JMLR: Workshop and Conference Proceedings 25:113–128, 2012 Asian Conference on Machine Learning

Online Learning of a Dirichlet Process Mixture of Generalized Dirichlet Distributions for Simultaneous Clustering and Localized Feature Selection

Wentao Fan WENTA_FA@ENCS.CONCORDIA.CA
Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada

Nizar Bouguila NIZAR.BOUGUILA@CONCORDIA.CA
The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

Online algorithms allow data instances to be processed in a sequential way, which is important for large-scale and real-time applications. In this paper, we propose a novel online clustering approach based on a Dirichlet process mixture of generalized Dirichlet (GD) distributions, which can be considered as an extension of the finite GD mixture model to the infinite case. Our approach is built on nonparametric Bayesian analysis where the determination of the number of clusters is sidestepped by assuming an infinite number of mixture components. Moreover, an unsupervised localized feature selection scheme is integrated with the proposed nonparametric framework to improve the clustering performance. By learning the proposed model in an online manner using a variational approach, all the involved parameters and features saliencies are estimated simultaneously and effectively in closed forms. The proposed online infinite mixture model is validated through both synthetic data sets and two challenging real-world applications namely text document clustering and online human face detection.

Keywords: online learning; clustering; Dirichlet process; nonparametric Bayesian; variational Bayes; generalized Dirichlet mixtures; localized feature selection.

1. Introduction

Many data mining, computer vision, pattern recognition and machine learning applications involve high-dimensional data. A crucial step to deal with high-dimensional data is dimensionality reduction (Kaski, 1998; Engebretsen et al., 2002) via extraction (e.g. principal components analysis, random projection (Achlioptas, 2001; Bingham and Mannila, 2001; Fradkin and Madigan, 2003)) or selection of features (filters or wrappers). In this paper we will focus on feature selection which has been the topic of extensive research in the past. This is mainly du to the importance of selecting relevant features to control model's complexity and then avoid over-fitting the data and improve generalization capabilities. Recently the use of mixture models has emerged as a principled method for simultaneous clustering and feature selection. The major advantage of mixture model is that it offers a formal approach to unsupervised learning. This fact has been widely detailed in the literature (see, for instance, (McLachlan and Peel, 2000)). Among various mixture models,

the Gaussian mixture has been a popular choice due to its simplicity and maturity (Constantinopoulos et al., 2006). The Gaussian assumption, however, is not realistic when the data clearly appear with a non-Gaussian structure. Several works have shown that other models such as the finite generalized Dirichlet (GD) mixture can be a better alternative to the Gaussian mixture in several applications, especially those involving proportional data, such as text and image modeling (Bouguila et al., 2007, 2009; Bouguila and Ziou, 2004, 2005, 2010). Thus, motivated by its flexibility and good performance obtained in these previous works, we shall focus in this paper on the GD mixture model for feature selection. Selecting the number of clusters that best describes the data without overfitting or underfitting it is one of the most challenging problems in finite mixture modeling. Traditionally, this problem is solved using maximum likelihood method in conjunction with model selection criteria (ex. MDL, BIC, MML, AIC, etc). However, this approach requires the evaluation of a given selection criterion for several numbers of mixture components which is highly computationally demanding. An alternative way to deal with the model selection problem is through a nonparametric Bayesian technique namely Dirichlet process (DP) (Korwar and Hollander, 1973; Ferguson, 1983) by assuming that there are an infinite number of mixture components. Indeed, the DP mixture model can be also viewed as an infinite mixture model, such that its complexity increases as the data set grows. As a result, the problem of underfitting is avoided by using a model with an unlimited complexity, and the trouble of overfiiting is tackled by adopting the Bayesian approach to compute or approximate the full posterior distributions of parameters. Thanks to the recent development of Markov chain Monte Carlo (MCMC) techniques (Robert and Casella, 1999), infinite mixture models based on Dirichlet processes have been widely used in various applications (Neal, 2000; Teh et al., 2004). The use of MCMC techniques, however, is often limited to small-scale problems in practice because of its high computational cost. A good alternative to the MCMC technique is a deterministic approximation technique known as variational inference (Attias, 1999; Jordan et al., 1999; Bishop, 2006), which only requires a modest amount of computational power and has provided promising performance in many applications involving mixture models. However, all these foregoing approaches work in a batch mode in which all the data instances need to be available in advance. Compared to batch algorithms, online learning algorithms are more efficient when dealing with massive and streaming data.

The main purpose of this paper is to develop a novel online unsupervised clustering approach based on a nonparametric Bayesian model learned in a variational way. Our contributions are listed as the following: First, we extend the finite GD mixture model to the infinite case using a stick-breaking construction such that the difficulty of choosing the appropriate number of clusters can be solved elegantly. Second, rather than using a global (i.e. produce a common feature subset for all the mixture components) unsupervised feature selection method as commonly used in many works (Law et al., 2004; Constantinopoulos et al., 2006; Boutemedjet et al., 2009), we integrate a localized feature selection scheme (Li et al., 2009) into our infinite mixture model where different feature subsets are associated with different mixture components. The motivation of this particular choice is based on the fact that recent works have shown that global feature selection may not be realistic in real life applications and that localized feature selection can generally provide better results (Li et al., 2009; Fan et al., 2011; Guan et al., 2011). Third, we develop an online learning algorithm for our model based on a natural gradient method. This is important for real-time applica-

tions, and also where large scale data sets are involved so that batch processing of all data points at once becomes infeasible. Finally, we propose a variational inference framework for learning the proposed model, such that the model parameters and the local features saliencies are estimated simultaneously in a closed form.
The rest of this paper is organized as follows: In Section 2, we present the infinite GD mixture model with localized feature selection scheme. In Section 3, an online variational framework is developed for learning the proposed model. Section 4 is devoted to the experimental results. Finally, conclusion is provided in Section 5.

