) = \boldsymbol{w}_c' \boldsymbol{f}(y_t, \boldsymbol{x}_t) \tag{21}$$

where we let $W = \{\boldsymbol{w}_c\}_{c=1}^{\infty}$, and denote as $\boldsymbol{f}(y, \boldsymbol{x})$ an ML-dimensional vector comprising L subvectors, with the yth one being equal to $\boldsymbol{x}$, and all the others equal to $\mathbf{0}$. Each one of the used discriminant functions $F(y_t, \boldsymbol{x}_t | s_t = c; W)$ captures complex non-linearities in a subpart of the observations space, related to an underlying subpattern in the modeled data, and associated with the corresponding (cth) latent model state. Regarding the model parameters $\boldsymbol{w}_c$, we choose to impose a standard Gaussian prior over them, of the form

$$p(\boldsymbol{w}_c) = \mathcal{N}(\boldsymbol{w}_c | \mathbf{0}, \boldsymbol{I}), \; \forall c \tag{22}$$

The above prescribed prior construction, given by Eqs. (11)-(22), defines the proposed iM^{2}EDM model. Then, if we consider a pair of input/output observation sequences $\{X, Y\}$, with $X = \{\boldsymbol{x}_t\}_{t=1}^T$ and $Y = \{y_t\}_{t=1}^T$, the overall discriminant function of iM^{2}EDM yields

$$\begin{aligned} F(Y, X) = & \sum_{i=1}^{\infty} q(s_1 = i) \, \langle \boldsymbol{w}_i' \rangle_{q(\boldsymbol{w}_i)} \, \boldsymbol{f}(y_1, \boldsymbol{x}_1) \\ & + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{t=2}^{T} q(s_t = j | s_{t-1} = i) \, \langle \boldsymbol{w}_j' \rangle_{q(\boldsymbol{w}_j)} \, \boldsymbol{f}(y_t, \boldsymbol{x}_t) \end{aligned} \tag{23}$$

Finally, similar to the discussions of Section 2.2, the prediction rule for our iM^{2}EDM model, with discriminant function (23), eventually yields

$$Y^* = \arg\max_{Y} F(Y, X) \tag{24}$$

3.2. Model Training

To learn the optimal (approximate) posterior distributions over the model parameters, i.e., W, $\boldsymbol{v}^{\pi}$, $\boldsymbol{v}^{\varpi}$, and $\boldsymbol{\alpha} = \{\{\alpha_i^{\varpi}\}_i, \alpha^{\pi}\}$, as well as over the latent variables of state allocation $S = \{s_t\}_{t=1}^T$, we resort to the MED framework. For this purpose, we need to introduce an appropriate loss function for the prediction rule (24) of our model. Following the discussions of Section 2.2, we introduce a hinge-loss function, yielding

$$\mathcal{R}(q(S, W)) \triangleq \max_{Y^*} \big[\sum_{t=1}^{T} \delta_t(y_t^*) + F(Y^*, X) - F(Y, X) \big] \tag{25}$$

where $\sum_{t=1}^T \delta_t(y_t^*)$ is equivalent to the *Hamming distance* between the estimated labels sequence Y^* and the correct one Y. Then, based on the principles of the MED framework, our learning problem reduces to solving the following entropic regularized risk minimization problem

$$\min_{q(\Psi)} \text{KL}\left(q(\Psi) || p(\Psi)\right)) + \gamma \, \mathcal{R}(q(S, W)) \tag{26}$$

where we denote $\Psi \triangleq \{W, S, \boldsymbol{v}^{\varpi}, \boldsymbol{v}^{\pi}, \boldsymbol{\alpha}\}$.

