Computers and Electrical Engineering 43 (2015) 48–65

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A hierarchical Dirichlet process mixture of generalized Dirichlet distributions for feature selection ☆

Wentao Fan [a], Hassen Sallay [b], Nizar Bouguila [c,*], Sami Bourouis [d]

[a] Department of Computer Science and Technology, Huaqiao University, China
[b] College of Computer and Information Systems, Umm Al-Qura University, Saudi Arabia
[c] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 1T7, Canada
[d] Taif University, Taif, Saudi Arabia

ARTICLE INFO

Article history:
Received 8 July 2014
Received in revised form 7 March 2015
Accepted 13 March 2015
Available online 31 March 2015

Keywords:
Clustering
Hierarchical Dirichlet process
Variational learning
Face detection
Facial expression recognition
Human gesture recognition

ABSTRACT

This paper addresses the problem of identifying meaningful patterns and trends in data via clustering (i.e. automatically dividing a data set into meaningful homogenous sub-groups such that the data within the same sub-group are very similar, and data in different sub-groups are very different). The clustering framework that we propose is based on the generalized Dirichlet distribution, which is widely accepted as a flexible modeling approach, and a hierarchical Dirichlet process mixture prior. A main advantage of the adopted hierarchical Dirichlet process is that it provides a principled elegant nonparametric Bayesian approach to model selection by supposing that the number of mixture components can go to infinity. In addition to capturing the structure of the data, the combination of hierarchical Dirichlet process and generalized Dirichlet distribution is shown to offer a natural efficient solution to the feature selection problem when dealing with high-dimensional data. We develop two variational learning approaches (i.e. batch and incremental) for learning the parameters of the proposed model. The batch algorithm examines the entire data set at once while the incremental one learns the model one step at a time (i.e. update the model's parameters each time new data are introduced). The utility of the proposed approach is demonstrated on real applications namely face detection, facial expression recognition, human gesture recognition, and off-line writer identification. The obtained results show clearly the merits of our statistical framework.

1. Introduction

In the last few years, we have seen an explosion of on-line data in digital form. It is crucial then to develop approaches and techniques to index and understand this content, and to represent it in compact forms. This is the case, for instance, of many forensics applications where one of the main goals is to automatically analyze or predict criminal incidents (see, for instance, [1]) which is vital for law enforcement agencies. Examples include intrusion detection, video-surveillance, user authentification (i.e. the task of confirming or denying the identity claimed by a given person) via face or fingerprint recognition to name a few [2]. Statistical approaches in general and graphical models in particular have been widely used for data

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Yi Wan.

* Corresponding author.
E-mail addresses: fwt@hqu.edu.cn (W. Fan), hmsallay@uqu.edu.sa (H. Sallay), nizar.bouguila@concordia.ca (N. Bouguila), s.bourouis@tu.edu.sa (S. Bourouis).

http://dx.doi.org/10.1016/j.compeleceng.2015.03.018
0045-7906/

(i.e. text, image, video) interpretation, understanding and modeling. Mixture models, as a particular case of graphical models, with their strong theoretical roots are known to be excellent statistical approaches for data analysis by offering a formal approach to unsupervised learning problems [3]. These models are at the heart of many challenging applications from different domains, but their deployment requires the resolution of some problems namely the choice of the components densities, the estimation of the parameters and model selection (i.e. determination of the number of components or model's complexity). Concerning the choice of components densities, apart from some exceptions most mixture-based approaches consider Gaussian distributions [4]. Recently, however, several works have shown that this choice is not appropriate in many real-life applications and that other distributions such as the generalized Dirichlet may offer better modeling capabilities especially when dealing with proportional data [5]. Thus, in this work we consider the generalized Dirichlet as the parent distribution of our models. Concerning model selection, several approaches have been proposed in the past and are deeply discussed in [5]. The majority of the used approaches consider information theory-based criteria such as minimum description length and minimum message length [5]. Yet, these approaches are computationally expensive since one needs to run the learning algorithm for every candidate number of components. Unlike these conventional approaches which assume a finite number of components, we do not confine ourselves to this assumption. Indeed, infinite mixtures are considered, where a new data point can be affected to a new cluster that was not previously seen, via the adoption of a nonparametric Bayesian approach based on hierarchical Dirichlet process prior. We are mainly motivated by the excellent results obtained recently thanks to Dirichlet processes which avoid elegantly model selection problems [6]. For a comprehensive treatment of nonparametric Bayesian approaches and Dirichlet process the reader is referred to [7] and references therein.