2. Model Specification

2.1. Finite GD Mixture with Localized Feature Selection

Suppose that we have a D-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ which is drawn from a finite mixture of generalized Dirichlet (GD) distributions with M components, such that (Bouguila et al., 2009):

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

where $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are the parameters of the GD distribution representing component j with $\boldsymbol{\alpha}_j = \{\alpha_{j1}, \ldots, \alpha_{jD}\}$ and $\boldsymbol{\beta}_j = \{\beta_{j1}, \ldots, \beta_{jD}\}$. $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_M\}$ denotes the mixing coefficients, subject to the following constraints: $0 \leq \pi_j \leq 1$, $\sum_{j=1}^{M} \pi_j = 1$. The GD distribution of $\boldsymbol{Y}$ with parameters $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ is given by

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

where $\sum_{l=1}^{D} Y_l < 1$ and $0 < Y_d < 1$ 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 $\gamma_{jD} = \beta_{jD} - 1$. Based on an interesting mathematical property of the GD distribution which is thoroughly discussed in Boutemedjet et al. (2009), we can transform the original data points into another D-dimensional space with independent features and rewrite the finite GD mixture model in the following form

$$p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_l|\alpha_{jl}, \beta_{jl}) \tag{3}$$

where $\boldsymbol{X} = (X_1, \ldots, X_D)$, $X_1 = Y_1$ and $X_l = Y_l/(1 - \sum_{k=1}^{l-1} Y_k)$ for $l > 1$, and $\mathrm{Beta}(X_l|\alpha_{jl}, \beta_{jl})$ is a Beta distribution defined with parameters $(\alpha_{jl}, \beta_{jl})$. Indeed, this property is important since the independence between the features now becomes a fact rather than an assumption as considered in previous unsupervised feature selection Gaussian mixture-based approaches (Law et al., 2004; Constantinopoulos et al., 2006) [1]. Next, we deploy a localized feature

1. It is well-known that the independence assumption rarely holds in real world applications and problems despite the fact that it achieves sometimes surprisingly good results (Keogh and Pazzani, 1999).

selection scheme (Li et al., 2009) which has been shown to outperform the global one. Thus, the distribution of each feature X_{il} can be approximated by

$$p(X_{il}) \simeq \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})^{\phi_{ijl}} \text{Beta}(X_{il}|\sigma_{jl}, \tau_{jl})^{1-\phi_{ijl}} \tag{4}$$

where ϕ_{ijl} is a binary latent variable and known as the feature relevance indicator, such that $\phi_{ijl} = 0$ if feature d of component j is irrelevant (i.e. noise) and follows a Beta distribution: $\text{Beta}(X_{il}|\sigma_{jl}, \tau_{jl})$. The prior distribution of $\boldsymbol{\phi}$ is defined as:

$$p(\boldsymbol{\phi}|\boldsymbol{\epsilon}) = \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{l=1}^{D} \epsilon_{jl_1}^{\phi_{ijl}} \epsilon_{jl_2}^{1-\phi_{ijl}} \tag{5}$$

where each ϕ_{ijl} is a Bernoulli variable such that $p(\phi_{ijl} = 1) = \epsilon_{jl_1}$ and $p(\phi_{ijl} = 0) = \epsilon_{jl_2}$. The vector $\boldsymbol{\epsilon}$ represents the features saliencies (i.e. the probabilities that the features are relevant) where $\boldsymbol{\epsilon}_{jl} = (\epsilon_{jl_1}, \epsilon_{jl_2})$ and $\epsilon_{jl_1} + \epsilon_{jl_2} = 1$. In addition, a Dirichlet prior distribution is placed over $\boldsymbol{\epsilon}$ with positive parameter $\boldsymbol{\varsigma}$:

$$p(\boldsymbol{\epsilon}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \text{Dir}(\boldsymbol{\epsilon}_{jl}|\boldsymbol{\varsigma}) \tag{6}$$

2.2. Infinite GD Mixture Models

In this subsection, we extent the finite GD mixture model to the infinite case by adopting a Dirichlet process (DP) mixture model, such that the obstacle of estimating the number of components can be circumvented. In this paper, the DP process is constructed by using a stick-breaking framework (Sethuraman, 1994; Blei and Jordan, 2005). That is, G is Dirichlet process distributed with a base distribution H and concentration parameter ψ (denoted as $G \sim \text{DP}(\psi, H)$), if the following requirements are met:

$$\lambda_j \sim \text{Beta}(1, \psi) \qquad \Omega_j \sim H \qquad \pi_j = \lambda_j \prod_{s=1}^{j-1}(1 - \lambda_s) \qquad G = \sum_{j=1}^{\infty} \pi_j \delta_{\Omega_j} \tag{7}$$

where δ_{Ω_j} represents the Dirac delta measure centered at Ω_j. The mixing weights π_j are defined by recursively breaking an unit length stick into an infinite number of pieces.
Assuming now that we have an infinite number of clusters and an observed data set $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$. First, a binary latent variable $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots)$ is placed over each vector $\boldsymbol{X}_i$, such that $Z_{ij} \in \{0, 1\}$ and $Z_{ij} = 1$ if $\boldsymbol{X}_i$ belongs to component j and 0, otherwise. Then, the likelihood function of the infinite GD mixture with localized feature selection can be written as

$$p(\mathcal{X}) = \prod_{i=1}^{N}\prod_{j=1}^{\infty} \left[\prod_{l=1}^{D} \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})^{\phi_{ijl}} \text{Beta}(X_{il}|\sigma_{jl}, \tau_{jl})^{1-\phi_{ijl}} \right]^{Z_{ij}} \tag{8}$$

The prior distribution of latent variables $\mathcal{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ is given by

$$p(\mathcal{Z}|\boldsymbol{\pi}) = \prod_{i=1}^{N}\prod_{j=1}^{\infty} \pi_j^{Z_{ij}} \tag{9}$$

According to the stick-breaking construction of DP as stated in (7), $\boldsymbol{\pi}$ is a function of $\boldsymbol{\lambda}$, then we have

$$p(\mathcal{Z}|\boldsymbol{\lambda}) = \prod_{i=1}^{N}\prod_{j=1}^{\infty}[\lambda_j \prod_{s=1}^{j-1}(1-\lambda_s)]^{Z_{ij}} \tag{10}$$

The prior distribution of $\boldsymbol{\lambda}$ is the specific Beta distribution given in (7):

$$p(\boldsymbol{\lambda}|\boldsymbol{\psi}) = \prod_{j=1}^{\infty}\text{Beta}(1,\psi_j) = \prod_{j=1}^{\infty}\psi_j(1-\lambda_j)^{\psi_j-1} \tag{11}$$