To solve (26), we further assume that the sought (approximate) posterior factorizes similar to the considered prior $p(\Psi)$: $q(\Psi) = q(W)q(S)q(\boldsymbol{v}^{\varpi})q(\boldsymbol{v}^{\pi})q(\boldsymbol{\alpha})$. Under this assumption, usually referred to as the mean-field approximation (Chandler, 1987; Winn & Bishop, 2005), our learning problem eventually becomes

$$\begin{aligned}\min_{q(W),q(S),q(\boldsymbol{v}^{\varpi}),q(\boldsymbol{v}^{\pi}),q(\boldsymbol{\alpha})} \Big\{ & \mathrm{KL}\left(q(W)||p(W)\right) \\ & + \langle \log q(S) \rangle_{q(S)} + \mathrm{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) \\ & + \langle \log q(\boldsymbol{v}^{\pi}) \rangle_{q(\boldsymbol{v}^{\pi})} + \langle \log q(\boldsymbol{v}^{\varpi}) \rangle_{q(\boldsymbol{v}^{\varpi})} \\ & - \langle \log p(S, \boldsymbol{v}^{\varpi}, \boldsymbol{v}^{\pi} | \boldsymbol{\alpha}) \rangle_{q(S),q(\boldsymbol{v}^{\pi}),q(\boldsymbol{v}^{\varpi}),q(\boldsymbol{\alpha})} \\ & + \gamma\, \mathcal{R}(q(S,W)) \Big\}\end{aligned} \tag{27}$$

Apparently, under the infinite dimensional setting of our model, optimization of the risk function (27) is intractable, as it entails an infinite number of optimized factors. To resolve this issue, and render our model training algorithm computationally feasible, we truncate the imposed stick-breaking priors (Blei & Jordan, 2006): we fix a value C and let the posteriors over the v_{ij}^{ϖ} and the v_i^{π} have the property $q(v_{iC}^{\varpi} = 1) = 1,\ \forall i$, and $q(v_C^{\pi} = 1) = 1$. In other words, we set $\pi_c(\boldsymbol{v}^{\pi})$ and $\varpi_{ic}(\boldsymbol{v}^{\varpi})$ equal to zero for $c > C$. Note that, under this setting, our iM^{2}EDM model still involves a full stick-breaking prior; truncation is not imposed on the model itself, but only on its (approximate) posterior distribution to allow for a tractable inference procedure. Hence, the truncation level C is a free parameter, and not part of the prior model specification.

Posteriors over the model parameters. Solving problem (27), the posterior over $\boldsymbol{w}_c$ yields

$$q(\boldsymbol{w}_c) = \mathcal{N}(\boldsymbol{w}_c | \boldsymbol{\mu}_c, \boldsymbol{I}) \tag{28}$$

where

$$\boldsymbol{\mu}_c = \sum_t q(s_t = c) \sum_y \lambda_t^y \left[\boldsymbol{f}(y_t, \boldsymbol{x}_t) - \boldsymbol{f}(y, \boldsymbol{x}_t) \right] \tag{29}$$

where the multipliers λ_t^y are computed by resolving the dual quadratic programming problem

$$\begin{aligned}\max_{\boldsymbol{\lambda}} \ & -\frac{1}{2} \sum_c \boldsymbol{\mu}_c' \boldsymbol{\mu}_c + \sum_t \sum_y \lambda_t^y \delta_t(y) \\ \text{s.t.} \ & 0 \leq \sum_y \lambda_t^y \leq \gamma,\ \forall t\end{aligned} \tag{30}$$

Similar, regarding the stick-breaking variables, we have

$$q(v_{ij}^{\varpi}) = \mathrm{Beta}(\tilde{\beta}_{ij}^{\varpi}, \hat{\beta}_{ij}^{\varpi}) \tag{31}$$

where

$$\tilde{\beta}_{ij}^{\varpi} = 1 + \sum_{t=2}^{T} q(s_t = j | s_{t-1} = i) \tag{32}$$

$$\hat{\beta}_{ij}^{\varpi} = \langle \alpha_i^{\varpi} \rangle_{q(\alpha_i^{\varpi})} + \sum_{\varrho=j+1}^{C} \sum_{t=2}^{T} q(s_t = \varrho | s_{t-1} = i) \tag{33}$$

and

$$q(v_i^{\pi}) = \mathrm{Beta}(\tilde{\beta}_i^{\pi}, \hat{\beta}_i^{\pi}) \tag{34}$$

where

$$\tilde{\beta}_i^{\pi} = 1 + q(s_1 = i) \tag{35}$$

$$\hat{\beta}_i^{\pi} = \langle \alpha^{\pi} \rangle_{q(\alpha^{\pi})} + \sum_{\varrho=i+1}^{C} q(s_1 = \varrho) \tag{36}$$

