Pattern Recognition 44 (2011) 295–306

Contents lists available at ScienceDirect

Pattern Recognition

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

A variational Bayesian methodology for hidden Markov models utilizing Student's-*t* mixtures

Sotirios P. Chatzis [a,*], Dimitrios I. Kosmopoulos [b]

[a] Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, South Kensington Campus, SW7 2BT, UK
[b] Institute of Informatics and Telecommunications, NSCR Dimokritos, P. Grigoriou & Neapoleos Str., 15310 Athens, Greece

ARTICLE INFO

Article history:
Received 20 December 2008
Received in revised form
27 July 2010
Accepted 1 September 2010

Keywords:
Hidden Markov models
Student's-*t* distribution
Variational Bayes
Speaker identification
Robotic task failure
Violence detection

ABSTRACT

The Student's-*t* hidden Markov model (SHMM) has been recently proposed as a robust to outliers form of conventional continuous density hidden Markov models, trained by means of the expectation–maximization algorithm. In this paper, we derive a tractable variational Bayesian inference algorithm for this model. Our innovative approach provides an efficient and more robust alternative to EM-based methods, tackling their singularity and overfitting proneness, while allowing for the automatic determination of the optimal model size without cross-validation. We highlight the superiority of the proposed model over the competition using synthetic and real data. We also demonstrate the merits of our methodology in applications from diverse research fields, such as human computer interaction, robotics and semantic audio analysis.

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). 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 [1]. The hidden observation densities associated with each state of a continuous HMM must be capable of approximating arbitrarily complex probability density functions. Finite Gaussian mixture models (GMMs) are the most common selection of emission distribution models in the continuous HMM literature [2]. Their popularity stems from the well-known capability of GMMs to successfully approximate unknown random distributions, including distributions with multiple modes, while also providing a simple and computationally efficient maximum-likelihood (ML) estimation framework using the expectation–maximization (EM) algorithm [3]. Nevertheless, GMMs do also suffer from a significant drawback concerning their parameters estimation procedure, which is well-known to be adversely affected by the presence of outliers in the datasets used for the model fitting.

To tackle these issues, we have proposed in [4] a novel form of continuous HMMs where the hidden state distributions are modeled using finite mixtures of multivariate Student's-*t* densities. The multivariate Student's-*t* distribution is a bell-shaped distribution with heavier tails compared to the Gaussian; as a consequence, Student's-*t* mixture models (SMMs) provide an alternative to GMMs means of probabilistic generative modeling with high robustness to training data outliers. The so-obtained Student's-*t* hidden Markov model (SHMM) has been considered in [4] under the ML paradigm using the EM algorithm; as it has been shown, the SHMM provides an effective, computationally efficient and application-independent means for outlier tolerant representation and classification of sequential data by means of continuous HMMs.

In this paper, we provide an alternative treatment of the SHMM under a *Bayesian* framework using a *variational approximation*, yielding the *variational Bayesian* SHMM (VB-SHMM). Variational Bayesian treatments of statistical models present significant advantages over ML-based alternatives: ML approaches have the undesirable property of being ill-posed since the likelihood function is unbounded from above [5–7]. This fact results in several very significant shortcomings. To begin with, a significant difficulty concerns the infinities which plague the likelihood function, associated with the collapsing of the bell-shaped component distributions onto individual data points and, hence, resulting in singular or near-singular covariance matrices

* Corresponding author.
E-mail addresses: soteri0s@me.com (S.P. Chatzis), dkosmo@iit.demokritos.gr (D.I. Kosmopoulos).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.001

[7]. Obviously, the adoption of a Bayesian model inference algorithm, providing posterior distributions over the model parameters instead of point-estimates, would allow for the natural resolution of these issues [5–7]. Another central issue ML treatments of generative models are confronted with concerns selection of the optimal model size. Maximum likelihood is unable to address this issue since it favors models of ever-increasing complexity, thus leading to overfitting [17,10].

