Knowl Inf Syst
DOI 10.1007/s10115-011-0467-4

REGULAR PAPER

A countably infinite mixture model for clustering and feature selection

Nizar Bouguila · Djemel Ziou

Received: 25 August 2009 / Revised: 8 October 2011 / Accepted: 8 December 2011

Abstract Mixture modeling is one of the most useful tools in machine learning and data mining applications. An important challenge when applying finite mixture models is the selection of the number of clusters which best describes the data. Recent developments have shown that this problem can be handled by the application of non-parametric Bayesian techniques to mixture modeling. Another important crucial preprocessing step to mixture learning is the selection of the most relevant features. The main approach in this paper, to tackle these problems, consists on storing the knowledge in a generalized Dirichlet mixture model by applying non-parametric Bayesian estimation and inference techniques. Specifically, we extend finite generalized Dirichlet mixture models to the infinite case in which the number of components and relevant features do not need to be known a priori. This extension provides a natural representation of uncertainty regarding the challenging problem of model selection. We propose a Markov Chain Monte Carlo algorithm to learn the resulted infinite mixture. Through applications involving text and image categorization, we show that infinite mixture models offer a more powerful and robust performance than classic finite mixtures for both clustering and feature selection.

Keywords Non-parametric Bayesian methods · Dirichlet process · Clustering · Feature selection · Mixture models · Generalized Dirichlet · MCMC · Categorization

1 Introduction

Categorization is an important problem in the case of natural language texts and images processing due to the increased generation of digital documents (images or text). There is a wide

N. Bouguila (✉)
CIISE, Concordia University, Montreal, QC H3G 1T7, Canada
e-mail: bouguila@ciise.concordia.ca

D. Ziou
Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
e-mail: djemel.ziou@usherbrooke.ca

Published online: 21 December 2011

body of literature in several fields concerned with the categorization problem which involves grouping objects into categories such that elements of a category are similar according to a certain number of features, and members of different categories are different. Learning techniques are frequently used for this task by building and training classifiers from a set of preclassified examples. Finite mixture models are one of the most used learning techniques and have several features, such as providing a rich tractable and flexible basis for statistical inference, that make them attractive for categorization and clustering problems [1]. One of the most challenging aspects, when using finite mixture models, is usually to estimate the number of clusters that best describes the data without over- or under-fitting it. For this purpose, many approaches have been suggested and have resulted in a large number of algorithms [1–3]. The majority of these approaches, however, are unsuitable for complex real-world data mining problems where data sets grow over time in the presence of uncertainty and incompleteness. Another important challenge is the large dimensionality of the data that we have to deal with [4]. Several features may be irrelevant and can have a profoundly negative effect upon model learning [5]. Thus, a crucial preprocessing step is the selection of the most relevant features (i.e, the features with higher probabilities to be informative) that best describe a given image or document [6].

In this paper (this is an extended and revised version of [7]), we are interested in Bayesian non-parametric approaches for modeling and selection using mixture of Dirichlet processes [8], which have been shown to be a powerful alternative to determine the number of clusters [9,10]. In contrast with classic Bayesian approaches, which suppose an unknown finite number of mixture components, non-parametric Bayesian approaches assume infinitely complex models (i.e., an infinite number of components) and have witnessed considerable theoretical and computational advances in recent years [11]. Indeed, non-parametric Bayesian approaches allow the increasing of the number of mixture components to infinity, which removes the problems underlying the determination of the number of clusters that can increase or decrease as new data arrive (i.e., as more data are seen, more details about the statistical model are revealed). Because of their simplicity, and thanks to the development of MCMC techniques, infinite mixture models based on Dirichlet processes are now widely used, as a new paradigm for unsupervised learning, in different domains and variety of applications [12].

The majority of the work with infinite mixture models makes the Gaussian assumption [13]. However, we have shown in previous works that other distributions such as the Dirichlet [2] and the generalized Dirichlet [3,14] can give better results in some applications and are more appropriate especially when modeling proportion vectors. It is noteworthy that, in contrast with some of our previous works where the generalized Dirichlet is used as a prior to the multinomial distribution [15], the generalized Dirichlet is used as the parent distribution in this paper to model directly the data and select its most relevant features. We propose then an infinite mixture model based on generalized Dirichlet distributions for text and image categorization using bag-of-words approach (visual words in the case of images). We propose also a MCMC procedure that makes use of both Gibbs and Metropolis-Hastings updates to learn this model via posteriors inference. The proposed infinite model can be viewed as a natural extension to the models proposed in [3] and [16] where expectation-maximization and Bayesian algorithms have been proposed to learn finite generalized Dirichlet mixtures.

The rest of the paper is structured as follows: Firstly, we review the generalized Dirichlet mixture (Sect. 2), and then we propose its extension to the infinite case for clustering and feature selection, and we give the complete learning algorithm (Sect. 3). Experimental results are illustrated in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Finite generalized Dirichlet mixture modeling and feature selection

2.1 Generalized Dirichlet mixture

Given an independent and identically distributed (iid) set of D-dimensional vectors $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$, representing images or text documents, from a generalized Dirichlet mixture, the associated likelihood function is

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} p(\vec{X}_i|\Theta) \tag{1}$$

where

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p(Z_i = j)p(\vec{X}_i|\theta_j) \tag{2}$$

Here, M is the number of clusters present on the set of vectors, and Z_i indicates from which cluster each vectors $\vec{X}_i$ arose (i.e., $Z_i = j$ means that $\vec{X}_i$ comes from component j). $p_j = p(Z_i = j)$ represents the a priori probability that the vector $\vec{X}_i$ was generated by component j, and it follows from Bayes' theorem that $p(Z_i = j|\vec{X}_i)$, the probability that vector i is in cluster j, conditional on having observed $\vec{X}_i$ is given by $p(Z_i = j|\vec{X}_i) \propto p_j p(\vec{X}_i|\theta_j)$. Of course, being probabilities, the p_j must satisfy: $p_j > 0$, $j = 1, \ldots, M$, and $\sum_{j=1}^{M} p_j = 1$. Finally, $\Theta = (\theta_1, \ldots, \theta_M, p_1, \ldots, p_M)$ is the set of all the model parameters and $p(\vec{X}_i|\theta_j)$ is our probability density function taken as a generalized Dirichlet with parameters $\theta_j = \{\alpha_{j1}, \beta_{j1}, \ldots, \alpha_{jD}, \beta_{jD}\}$:

$$p(\vec{X}_i|\theta_j) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_{jd} + \beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})} X_{id}^{\alpha_{jd}-1} \left(1 - \sum_{k=1}^{d} X_{ik}\right)^{\gamma_{jd}} \tag{3}$$

where $\sum_{d=1}^{D} X_{id} < 1$ 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 $\gamma_{jD} = \beta_{jD} - 1$. Note that the generalized Dirichlet distribution is reduced to a Dirichlet distribution [2] when $\beta_{jd} = \alpha_{jd+1} + \beta_{jd+1}$, $d = 1, \ldots, D-1$.

The main approach to estimate the parameters of a mixture model is to maximize the likelihood through the Expectation-Maximization (EM) algorithm which is the theoretical framework first introduced in the seminal paper by Dempster et al. [17]. In [18], the authors proposed an EM-based algorithm for the estimation of the generalized Dirichlet mixture parameter. This estimation is based on an interesting property of the generalized Dirichlet distribution that we shall exploit in this paper. Indeed, if a vector $\vec{X}_i$ has a generalized Dirichlet distribution with parameters $(\alpha_1, \beta_1, \ldots, \alpha_D, \beta_D)$, then we can construct a vector $\vec{Y}_i = (Y_{i1}, \ldots, Y_{iD})$ using the following geometric transformation: $Y_{id} = X_{id}$ if $d = 1$ and $Y_{id} = \frac{X_{id}}{(1-X_{i1}-\ldots-X_{i,D-1})}$ for $d = 2, 3, \ldots, D$. In this vector $\vec{Y}_i$, each Y_{id}, $d = 1, \ldots, D$ has a Beta distribution with parameters α_d and β_d. Boutemedjet et al. [19] have shown that this interesting property allows the factorization of $p(Z_i = j|\vec{X}_i)$ as

$$p(Z_i = j|\vec{X}_i) \propto p_j \prod_{d=1}^{D} \frac{\Gamma(\alpha_{jd} + \beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})} Y_{id}^{\alpha_{jd}-1}(1 - Y_{id})^{\beta_{jd}-1} \tag{4}$$