One major limitation of the implementation of mixture models in real-life scenarios is dealing with high-dimensional data which causes sparsely populated spaces. Feature selection techniques have been widely used in the past to tackle this problem and allow generally to attain significant generalization improvement, to unseen data, in many applications such as text classification, image retrieval and annotation. Indeed, it is crucial to take into account the fact that features are generally not equally important (or relevant) for a given task (e.g. clustering) and that irrelevant features may even compromise modeling capabilities (e.g. by blurring clusters). Feature selection is a difficult problem especially in unsupervised settings, characterized by the absence of labeled data that could guide the selection process, which is actually our case since we are considering mixture models. In this context, features have been generally assumed to be independent (e.g. by considering Gaussians with diagonal covariance matrices) [4], to reduce complexity and the number of parameters to learn, which is not the case in several real-life applications [8]. However, the authors in [8] have shown that this assumption could be avoided in the case of the generalized Dirichlet distribution thanks to some interesting mathematical properties that allow the independence between features to become a fact, via a simple geometric transformation, as it will be explained in the next section.

A main drawback each time nonparametric Bayesian approaches have been considered is the fact that learning has been typically performed using Markov Chain Monte Carlo (MCMC) techniques namely Gibbs sampling and Metropolis–Hastings which is clearly time consuming [9]. There is a rich tradition in machine learning of studying inference techniques. The design of a good inference algorithm is particularly difficult. Recently, the variational approach has received a lot of attention by offering a compromise between deterministic approaches and purely Bayesian ones. Indeed, it can be viewed as deterministic approximation to Bayesian inference that allows to incorporate prior information in a principled way. The main idea is to approximate the model posterior distribution by minimizing the Kullback–Leibler divergence between the exact (or true) posterior and an approximating distribution which have had some impressive successes in learning complex statistical models (see, for instance, [10–12] and references therein). Thus, an important contribution of this work is the development of a variational algorithm to learn our hierarchical Dirichlet process mixture of generalized Dirichlet distributions. In order to face the fact that data extracted from real-life applications are generally dynamic, we extend our batch learning algorithm to online setting where data are supposed to arrive sequentially which is crucial in several real-life applications.

The rest of this paper is structured as follows. The paper begins with a presentation of the hierarchical Dirichlet process mixture of GD Distributions in Section 2, followed by the development of our batch and online learning algorithms in Section 3. Section 4 presents the experimental results in the context of four challenging real-life applications namely face detection, facial expression recognition, human gesture recognition, and off-line writer identification. Section 5 concludes the paper with some final remarks.

2. Hierarchical Dirichlet process mixture of GD distributions

The hierarchical Dirichlet process framework is particularly useful in problems for modeling grouped data where observations are organized into groups allowed to remain statistically linked by sharing mixture components [13]. In this section, we develop a hierarchical Dirichlet process mixture model of generalized Dirichlet distributions with an unsupervised feature selection scheme.

2.1. Hierarchical Dirichlet process mixture model

The formal definition of the Dirichlet process [14] is as following: let H be a distribution over some probability space Θ and γ be a positive real number, then a random distribution G is distributed according to a Dirichlet process with the base distribution H and concentration parameter γ, denoted as $G \sim \mathrm{DP}(\gamma, H)$, if