Finally, the innovation parameters yield

$$q(\alpha_i^{\varpi}) = \mathcal{G}(\alpha_i^{\varpi} | \tilde{\eta}_i^{\varpi}, \hat{\eta}_i^{\varpi}) \tag{37}$$

where

$$\tilde{\eta}_i^{\varpi} = \eta_1 + C - 1 \tag{38}$$

$$\hat{\eta}_i^{\varpi} = \eta_2 - \sum_{j=1}^{C-1} \left[\psi(\hat{\beta}_{ij}^{\varpi}) - \psi(\tilde{\beta}_{ij}^{\varpi} + \hat{\beta}_{ij}^{\varpi}) \right] \tag{39}$$

and

$$q(\alpha^{\pi}) = \mathcal{G}(\alpha^{\pi} | \tilde{\varepsilon}^{\pi}, \hat{\varepsilon}^{\pi}) \tag{40}$$

where

$$\tilde{\varepsilon}^{\pi} = \varepsilon_1 + C - 1 \tag{41}$$

$$\hat{\varepsilon}^{\pi} = \varepsilon_2 - \sum_{i=1}^{C-1} \left[\psi(\hat{\beta}_i^{\pi}) - \psi(\tilde{\beta}_i^{\pi} + \hat{\beta}_i^{\pi}) \right] \tag{42}$$

Posteriors over the latent variables. Minimizing the criterion (27) w.r.t. $q(S)$, we obtain

$$q(S) = \frac{1}{Q} \pi_{s_1}^* \prod_{t=1}^{T-1} \varpi_{s_t s_{t+1}}^* \prod_{t=1}^{T} p^*(y_t, \boldsymbol{x}_t | s_t) \tag{43}$$

where

$$\pi_c^* \triangleq \exp\left[\langle \log \pi_c(\boldsymbol{v}^{\pi}) \rangle_{q(\boldsymbol{v}^{\pi})} \right] \tag{44}$$

$$\varpi_{s_t s_{t+1}}^* \triangleq \exp\left[\langle \log \varpi_{s_t s_{t+1}}(\boldsymbol{v}^{\varpi}) \rangle_{q(\boldsymbol{v}^{\varpi})} \right] \tag{45}$$

$$p^*(y_t, \boldsymbol{x}_t | s_t = c) \triangleq \exp\Big(\sum_y \lambda_t^y \boldsymbol{\mu}_c' \left[\boldsymbol{f}(y_t, \boldsymbol{x}_t) - \boldsymbol{f}(y, \boldsymbol{x}_t) \right] \Big) \tag{46}$$

and Q is a normalizing constant. From (46), it follows that the expression of $q(S)$ for our model is analogous to the expression of $q(S)$ pertaining to a first-order hidden Markov model (HMM) (McLachlan & Peel, 2000), with initial state prior probabilities equal to π_c^*, state-transition priors equal to ϖ_{ij}^*, and state-conditional

likelihoods equal to $p^*(y_t, \boldsymbol{x}_t|s_t = c)$. Therefore, computation of both the state-transition posteriors $q(s_t = j|s_{t-1} = i)$, $\forall t > 1$, and $\forall i, j$, as well as the state assignment posteriors $q(s_t = c), \forall c, t$, of the iM^{2}EDM can be performed by considering the analogous (proxy) HMM described previously, and running the *forward-backward algorithm* (Rabiner, 1989) for that HMM. This way, we obtain the sought latent variable posteriors for our model in an elegant and computationally efficient manner.