Last, we need to introduce conjugate priors over parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ of Beta distributions. Here, as proposed by Ma and Leijon (2011), we assume that these Beta parameters are statistically independent and Gamma priors $\mathcal{G}(\cdot)$ are adopted to approximate the conjugate priors. Thus, the prior distribution for parameter $\boldsymbol{\alpha}$ is given by

$$p(\boldsymbol{\alpha}) = \mathcal{G}(\boldsymbol{\alpha}|\boldsymbol{u},\boldsymbol{v}) = \prod_{j=1}^{\infty}\prod_{l=1}^{D}\frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})}\alpha_{jl}^{u_{jl}-1}e^{-v_{jl}\alpha_{jl}} \tag{12}$$

Similarly, we have $p(\boldsymbol{\beta}) = \mathcal{G}(\boldsymbol{\beta}|\boldsymbol{p},\boldsymbol{q})$, $p(\boldsymbol{\sigma}) = \mathcal{G}(\boldsymbol{\sigma}|\boldsymbol{g},\boldsymbol{h})$ and $p(\boldsymbol{\tau}) = \mathcal{G}(\boldsymbol{\tau}|\boldsymbol{s},\boldsymbol{k})$.

3. Online Variational Model Learning

In this section, following the online learning framework proposed by Sato (2001), we develop an online variational inference framework for learning the infinite GD mixture model with localized feature selection. To simplify the notation, we define $\Theta = \{\mathcal{W},\Lambda\}$, where $\mathcal{W} = \{\mathcal{Z},\boldsymbol{\phi}\}$ and $\Lambda = \{\boldsymbol{\lambda},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\tau}\}$. In variational inference, the idea is to find an approximation $Q(\Theta)$ for the posterior distribution $p(\Theta|\mathcal{X})$. This is done by maximizing the lower bound of $\ln p(\mathcal{X})$, which is defined by

$$\mathcal{L}(Q) = \int Q(\Theta)\ln[p(\mathcal{X},\Theta)/Q(\Theta)]d\Theta \tag{13}$$

In this work, we adopt the factorial approximation for the variational inference to factorize $Q(\Theta)$ into disjoint tractable distributions. In online learning, let t denotes the actual amount of observed data. Then, the current lower bound for the observed data is given by

$$\mathcal{L}^{(t)}(Q) = \frac{N}{t}\sum_{i=1}^{t}\int Q(\Lambda)\mathrm{d}\Lambda\sum_{\boldsymbol{W}_i}Q(\boldsymbol{W}_i)\ln\left[\frac{p(\boldsymbol{X}_i,\boldsymbol{W}_i|\Lambda)}{Q(\boldsymbol{W}_i)}\right] + \int Q(\Lambda)\ln\left[\frac{p(\Lambda)}{Q(\Lambda)}\right]\mathrm{d}\Lambda \tag{14}$$

where $\mathcal{W} = (\boldsymbol{W}_1,\ldots,\boldsymbol{W}_N)$ with $\boldsymbol{W}_i = \{\boldsymbol{Z}_i,\boldsymbol{\phi}_i\}$.
The core idea of the online variational algorithm is to successively maximize the current variational lower bound (14). Assume that we have already observed the data set $\{X_1,\ldots X_{(t-1)}\}$. For a new observation X_t, we can maximize the current lower bound $\mathcal{L}^{(t)}(Q)$ with respect to $Q(\boldsymbol{\phi}_t)$, while other variational factors are fixed to $Q(Z_{(t-1)})$, $Q^{(t-1)}(\boldsymbol{\lambda})$, $Q^{(t-1)}(\boldsymbol{\epsilon})$, $Q^{(t-1)}(\boldsymbol{\alpha})$, $Q^{(t-1)}(\boldsymbol{\beta})$, $Q^{(t-1)}(\boldsymbol{\sigma})$ and $Q^{(t-1)}(\boldsymbol{\tau})$. Moreover, we adopt a truncation technique proposed by Blei and Jordan (2005) to truncate the variational distributions at

a value M, such that $\lambda_M = 1$, $\sum_{j=1}^{M} \pi_j = 1$, and $\pi_j = 0$ when $j > M$. Notice that, the truncation level M is a variational parameter which can be freely initialized and will be optimized automatically during the learning process. Thus, the variational solution to $Q(\boldsymbol{\phi}_t)$ can be obtained by

$$Q(\boldsymbol{\phi}_t) = \prod_{j=1}^{M} \prod_{l=1}^{D} f_{tjl}^{\phi_{tjl}} (1 - f_{tjl})^{(1-\phi_{tjl})} \tag{15}$$

where

$$f_{tjl} = \frac{\widetilde{f}_{tjl}}{\widetilde{f}_{tjl} + \widehat{f}_{tjl}}, \quad \widetilde{f}_{tjl} = \exp\left[r_{(t-1)j}\vartheta + \langle \ln \epsilon_{jl_1}^{(t-1)} \rangle\right]$$

$$\widehat{f}_{tjl} = \exp\left[r_{(t-1)j}\xi + \langle \ln \epsilon_{jl_2}^{(t-1)} \rangle\right]$$

$$\vartheta = \widetilde{\mathcal{R}}_{jl}^{(t-1)} + (\bar{\alpha}_{jl}^{(t-1)} - 1) \ln X_{tl} + (\bar{\beta}_{jl}^{(t-1)} - 1) \ln(1 - X_{tl})$$

$$\xi = \widetilde{\mathcal{F}}_{jl}^{(t-1)} + (\bar{\sigma}_{jl}^{(t-1)} - 1) \ln X_{tl} + (\bar{\tau}_{jl}^{(t-1)} - 1) \ln(1 - X_{tl})$$

$$\langle \ln \epsilon_{jl_1} \rangle = \Psi(\varsigma_1^*) - \Psi(\varsigma_1^* + \varsigma_2^*), \quad \langle \ln \epsilon_{jl_2} \rangle = \Psi(\varsigma_2^*) - \Psi(\varsigma_1^* + \varsigma_2^*)$$