This factorization means that clustering $\mathcal{X}$ using a finite generalized Dirichlet mixture model can be done via the clustering of $\mathcal{Y} = (\vec{Y}_1, \ldots, \vec{Y}_N)$ using the following mixture model with conditionally independent features

$$p(\vec{Y}_i|\xi, P) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} p_{Beta}(Y_{id}|m_{jd}, s_{jd}) \tag{5}$$

where

$$p_{Beta}(Y_{id}|m_{jd}, s_{jd}) = \frac{\Gamma(s_{jd})}{\Gamma(s_{jd}m_{jd})\Gamma(s_{jd}(1-m_{jd}))} Y_{id}^{s_{jd}m_{jd}-1}(1-Y_{id})^{s_{jd}(1-m_{jd})-1} \tag{6}$$

$s_{jd} = \alpha_{jd} + \beta_{jd}$ and $m_{jd} = \frac{\alpha_{jd}}{s_{jd}}$ and can be viewed as scales and locations, respectively. Note that this alternative parametrization was also adopted in the case of the Bayesian estimation of finite Beta mixtures, in [20], since it provides interpretable parameters. $\xi = (\xi_1, \ldots, \xi_M)$, $\xi_j = (\vec{s}_j, \vec{m}_j)$, $\vec{s}_j = (s_{j1}, \ldots, s_{jD})$, and $\vec{m}_j = (m_{j1}, \ldots, m_{jD})$ are the set of parameters defining the j-th component, and $P = (p_1, \ldots, p_M)$. This means that generalized Dirichlet mixture models have the ability to reduce complex multidimensional clustering problems to a sequence of one-dimensional ones [18].

2.2 Feature selection

It is largely recognized that a good selection/weighting of features is an important requirement for successful machine learning models. Indeed, the model structure is generally contained in only a part of the available features. Feature selection generally improves learning capabilities. For instance, the learning can be performed with less training data. Various feature selection approaches have been proposed throughout the years. However, the majority of these approaches have been devoted to the supervised case. This has motivated recent research to revisit feature selection in the case of mixture models (see, for instance, [19] and references therein) where the problem can be approached as following

$$p(\vec{Y}_i|\Xi) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \left[\rho_d p_{Beta}(Y_{id}|m_{jd}, s_{jd}) + (1-\rho_d) p_{Beta}(Y_{id}|m_d^{irr}, s_d^{irr})\right] \tag{7}$$

where $\Xi = \{\xi, P, \rho, \xi^{irr}\}$ is the set of all the model parameters, $\rho = (\rho_1, \ldots, \rho_D)$, $\xi^{irr} = (\vec{m}^{irr}, \vec{s}^{irr})$, $\vec{m}^{irr} = (m_1^{irr}, \ldots, m_D^{irr})$, $\vec{s}^{irr} = (s_1^{irr}, \ldots, s_D^{irr})$.

$p_{Beta}(Y_{id}|m_d^{irr}, s_d^{irr})$ can be viewed as a background distribution, common to all mixture components, to explain irrelevant features. $\rho_d = p(z_{id} = 1)$ represents the probability that the dth feature is relevant for clustering where z_{id} is a hidden variable equal to 1 if the dth feature of $\vec{X}_i$ is relevant (i.e., generated from the component $p_{Beta}(Y_{id}|m_{jd}, s_{jd})$) and 0 (i.e., generated from the background component $p_{Beta}(Y_{id}|m_d^{irr}, s_d^{irr})$), otherwise. Notice that the model in Eq. 7 allows to include the feature selection as a part of the clustering and is reduced to the mixture model in Eq. 5 when $\rho_d = 0, d = 1, \ldots, D$. A graphical representation of the previous model is shown in Fig. 1 where $z_i = (z_{i1}, \ldots, z_{iD})$.

2.3 Bayesian modeling

The EM algorithm and its several extensions try to estimate a single "best" model that is not generally realistic, since the data may suggest many "good" models. A better approach is to consider the average result computed over several models, which can be done through Bayesian approaches [21]. The major difference between Bayesian and deterministic (or standard likelihood) approaches is that with Bayesian inference, we modify the likelihood into a posterior distribution as following

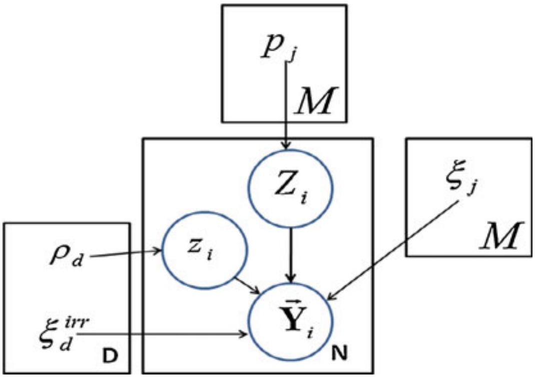


Fig. 1 Graphical model representation for the model in Eq. 7 based on finite generalized Dirichlet mixture for feature selection. *Nodes* in this graph represent random variables, *boxes* indicate repetition (with the number of repetitions in the *lower right*) and arcs describe conditional dependencies between variables

$$p(\Xi|\mathcal{Y}) \propto p(\mathcal{Y}|\Xi)p(\Xi) \tag{8}$$

where $p(\Xi)$ is the prior distribution. This means that the mixture parameters will themselves be considered as random variables, and then the ultimate goal will be the estimation of a distribution over the parameters rather than a single set of parameters. The joint distribution of all variables in the model can be written as

$$\begin{aligned} p(P, Z, \rho, z, \xi, \xi^{irr}, \mathcal{Y}) &= p(P)p(Z|P)p(\rho|P, Z)p(z|\rho, P, Z)p(\xi|P, Z, \rho, z) \\ &\quad \times p(\xi^{irr}|P, Z, \rho, z, \xi) \times p(\mathcal{Y}|P, Z, \rho, z, \xi, \xi^{irr}) \end{aligned} \tag{9}$$

where $z = (z_1, \ldots, z_N)$. It is worth mentioning that if we condition on Z and z, the distribution of $\mathcal{Y}$ is simply given by

$$\begin{aligned} p(\mathcal{Y}|P, Z, \rho, z, \xi, \xi^{irr}) &= p(\mathcal{Y}|\xi, \xi^{irr}, Z, z) \\ &= \prod_{i=1}^{N}\prod_{d=1}^{D}\left[\left(p_{Beta}(Y_{id}|m_{Z_i d}, s_{Z_i d})\right)^{z_{id}}\left(p_{Beta}(Y_{id}|m_d^{irr}, s_d^{irr})\right)^{1-z_{id}}\right] \end{aligned} \tag{10}$$

Besides, we impose commonly used further conditional independencies, so that $p(\rho|P, Z) = p(\rho)$, $p(z|\rho, P, Z) = p(z|\rho)$, $p(\xi|P, Z, \rho, z) = p(\xi)$, $p(\xi^{irr}|P, Z, \rho, z, \xi) = p(\xi^{irr})$, $p(\xi|Z, P) = p(\xi)$, thus

$$p(P, Z, \rho, z, \xi, \xi^{irr}, \mathcal{Y}) = p(P)p(Z|P)p(\rho)p(z|\rho)p(\xi)p(\xi^{irr})p(\mathcal{Y}|\xi, \xi^{irr}, Z, z)$$

The selection of the prior distribution is an important issue in Bayesian statistics. In our case, we suppose that ξ, ξ^{irr}, ρ, and P follow priors depending on hyperparameters, drawn from independent hyperpriors, Λ, Λ^{irr}, δ and η, respectively. In addition, our prior distributions are that $\vec{s}_j$, $\vec{m}_j$, $\vec{s}^{irr}$, and $\vec{m}^{irr}$ are all drawn independently:

$$p(\xi|\Lambda) = \prod_{j=1}^{M} p(\vec{s}_j|\Lambda_{js})p(\vec{m}_j|\Lambda_{jm}) \tag{11}$$

$$p(\xi^{irr}|\Lambda^{irr}) = p(\vec{s}^{irr}|\Lambda_s^{irr})p(\vec{m}^{irr}|\Lambda_m^{irr}) \tag{12}$$

where $\Lambda = (\Lambda_1, \ldots, \Lambda_M)$, $\Lambda_j = (\Lambda_{js}, \Lambda_{jm})$, and $\Lambda^{irr} = (\Lambda_s^{irr}, \Lambda_m^{irr})$. A standard choice as a prior for P is the Dirichlet distribution which is justified by the fact that the Dirichlet is conjugate to the multinomial [22]:

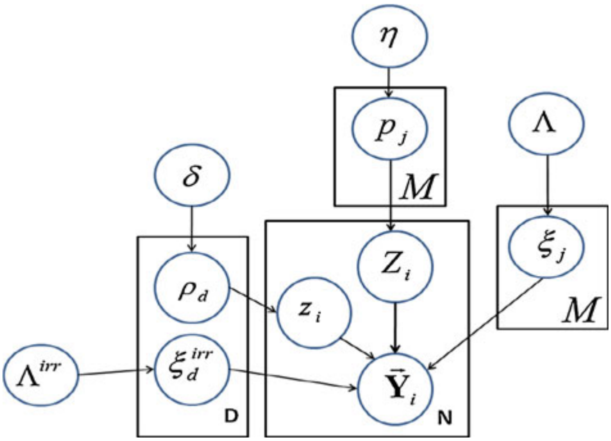


Fig. 2 Graphical representation for the Bayesian model

$$p(P|\eta_1,\ldots,\eta_M) = \frac{\Gamma(\sum_{j=1}^M \eta_j)}{\prod_{j=1}^M \Gamma(\eta_j)} \prod_{j=1}^M p_j^{\eta_j - 1} \tag{13}$$

where $(\eta_1,\ldots,\eta_M) \in \mathbb{R}^{+M}$ are the parameters of the Dirichlet. By taking $\eta_j = \frac{\eta}{M}$, $j = 1,\ldots,M$, where $\eta \in \mathbb{R}^+$, we obtain

$$p(P|\eta) = \frac{\Gamma(\eta)}{\Gamma(\frac{\eta}{M})^M} \prod_{j=1}^M p_j^{\eta - 1} \tag{14}$$

To add more flexibility to the model, it is common to assume that the hyperparameters $\eta, \delta, \Lambda^{irr}$, and Λ themselves follow certain distributions $p(\eta)$, $p(\delta)$, $p(\Lambda^{irr})$, and $p(\Lambda)$, respectively. A graphical model representing our Bayesian model is shown in Fig. 2. The joint distribution of all our model's variables is then expressed by the following factorization

$$\begin{aligned} p(P, Z, \rho, z, \xi, \xi^{irr}, \eta, \delta, \Lambda, \Lambda^{irr}, \mathcal{Y}) = {} & p(\eta)p(\delta)p(\rho|\delta)p(\Lambda)p(\Lambda^{irr})p(P|\eta)p(Z|P) \\ & \times p(\rho|\delta)p(z|\rho)p(\mathcal{Y}|\xi, \xi^{irr}, Z, z)p(\vec{s}^{irr}|\Lambda_s^{irr}) \\ & \times p(\vec{m}^{irr}|\Lambda_m^{irr}) \prod_{j=1}^M \Big(p(\vec{s}_j|\Lambda_{js})p(\vec{m}_j|\Lambda_{jm}) \Big) \end{aligned}$$

By applying Bayes rule and conditioning on the observed data, we can easily observe that the posterior distribution is proportional to the joint distribution

$$p(P, Z, \rho, z, \xi, \xi^{irr}, \eta, \delta, \Lambda, \Lambda^{irr}|\mathcal{Y}) \propto p(P, Z, \rho, z, \xi, \xi^{irr}, \eta, \delta, \Lambda, \Lambda^{irr}, \mathcal{Y}) \tag{15}$$

Having our posterior distribution in hand, we can simulate the model parameters using MCMC methods. For a self-contained entry into practical and computational Bayesian statistics using MCMC, the reader is referred to accessible expositions in [22].

3 Infinite mixture model

3.1 The infinite model

In the quest for creating more realistic and flexible models, several non-parametric Bayesian approaches have been proposed in the last few years. These approaches assume that the models, from which the data have been generated, may have an infinite number of parameters (i.e., the observations are assumed to belong to one of a potentially infinite number of clusters). There is a rich literature on infinite models (see, for instance, [10,13]). This interest is mainly driven by several applications and settings where data are dynamic. In this specific case, the arguable assumption of finite number of components is very restrictive, since the number of clusters may increase or decrease as new data are introduced. Infinite models generally circumvent these limitations by taking into account the model uncertainty and by introducing the more accurate assumption that the data come from a potentially infinite number of clusters of which only a finite subset is observed and can be updated as new data arrive. For example, in the case of knowledge discovery from Internet, some data pertain to known categories, while other newly extracted knowledge may correspond to unknown categories that should be created. This is the case in some scientific applications also where new discovered species cannot be assigned to already known categories.

In the following, we present the main idea behind infinite mixture models that consists actually on the possibility to create or remove clusters as new data arrive. According to Eq. 15, the only terms that involve P whose dimensionality is M are $p(Z|P)$ and $p(P|\eta)$. Recall that $p_j = p(Z_i = j)$, $j = 1, \ldots, M$, thus

$$p(Z|P) = \prod_{j=1}^{M} p_j^{n_j} \tag{16}$$

where $n_j = \sum_{i=1}^{N} \mathbb{I}_{Z_i=j}$ is the number of vector in cluster j. Because the Dirichlet is a conjugate prior to the multinomial, we can marginalize out P from Eq. 15

$$\begin{aligned} p(Z|\eta) &= \int_P p(Z|P)p(P|\eta)dP = \frac{\Gamma(\eta)}{\prod_{j=1}^{M}\Gamma\left(\frac{\eta}{M}\right)} \int_P \prod_{j=1}^{M} p_j^{n_j+\frac{\eta}{M}-1} dP \\ &= \frac{\Gamma(\eta)}{\Gamma(\eta+N)} \prod_{j=1}^{M} \frac{\Gamma\left(\frac{\eta}{M}+n_j\right)}{\Gamma(\frac{\eta}{M})} \end{aligned}$$

which can be considered as a prior on Z. We have also

$$p(P|Z,\eta) = \frac{p(Z|P)p(P|\eta)}{p(Z|\eta)} = \frac{\Gamma(\eta+N)}{\prod_{j=1}^{M}\Gamma\left(\frac{\eta}{M}+n_j\right)} \prod_{j=1}^{M} p_j^{n_j+\frac{\eta}{M}-1} \tag{17}$$

which is a Dirichlet distribution with parameters $(n_1 + \frac{\eta}{M}, \ldots, n_M + \frac{\eta}{M})$ from which we can show that:

$$p(Z_i = j|\eta, Z_{-i}) = \frac{n_{-i,j} + \frac{\eta}{M}}{N - 1 + \eta} \tag{18}$$

where $Z_{-i} = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_N\}$, $n_{-i,j}$ is the number of vectors, excluding $\vec{Y}_i$, in cluster j. Letting $M \to \infty$ in Eq. 18, the conditional prior gives the following limits [10,13]

$$p(Z_i = j|\eta, Z_{-i}) = \begin{cases} \frac{n_{-i,j}}{N-1+\eta} & \text{if } n_{-i,j} > 0 \text{ (cluster } j \in \mathcal{R}) \\ \frac{\eta}{N-1+\eta} & \text{if } n_{-i,j} = 0 \text{ (cluster } j \in \mathcal{U}) \end{cases} \tag{19}$$

where $\mathcal{R}$ and $\mathcal{U}$ are the sets of represented and unrepresented clusters, respectively. The previous equation describes actually Dirichlet processes that are sometimes described by analogy to Chinese restaurant processes [23], because it can be viewed as a sequential restaurant "seat arrangement" where customers arrive sequentially at a Chinese restaurant and then assigned, according to the current seating arrangement of previous customers, to an infinite number of tables having infinite number of seats. Note that despite the fact that the number of clusters is supposed to be infinite, the number of represented (i.e., non-empty) clusters is finite and should be between 1 and N. From Eq. 19, we can note also that if a cluster is represented, its conditional prior will depend on the number of observations assigned to it. Conditional prior for unrepresented clusters depends on η and N. Indeed, according to Eq. 19, a new cluster may appear with a probability $\frac{\eta}{N-1+\eta}$. Thus, the average number of clusters M is given by $\sum_{i=1}^{N} \frac{\eta}{\eta+i-1} \in \mathcal{O}(\eta \log N)$ [12], which shows that the number of clusters increases only in a logarithmic manner in the number of observations.