3.3. Prediction Generation

Given a trained iM^{2}EDM, when a new test sequence $X = \{\boldsymbol{x}_t\}_{t=1}^T$ is provided, the prediction task consists in computing the optimal corresponding class labels sequence $Y = \{y_t\}_{t=1}^T$, by application of the prediction rule (24).

For this to happen, we need to compute the initial state posteriors $q(s_1 = i)$ and the state-transition posteriors $q(s_t = j|s_{t-1} = i)$ for the test sequence X'. This can be conducted in a fashion similar to the MED training algorithm of our model: specifically, to obtain the sought posteriors, we consider the regularized risk minimization problem

$$\min_{q(S)} \mathrm{KL}\left(q(S)||p(S)\right)) + \gamma\, \mathcal{R}(q(S, W)) \tag{47}$$

which is clearly analogous to the approach we used for model training, and yields exactly the same posterior expressions for $q(S)$ as those obtained in Section 3.2.2.

However, a careful inspection of Eqs. (43) and (46) shows that computation of these posteriors, $q(S)$, requires knowledge of the class labels Y, which are the unknown sought quantities of our prediction algorithm. Therefore, application of the prediction rule (24) of the iM^{2}EDM yields a computationally cumbersome dynamic programing optimization procedure.

To ameliorate these drawbacks, we devise an alternative approximate algorithm for prediction generation under the proposed iM^{2}EDM method. Our approximation, inspired by the mean-field principle (Winn & Bishop, 2005; Chandler, 1987), and the point-pseudo-likelihood technique of Qian & Titterington (1991a;b), is an iterative algorithm comprising the following steps:

1. Initially, an approximate estimate of Y is obtained by resolving (24), with the posteriors $q(S)$ in (23) replaced with the corresponding posterior expectations of the prior configurations $\pi_c(\boldsymbol{v}^\pi)$ and $\varpi_{ij}(\boldsymbol{v}^\varpi)$. In other words, we optimize (24) by approximating $F(Y, X)$ with

$$F(Y,X) \approx \sum_{i=1}^{\infty} \langle \pi_i(\boldsymbol{v}^\pi) \rangle_{q(\boldsymbol{v}^\pi)} \langle \boldsymbol{w}'_i \rangle_{q(\boldsymbol{w}_i)} \boldsymbol{f}(y_1, \boldsymbol{x}_1) + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{t=2}^{T} \langle \varpi_{ij}(\boldsymbol{v}^\varpi) \rangle_{q(\boldsymbol{v}^\varpi)} \langle \boldsymbol{w}'_j \rangle_{q(\boldsymbol{w}_j)} \boldsymbol{f}(y_t, \boldsymbol{x}_t) \tag{48}$$

   Note that (48) yields a greedy optimization procedure.

2. Using the so-obtained estimates Y, we compute the posteriors $q(S)$ using the proxy HMM described previously.

3. We run the optimization process (24), with the $F(Y, X)$ given now by the exact expression (23), and considering the posteriors $q(S)$ as known quantities (obtained in the previous step). This way, the $q(S)$ are removed from the optimization of (23); hence, the initial dynamic programming problem reduces to a simple greedy optimization, similar to the one induced by (48).

4. We repeat steps 2 and 3 until convergence.

3.4. Relation to existing approaches

Our method follows the MED paradigm, which combines the ideas of large-margin classification and Bayesian inference techniques. It also exploits the merits of Bayesian nonparametrics, to allow for automatic model order determination. In that sense, our approach is related to the iSVM approach (Zhu et al., 2011); iSVM is an infinite mixture model of large-margin (SVM) classifiers. The inference algorithm of our model shares several common steps with the variational algorithm of iSVM. On the other hand, iSVM employs a likelihood model, while the proposed model strictly follows the MED framework, without a data likelihood model. In this latter aspect, our method shares similar concept with the nonparametric MED model for matrix factorization of Xu et al. (2012).