In our work, we conduct a *Bayesian treatment* of the SHMM, overcoming the problems of ML approaches elegantly, by marginalizing over the model parameters with respect to appropriate priors. The resulting model (marginal) likelihood can then be maximized with respect to the model size, in case one aims at optimal model selection, or combined with a prior over the model size if the goal is model averaging [17,16]. Our novel approach is based on *variational approximation* methods [8], which have recently emerged as a deterministic alternative to Markov chain Monte-Carlo (MCMC) algorithms for doing Bayesian inference for probabilistic generative models [9,10], with better scalability in terms of computational cost [11]. Variational Bayesian inference has previously been applied to relevance vector machines [12], GMMs [13], autoregressive models [14,15], SMMs [16,17], mixtures of factor analyzers [18–20], discrete HMMs [21], Gaussian HMMs [22], as well as HMMs with Poisson and autoregressive observation models [23], thereby ameliorating the singularity and overfitting problems of ML approaches.

The remainder of this paper is organized as follows: In Section 2, a brief review of the SHMM is provided. In Section 3, the proposed variational Bayesian treatment of the SHMM is carried out, yielding the variational Bayesian SHMM algorithm. In Section 4, the experimental evaluation of the proposed algorithm is conducted, considering a series of data modeling and classification applications and using real-world datasets. In the final section, our results are summarized and discussed.

2. The Student's-*t* HMM

Let us suppose an N-state HMM where the hidden emission density of each state is modeled by a K-component finite mixture model. Considering that the component distributions of the K-component finite mixture models modeling the HMM state densities are multivariate Student's-t distributions, the definition of the Student's-t HMM is obtained. The pdf of a d-dimensional Student's-t distribution with mean $\boldsymbol{\mu}$, precision $\boldsymbol{R}$, and ν degrees of freedom is given by

$$t(\boldsymbol{x}_t|\boldsymbol{\mu},\boldsymbol{R},\nu)=\frac{\Gamma\left(\frac{\nu+d}{2}\right)|\boldsymbol{R}|^{1/2}(\pi\nu)^{-d/2}}{\Gamma(\nu/2)\{1+\mathrm{MD}(\boldsymbol{x}_t,\boldsymbol{\mu}|\boldsymbol{R}^{-1})/\nu\}^{(\nu+d)/2}} \tag{1}$$

where $\mathrm{MD}(\boldsymbol{x}_t,\boldsymbol{\mu}|\boldsymbol{R}^{-1})$ is the squared Mahalanobis distance between $\boldsymbol{x}_t,\boldsymbol{\mu}$ with covariance matrix (inverse precision) $\boldsymbol{R}^{-1}$ [24] and $\Gamma(\cdot)$ is the Gamma function.

The SHMM can be modeled by the set of parameters $\Psi=\{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{C},\boldsymbol{\Theta},\boldsymbol{\nu}\}$, where $\boldsymbol{\pi}=(\pi_i)_{i=1}^N$ is the initial-state probability vector, $\boldsymbol{A}=(a_{ij})_{i,j=1}^N$ is the $N\times N$ one-step transition matrix, $\boldsymbol{C}=(c_{ij})_{i,j=1}^{N,K}$ is the $N\times K$ mixture coefficient matrix, with c_{ij} denoting the mixing proportion of the jth component density of the hidden emission distribution of the ith SHMM state, $\boldsymbol{\Theta}$ is the $N\times K$ parameter matrix that comprises the means $\boldsymbol{\mu}_{ij}$ and the precisions $\boldsymbol{R}_{ij}$ of the constituent Student's-t densities of the model, that is $\boldsymbol{\Theta}=(\theta_{ij})_{i,j=1}^{N,K}$ where $\theta_{ij}=\{\boldsymbol{\mu}_{ij},\boldsymbol{R}_{ij}\}$, and $\boldsymbol{\nu}=(\nu_{ij})_{i,j=1}^{N,K}$ is the NK vector of the degrees of freedom of the model component densities.