3.2 Model learning

The model learning is based on building estimates of the posterior distribution for our infinite generalized Dirichlet mixture model using MCMC methods. Posterior distributions in the case of mixture models are known to have intractable forms. A suitable technique to generate samples from such distributions is Gibbs sampling by updating each parameter in turn from its conditional posterior distribution, given all the rest of variables in the model. Interesting discussions about some recent inference approaches for Dirichlet process models can be found in [24]. In the following, we focus on the development of the conditional posteriors that we will use to perform Gibbs sampling.

3.2.1 Conditional posterior distributions of $\vec{m}_j$ and $\vec{m}^{irr}$

We know that each location m_{jd} is defined in the compact support [0, 1], then an appealing flexible choice as a prior is the Beta distribution, with location ε and scale ζ common to all components, which was found flexible in real applications. Thus, $\vec{m}_j$ for each cluster is given the following prior:

$$p(\vec{m}_j|\zeta, \varepsilon) \sim \prod_{d=1}^{D} \frac{\Gamma(\zeta) m_{jd}^{\zeta\varepsilon-1}(1-m_{jd})^{\zeta(1-\varepsilon)-1}}{\Gamma(\zeta\varepsilon)\Gamma(\zeta(1-\varepsilon))} \tag{20}$$

where $\vec{m}_j = (m_{j1}, \ldots, m_{jD})$. As for $\vec{m}^{irr}$, we consider the following prior with location ε^{irr} and scale ζ^{irr} common to all dimensions

$$p(\vec{m}^{irr}|\zeta^{irr}, \varepsilon^{irr}) \sim \prod_{d=1}^{D} \frac{\Gamma(\zeta^{irr}) {m_d^{irr}}^{\zeta^{irr}\varepsilon^{irr}-1}(1-{m_d}^{irr})^{\zeta^{irr}(1-\varepsilon^{irr})-1}}{\Gamma(\zeta^{irr}\varepsilon^{irr})\Gamma(\zeta^{irr}(1-\varepsilon^{irr}))} \tag{21}$$

So, the generic hyperparameters Λ_{jm} and Λ_m^{irr} become (ζ, ε) and $(\zeta^{irr}, \varepsilon^{irr})$, respectively. Thus, according to the previous two equation and our joint distribution in Eq. 15, the full conditional posterior distributions for $\vec{m}_j$ and $\vec{m}^{irr}$, giving the rest of the parameters, are as follows:

$$
\begin{aligned}
p(\vec{m}_j|\ldots) &\propto p(\vec{m}_j|\Lambda_{jm})p(\mathcal{Y}|\xi,\xi^{irr},Z,z)\\
&\propto \prod_{(Z_i,z_{id})=(j,1)} \frac{\Gamma(s_{jd})}{\Gamma(s_{jd}m_{jd})\Gamma(s_{jd}(1-m_{jd}))} Y_{id}^{s_{jd}m_{jd}-1}(1-Y_{id})^{s_{jd}(1-m_{jd})-1}\\
&\times \prod_{d=1}^{D} \frac{\Gamma(\zeta)}{\Gamma(\zeta\varepsilon)\Gamma(\zeta(1-\varepsilon))} m_{jd}^{\zeta\varepsilon-1}(1-m_{jd})^{\zeta(1-\varepsilon)-1}
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
p(\vec{m}^{irr}|\ldots) &\propto p(\vec{m}^{irr}|\Lambda_m^{irr})p(\mathcal{Y}|\xi,\xi^{irr},Z,z)\\
&\propto \prod_{z_{id}=0} \frac{\Gamma(s_d^{irr})}{\Gamma(s_d^{irr}m_d^{irr})\Gamma(s_d^{irr}(1-m_d^{irr}))} Y_{id}^{s_d^{irr}m_d^{irr}-1}(1-Y_{id})^{s_d^{irr}(1-m_d^{irr})-1}\\
&\times \prod_{d=1}^{D} \frac{\Gamma(\zeta^{irr})}{\Gamma(\zeta^{irr}\varepsilon^{irr})\Gamma(\zeta^{irr}(1-\varepsilon^{irr}))} m_d^{irr^{\zeta^{irr}\varepsilon^{irr}-1}}(1-m_d^{irr})^{\zeta^{irr}(1-\varepsilon^{irr})-1}
\end{aligned}
\tag{23}
$$

The hyperparameters ε, ε^{irr}, ζ, and ζ^{irr} associated with the m_{jd} and m_d^{irr} are given uniform and inverse Gamma priors as follows:

$$
p(\varepsilon) \sim \mathcal{U}_{[0,1]} \qquad p(\varepsilon^{irr}) \sim \mathcal{U}_{[0,1]} \tag{24}
$$

$$
p(\zeta|\varphi,\varrho) \sim \frac{\varrho^{\varphi}\exp(-\varrho/\zeta)}{\Gamma(\varphi)\zeta^{irr^{\varphi+1}}} \quad p(\zeta^{irr}|\varphi,\varrho) \sim \frac{\varrho^{\varphi}\exp(-\varrho/\zeta^{irr})}{\Gamma(\varphi)\zeta^{irr^{\varphi+1}}} \tag{25}
$$

Thus, according to Eqs. 15, 20, 21, and 24, we have

$$
p(\varepsilon|\ldots) \propto p(\varepsilon)\prod_{j=1}^{M} p(\vec{m}_j|\zeta,\varepsilon) \tag{26}
$$

$$
p(\varepsilon^{irr}|\ldots) \propto p(\varepsilon^{irr})p(\vec{m}^{irr}|\zeta^{irr},\varepsilon^{irr}) \tag{27}
$$

And according to Eqs. 15, 20, 21, and 25, we have

$$
p(\zeta|\ldots) \propto p(\zeta|\varphi,\varrho)\prod_{j=1}^{M} p(\vec{m}_j|\zeta,\varepsilon) \tag{28}
$$

$$
p(\zeta^{irr}|\ldots) \propto p(\zeta^{irr}|\varphi,\varrho)p(\vec{m}^{irr}|\zeta^{irr},\varepsilon^{irr}) \tag{29}
$$

3.2.2 Conditional posterior distribution of $\vec{s}_j$ and $\vec{s}^{irr}$

Since the scale s_{jd} controls the dispersion of the distributions, a common choice as a prior is an inverse gamma with shape σ and scale ϖ common to all components [25], then

$$
p(\vec{s}_j|\sigma,\varpi) \sim \prod_{d=1}^{D} \frac{\varpi^{\sigma}\exp(-\varpi/s_{jd})}{\Gamma(\sigma)s_{jd}^{\sigma+1}} \tag{30}
$$

As for $\vec{s}^{irr}$, we consider the following prior with shape σ^{irr} and scale ϖ^{irr} common to all dimensions

$$
p(\vec{s}^{irr}|\sigma^{irr},\varpi^{irr}) \sim \prod_{d=1}^{D} \frac{\varpi^{irr^{\sigma^{irr}}}\exp(-\varpi^{irr}/s_d^{irr})}{\Gamma(\sigma^{irr})s_d^{irr^{\sigma^{irr}+1}}} \tag{31}
$$

Having this priors, the full conditional posteriors for $\vec{s}_j$ and $\vec{s}^{irr}$ are

$$p(\vec{s}_j|\ldots) \propto p(\vec{s}_j|\Lambda_{js})p(\mathcal{Y}|\xi, \xi^{irr}, Z, z)$$
$$\propto \prod_{(Z_i, z_{id})=(j,1)} \frac{\Gamma(s_{jd})Y_{id}^{s_{jd}m_{jd}-1}(1-Y_{id})^{s_{jd}(1-m_{jd})-1}}{\Gamma(s_{jd}m_{jd})\Gamma(s_{jd}(1-m_{jd}))} \prod_{d=1}^{D} \frac{\varpi^{\sigma}\exp(-\varpi/s_{jd})}{\Gamma(\sigma)s_{jd}^{\sigma+1}} \tag{32}$$

$$p(\vec{s}^{irr}|\ldots) \propto p(\vec{s}^{irr}|\Lambda_s^{irr})p(\mathcal{Y}|\xi, \xi^{irr}, Z, z)$$
$$\propto \prod_{z_{id}=1} \frac{\Gamma(s_d^{irr})Y_{id}^{s_d^{irr}m_d^{irr}-1}(1-Y_{id})^{s_d^{irr}(1-m_d^{irr})-1}}{\Gamma(s_d^{irr}m_d^{irr})\Gamma(s_d^{irr}(1-m_d^{irr}))} \prod_{d=1}^{D} \frac{{\varpi^{irr}}^{\sigma^{irr}}\exp(-\varpi^{irr}/s_d^{irr})}{\Gamma(\sigma^{irr}){s_d^{irr}}^{\sigma^{irr}+1}} \tag{33}$$