4. Experiments

In the following, we experimentally evaluate our approach using real-world datasets. We compare the performance of our approach to large-margin HMMs (LM) (Sha & Saul, 2007), moderate-order CRFs of 5th order (5-CRF) (Ye et al., 2009), the hidden Markov support vector machine (HMSVM) approach of Altun et al. (2004), and the iSVM approach of Zhu et al. (2011) with RBF kernels. Our source codes were developed in MATLAB R2012a.

Table 1. Sports Video Mining: Recognition rates (%).

Class	5-CRF	iSVM	LM	HMSVM	iM^{2}EDM
long play	71.46	64.19	70.62	72.65	74.58
short play	74.63	61.88	70.28	73.05	77.71
kick	73.88	66.36	75.77	76.44	79.55
field goal play	74.35	69.12	71.39	73.17	77.49
central-view	73.12	68.35	71.68	73.40	78.44
left-view	73.83	66.40	71.24	70.09	76.09
right-view	74.09	64.84	70.43	72.43	77.62
end-zone-view	75.55	70.01	75.06	74.66	79.03

Figure 1. Sports Video Mining: Few example frames.

4.1. Sports Video Mining

In this experiment, we consider the problem of sports video mining in football videos. We follow the experimental setup of (Ding & Fan, 2009). We detect four camera view classes, namely central, left, right, and end-zone, and four play types, namely long play, short play, kick, and field goal play.

Feature Extraction. To capture camera view information, we use the color distribution and the yard line angle (Dingand & Fan, 2007). For this purpose, we estimate the spatial color distribution, and perform edge detection using the Canny algorithm, which we combine with the Hough transform to detect the yard lines and to compute their angles. Regarding play type information extraction, we utilize for this purpose camera motion information (panning and tilting), as this information is sufficient to characterize different play types: strong panning is usually associated with a long play, while a weak panning effect is usually associated with short plays (Ding & Fan, 2009). To compute the two kinds of camera motion, we choose the optical flow-based method of Srinivasan et al. (1997).

Experimental Setup and Results. We evaluate the efficacy of the proposed approach by using a database comprising twelve 30-min NFL American football games. The videos are of 720 × 576 resolution, and were preprocessed so as to remove commercials and replays. As such, from each video, a series of play shots was obtained, with each video being typically segmented into 138–189 shots, and each shot comprising 600–900 frames. We provide few example frames of the used videos in Fig. 1.

Table 2. Depth image sequence segmentation experiments: Error rates obtained by the evaluated methods.

Method	Mean Error Rate (%)	p-value
5-CRF	27.13	10^{-9}
LM	27.56	10^{-9}
HMSVM	26.80	10^{-8}
iSVM	30.41	10^{-9}
iM^{2}EDM	23.16	

From this raw data, we extract the feature descriptors presented previously, and use them to train and evaluate the considered models. We use cross-validation in the following fashion: in each cycle, we use 75% of the available shots for training, and the rest for testing. We run the same experiment 10 times, using each time different splits of the available video shots into training and test sets, to account for the effect of random selection of samples.

In Table 1, we provide the obtained performances of the evaluated algorithms (average results per detected class over the conducted 10 repetitions of our experiments). As we observe, our method works clearly better than all the considered approaches. Further, exploiting the availability of multiple performance measurements for the evaluated algorithms (over 10 experiment repetitions), we evaluate the statistical significance of the obtained average performance differences using the Student's-t test. Generated p-values of the Student's-t test below 0.05 strongly indicate that the average performance statistics of two compared methods provide a very good assessment of their actual performance difference. Running the Student's-t test, we obtained p-values ranging from 10^{-4} in the case of the 5-CRF method (compared against our method) to 10^{-9} in the case of the iSVM; thus, the Student's-t test found that the obtained performance differences between our method and its competitors are strongly statistically significant in all cases.