In the above equations, $\Psi(\cdot)$ is the digamma function, $\bar{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{u_{jl}^*}{v_{jl}^*}$, $\bar{\beta}_{jl} = \frac{p_{jl}^*}{q_{jl}^*}$, $\bar{\sigma}_{jl} = \frac{g_{jl}^*}{h_{jl}^*}$ and $\bar{\tau}_{jl} = \frac{s_{jl}^*}{k_{jl}^*}$. Notice that, $\widetilde{\mathcal{R}}$ and $\widetilde{\mathcal{F}}$ are the lower bounds of $\mathcal{R} = \left\langle \ln \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right\rangle$ and $\mathcal{F} = \left\langle \ln \frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)} \right\rangle$, respectively. Since these expectations are intractable, we use the second-order Taylor series expansion to find their lower bounds as proposed by Ma and Leijon (2011). Next, the current lower bound $\mathcal{L}^{(t)}(Q)$ is maximized with respect to $Q(\boldsymbol{Z}_t)$, while $Q(\boldsymbol{\phi}_t)$ is fixed and other variational factors remain to their $(t-1)$th values. Therefore, we can obtain

$$Q(\boldsymbol{Z}_t) = \prod_{j=1}^{M} r_{tj}^{Z_{tj}} \tag{16}$$

where

$$r_{tj} = \frac{\tilde{r}_{tj}}{\sum_{j=1}^{M} \tilde{r}_{tj}}$$

$$\tilde{r}_{tj} = \exp\left\{ \sum_{l=1}^{D} f_{tjl}\vartheta + \sum_{l=1}^{D} (1 - f_{tjl})\xi + \langle \ln \lambda_j^{(t-1)} \rangle + \sum_{s=1}^{j-1} \langle \ln(1 - \lambda_s^{(t-1)}) \rangle \right\}$$

$$\langle \ln \lambda_j \rangle = \Psi(c_j) - \Psi(c_j + d_j), \quad \langle \ln(1 - \lambda_j) \rangle = \Psi(d_j) - \Psi(c_j + d_j)$$

In the following step, we maximize the current lower bound $\mathcal{L}^{(t)}(Q)$ with respect to $Q^{(t)}(\boldsymbol{\lambda})$ and $Q^{(t)}(\boldsymbol{\epsilon})$ while holding other variational factors fixed. Then, the variational solutions to $Q^{(t)}(\boldsymbol{\lambda})$ and $Q^{(t)}(\boldsymbol{\epsilon})$ can be given by

$$Q^{(t)}(\boldsymbol{\lambda}) = \prod_{j=1}^{M} \text{Beta}(\lambda_j^{(t)} | c_j^{(t)}, d_j^{(t)}) \tag{17}$$

$$Q^{(t)}(\boldsymbol{\epsilon}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \text{Dir}(\boldsymbol{\epsilon}_{jl}^{(t)} | \boldsymbol{\varsigma}^{*(t)}) \tag{18}$$

where the hyperparameters are defined by
$c_j^{(t)} = c_j^{(t-1)} + \rho_t \Delta c_j^{(t)}, d_j^{(t)} = d_j^{(t-1)} + \rho_t \Delta d_j^{(t)}, \varsigma_1^{*(t)} = \varsigma_1^{*(t-1)} + \rho_t \Delta \varsigma_1^{*(t)}, \varsigma_2^{*(t)} = \varsigma_2^{*(t-1)} + \rho_t \Delta \varsigma_2^{*(t)}$
where ρ_t is the learning rate and is defined by $\rho_t = (\eta_0 + t)^{-a}$ with the constraints: $a \in (0.5, 1]$ and $\eta_0 \geq 0$ (Hoffman et al., 2010; Wang et al., 2011). The role of ρ_t is to forget the earlier inaccurate estimation effects that contributed to the lower bound and accelerate the convergence of the learning process. In the above equations, $\Delta c_j^{(t)}$, $\Delta d_j^{(t)}$, $\Delta \varsigma_1^{*(t)}$ and $\Delta \varsigma_2^{*(t)}$ are the natural gradients of the corresponding hyperparameters. This is motivated by the fact that variational algorithm can be performed as a natural gradient method (Amari, 1998) which can then be easily adapted to online inference. The natural gradient of a parameter is obtained by multiplying the gradient by the inverse of Riemannian metric, which cancels the coefficient matrix for the posterior parameter distribution. Thus, the following natural gradients can be obtained

$$\Delta c_j^{(t)} = c_j^{(t)} - c_j^{(t-1)} = 1 + N r_{tj} - c_j^{(t-1)} \tag{19}$$

$$\Delta d_j^{(t)} = d_j^{(t)} - d_j^{(t-1)} = \psi_j + N \sum_{s=j+1}^{M} r_{ts} - d_j^{(t-1)} \tag{20}$$

$$\Delta \varsigma_1^{*(t)} = \varsigma_1^{*(t)} - \varsigma_1^{*(t-1)} = \varsigma_1 + N f_{tjl} - \varsigma_1^{*(t-1)} \tag{21}$$

$$\Delta \varsigma_2^{*(t)} = \varsigma_2^{*(t)} - \varsigma_2^{*(t-1)} = \varsigma_2 + N(1 - f_{tjl}) - \varsigma_2^{*(t-1)} \tag{22}$$

Last, the current lower bound $\mathcal{L}^{(t)}(Q)$ is maximized with respect to $Q^{(t)}(\boldsymbol{\alpha})$, $Q^{(t)}(\boldsymbol{\beta})$, $Q^{(t)}(\boldsymbol{\sigma})$ and $Q^{(t)}(\boldsymbol{\tau})$ while other variational factors remain to their current values. By applying the natural gradient method again, the variational solution to $Q^{(t)}(\boldsymbol{\alpha})$ is obtained by

$$Q^{(t)}(\boldsymbol{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}^{(t)} | u_{jl}^{*(t)}, v_{jl}^{*(t)}) \tag{23}$$

where $u_{jl}^{*(t)} = u_{jl}^{*(t-1)} + \rho_t \Delta u_{jl}^{*(t)}$ and $v_{jl}^{*(t)} = v_{jl}^{*(t-1)} + \rho_t \Delta v_{jl}^{*(t)}$. The corresponding natural gradients are given by

$$\begin{aligned} \Delta u_{jl}^{*(t)} = u_{jl}^{*(t)} - u_{jl}^{*(t-1)} \quad = \quad & u_{jl} + N r_{tj} f_{tjl} \bar{\alpha}_{jl} \big[\Psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \Psi(\bar{\alpha}_{jl}) \\ & + \bar{\beta}_{jl} \Psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) (\langle \ln \beta_{jl} \rangle - \ln \bar{\beta}_{jl}) \big] - u_{jl}^{*(t-1)} \end{aligned} \tag{24}$$

$$\Delta v_{jl}^{*(t)} = v_{jl}^{*(t)} - v_{jl}^{*(t-1)} = v_{jl} - N r_{tj} f_{tjl} \ln X_{tl} - v_{jl}^{*(t-1)} \tag{25}$$