The hyperparameters, σ, σ^{irr}, ϖ, and ϖ^{irr}, associated with the s_{jd} and s_d^{irr} are given inverse Gamma and exponential priors:

$$p(\sigma|\lambda, \mu) \sim \frac{\mu^{\lambda}\exp(-\mu/\sigma)}{\Gamma(\lambda)\sigma^{\lambda+1}} \qquad p(\sigma^{irr}|\lambda, \mu) \sim \frac{\mu^{\lambda}\exp(-\mu/\sigma^{irr})}{\Gamma(\lambda){\sigma^{irr}}^{\lambda+1}} \tag{34}$$

$$p(\varpi|\phi) \sim \phi\exp(-\phi\varpi) \qquad p(\varpi^{irr}|\phi) \sim \phi\exp(-\phi\varpi^{irr}) \tag{35}$$

Thus, according to Eqs. 15, 30, 31, and 34, we have

$$p(\sigma|\ldots) \propto p(\sigma|\lambda, \mu)\prod_{j=1}^{M} p(\vec{s}_j|\sigma, \varpi) \tag{36}$$

$$p(\sigma^{irr}|\ldots) \propto p(\sigma^{irr}|\lambda^{irr}, \mu^{irr})p(\vec{s}^{irr}|\sigma^{irr}, \varpi^{irr}) \tag{37}$$

And according to Eqs. 15, 30, 31, and 35, we have

$$p(\varpi|\ldots) \propto p(\varpi|\phi)\prod_{j=1}^{M} p(\vec{s}_j|\sigma, \varpi) \tag{38}$$

$$p(\varpi^{irr}|\ldots) \propto p(\varpi^{irr}|\phi^{irr})p(\vec{s}^{irr}|\sigma^{irr}, \varpi^{irr}) \tag{39}$$

3.2.3 Conditional posterior distribution of ρ

We know that each ρ_d is defined in the compact support [0, 1], thus we consider for it a Beta distribution, with location δ_1 and scale δ_2 common to all dimensions, as a prior, which give us

$$p(\rho|\delta) = \left[\frac{\Gamma(\delta_2)}{\Gamma(\delta_1\delta_2)\Gamma(\delta_2(1-\delta_1))}\right]^D \prod_{d=1}^{D} \rho_d^{\delta_1\delta_2-1}(1-\rho_d)^{\delta_2(1-\delta_1)-1} \tag{40}$$

Recall that $\rho_d = p(z_{id} = 1)$ and $1 - \rho_d = p(z_{id} = 0)$, $d = 1, \ldots, D$, thus each z_i follows a D-variate Bernoulli distribution, and we have

$$p(z|\rho) = \prod_{i=1}^{N}\prod_{d=1}^{D} \rho_d^{z_{id}}(1-\rho_d)^{1-z_{id}} = \prod_{d=1}^{D} \rho_d^{f_d}(1-\rho_d)^{N-f_d} \tag{41}$$

where $f_d = \sum_{i=1}^{N} \mathbb{I}_{z_{id}=1}$. Then, according to Eqs. 15, 40, and 41, we have

$$p(\rho|\ldots) \propto p(\rho|\delta)p(z|\rho) \propto \prod_{d=1}^{D} \rho_d^{\delta_1\delta_2+f_d-1}(1-\rho_d)^{\delta_2(1-\delta_1)+N-f_d-1} \tag{42}$$

The hyperparameters δ_1 and δ_2 are given uniform $p(\delta_1) \sim \mathcal{U}_{[0,1]}$ and inverse Gamma $p(\delta_2|\varphi_\delta, \varrho_\delta)$ priors that give us the following posteriors

$$p(\delta_1|\ldots) \propto p(\delta_1)p(\rho|\delta) \tag{43}$$

$$p(\delta_2|\ldots) \propto p(\delta_2|\varphi_\delta, \varrho_\delta)p(\rho|\delta) \tag{44}$$

3.2.4 Conditional posterior distribution of Z

Having the conditional priors in Eq. 19, the conditional posteriors are obtained by combining these priors with the likelihood of the data [10,13]

$$p(Z_i = j|\ldots) = \begin{cases} \frac{n_{-i,j}}{N-1+\eta} p(\vec{Y}_i|\vec{m}_j, \vec{s}_j, \xi^{irr}, z_i) & \text{if } j \in \mathcal{R} \\ \int \frac{\eta p(\vec{Y}_i|\vec{m}_j, \vec{s}_j, \xi^{irr}, z_i) p(\vec{m}_j, \vec{s}_j, \xi^{irr})}{N-1+\eta} d\vec{m}_j d\vec{s}_j d\xi^{irr} & \text{if } j \in \mathcal{U} \end{cases} \tag{45}$$

where we have $p(\vec{m}_j, \vec{s}_j, \xi^{irr}) = p(\vec{m}_j|\zeta, \varepsilon)p(\vec{s}_j|\sigma, \varpi)p(\vec{m}^{irr}|\zeta^{irr}, \varepsilon^{irr})p(\vec{s}^{irr}|\sigma^{irr}, \varpi^{irr})$, $p(\vec{Y}_i|\vec{m}_j, \vec{s}_j, \xi^{irr}, z_i) = \prod_{d=1}^{D} p_{Beta}(Y_{id}|m_{Z_i d}, s_{Z_i d})^{z_{id}} p_{Beta}(Y_{id}|m_d^{irr}, s_d^{irr})^{1-z_{id}}$.

Equation 45 can be explained as the following: Each vector has a certain probability to be assigned to a represented component and a certain probability to be associated with an unrepresented component. If a vector is assigned to an unrepresented cluster, a new represented cluster is generated. In contrast, if all data vectors of a represented cluster are affected to other clusters during a sampling iteration, this cluster becomes empty and then unrepresented. It is important to note that all the unrepresented clusters have identical parameters, and then we do not need to differentiate them [13]. The choice of η is crucial in our model. In fact, the number of clusters is directly related to η, which controls the generation frequency of new clusters. Then, we have chosen an inverse gamma prior for the concentration parameter η

$$p(\eta|\chi, \kappa) \sim \frac{\kappa^\chi \exp(-\kappa/\eta)}{\Gamma(\chi)\eta^{\chi+1}} \tag{46}$$

which gives with Eq. 19 the following posterior (for more details, see [13])

$$p(\eta|\ldots) \propto \frac{\kappa^\chi \exp(-\kappa/\eta)}{\Gamma(\chi)\eta^{\chi+1}} \frac{\eta^M \Gamma(\eta)}{\Gamma(N+\eta)} \tag{47}$$

The hyperparameters χ and κ are given inverse Gamma $p(\chi|\lambda_\chi, \mu_\chi)$ and exponential $p(\kappa|\phi_\kappa)$ priors, respectively, which give us the following posteriors:

$$p(\chi|\ldots) = p(\eta|\chi, \kappa)p(\chi|\lambda_\chi, \mu_\chi) \tag{48}$$

$$p(\kappa|\ldots) = p(\eta|\chi, \kappa)p(\kappa|\phi_\kappa) \tag{49}$$

A graphical model representing our infinite generalized Dirichlet mixture with feature selection is shown in Fig. 3.

3.3 Complete algorithm

Having all the posteriors, we can employ a Gibbs sampler, and each iteration will be based on the following steps:

- Generate Z_i from Eq. 45 and then update n_j, $j = 1, \ldots, M, i = 1, \ldots, N$.
- Update the number of represented components M.
- $p_j = \frac{n_j}{N+\eta}$, $j = 1, \ldots, M$ and the mixing parameters of unrepresented components $p_U = \frac{\eta}{\eta+N}$.

