Pattern Recognition 45 (2012) 3985–3996

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A reservoir-driven non-stationary hidden Markov model

Sotirios P. Chatzis *, Yiannis Demiris

Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, South Kensington Campus, London SW7 2BT, United Kingdom

ARTICLE INFO

Article history:
Received 30 December 2011
Received in revised form
14 April 2012
Accepted 18 April 2012
Available online 3 May 2012

Keywords:
Hidden Markov model
Dirichlet process
Reservoir

ABSTRACT

In this work, we propose a novel approach towards sequential data modeling that leverages the strengths of hidden Markov models and echo-state networks (ESNs) in the context of non-parametric Bayesian inference approaches. We introduce a *non-stationary* hidden Markov model, the time-dependent state transition probabilities of which are driven by a high-dimensional signal that encodes the whole history of the modeled observations, namely the state vector of a postulated observations-driven ESN reservoir. We derive an efficient inference algorithm for our model under the variational Bayesian paradigm, and we examine the efficacy of our approach considering a number of sequential data modeling applications.

1. Introduction

The hidden Markov model (HMM) is increasingly being adopted in applications since it provides a convenient way of modeling observations appearing in a sequential manner and tending to cluster or to alternate between different possible components (subpopulations) [1]. Specifically, HMMs with continuous observation densities have been used in a wide spectrum of applications in ecology, encryption, image understanding, speech recognition, and machine vision applications [2].

Hidden Markov models are based on the assumption that each data point in a sequence of observations is generated by a latent (hidden) model state. Usually, a first-order hidden Markov chain is postulated, thus limiting the considered state dependencies only to successive observations. Longer dependencies between data over time may be also considered, by postulating the higher-order hidden Markov chains; however, such a selection may also give rise to an overwhelming increase in the computational complexity of the model, rendering it unattractive in most practical applications [2].

Echo-state networks are a groundbreaking and surprisingly efficient network structure for recurrent neural network (RNN) training [3–6]. ESNs avoid the shortcomings of typical, gradient-descent-based RNN training by randomly creating a recurrent neural network which remains unchanged during training. This RNN is called the *reservoir*. It is passively excited by the input signal and maintains in its state a non-linear transformation of the input history. Indeed, the function of the reservoir in ESNs can be compared to that of the kernel function in kernel machine approaches (e.g., support vector machines [7], relevance vector machines [8], and their variants [9]): input signals drive the non-linear reservoir and produce a high-dimensional dynamical "echo response", which is used as a non-orthogonal basis to reconstruct the desired outputs. The obtained reservoir state values of the ESN networks capture long-term dependencies between the modeled data, by encoding the history of the observed values of their driving signals.

Motivated by these advances, in this paper we exploit the merits of ESN reservoirs in order to provide a novel *non-stationary HMM* formulation for sequential data modeling. The proposed model is based on the fundamental assumption that the probabilities of HMM state transition are not stationary, but instead they depend on time, and specifically on the whole history of observed data, as encoded in the state vectors of an echo-state network reservoir. That is, an HMM with reservoir-driven non-stationary state transition probabilities is essentially introduced. The main advantage of the proposed approach is that it allows to model longer temporal dependencies compared to conventional HMMs, by introducing the dynamic information captured from the postulated ESN reservoirs into the state transition mechanics of the latent Markov chain. Derivation of our model is conducted under a non-parametric Bayesian approach to allow for automatic data-driven determination of the appropriate model size.

Non-parametric Bayesian modeling techniques, especially Dirichlet process (DP) prior-based models, have become very popular in statistics over the last few years, for performing non-parametric density estimation [10–12]. 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.). This theory is based on the observation that an infinite number of component distributions in an ordinary finite mixture model tends on the limit to a Dirichlet

* Corresponding author. Tel.: +44 7757560759.
E-mail address: soteri0s@me.com (S.P. Chatzis).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.04.018

process prior [11,13]. Exploitation of the merits of non-parametric Bayesian statistics has allowed for coming up with computationally efficient formulations of HMMs that allow for doing inference over the number of model states, thus obviating the need of model order selection. For example, in [14], an infinite HMM was proposed, based on the introduction of a hierarchical Dirichlet process (HDP) prior over the model state transition probabilities. In [15], hierarchical stick-breaking priors were imposed over the model state transition probabilities instead of the HDP, to allow for more efficient model inference by means of a truncated variational Bayesian inference technique.

As we shall discuss in the following sections, the formulation of our model consists in introduction of a joint stick-breaking and ESN reservoir-driven prior over the model state transition probabilities, which gives rise to an elaborate reservoir-driven HMM in the context of a non-parametric Bayesian inference setting. We derive an efficient truncated algorithm for model inference based on the variational Bayesian paradigm, and we experimentally demonstrate the efficacy of our approach. We dub the resulting model the echo-state stick-breaking HMM (ES-SB-HMM).

Indeed, our approach towards non-stationary HMMs with observation-driven state transitions is related to the approach taken by conditional random fields (CRFs). A CRF is simply a log-linear model representing the conditional distribution of the model states given the observed data with an associated graphical structure. In other words, they explicitly model data-driven transitions. Because the model is conditional, dependencies among the observed variables do not need to be explicitly represented, affording the use of rich, global features of the input [16]. A drawback of CRFs is that they cannot be used for the classification of whole sequences into a number of learned classes. The hidden CRF (HCRF) [17] is a discriminative model that caters to these needs, by modeling the class labels of whole sequences of observations conditional on the observed sequential data, considering that each observation is also assigned a latent label variable which is optimized as model parameter.