where $\langle \ln \beta_{jl} \rangle = \Psi(p_{jl}) - \ln q_{jl}$. The solutions to the hyperparameters of $Q^{(t)}(\boldsymbol{\beta})$, $Q^{(t)}(\boldsymbol{\sigma})$ and $Q^{(t)}(\boldsymbol{\tau})$ can be computed similarly as for $u_{jl}^{*(t)}$ and $v_{jl}^{*(t)}$.
It is worth mentioning that the online variational algorithm can be defined as a stochastic approximation method (Kushner and Yin, 1997) for estimating the expected lower bound and the convergence is guaranteed if the learning rate satisfies the following conditions (Sato, 2001): $\sum_{t=1}^{\infty} \rho_t = \infty$, and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$. The algorithm of online variational inference for infinite GD mixture model with localized feature selection is summarized in Algorithm 1. After convergence, we can compute $\langle \lambda_j \rangle = \frac{c_j}{c_j + d_j}$ and substitute it into (7) to obtain the estimated values of mixing coefficients π_j. The number of components is determined by removing the components with small mixing coefficients close to 0 after convergence.

Algorithm 1

Choose the initial truncation level M and the initial values for hyperparameters u_{jl}, v_{jl}, p_{jl}, q_{jl}, g_{jl}, h_{jl}, s_{jl}, k_{jl}, ψ_j, ς_1 and ς_2.
for $t = 1 \rightarrow N$ **do**
  The variational E-step:
  Update the variational solutions to $Q(\boldsymbol{\phi}_t)$ and $Q(\boldsymbol{Z}_t)$ using (15) and (16).
  The variational M-step:
  Compute learning rate $\rho_t = (\eta_0 + t)^{-a}$.
  Calculate the following natural gradients: $\Delta u_{jl}^{*(t)}$, $\Delta v_{jl}^{*(t)}$, $\Delta p_{jl}^{*(t)}$, $\Delta q_{jl}^{*(t)}$, $\Delta g_{jl}^{*(t)}$, $\Delta h_{jl}^{*(t)}$, $\Delta s_{jl}^{*(t)}$ and $\Delta k_{jl}^{*(t)}$, $\Delta c_j^{(t)}$, $\Delta d_j^{(t)}$, $\Delta \varsigma_1^{*(t)}$ and $\Delta \varsigma_2^{*(t)}$.
  Update the variational solutions to $Q^{(t)}(\boldsymbol{\lambda})$, $Q^{(t)}(\boldsymbol{\epsilon})$, $Q^{(t)}(\boldsymbol{\alpha})$, $Q^{(t)}(\boldsymbol{\beta})$, $Q^{(t)}(\boldsymbol{\sigma})$ and $Q^{(t)}(\boldsymbol{\tau})$.
  Repeat the variational *E-step* and *M-step* until new data is observed.
end for

4. Experimental Results

In this section, we evaluate the effectiveness of the proposed online infinite GD mixture model with localized feature selection (noted as *OIGDLFs*) using both synthetic data and two challenging applications involving online text document clustering and online human face detection. In our experiments, the initial truncation level M is set to 15. Initial values of hyperparameters u, p, g and s of the Gamma priors are set to 1, and v, q, h, k are set to 0.01. The hyperparameters ψ, ς_1 and ς_2 are set to 0.1. The parameters a and η_0 of the learning rate are set to 0.8 and 64, respectively. Our simulations have supported these specific choices.

4.1. Synthetic Data

The goal of synthetic data is to evaluate the performance of the proposed *OIGDLFs* in terms of estimation (estimating the model's parameters) and selection (selecting the number of components of the mixture model), on two 10-dimensional synthetic data sets (two relevant features and eight irrelevant features). We ran the proposed algorithm 10 times, the actual and average estimated parameters of the distributions representing the relevant features for each data set using the proposed online algorithm are shown in Table 1. According to this table, the parameters of the model, representing relevant features, and its mixing coefficients are accurately estimated by the *OIGDLFs*. Although we do not show the estimated values of the parameters of the irrelevant features (the eight remaining features), accurate results (in terms of parameters estimation) were obtained by adopting the proposed algorithm as well. Figure 1 gives the feature saliencies of all the 10 features for each data set. It obviously shows that features 1 and 2 have been assigned a high degree of relevance, which matches the ground-truth. Furthermore, we have also investigated the learning time of *OIGDLFs* and compared it to its batch version (denoted as *IGDLFs*) on the same data sets using a computer with Intel's Core2 Duo processor 2.00 GHz. As we can see from Table 2, *OIGDLFs* is more than three times faster than *IGDLFs* for each data set. This phenomenon becomes more obvious as the amount of data increases.

Table 1: Parameters of the synthetic data. N denotes the total number of elements, N_j denotes the number of elements in cluster j. α_{j1}, α_{j2}, β_{j1}, β_{j2} and π_j are the real parameters. $\hat{\alpha}_{j1}$, $\hat{\alpha}_{j2}$, $\hat{\beta}_{j1}$, $\hat{\beta}_{j2}$ and $\hat{\pi}_j$ are the average estimated parameters using the proposed algorithm.

	N_j	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	π_j	$\hat{\alpha}_{j1}$	$\hat{\beta}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\beta}_{j2}$	$\hat{\pi}_j$
Data set 1	200	1	10	15	23	12	0.50	10.38	15.46	22.04	11.58	0.503
($N = 400$)	200	2	20	27	30	25	0.50	21.15	28.29	31.52	24.37	0.497
Data set 2	200	1	10	15	23	12	0.25	9.26	14.51	24.23	11.64	0.247
($N = 800$)	200	2	20	27	30	25	0.25	20.98	26.17	28.96	24.39	0.255
	400	3	18	35	8	22	0.50	17.45	33.38	7.69	21.58	0.498
Data set 3	250	1	10	15	23	12	0.25	9.31	15.82	24.55	12.73	0.252
($N = 1000$)	250	2	20	27	30	25	0.25	19.41	27.68	31.72	26.25	0.246
	250	3	18	35	8	22	0.25	18.56	36.19	8.49	22.81	0.257
	250	4	37	18	43	10	0.25	38.22	17.73	41.96	10.18	0.259
Data set 4	400	1	10	15	23	12	0.20	10.45	14.33	23.72	12.51	0.198
($N = 2000$)	400	2	20	27	30	25	0.20	19.26	26.53	29.40	25.87	0.203
	400	3	18	35	8	22	0.20	17.85	34.06	7.28	23.19	0.205
	400	4	37	18	43	10	0.20	36.54	18.67	44.31	9.49	0.196
	400	5	16	25	40	33	0.20	16.95	24.18	39.05	34.34	0.198

Table 2: Runtime (in seconds) comparison for different data sets using the online and the batch algorithms.