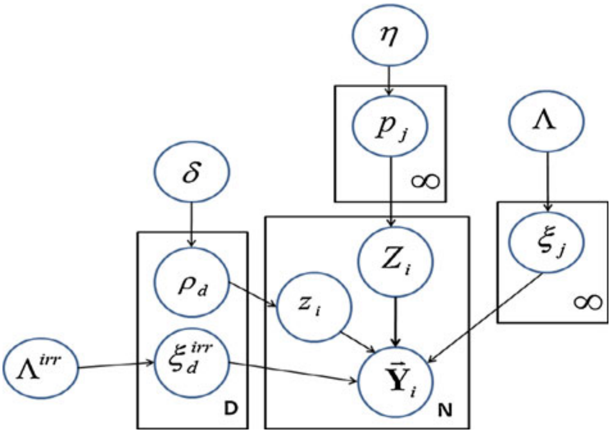


Fig. 3 Graphical representation for the infinite model

- Generate z_i from a D-variate Bernoulli distribution with parameters $(\hat{z}_{i1}, \ldots, \hat{z}_{iD})$, where $\hat{z}_{id} = \frac{\rho_d p_{Beta}(Y_{id}|m_{jd}, s_{jd})}{\rho_d p_{Beta}(Y_{id}|m_{jd}, s_{jd}) + (1-\rho_d) p_{Beta}(Y_{id}|m_d^{irr}, s_d^{irr})}$ denotes the expectation for z_{id}.
- Generate ρ_d from Eq. 42, $d = 1, \ldots, D$.
- Generate $\vec{m}_j$ from Eq. 22 and $\vec{s}_j$ from Eq. 32, $j = 1, \ldots, M$.
- Generate $\vec{m}^{irr}$ from Eq. 23 and $\vec{s}^{irr}$ from Eq. 33,
- Update the hyperparameters: Generate ε, ζ, ε^{irr}, ζ^{irr}, σ, ϖ, σ^{irr}, ϖ^{irr}, δ_1, δ_2, η, χ, and κ from Eqs. 26, 28, 27, 29, 36, 38, 37, 39, 43, 44, 47, 48, and 49, respectively.

Note that in the initialization step, the algorithm started by assuming that all the vectors are in the same cluster, and the initial parameters are generated as random samples from their prior distribution. The distributions given by Eqs. 26, 28, 27, 29, 36, 38, 37, 39, 43, 44, 47, 48, and 49 are not of standard forms. However, it is possible to show that they are log-concave (i.e., it is straightforward to show that the second derivatives of the logarithms of these functions are negative), then the sample generation is based on the adaptive rejection sampling (ARS) [26] to obtain the values of ε, ζ, ε^{irr}, ζ^{irr}, σ, ϖ, σ^{irr}, ϖ^{irr}, δ_1, δ_2, η, χ, and κ. The sampling of the vectors Z_i requires the evaluation of the integral in Eq. 45, which is not analytically tractable. Thus, we have used an approach, originally proposed in [10] and used with success in [13], which consists on approximating this integral by generating a Monte Carlo estimate by sampling from the priors of $\vec{m}_j$, $\vec{m}^{irr}$, $\vec{s}_j$, and $\vec{s}^{irr}$. More details about this sampling method are given in [10]. The sampling of $\vec{m}_j$, $\vec{s}_j$, $\vec{m}^{irr}$, and $\vec{s}^{irr}$ is more complex, since the posteriors given by Eqs. 22, 32, 23, and 33 do not have known forms. Thus, we have used the Metropolis-Hastings algorithm (M-H) with log-normal proposal for $\vec{s}_j$ and $\vec{s}^{irr}$, and Beta proposals for $\vec{m}_j$ and $\vec{m}^{irr}$ [27].

4 Experimental results

In this section, we demonstrate the utility of the infinite generalized Dirichlet mixture model and feature selection in performing data mining and knowledge discovery applications. We start by an application involving text mining experiments where we focus on text documents classification and spam email filtering. The second application concerns image databases categorization using visual features. Moreover, we compare the proposed approach in this

Table 1 Classification results for the industry sector and 20 newsgroups data sets

	Industry sector	20 newsgroups
Infinite mixture	0.87 ± 0.02	0.84 ± 0.02
Finite mixture	0.85 ± 0.03	0.81 ± 0.03
Bayesian finite mixture	0.86 ± 0.01	0.83 ± 0.02
Infinite mixture + FS	0.90 ± 0.01	0.88 ± 0.01
Finite mixture + FS	0.88 ± 0.02	0.85 ± 0.01
Bayesian finite mixture + FS	0.88 ± 0.01	0.86 ± 0.01

paper, for estimation and selection, with the approach in [18,28] and the Bayesian approach in [16]. In these applications, the values of the hyperparameters have been set experimentally as following $(\varphi, \varrho, \lambda, \mu, \phi, \varphi_\delta, \varrho_\delta, \lambda_\chi, \mu_\chi, \phi_\kappa) = (1, 1, 1, 1, 1/8, 1, 1, 1, 1, 1)$. These choices have been found reasonable according to our experimental results.

4.1 Text mining

In the first experiment, we test our model for the classification of three well-known data sets consisting of a set of documents and which were used, for instance, in [29]: the industry sector, the Mod Apte split of the Reuters-21578 document collection, and the 20 newsgroups data sets. The industry sector data set[1] has a vocabulary of 55,055 words and is composed of 104 classes containing 9,555 documents having an average of 606 words. The first half of this data set is used for training and the second one for testing. The Mod Apte split data set is a subset of the well-known corpus Reuters-21578,[2], has a vocabulary of 15,996 words, and is composed of 90 classes containing 7,770 training documents and 3,019 test documents. The documents in this data set have an average of 70 words and are multi-labeled. The 20 newsgroups data set[3] contains 18,828 documents grouped in 20 classes (80% of the documents is used for training and the rest for testing) with a vocabulary of 61,298 words and average document length of 116 words.

The first step in our experiment was processing the different documents using the Rainbow toolbox [30] with word stemming to extract the count vectors, and the 500 most common words were removed from the vocabularies. After this first preprocessing, each document was represented by a vector of counts. After normalizing these vectors, each document was represented by a vector of proportions. The proportions vectors in the different training sets were then modeled by infinite mixtures, using the algorithm in the previous section, and each test vector was affected to the class that gives the highest likelihood. Following [29], the performance of our model for the industry sector and newsgroups data sets was based on the precision measure. Table 1 shows the classification results, averaged over 20 random selection of the training and test sets, for the industry sector and 20 newsgroups data sets (20 categories) with and without feature selection (FS). According to this table, the infinite model produces better results than the finite one estimated either using the EM-based or the Bayesian methods. Moreover, we can see that feature selection improves the classification results.

For the Mod Apte split data set, however, the precision alone is not a sufficient measure, since the data is multi-labeled, then we have also considered the recall measure. The precision

[1] http://www.cs.umass.edu/~mccallum/code-data.html.

[2] The corpus is from http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

[3] We have used the third version ("18828") of the Newsgroup-20 data set which does not include cross-posts and includes only the "From" and "Subject" headers (people.csail.mit.edu/people/jrennie/20Newsgroups).

Table 2 Classification results for the Mod Apte split data set

	Macro	Micro
Infinite mixture	0.56	0.75
Finite mixture	0.53	0.73
Bayesian finite mixture	0.54	0.74
Infinite mixture + FS	0.61	0.80
Finite mixture + FS	0.58	0.77
Bayesian finite mixture + FS	0.58	0.78

Table 3 Description of the tested data sets

Dataset	Ham (%)	Spam (%)	Total	
Spambase	59.5	39.5	4,601	
ECML/PKDD challenge	50	50	Task A: 6,500	Task B: 500

and the recall are then combined using the break-even point defined by the micro or macro averaging. Table 2 shows the classification results for the Mod Apte split data set.[4]

We have also tested our algorithm on the challenging problem of e-mail classification for which a variety of approaches have been proposed. In particular, we focus on the specific problem of spam filtering. In our experiments, we have used three large publicly available data sets. The first data set is spambase,[5] the second data set is 2006 ECML/PKDD Discovery Challenge.[6] The two data sets represent collected email messages including header, subject, and body text without any alteration, converted to feature vector format indicating the probability of occurrences of a given feature (word) in messages. Moreover, the 2006 ECML Discovery Challenge data set eliminates the strings that occur less than four times in the corpus. Table 3 shows a description of the three tested data sets.

Tables 4, 5, and 6 show the obtained results, averaged over 20 random selection of the training (50%) and test (50%) sets, for the different data sets. According to the results, we can see clearly that the infinite model applied with feature selection is preferred and gives the best results, in terms of precision, recall, and accuracy, compared with the finite model. Moreover, compared with SVM, applied with several kernels namely polynomial, Gaussian, χ^2, and sigmoid, our model was found to generally reach better or comparable results.

4.2 Images organization by content

With the ease of creation of multimedia data such as images, an increasing amount of visual information is now available. An important challenge is the development of models and algorithms for the categorization of such content which facilitates, for instance, the search over image databases and recommendation [31–33]. Statistical modeling is a crucial step for this problem where models are used to represent the stochastic structure of images and extract useful information from image data. The problem of images categorization is generally approached by extracting low-level features such as color and texture. In our case, we

[4] As mentioned in [29], there is only a single standard training/test set split for this data set, thus the standard deviation is not given.