The remainder of this work is organized as follows: In Section 2, we provide a brief overview of echo-state networks and the DP prior. In Section 3, we introduce the ES-SB-HMM and derive an efficient truncated variational Bayesian algorithm for model inference. In Section 4, we evaluate our approach considering a number of applications from diverse domains, using benchmark datasets, and we compare it to CRFs, HCRFs, and SB-HMMs. Finally, in the last section, we summarize our results and draw our conclusions.

2. Theoretical background

2.1. Echo-state networks

As already discussed, the basic component of ESNs is a discrete-time RNN, called the reservoir. Let us consider an ESN comprising N reservoir neurons. ESN function is described by the following reservoir state update equation:

$$\boldsymbol{\varsigma}_{t+1} = (1-\gamma)h(\boldsymbol{W}\boldsymbol{\varsigma}_t + \boldsymbol{W}_{in}\boldsymbol{x}_{t+1}) + \gamma\boldsymbol{\varsigma}_t \tag{1}$$

where $\boldsymbol{\varsigma}_t$ is the reservoir state at time t (an N-dimensional vector of real numbers), $\boldsymbol{W}$ is the reservoir weight matrix, that is, the matrix of the weights of the synaptic connections between the reservoir neurons, $\boldsymbol{x}_t$ is the observed signal fed to the network at time t, $\gamma \geq 0$ is the *retainment rate* of the reservoir (with $\gamma > 0$ if leaky integrator neurons are considered), $\boldsymbol{W}_{in}$ are the weights of $\boldsymbol{x}_t$, and $h(\cdot)$ is the activation function of the reservoir. All the weight matrices to the reservoir ($\boldsymbol{W}$,$\boldsymbol{W}_{in}$) are initialized randomly. The initial state of the reservoir is usually set to zero, $\boldsymbol{\varsigma}_0 = \boldsymbol{0}$.

An extensively studied subject in the field of ESNs concerns the introduction of appropriate *goodness* measures of the reservoir structure. Indeed, the classical feature that reservoirs should possess is the echo-state property. This property essentially states that the effect of a previous reservoir state and a previous input on a future state should vanish gradually as time passes, and not persist or even get amplified. However, for most practical purposes, the echo-state property can be easily satisfied by merely ensuring that the *reservoir weight matrix* $\boldsymbol{W}$ is *contractive*, i.e., by scaling the reservoir weight matrix so that its *spectral radius* $\rho(\boldsymbol{W})$ (that is, its largest absolute eigenvalue) is less than one [18]. Indeed, this condition has been proved to be sufficient in the practical applications of ESNs; nevertheless, various researchers have also provided more rigorous global asymptotic stability conditions, providing better theoretical guarantees for ESNs to perform well on a physical system (see, e.g. [19]). It has been shown that the maximum possible short-term memory length of an ESN reservoir comprising N neurons is N time points [3].

2.2. Dirichlet process models

Dirichlet process (DP) models were first introduced by Ferguson [20]. 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

$$G|\{G_0,\alpha\} \sim \mathrm{DP}(G_0,\alpha) \tag{2}$$

$$\Theta_n^*|G \sim G, \quad n = 1,\ldots,N \tag{3}$$

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 $\alpha/(\alpha+N-1)$ 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 [21]. Let $\{\Theta_c\}_{c=1}^K$ be the set of distinct values taken by the variables $\{\Theta_n^*\}_{n=1}^{N-1}$. Denoting as f_c^{N-1} the number of values in $\{\Theta_n^*\}_{n=1}^{N-1}$ that equal to Θ_c, the distribution of Θ_N^* given $\{\Theta_n^*\}_{n=1}^{N-1}$ can be shown to be of the form [21]

$$p(\Theta_N^*|\{\Theta_n^*\}_{n=1}^{N-1},G_0,\alpha) = \frac{\alpha}{\alpha+N-1}G_0 + \sum_{c=1}^{K}\frac{f_c^{N-1}}{\alpha+N-1}\delta_{\Theta_c} \tag{4}$$

where δ_{Θ_c} denotes the distribution concentrated at a single point Θ_c. These results illustrate two key properties of the DP scheme. First, the innovation parameter α plays a key-role in determining the number of distinct parameter values. A larger α induces a higher tendency of drawing new parameters from the base distribution G_0; indeed, as $\alpha \to \infty$ we get $G \to G_0$. On the contrary, as $\alpha \to 0$ all $\{\Theta_n\}_{n=1}^N$ tend to cluster to a single random variable. Second, the more often a parameter is shared, the more likely it will be shared in the future.

A characterization of the (unconditional) distribution of the random variable G drawn from a Dirichlet process $\mathrm{DP}(G_0,\alpha)$ is provided by the *stick-breaking construction* of Sethuraman [22]. 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 $\mathrm{Beta}(1,\alpha)$, and the Θ_c are independently drawn from the base distribution G_0. The stick-breaking representation of G is then given by [22]

$$G = \sum_{c=1}^{\infty}\pi_c(\boldsymbol{v})\delta_{\Theta_c} \tag{5}$$