Method	*OIGDLFs*	*IGDLFs*
Data set 1	5.76	17.13
Data set 2	11.52	39.86
Data set 3	14.28	47.18
Data set 4	32.64	121.01

4.2. Online Text Document Clustering

Text Document clustering is the process of grouping similar unlabeled documents together into a set of categories. During the last decade, the problem of text clustering has been the topic of extensive research (Lewis et al., 2004; Kim et al., 2005). It is a crucial step in various applications such as text retrieval, news filtering, e-mails classification, and browsing document collections. In this experiment, we focus on online text clustering. We test the performance of the proposed approach on five well-known data sets that are exten-

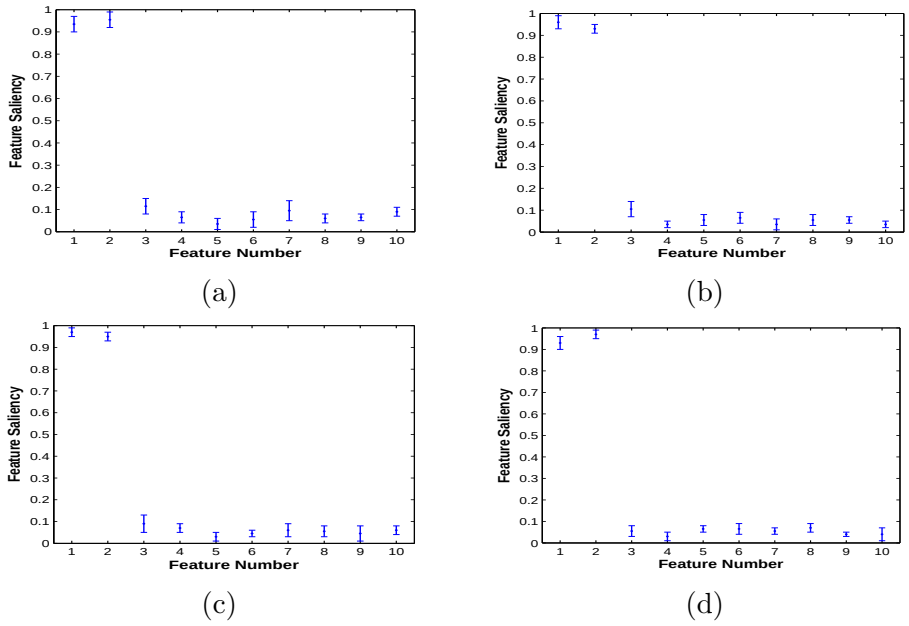

Figure 1: Feature saliency for synthetic data sets with one standard deviation over ten runs. (a) Data set 1, (b) Data set 2, (c) Data set 3, (d) Data set 4.

sively used in the information retrieval literature: CSTR, WebKB[2], WebKB4, Reuters10 and 20Newsgroups[3]. The CSTR data set contains 476 abstracts of technical reports published by the computer science department of university of Rochester from year 1991 to 2002. These documents are divided into four research areas: Natural Language Processing(NLP), Robotics/Vision, Systems, and Theory. The WebKB data set consists of 8300 web pages collected from university computer science departments. These documents are divided into seven categories: student, faculty, staff, course, project, department and other. The WebKB4 data set, which is a subset of the WebKB data set, containing 4,199 web pages and limited to the four most common categories: Course, Faculty, Project, and Student. The well-known corpus Reuters-21578, which contains 135 categories, is composed of documents collected from the Reuters newswire in 1987. In our work, we use a subset of the Reuters-21578 which includes the 10 most frequent categories among the 135 topics and we call it Reuters10. The 20Newsgroups data set contains approximately 20,000 newsgroup documents that are evenly partitioned across 20 different newsgroups. The characteristics of these data sets are summarized in Table 3.

The methodology of our online text clustering approach can be described as following: First, the Rainbow package[4] is used in a preprocessing step to select the top 500 words by removing the rare (occurred less than 30 times) and stop words (such as "the", "and", "or", etc.). Next, each document is represented by a vector of counts (i.e. a histogram that containing the frequency of occurrence of each word in its vocabulary). After apply-

2. Available at: http://www.cs.cmu.edu/∼textlearning/
3. Available at: http://qwone.com/∼jason/20Newsgroups/
4. http://www.cs.cmu.edu/∼mccallum/bow/

Table 3: The description of text data sets

Data sets	No. of documents	No. of classes
CSTR	476	4
WebKB4	4199	4
WebKB	8280	7
Reuters10	2900	10
20Newsgroups	20000	20

ing the geometric transformation presented in Section 2, these vectors are then modeled by our online infinite mixture model using the algorithm proposed in the previous section. Finally, the classification is performed by applying Bayes' decision rule. We run the algorithm 20 times to investigate its performance. Moreover, we investigate the advantages of the proposed *OIGDLFs* approach by comparing it to: the online infinite GD mixture model with global feature selection (*OIGDGFs*), the online infinite GD mixture model without feature selection (*OIGD*), the online finite GD mixture model with localized feature selection (*OFGDLFs*) and the online infinite Gaussian mixture model with localized feature selection (*OIGMLFs*). To make a fair comparison, all of these methods are learned by variational inference. Since our target is to obtain a low misclassification error and high discrimination among different classes in a data set and the document data sets used in our experiments are relatively balanced, we use classification accuracy as the evaluation measure. The average performance of text clustering is illustrated in Table 4 and Figure 2

Table 4: Text clustering results (average and standard deviation over 20 runs)