Let $X=\{\boldsymbol{x}_t\}_{t=1}^T$ be an observed data sequence, with $\boldsymbol{x}_t\in\mathcal{X}\subseteq\mathbb{R}^d$, modeled by an SHMM. The latent (unobserved) data associated with this sequence comprise the corresponding state sequence $S=\{s_t\}_{t=1}^T$, where $s_t=1,\ldots,N$ is the indicator of the state the tth observation is emitted from, and the sequence of the corresponding mixture component indicators $L=\{l_t\}_{t=1}^T$, where $l_t=1,\ldots,K$ indicates the mixture component density that generated the tth observation. The likelihood of the parameters set Ψ of the SHMM given the observable data X is, then, given by

$$p(X|\Psi)=\sum_{S,L}\pi_{s_1}\left[\prod_{t=1}^{T-1}a_{s_ts_{t+1}}\right]\left[\prod_{t=1}^{T}c_{s_tl_t}p(\boldsymbol{x}_t|\theta_{s_tl_t},\nu_{s_tl_t})\right] \tag{2}$$

As it has been discussed in [24], there is no closed-form solution for likelihood maximization of a Student's-t distribution. However, a computationally elegant solution can be obtained [16,17] by exploiting the property of the Student's-t distribution [24]

$$t(\boldsymbol{x}_t|\boldsymbol{\mu},\boldsymbol{R},\nu)=\int_0^\infty\mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu},u_t\boldsymbol{R})\mathcal{G}(u_t|\nu/2,\nu/2)\,\mathrm{d}u_t \tag{3}$$

which implies that a Student's-t density can be viewed as an infinite sum of Gaussians with the same mean and scaled precisions, where the precision scalars are Gamma-distributed latent variables depending on the degrees of freedom of the Student's-t density. Let us denote as $U=\{u_{s_tl_t}\}$ the sequence of the (latent) precision scalars associated with the observed data, depending on the corresponding unobserved state sequence and mixture component indicator sequence. Then, we have that

$$\boldsymbol{x}_t\sim t(\boldsymbol{\mu}_{s_tl_t},\boldsymbol{R}_{s_tl_t},\nu_{s_tl_t}) \tag{4}$$

is equivalent to

$$\boldsymbol{x}_t|u_{s_tl_t}\sim\mathcal{N}(\boldsymbol{\mu}_{s_tl_t},u_{s_tl_t}\boldsymbol{R}_{s_tl_t}) \tag{5}$$

where

$$p(u_{s_tl_t}|\nu_{s_tl_t})=\mathcal{G}(u_{s_tl_t}|\nu_{s_tl_t}/2,\nu_{s_tl_t}/2) \tag{6}$$

Under this regard, and using (3), the likelihood of the SHMM (2) eventually becomes

$$p(X|\Psi)=\sum_{S,L}\pi_{s_1}\int\mathrm{d}u_{s_tl_t}\left[\prod_{t=1}^{T-1}a_{s_ts_{t+1}}\right]\left[\prod_{t=1}^{T}c_{s_tl_t}p(\boldsymbol{x}_t|\theta_{s_tl_t},u_{s_tl_t})p(u_{s_tl_t}|\nu_{s_tl_t})\right] \tag{7}$$

3. Variational Bayesian inference for the SHMM

Variational Bayesian inference for the SHMM comprises introduction of a set of prior distributions over the model parameters and further maximization of the log marginal likelihood (log evidence) of the resulting model. For convenience, we choose priors conjugate to the considered observable and latent data, as this selection greatly simplifies inference and interpretability [8]. This way, the prior for the initial-state probabilities vector is chosen to follow a Dirichlet distribution

$$p(\boldsymbol{\pi})=\mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\phi}^\pi)=\mathcal{D}(\pi_1,\ldots,\pi_N|\phi_1^\pi,\ldots,\phi_N^\pi) \tag{8}$$

In the same fashion, we choose

$$p(\boldsymbol{A})=\prod_{i=1}^N\mathcal{D}(a_{i1},\ldots,a_{iN}|\phi_{i1}^A,\ldots,\phi_{iN}^A) \tag{9}$$

$$p(\boldsymbol{C})=\prod_{i=1}^N\mathcal{D}(c_{i1},\ldots,c_{iK}|\phi_{i1}^C,\ldots,\phi_{iK}^C) \tag{10}$$

Under the equivalent expression (5) of the Student's-t distribution, we let the joint (conjugate exponential) prior on the means