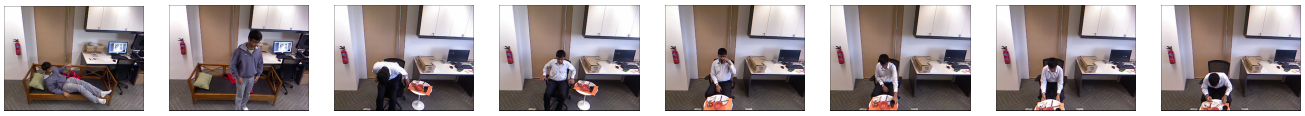

Figure 2. Depth image sequence segmentation experiments: Some characteristic frames.

4.2. Activity recognition in depth image sequences

In this experiment, we evaluate our method in segmenting and classifying depth image sequences, which depict humans performing actions in an assistive living environment. More specifically, our experiments are based on the dataset described in Ni et al. (2011). This dataset includes several actions from which we have selected the following: (1) get up from bed, (2) go to bed, (3) sit down, (4) eat meal, and (5) drink water. Some example frames from the considered dataset are depicted in Fig. 2. We seek to recognize these actions (1)-(5), using as our observable input the sequence of vectors $\boldsymbol{x}$ extracted as described next.

From this dataset, we extract features similar to Ni et al. (2011), by computing motion history images along the depth change directions. To calculate depth change, we use depth maps computed by a Kinect$^{\text{TM}}$ device. Kinect depth maps, however, contain a significant amount of noise. After frame differencing and thresholding, we noticed that motion was encoded even in areas with only still objects. To tackle this problem, we use median filtering. In the temporal domain, each pixel value is replaced by the minimum of its neighbors. Eventually, from these motion history images, we extract the first 12 complex Zernike coefficients (both norm and angle) (Kosmopoulos & Chatzis, 2010), and use them as our feature vectors.

In our experiments, each action is contained in 35 video sequences. Each one of these sequences, derived from the dataset of Ni et al. (2011), contains at least two of the considered actions (sequentially appearing). This setting enables us to assess the capacity of the evaluated algorithms to recognize these actions in real-world activities (in an assistive living environment). We subsample these video sequences by a factor of 2, similar to Ni et al. (2011). We use cross-validation in the following fashion: in each cycle, we use 15 randomly selected video sequences to perform training, and keep the rest 20 for testing. We run the same experiment 50 times to account for the effect of random selection of samples. Recognition consists in assigning each feature vector to a corresponding action class. We provide the obtained average performance results (mean obtained error) over the conducted experiment repetitions in Table 2, where we also illustrate the obtained p-values of the Student's-t test. As we observe, our method obtains competitive results, yielding statistically-significant performance differences over the considered alternative methods.

4.3. Discussion on Computational Complexity

The computational costs of both the training and prediction algorithms of our model are comparable to those of iSVM. This is due to the very efficient nature of the forward-backward algorithm used in model training, and the approximation of the original predictive functional of our model by treating $q(S)$ as a known, iteratively updated quantity. Note also that, in all our experiments, the iterative prediction algorithm of our model converged in less than 10 repetitions.

5. Conclusions and Future Work

In this paper, we presented a Markov switching model comprising an infinite set of component large-margin classifiers for sequential data. Our model is capable of capturing subtle temporal patterns underlying sequential data observations; further, by leveraging the strengths of Bayesian nonparametrics, specifically stick-breaking priors, it allows for data-driven determination of the appropriate number of component large-margin classifiers. Model training and inference was made possible by utilizing the MED framework in the context of an efficient truncated representation of the stick-breaking process. We illustrated the efficacy of our approach using two real-wold datasets, and comparing its performance to state-of-the-art alternatives.