Method	CSTR	WebKB4	WebKB	Reuters10	20Newsgroups
OIGDLFs	86.28 (1.33)	83.27 (0.96)	76.38 (1.07)	79.51 (1.19)	89.79 (0.89)
OIGDGFs	84.13 (1.52)	81.45 (1.21)	73.82 (1.28)	76.67 (1.14)	87.11 (0.96)
OIGD	81.57 (1.89)	79.37 (1.43)	71.29 (1.21)	74.35 (1.31)	85.36 (0.85)
OFGDLFs	79.62 (1.77)	78.20 (1.37)	68.59 (1.45)	73.41 (1.29)	83.72 (1.15)
OIGMLFs	81.97 (1.65)	80.25 (1.19)	70.95 (1.33)	73.89 (1.22)	84.13 (0.92)

using different approaches. We have also shown the estimated number of classes for each data sets in Table 5. As we can see from these results, the proposed *OIGDLFs* obtains the best performance in terms of the highest average classification accuracy rate and the most accurate estimated number of classes for all the data sets. Clearly, the approach with local feature selection (*OIGDLFs*) performs better than both the one with global feature selection (*OIGDGFs*) and the one without feature selection (*OIGD*). We may also notice that, the finite mixture model approach *OFGDLFs* performs the worst in comparison with infinite approaches, especially in detecting the correct number of components as shown in Table 5, which demonstrates the advantage of using infinite mixture models. In addition, in Table 4, we can observe that *OIGDLFs* outperforms *OIGMLFs* approach, which verifies

the fact that GD mixture models have better modeling capability than Gaussian mixtures for proportional data.

Table 5: Estimated number of classes (average and standard deviation over 20 runs)

Method	CSTR	WebKB4	WebKB	Reuters10	20Newsgroups
OIGDLFs	3.31 (0.61)	3.43 (0.53)	6.21 (0.59)	8.77 (0.72)	18.12 (0.64)
OIGDGFs	3.14 (0.55)	3.23 (0.62)	6.15 (0.62)	8.69 (0.77)	17.37 (0.79)
OIGD	3.05 (0.69)	3.18 (0.68)	6.06 (0.54)	8.61 (0.68)	17.15 (0.57)
OFGDLFs	2.97 (0.58)	2.91 (0.72)	5.98 (0.71)	8.43 (0.82)	16.83 (0.83)
OIGMLFs	3.23 (0.66)	3.26 (0.55)	6.11 (0.56)	8.65 (0.76)	17.67 (0.59)

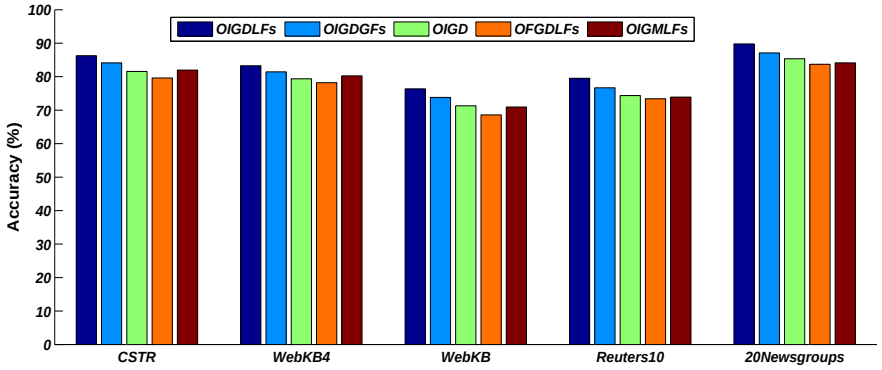

Figure 2: Performance comparison on text data sets using different approaches.

Figure 3: Sample images of the Caltech face and the Caltech background data sets: the first row contains face samples, the second row represents background samples.

4.3. Online Human Face Detection

Face detection is a important task in computer vision and has been applied in various applications such as video surveillance, image database management and human-computer interfaces. The goal of face detection is to distinguish the images that contain human faces from non-face ones. It is also a common preprocessing step for a facial recognition system (see, for instance, (Li et al., 2002; Amit and Trouvé, 2007)). In order to develop an effective face detection system, two important aspects have to be considered: facial representation and classifier design. The aim of facial representation is to extract discriminative low level features from raw face images. In our work, we adopt the local binary patterns (LBP)[5] feature proposed by Hadid et al. (2004) which has shown promising results in the filed of face detection. The proposed *OIGDLFs* is employed as the classifier in this experiment for discriminating face images from non-face ones.
The data set that we have used for face detection is the Caltech face data set[6]. It contains 450 front human face images which are recorded under natural conditions (i.e. varying illumination, expressions and complex background). The Caltech background data set (450 images) was adopted for non-face images [7]. Sample images from the Caltech face and the Caltech background data sets are displayed in Figure 3. The first step in this experiment is to extract LBP features from raw images by encoding both local and global facial characteristics into a compact feature histogram. As a result, each image was represented by a 203-dimensional histogram vector. Then, we perform our approach directly (i.e., we do not separate the data set into training and test sets) as a classifier to detect human faces by assigning the sequential arriving image to the group (face or non-face) that most likely generated it. We evaluated the detection performance of the proposed algorithm by running it 20 times. Moreover, we compare our approach with the following approaches: *OIGDGFs*, *OIGD*, *OFGDLFs* and *OIGMLFs*. The average performance of images categorization is illustrated in Table 6 for different approaches. According to this table, the proposed *OIGDLFs* outperforms the other four approaches in terms of the highest average classification rate (91.69%). The corresponding feature saliencies of the 203-dimensional

Table 6: The average classification accuracy rate (%) and the corresponding standard deviation obtained over 20 runs using different methods.

Methods	*OIGDLFs*	*OIGDGFs*	*OIGD*	*OFGDLFs*	*OIGMLFs*
Accuracy (%)	91.69 (1.08)	89.72 (1.26)	87.04 (1.17)	86.35 (1.34)	88.51 (1.29)

histogram vector obtained by *OIGDLFs* are illustrated in Figure 4. According to this figure, it is clear that the different features do not contribute equally in the classification, since they have different relevance degrees.

5. Source code available at: http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab
6. http://www.robots.ox.ac.uk/∼vgg/data.html.
7. http://www.vision.caltech.edu/html-files/archive.html.

