Pattern Recognition 46 (2013) 2754–2769

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection

Wentao Fan [a], Nizar Bouguila [b,*]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada H3G 1T7
[b] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada H3G 1T7

ARTICLE INFO

Article history:
Received 25 April 2012
Received in revised form
27 February 2013
Accepted 26 March 2013
Available online 17 April 2013

Keywords:
Infinite mixture models
Dirichlet process
Generalized Dirichlet
Feature selection
Clustering
Images categorization
Image auto-annotation

ABSTRACT

This paper introduces a novel enhancement for unsupervised feature selection based on generalized Dirichlet (GD) mixture models. Our proposal is based on the extension of the finite mixture model previously developed in [1] to the infinite case, via the consideration of Dirichlet process mixtures, which can be viewed actually as a purely nonparametric model since the number of mixture components can increase as data are introduced. The infinite assumption is used to avoid problems related to model selection (i.e. determination of the number of clusters) and allows simultaneous separation of data in to similar clusters and selection of relevant features. Our resulting model is learned within a principled variational Bayesian framework that we have developed. The experimental results reported for both synthetic data and real-world challenging applications involving image categorization, automatic semantic annotation and retrieval show the ability of our approach to provide accurate models by distinguishing between relevant and irrelevant features without over- or under-fitting the data.

1. Introduction

As the amount of multimedia information available increases, powerful approaches for analyzing, managing and categorizing these data become crucial. Clustering plays an important role in exploratory analysis of data. It provides principled means of discovering heterogenous groupings (i.e. clusters) in data and has been the topic of extensive research in the past [2–7]. Data clustering is known to be a challenging task in modern knowledge discovery and data mining. This is especially true in high-dimensional spaces mainly because of data sparsity [8,9] and a crucial step in this case is the selection of relevant features [10–12,1]. Finite mixture models are well suited for clustering due to their simple structure and flexibility which offer a principled formal approach to unsupervised learning [13,14]. In the classic approach to mixture models implementation, the density components are usually chosen as Gaussian and the number of components is supposed to be finite. Many methods for selecting the optimal number of clusters can be found in the literature (see, for instance, [15–17]). These approaches can be classified into two groups namely deterministic and Bayesian. The majority of both deterministic and Bayesian previous model selection approaches have to consider all possible values of the number of mixture components up to a certain maximum value and then choose the optimal one according to a certain criterion which is unfortunately computationally prohibitive (i.e. the learning algorithm have to be run for different choices of the number of mixture components) and may cause over- and under-fitting problems. A significant contribution that overcomes these drawbacks was made in [18] through the development of infinite mixture models which constitute an interesting extension of the typical finite mixture models approach by allowing the number of mixture components to increase as new data arrive. Infinite mixture models are based on the notion of Dirichlet processes which is one of the most popular Bayesian nonparametric models and is defined as a distribution over distributions [19–21].

Thanks to the recent development of Markov Chain Monte Carlo (MCMC) techniques [22], infinite mixture models have been widely and successfully used in various applications (see, for instance, [23–28]) by embodying the well-known Occam's Razor principle [29]. Concerning feature selection, although a lot of attention has been devoted to supervised feature selection (see, for instance, [30–32]), some unsupervised feature selection techniques have been proposed recently [33–39]. And some of these unsupervised techniques have been based on finite mixture models, but generally suppose that each per-component density is Gaussian with diagonal covariance matrix (i.e. the features are supposed independent)[1] [41–43].

* Corresponding author. Tel.: +1 5148482424; fax: +1 5148483171.
E-mail addresses: wenta_fa@encs.concordia.ca (W. Fan), nizar.bouguila@concordia.ca (N. Bouguila).

[1] However, it is well-known that the independence assumption is infrequently met in practice [40].

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.03.026

Recently a nonparametric Bayesian unsupervised feature selection approach has been proposed in [44]. The main idea was the consideration of the infinite generalized Dirichlet (GD) mixture model, which offers high flexibility and ease of use, for simultaneous clustering and feature selection. One of the main advantages of this approach is that the structural properties of the GD allows it to be defined in a space where the independence of the features becomes a fact and not an assumption as shown for instance in [1]. The authors in [44] have proposed a fully Bayesian treatment of the unsupervised feature selection approach that they have previously introduced in [1] in order to overcome problems related to deterministic learning. The learning approach in [44] was based on the introduction of prior distributions over the mixture parameters. These parameters have been then estimated using a typical MCMC approach based on both Gibbs sampling and Metropolis–Hastings algorithms. MCMC techniques are effective for parameters estimation, but are unfortunately computationally very demanding and it can be very hard to diagnose their convergence. This is especially true in the case of high-dimensional data which involve the integration over a large number of model parameters. The accurate evaluation of such high-dimensional integrals has been the topic of extensive research. Recently, variational approaches, known also as ensemble learning [45–47], have been proposed as an efficient alternative to MCMC techniques. Motivated by the good results obtained recently using variational techniques for modeling mixture models, in this paper we extend the learning approach in [44] by developing a variational alternative. The contribution of this paper is three-fold. First, we extend the finite GD mixture model with feature selection to the infinite case using a stick-breaking construction [48] such that the difficulty of choosing the appropriate number of clusters can be solved elegantly. Second, we propose a variational inference framework for learning the proposed model, such that the model parameters and features saliencies are estimated simultaneously in a closed form. In particular, conjugate priors are developed for all the involved parameters. Last, we apply the proposed approach to solve two challenging problems involving visual scenes categorization, and image automatic semantic annotation and retrieval. An appealing feature of the proposed variational approach is that it allows avoiding over-fitting by finding a compromise between generality and the number of parameters by implicitly providing a model order selection criterion [49,46,50]. Readers unfamiliar with Bayesian learning and the variational Bayes framework are referred to [45,51].

