Signal Processing 93 (2013) 1531–1546

Contents lists available at SciVerse ScienceDirect

Signal Processing

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

Simultaneous Bayesian clustering and feature selection using RJMCMC-based learning of finite generalized Dirichlet mixture models

Tarek Elguebaly [a], Nizar Bouguila [b,*]

[a] Electrical and Computer Engineering Department, Faculty of Engineering and Computer Science, Concordia University, Montreal, Qc, Canada H3G 2W1

[b] Concordia Institute for Information Systems Engineering, Faculty of Engineering and Computer Science, Concordia University, Montreal, Qc, Canada H3G 2W1

ARTICLE INFO

Article history:
Received 2 March 2012
Received in revised form
19 July 2012
Accepted 21 July 2012
Available online 18 August 2012

Keywords:
Clustering
Finite mixture models
Generalized Dirichlet
Feature selection
Bayesian analysis
RJMCMC
Gibbs sampling
Metropolis–Hastings
Video categorization
Pedestrian detection
Face recognition

ABSTRACT

Selecting relevant features in multidimensional data is important in several pattern analysis and image processing applications. The goal of this paper is to propose a Bayesian approach for identifying clusters of proportional data based on the selection of relevant features. More specifically, we consider the problem of selecting relevant features in unsupervised settings when generalized Dirichlet mixture models are considered to model and cluster proportional data. The learning of the proposed statistical model, to formulate the unsupervised feature selection problem, is carried out using a powerful reversible jump Markov chain Monte Carlo (RJMCMC) technique. Experiments involving the challenging problems of human action videos categorization, pedestrian detection and face recognition indicate that the proposed approach is efficient.

1. Introduction

Larger and larger multimedia data are collected and stored everyday presenting enormous opportunities and challenges by increasing the need for efficient modeling, knowledge extraction, categorization and clustering approaches [1,2]. A fundamental task in many of these approaches is the learning of appropriate statistical models. Mixture models are now among the most widely used statistical approaches in many areas of application and allow a formal approach for unsupervised learning [3]. In such context, classic interest is often in the determination of the number of clusters (i.e. model selection) and estimation of the mixture's parameters (see, for instance, [4,5]). Another essential issue in the case of statistical modeling in general and finite mixtures in particular is feature selection (i.e. identification of the relevant or discriminative features describing the data) especially in the case of high-dimensional data which analysis has been the topic of extensive research in the past [6–8]. Indeed, feature selection has been shown to be a crucial step in several image processing, computer vision and pattern recognition applications such as object detection [9,10], handwriting separation [11], image retrieval, categorization and recognition [12,13]. Feature selection is a major concern not only because it speeds up learning but also because it improves model accuracy and generalization. Moreover, the learning of the

* Corresponding author.
E-mail addresses: t_elgue@encs.concordia.ca (T. Elguebaly), bouguila@ciise.concordia.ca, nizar.bouguila@concordia.ca (N. Bouguila).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.07.037

mixture parameters (i.e. both model selection and parameters estimation) is greatly affected by the quality of the features used as shown for instance in [14] which has given renewed attention to the feature selection problem especially in unsupervised settings. Like many other model-based feature selection approaches (see, for instance, [15,16]) this work has been based on the Gaussian assumption by assuming diagonal covariance matrices [16] for all clusters (i.e. all the features are assumed independent). Although the normality assumption has been taken for granted in general, several works have shown that this assumption is not sound and not realistic in several applications [17,18] since per-class distributions of real data often deviate from the Gaussian distribution. This is especially true in the case of proportional data which arise in many fields of application and for which the generalized Dirichlet (GD) mixture has been shown to be an efficient modeling choice as clearly shown, for instance, in [19,5,13]. In particular, the authors in [13] have proposed a mixture-based feature selection approach relying on GD distribution and benefiting from its interesting mathematical properties and flexibility.

The unsupervised feature selection model in [13] has been trained using a minimum message length (MML) [20] objective function with the expectation–maximization (EM) [21] algorithm. This algorithm is, however, prone to initialization errors and converges either to a local maximum or to a saddle point solution which may compromise the modeling capabilities. More recently, with the increase on power of computing resources, a number of more advanced approaches based on Bayesian inference have been proposed, along with associated learning algorithms to design them in order to overcome drawbacks related with the EM framework [22]. Indeed, Bayesian inference is enjoying increasing attention in the statistical learning literature and allows to avoid over-fitting and suboptimal generalization performance. The main idea is to consider an ensemble of models described by a probability distribution over all possible parameter values, rather than considering a single model as in EM-based learning, in order to take into account model uncertainty [23]. The implementation of Bayesian inference-based approaches is based on MCMC approximations (e.g. Gibbs sampler, Metropolis–Hastings) which are important tools for statistical inference in signal and image processing applications especially in non-Gaussian settings [24,25].