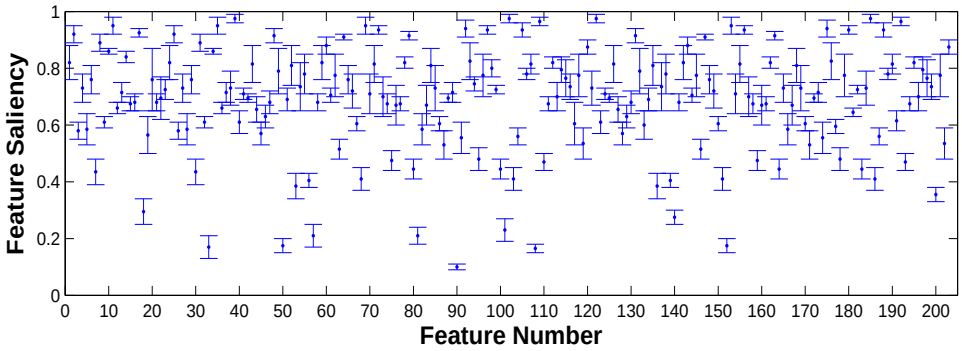

Figure 4: Feature saliencies obtained using *OIGDLFs* over 20 runs.

5. Conclusion

In this paper, we have proposed a novel online approach for simultaneous clustering and localized feature selection based on variational learning of infinite GD mixture models. Within this framework, the model parameters and features saliencies are estimated simultaneously and effectively while the difficulty of determining the number of clusters is solved in an elegant way. The effectiveness of the proposed approach has been evaluated on both synthetic data sets and two real applications regarding online text clustering and online face detection.

Acknowledgments

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

D. Achlioptas. Database-Friendly Random Projections. In *Proc. of the twentieth ACM symposium on Principles of database systems (PODS)*, pages 274–281. ACM, 2001.

S. Amari. Natural Gradient Works Efficiently in Learning. *Neural Comput.*, 10:251–276, 1998.

Y. Amit and A. Trouvé. POP: Patchwork of Parts Models for Object Recognition. *International Journal of Computer Vision*, 75:267–282, 2007.

H. Attias. A Variational Bayes Framework for Graphical Models. In *Proc. of Neural Information Processing Systems (NIPS)*, pages 209–215, 1999.

E. Bingham and H. Mannila. Random Projection in Dimensionality Reduction: Applications To Image and Text Data. In *Proc. of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 245–250, 2001.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D. M. Blei and M. I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1:121–144, 2005.

N. Bouguila and D. Ziou. A Powreful Finite Mixture Model Based on the Generalized Dirichlet Distribution: Unsupervised Learning and Applications. In *ICPR (1)*, pages 280–283, 2004.

N. Bouguila and D. Ziou. MML-Based Approach for High-Dimensional Learning using the Generalized Dirichlet Mixture. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, page 53. IEEE Computer Society, 2005.

N. Bouguila and D. Ziou. A Dirichlet Process Mixture of Generalized Dirichlet Distributions for Proportional Data Modeling. *IEEE Transactions on Neural Networks*, 21(1):107–122, 2010.

N. Bouguila, D. Ziou, and R. I. Hammoud. A Bayesian Non-Gaussian Mixture Analysis: Application to Eye Modeling. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

N. Bouguila, D. Ziou, and R. I. Hammoud. On Bayesian Analysis of a Finite Generalized Dirichlet Mixture Via a Metropolis-within-Gibbs Sampling. *Pattern Analysis and Applications*, 12(2):151–166, 2009.

S. Boutemedjet, N. Bouguila, and D. Ziou. A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1429–1443, 2009.

C. Constantinopoulos, M.K. Titsias, and A. Likas. Bayesian Feature And Model Selection for Gaussian Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (6):1013–1018, 2006.

L. Engebretsen, P. Indyk, and R. O'Donnell. Derandomized Dimensionality Reduction With Applications. In *Proc. of the thirteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 705–712, 2002.

W. Fan, N. Bouguila, and D. Ziou. Unsupervised Anomaly Intrusion Detection via Localized Bayesian Feature Selection. In *Proc. of International Conference on Data Mining (ICDM)*, pages 1032–1037, 2011.

T. S. Ferguson. Bayesian Density Estimation by Mixtures of Normal Distributions. *Recent Advances in Statistics*, 24:287–302, 1983.

D. Fradkin and D. Madigan. Experiments With Random Projections For Machine Learning. In *Proc. of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 517–522, 2003.

Y. Guan, J. G. Dy, and M. I. Jordan. A Unified Probabilistic Model for Global and Local Unsupervised Feature Selection. In *Proc. of International Conference on Machine Learning (ICML)*, pages 1073–1080, 2011.

A. Hadid, M. Pietikainen, and T. Ahonen. A Discriminative Feature Space for Detecting and Recognizing Faces. In *Proc. of CVPR 2004*, volume 2, pages 797–804, 2004.

M. D. Hoffman, D. M. Blei, and F. Bach. Online Learning for Latent Dirichlet Allocation. In *Proc. of Neural Information Processing Systems (NIPS)*, pages 856–864, 2010.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

S. Kaski. Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering. In *Proc. of International Conference on Neural Networks (IJCNN)*, pages 413–418, 1998.

E. Keogh and M. Pazzani. Learning Augmented Bayesian Classifiers: A Comparison of Distribution-Based and Classification-Based Approaches. In *Proc. of the seventh international workshop on artificial intelligence and statistics (AISTAT)*, pages 225–230, 1999.

H. Kim, P. Howland, and H. Park. Dimension Reduction in Text Classification with Support Vector Machines. *Journal of Machine Learning Research*, 6:37–53, 2005.

R. M. Korwar and M. Hollander. Contributions to the theory of Dirichlet processes. *The Annals of Probability*, 1:705–711, 1973.

H.J. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*. Applications of mathematics. Springer, 1997.

M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9): 1154–1166, 2004.

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.

S. Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical Learning of Multi-view Face Detection. In *Proc. of the 7th European Conference on Computer Vision (ECCV)*, pages 67–81, 2002.

Y. Li, M. dong, and J. Hua. Simultaneous Localized Feature Selection and Model Detection for Gaussian Mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:953–960, 2009.

Z. Ma and A. Leijon. Bayesian Estimation of Beta Mixture Models with Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2160 – 2173, 2011.

G.J. McLachlan and D. Peel. *Finite Mixture Models*. New York: Wiley, 2000.

R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

M. Sato. Online Model Selection Based on the Variational Bayes. *Neural Comput.*, 13:1649–1681, 2001.

J. Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:705–711, 2004.

C. Wang, J. W. Paisley, and D. M. Blei. Online Variational Inference for the Hierarchical Dirichlet Process. *Journal of Machine Learning Research - Proceedings Track*, 15:752–760, 2011.