Future goals in this line of research comprise imposing kernel functions on the input observations $\boldsymbol{x}$, instead of the linear construction of the component-wise discriminant functions $F(y, \boldsymbol{x}|s)$, implied by the way we have defined the auxiliary functions $f(y, \boldsymbol{x})$ in Section 3.1. This development will allow for our method to handle cases where the nature of the modeled sequential observations is not vectorial (feature vectors), but graphs, trees, or other types of structured input.

We shall provide demos of our method at: http://www.cut.ac.cy/eecei/staff//sotirios.chatzis.

References

Altun, Yasemin, Tsochantaridis, Ioannis, and Hofmann, Thomas. Hidden Markov support vector machines. In *Proc. ICML*, 2004.

Antoniak, C. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

Blackwell, D. and MacQueen, J. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Blei, David M. and Jordan, Michael I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

Chandler, D. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, 1987.

Collobert, R., Bengio, S., and Bengio, Y. A parallel mixture of SVMs for very large scale problems. In *Proc. NIPS*, 2002.

Ding, Y. and Fan, G. Sports video mining via multichannel segmental hidden markov models. *IEEE Trans. on Multimedia*, 11(7):1301–1309, 2009.

Dingand, Y. and Fan, G. Segmental hidden Markov models for view-based sport video analysis. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2007.

Ferguson, T. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

Fu, Z., Robles-Kelly, A., and Zhou, J. Mixing linear SVMs for nonlinear classification. *IEEE Trans. on Nueral Networks*, 21(12):1963 – 1975, 2010.

Jaakkola, T., Meila, M., and Jebara, T. Maximum entropy discrimination. In *Proc. NIPS*, 1999.

Kosmopoulos, D. and Chatzis, S.P. Robust visual behavior recognition. *Signal Processing Magazine, IEEE*, 27(5):34 –45, sept. 2010.

Lewis, D., Jebara, T., and Noble, W. Nonstationary kernel combination. In *Proc. ICML*, 2006.

McLachlan, G. and Peel, D. *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, 2000.

Muller, P. and Quintana, F. Nonparametric Bayesian data analysis. *Statist. Sci.*, 19(1):95–110, 2004.

Neal, R. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.*, 9:249–265, 2000.

Ni, Bingbing, Wang, Gang, and Moulin, Pierre. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *ICCV Workshops*, pp. 1147–1153, 2011.

Qian, W. and Titterington, D.M. Estimation of parameters in hidden Markov models. *Philos. Trans. R. Soc. London Ser. A*, 337:407–428, 1991a.

Qian, W. and Titterington, D.M. Stochastic relaxations and EM algorithms for Markov random fields. *J. Statist. Comput. Simul.*, 40:55–69, 1991b.

Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:245–255, 1989.

Sethuraman, J. A constructive definition of the Dirichlet prior. *Statistica Sinica*, 2:639–650, 1994.

Sha, Fei and Saul, Lawrence K. Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pp. 313–316, 2007.

Srinivasan, M., Venkatesh, S., and Hosie, R. Qualitative estimation of camera motion parameters from video sequences. *Pattern Recognit.*, 30:593–606, 1997.

Walker, S., Damien, P., Laud, P., and Smith, A. Bayesian nonparametric inference for random distributions and related functions. *J. Roy. Statist. Soc. B*, 61(3):485–527, 1999.

Winn, J. and Bishop, C.M. Variational message passing. *J. Machine Learning Research*, 6:661–694, 2005.

Xu, Minjie, Zhu, Jun, and Zhang, Bo. Nonparametric max-margin matrix factorization for collaborative prediction. In *NIPS*, 2012.

Ye, Nan, Lee, Wee Sun, Chieu, Hai Leong, and Wu, Dan. Conditional random fields with high-order features for sequence labeling. In *Proc. NIPS*, 2009.

Zhu, J. and Xing, E. Maximum entropy discrimination Markov networks. *J. Machine Learning Research*, 10:2531–2569, 2009.

Zhu, Jun, Chen, Ning, and Xing, Eric P. Infinite SVM: a Dirichlet process mixture of large-margin kernel machines. In *Proc. ICML*, 2011.