[5] http://archive.ics.uci.edu/ml/.

[6] http://www.ecmlpkdd2006.org/challenge.html.

Table 4 Spambase corpus results

	Precision	Recall	Accuracy
Infinite mixture	0.800	0.798	0.831
Finite mixture	0.781	0.779	0.798
Bayesian finite mixture	0.795	0.783	0.810
Infinite mixture + FS	0.829	0.803	0.851
Finite mixture + FS	0.799	0.797	0.842
Bayesian finite mixture + FS	0.818	0.800	0.846
SVM (Polynomial)	0.758	0.730	0.793
SVM (Gaussian)	0.515	0.707	0.724
SVM (χ^2)	0.819	0.795	0.845
SVM (Sigmoid)	0.607	0.713	0.748

Table 5 ECML/PKDD (Task A) corpus results

	Precision	Recall	Accuracy
Infinite mixture	0.910	0.702	0.810
Finite mixture	0.797	0.689	0.788
Bayesian finite mixture	0.891	0.697	0.798
Infinite mixture + FS	0.941	0.799	0.821
Finite mixture + FS	0.837	0.773	0.793
Bayesian finite mixture + FS	0.921	0.779	0.812
SVM (Polynomial)	0.805	0.784	0.792
SVM (Gaussian)	0.684	0.590	0.604
SVM (χ^2)	0.815	0.790	0.814
SVM (Sigmoid)	0.935	0.676	0.743

Table 6 ECML/PKDD (Task B) corpus result

	Precision	Recall	Accuracy
Infinite mixture	0.949	0.757	0.821
Finite mixture	0.902	0.701	0.789
Bayesian finite mixture	0.921	0.721	0.795
Infinite mixture + FS	0.952	0.823	0.834
Finite mixture + FS	0.913	0.812	0.819
Bayesian finite mixture + FS	0.933	0.820	0.827
SVM (Polynomial)	0.690	0.985	0.842
SVM (Gaussian)	0.740	0.986	0.865
SVM (χ^2)	0.970	0.520	0.537
SVM (Sigmoid)	0.500	0.757	0.670

have used an approach based on the adaptation of the bag-of-words approach for computer vision applications, which has shown impressive levels of performance [34]. This approach is based on four steps: (1) Automatic detection of points of interest in images, (2) Computation of local descriptors over those points, (3) Quantization of the descriptors into words to construct the visual vocabulary, and (4) Computation of the occurrences of each visual

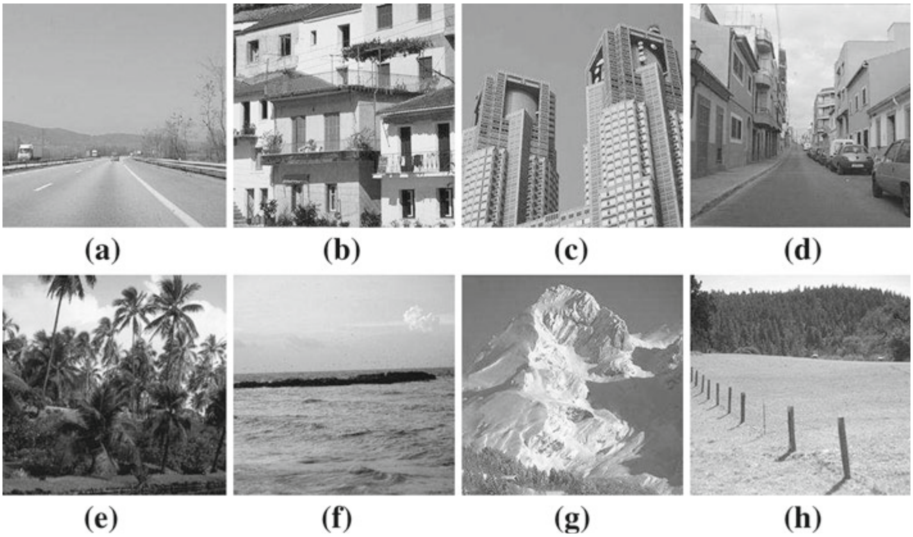


Fig. 4 Sample images from each group in the first data set. **a** Highway, **b** inside of cities, **c** tall buildings, **d** streets, **e** forest, **f** coast, **g** mountain, **h** open country

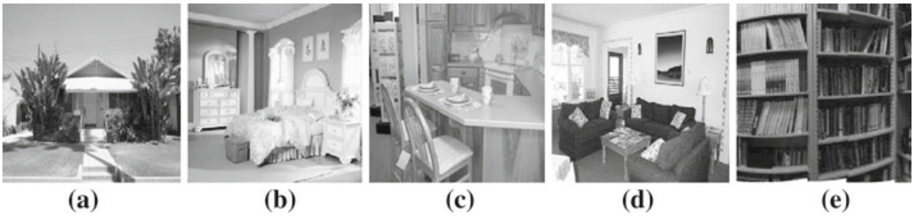


Fig. 5 Additional categories in the second data set. **a** Suburb residence, **b** bedroom, **c** kitchen, **d** livingroom, **e** office

word in the vocabulary which results in the representation of a given image by a vector of frequencies (i.e., histogram of visual words).

Our evaluation was based on three different image data sets. The first data set is composed of eight categories [35]: coasts (360 images), forest (328 images), mountain (374 images), open country (410 images), highway (260 images), inside of cities (308 images), tall buildings (356 images), and streets (292 images). Figure 4 shows examples of these images that have an average size of approximately 250 × 250 pixels. Note that this data set is actually composed of two subsets (natural and man-made images) of four categories each. The second data set contains 13 categories of natural scenes [36] and consists of the eight categories in the first data set (2688 images) plus the following categories: suburb residence (241 images), bedroom (174 images), kitchen (151 images), livingroom (289 images), and office (216 images). Figure 5 shows examples of images from these additional categories that have an average size of approximately 250 × 300 pixels. The third data set contains 15 categories [37] and consists of the 13 categories of the second data set plus 626 other images that have an average size of 250 × 300 pixels and divided into two categories (see Fig. 6): store (315 images) and industrial (311 images). We divided each of this data sets 10 times randomly into two separate halves, half for training and half for testing. From the training

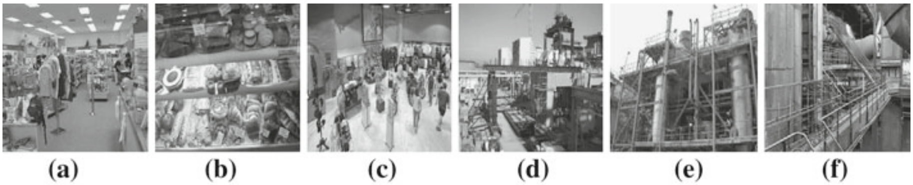


Fig. 6 Additional categories in the third data set. **a**–**c** Store, **d**–**f** industrial

Table 7 Classification performance (%) obtained for the different data sets using different approaches without feature selection

	Infinite mixture	Finite mixture	Bayesian finite mixture
Data set 1 (all categories)	73.02	72.14	72.81
Data set 1 (natural categories)	74.14	73.07	73.72
Data set 1 (man-made categories)	75.77	74.82	75.08
Data set 2	74.21	65.88	68.23
Data set 3	70.53	64.81	67.98

sets, about 30 images from each category were taken, randomly, to construct the visual vocabulary by detecting interest points from these images using the difference-of-Gaussians point detector, since it has shown excellent performance [38]. Then, we used SIFT descriptor [38], computed on detected keypoints of all images and giving 128-dimensional vector for each keypoint. Moreover, extracted vectors were clustered using the K-Means algorithm providing 800 visual words (we have tested several vocabulary sizes, and the best classification results were obtained with 800 visual words). Each image in the data sets was then represented by a 800-dimensional vector describing the probabilities of a set of visual words, provided from the constructed visual vocabulary. Our categorization approach is based on a classifier. The inputs to the classifier are the 800-dimensional vectors extracted from the different database classes. These vectors are separated into the unknown or test set of vectors, whose class is unknown, and the training set of vectors, whose class is known. The training set is necessary to adapt the classifier to each possible class before the unknown set is submitted to the classifier. Then, we apply our algorithm, presented in Sect. 3.3, to the training vectors in each class. After this stage, each class in the database is represented by a statistical model (Eq. 7). Finally, in the classification stage, each unknown image is assigned to the class increasing more its log likelihood.