The aim of this paper is to extend the unsupervised feature selection approach previously proposed in [13] by reformulating it within a fully Bayesian framework. We present a learning algorithm based on RJMCMC technique [26] which has been widely studied and applied since its introduction by offering an alternative to MCMC that includes the automatic determination of the appropriate number of mixture components and the quantification of the uncertainty (see, for instance, [27–31]). We are mainly motivated by the good results obtained recently using Bayesian learning techniques in machine learning applications in general and for the unsupervised feature selection problem in particular [15]. The remainder of the paper is structured as follows: In Section 2 we review the feature selection model based on the generalized Dirichlet mixture. In Section 3 we propose a Bayesian extension to this model and we present the complete learning algorithm where inference on the parameters is made by constructing a Gibbs sampling technique. In Section 4 we provide experimental results and we conclude with a discussion and a summary of the work in Section 5.

2. The unsupervised feature selection model

In this section, we briefly describe the unsupervised feature selection approach based on the finite GD mixture model. Although this paper is self-contained, we refer the interested reader to [13] for detailed discussions and analysis.

2.1. The mixture model

Mixture models have recently drawn a great deal of interest, being recognized as a powerful framework for probabilistic inference. One of the cited advantages of finite mixture models is that they allow for principled methods for reasoning with incomplete data. Let $\{\vec{X}_1,\ldots,\vec{X}_N\}$ be an unlabeled dataset where each vector $\vec{X}_i$ is composed of a set of continuous features representing a given object (e.g. image, video, document, etc.). Here we assume that each vector follows a mixture of generalized Dirichlet distributions of which each generalized Dirichlet has parameters θ_j, $j=1,\ldots,M$ and the mixing weights, which are positive and sum to one, of the different components are $\vec{P}=(p_1,\ldots,p_M)$:

$$p(\vec{X}_i|\Theta_M)=\sum_{j=1}^{M}p_j p(\vec{X}_i|\theta_j) \tag{1}$$

where M is the number of components which determines the structure of the model, $\Theta_M=(\vec{P},\vec{\theta})$, $\vec{\theta}=(\theta_1,\ldots,\theta_M)$ and $p(\vec{X}_i|\theta_j)$ are the components distributions which we take as generalized Dirichlet. In dimension D, the generalized Dirichlet density is defined by

$$p(\vec{X}_i|\theta_j)=\prod_{d=1}^{D}\frac{\Gamma(\alpha_{jd}+\beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})}X_{id}^{\alpha_{jd}-1}\left(1-\sum_{l=1}^{d}X_{il}\right)^{\gamma_{jd}} \tag{2}$$

where $\theta_j=(\vec{\alpha}_j,\vec{\beta}_j)$, $\vec{\alpha}_j=(\alpha_{j1},\ldots,\alpha_{jD})$ and $\vec{\beta}_j=(\beta_{j1},\ldots,\beta_{jD})$; $\sum_{d=1}^{D}X_{id}<1$ and $0<X_{id}<1$ for $d=1,\ldots,D$; $\gamma_{jd}=\beta_{jd}-\alpha_{jd+1}-\beta_{jd+1}$ for $d=1,\ldots,D-1$ and $\gamma_{jD}=\beta_{jD}-1$.

The Generalized Dirichlet distribution has an interesting property, previously shown in [13]. Indeed, if a vector $\vec{X}_i$ has a generalized Dirichlet distribution with parameters $(\vec{\alpha}_j,\vec{\beta}_j)$, then we can construct a vector $\vec{Y}_i$ using the following geometric transformation: $Y_{i1}=X_{i1}$ and $Y_{id}=X_{id}/(1-X_{i1}-\cdots-X_{id-1})$ for $d=2,3,\ldots,D$ such that each Y_{id} has a Beta distribution with parameters α_{jd} and β_{jd}. This transformation means that the generalized Dirichlet mixture model can be transformed to a multidimensional Beta mixture model with conditionally independent features:

$$p(\vec{Y}_i|\Theta_M)=\sum_{j=1}^{M}p_j\prod_{d=1}^{D}p_b(Y_{id}|\alpha_{jd},\beta_{jd}) \tag{3}$$