The paper is organized as follows. In Section 2 we present our infinite feature selection model. In Section 3 we develop a practical variational approach to learn the parameters of this model. Section 4 is devoted to experimental results of using our approach. This is followed, in Section 5, by a discussion of our findings and conclusions.

2. The infinite GD mixture model for feature selection

In this section, we describe our main unsupervised infinite feature selection model. We start by a brief overview of the finite GD mixture model. Then, the extension of this model to the infinite case and the integration of feature selection are proposed. Finally, we present the conjugate priors that we will consider for the resulting model learning.

2.1. The finite GD mixture model

Consider a random vector $\vec{Y} = (Y_1, \ldots, Y_D)$, drawn from a finite mixture of GD Distributions with M components [52] as

$$p(\vec{Y} | \vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^{M} \pi_j \mathrm{GD}(\vec{Y} | \vec{\alpha}_j, \vec{\beta}_j) \tag{1}$$

where $\vec{\alpha} = \{\vec{\alpha}_1, \ldots, \vec{\alpha}_M\}$, $\vec{\beta} = \{\vec{\beta}_1, \ldots, \vec{\beta}_M\}$, $\vec{\alpha}_j$ and $\vec{\beta}_j$ are the parameters of the GD distribution representing component j with $\vec{\alpha}_j = \{\alpha_{j1}, \ldots, \alpha_{jD}\}$ and $\vec{\beta}_j = \{\beta_{j1}, \ldots, \beta_{jD}\}$, and $\vec{\pi} = \{\pi_1, \ldots, \pi_M\}$ represents the mixing coefficients which are positive and sum to one. A GD distribution is defined as

$$\mathrm{GD}(\vec{Y} | \vec{\alpha}_j, \vec{\beta}_j) = \prod_{l=1}^{D} \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} Y_l^{\alpha_{jl}-1} \left(1 - \sum_{k=1}^{l} Y_k\right)^{\gamma_{jl}} \tag{2}$$

where $\sum_{l=1}^{D} Y_l < 1$ and $0 < Y_l < 1$ for $l = 1, \ldots, D$, $\alpha_{jl} > 0$, $\beta_{jl} > 0$, $\gamma_{jl} = \beta_{jl} - \alpha_{jl+1} - \beta_{jl+1}$ for $l = 1, \ldots, D-1$, and $\gamma_{jD} = \beta_{jD} - 1$.

Now, let us consider a set of N independent identically distributed vectors $\mathcal{Y} = (\vec{Y}_1, \ldots, \vec{Y}_N)$ assumed to arise from a finite GD mixture. Following the Bayes' theorem, the probability that vector i is in cluster j conditional on having observed $\vec{Y}_i$ (also known as *responsibilities*) can be written as

$$p(j | \vec{Y}_i) \propto \pi_j \mathrm{GD}(\vec{Y}_i | \vec{\alpha}_j, \vec{\beta}_j) \tag{3}$$

In our work, we exploit an interesting mathematical property of the GD distribution previously discussed in [52,1] to redefine the responsibilities as

$$p(j | \vec{Y}_i) \propto \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_{il} | \alpha_{jl}, \beta_{jl}) \tag{4}$$

where $X_{i1} = Y_{i1}$ and $X_{il} = Y_{il} / (1 - \sum_{k=1}^{l-1} Y_{ik})$ for $l > 1$ and $\mathrm{Beta}(X_{il} | \alpha_{jl}, \beta_{jl})$ is a Beta distribution defined with parameters $(\alpha_{jl}, \beta_{jl})$. Thus, the clustering structure for a finite GD mixture model underlying data set $\mathcal{Y}$ can be represented by a new data set $\mathcal{X} = (\vec{X}_1, \ldots \vec{X}_N)$ using the following mixture model with conditionally independent features

$$p(\vec{X}_i | \vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_{il} | \alpha_{jl}, \beta_{jl}) \tag{5}$$

It is noteworthy that this property plays a critical role for the GD mixture model, since the independence between the features becomes a fact rather than an assumption as considered in previous unsupervised feature selection Gaussian mixture-based approaches [41,42].

2.2. Infinite GD mixture model with feature selection

The Dirichlet process (DP) [20] is a stochastic process whose sample paths are probability measures with probability one. It can be considered as a distribution over distributions. The infinite GD mixture model with feature selection proposed in this paper is constructed using the DP with a stick-breaking representation. Stick-breaking representation is an intuitive and straightforward constructive definition of the DP [48,53,54]. It is defined as follows: given a random distribution G, it is distributed according to a DP: $G \sim DP(\psi, H)$ if the following conditions are satisfied:

$$\lambda_j \sim \mathrm{Beta}(1, \psi), \quad \Omega_j \sim H, \pi_j = \lambda_j \prod_{s=1}^{j-1} (1 - \lambda_s), \quad G = \sum_{j=1}^{\infty} \pi_j \delta_{\Omega_j} \tag{6}$$

where δ_{Ω_j} denotes the Dirac delta measure centered at Ω_j, and ψ is a positive real number. The mixing weights π_j are obtained by recursively breaking an unit length stick into an infinite number of pieces.

Assuming now that the observed data set is generated from a GD mixture model with a countably infinite number of components. Thus, (5) can be rewritten as

$$p(\vec{X}_i | \vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^{\infty} \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_{il} | \alpha_{jl}, \beta_{jl}). \tag{7}$$

Then, for each vector $\vec{X}_i$, we introduce a binary latent variable $\vec{Z}_i = (Z_{i1}, Z_{i2}, \ldots)$, such $Z_{ij} \in \{0, 1\}$ and $Z_{ij} = 1$ if $\vec{X}_i$ belongs to