Three classification tasks were considered for the first data set. The first one is classification into eight categories. The second and third ones were previously considered in [35] and are classification within the first subset composed of four categories of natural images, and classification within the second subset composed of four categories of man-made images. A summary of the classification results, measured by the average values of the diagonal entries of the confusion matrices obtained for the different classification tasks, is shown in Tables 7 and 8 without and with feature selection, respectively. During our experiments, we noticed that an important part of the misclassification errors occurs among the categories: bedroom, livingroom, kitchen and office which is the same conclusion reached in [36]. The tables shows the clear dominance of the infinite model. The results can be explained by the fact that the infinite model outperforms by its ability to incorporate uncertainties related to the selection of the correct number of clusters. We can clearly notice, also, that introducing feature selection improves further the results.

Table 8 Classification performance (%) obtained for the different data sets using different approaches with feature selection

	Infinite mixture	Finite mixture	Bayesian finite mixture
Data set 1 (all categories)	74.84	72.73	73.96
Data set 1 (natural categories)	76.01	74.03	74.08
Data set 1 (man-made categories)	77.11	75.23	75.69
Data set 2	75.03	67.89	69.85
Data set 3	71.42	66.14	68.37

5 Conclusion

We have addressed the problem of learning infinite generalized Dirichlet mixture models from data by taking a non-parametric Bayesian approach in order to provide more control over the number of mixture distribution components and by taking into account the fact that only a subset of features may influence the final model structure. We have developed a MCMC algorithm to sample from the posterior distributions associated with the selected priors for the different model parameters. The usefulness of the developed framework has been shown via applications involving image and text categorization. We found that assuming that the number of mixture components is unbounded and removing irrelevant features often improved classification performance. Our model has several natural extensions such as the development of a variational estimation approach which may improve further the results.

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

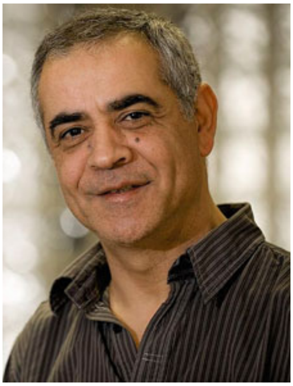
References

1. McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
2. Bouguila N, Ziou D (2006) Unsupervised selection of a finite Dirichlet mixture model: an MML-based approach. IEEE Trans Knowl Data Eng 18(8):993–1009
3. Bouguila N, Ziou D (2007) High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Trans Pattern Anal Mach Intell 29(10):1716–1731
4. Zhang M, Alhajj R (2010) Effectiveness of NAQ-tree as index structure for similarity search in high-dimensional metric space. Knowl Inf Syst 22:1–21
5. Moise G, Zimek A, Kröger P, Kriegel H-P, Sander J (2009) Subspace and projected clustering: experimental evaluation and analysis. Knowl Inf Syst 21:299–326
6. Lu J, Li R, Zhang Y, Zhao T, Lu Z (2010) Image annotation technique based on feature selection for class-pairs. Knowl Inf Syst 24(2):325–337
7. Bouguila N, Ziou D (2009) A non-parametric Bayesian learning model: application to text and image categorization. In: Proceedings of the 13th Pacific-Asia conference on advances in knowledge discovery and data mining (PAKDD). Springer, LNAI 5476, pp 463–474
8. Ferguson TS (1983) Bayesian density estimation by mixtures of normal distributions. In: Rizvi H, Rustagi J (eds) Recent advances in statistics. Academic Press, New York, pp 287–302
9. Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588
10. Neal RM (2000) Markov Chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9:249–265
11. Ghosh JK, Ramamoorthi RV (2003) Bayesian nonparametrics. Springer, Berlin

12. Teh YW, Jordan MI, Beal MI, Matthew J, Blei DM (2006) Hierarchical Dirichlet processes. J Am Stat Assoc 101(476):1566–1581
13. Rasmussen CE (2000) The infinite gaussian mixture model. In: Advances in neural information processing systems (NIPS), pp 554–560
14. Bouguila N, Ziou D (2004) A powerful finite mixture model based on the generalized Dirichlet distribution: unsupervised learning and applications. In: Proceedings of the 17th international conference on pattern recognition (ICPR), pp 280–283
15. Bouguila N (2008) Clustering of count data using generalized Dirichlet multinomial distributions. IEEE Trans Knowl Data Eng 20(4):462–474
16. Bouguila N, Ziou D, Hammoud RI (2009) On Bayesian analysis of a finite generalized Dirichlet mixture via a metropolis-within-gibbs sampling. Pattern Anal Appl 12(2):151–166
17. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Royal Stat Soc B 39:1–38
18. Bouguila N, Ziou D (2006) A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture. IEEE Trans Image Process 15(9):2657–2668
19. Boutemedjet S, Bouguila N, Ziou D (2009) A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. IEEE Trans Pattern Anal Mach Intell 31(9):1429–1443
20. Bouguila N, Ziou D, Monga E (2006) Practical Bayesian estimation of a finite beta mixture through gibbs sampling and its applications. Stat Comput 16(2):215–225
21. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman & Hall/CRC, London
22. Marin J-M, Robert CP (2007) Bayesian core: a practical approach to computational Bayesian statistics. Springer, Berlin
23. Ishwaran H, James LF (2003) Generalized weighted chinese restaurant processes for species sampling mixture models. Stat Sinica 13:1211–1235
24. Papaspiliopoulos O, Roberts GO (2008) Retrospective Markov Chain Monte Carlo methods for Dirichlet process hierarchical models. Stat Sinica 95(1):169–186
25. Carlin BP, Louis TA (2000) Bayes and empirical Bayes methods for data analysis, second edition. Chapman & Hall/CRC, London
26. Gilks WR, Wild P (1993) Algorithm aS 287: adaptive rejection sampling from log-concave density functions. Appl Stat 42(4):701–709
27. Chib S, Greenberg E (1995) Understanding the metropolis-hastings algorithm. Am Stat 49(4):327–335
28. Bouguila N, Ziou D (2004) Dirichlet-based probability model applied to human skin detection. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 521–524
29. Madsen RE, Kauchak D, Elkan C (2005) Modeling word burstiness using the Dirichlet distribution. In: Proceedings of the 22nd international conference on machine learning (ICML), pp 545–552
30. McCallum AK (1996) Bow: a toolkit for statistical language modeling, text retrieval, classification and clustering. Technical report
31. Gong Z, Liu Q (2009) Improving keyword based web image search with visual feature distribution and term expansion. Knowl Inf Syst 21:113–132
32. Bartolini I, Ciaccia P, Patella M (2009) Query processing issues in region-based image databases. Knowl Inf Syst. In press
33. Bouguila N, Ziou D, Vaillancourt J (2003) Novel mixtures based on the Dirichlet distribution: application to data and image classification. In: Machine learning and data mining in pattern recognition (MLDM), LNAI 2734. pp 172–181
34. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, 8th European conference on computer vision (ECCV)
35. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
36. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR), pp 524–531
37. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2169–2178
38. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110

Author Biographies

Nizar Bouguila received the engineer degree from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D degrees in computer science from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently an Associate Professor with the Concordia Institute for Information Systems Engineering (CIISE) at Concordia University, Montreal, QC, Canada. His research interests include image processing, machine learning, data mining, 3D graphics, computer vision, and pattern recognition. Prof. Bouguila received the best Ph.D Thesis Award in Engineering and Natural Sciences from Sherbrooke University in 2007. He was awarded the prestigious Prix d'excellence de l'association des doyens des études supèrieures au Québec (best Ph.D Thesis Award in Engineering and Natural Sciences in Québec), and was a runner-up for the prestigious NSERC doctoral prize.



Djemel Ziou received the B.Eng. degree in Computer Science from the University of Annaba (Algeria) in 1984, and Ph.D degree in Computer Science from the Institut National Polytechnique de Lorraine (INPL), France in 1991. From 1987 to 1993 he served as lecturer in several universities in France. During the same period, he was a researcher in the Centre de Recherche en Informatique de Nancy (CRIN) and the Institut National de Recherche en Informatique et Automatique (INRIA) in France. Presently, he is full Professor at the department of computer science, Université de Sherbrooke, QC, Canada. He is holder of the NSERC/Bell Canada Research Chair in personal imaging. He has served on numerous conference committees as member or chair. He heads the laboratory MOIVRE and the consortium CoRIMedia which he founded. His research interests include image processing, information retrieval, computer vision and pattern recognition.