Entropy-Based Variational Learning of Finite Generalized Inverted Dirichlet Mixture Model

Mohammad Sadegh Ahmadzadeh[1], Narges Manouchehri[2(✉)], Hafsa Ennajari[2], Nizar Bouguila[2], and Wentao Fan[3]

[1] Department of Electrical Engineering, Concordia University, Montreal, Canada
m_hmadza@encs.concordia.ca
[2] Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
narges.manouchehri@mail.concordia.ca, h_ennaja@encs.concordia.ca, nizar.bouguila@concordia.ca
[3] Department of Computer Science and Technology, Huaqiao University, Xiamen, China
fwt@hqu.edu.cn

Abstract. Mixture models are considered as a powerful approach for modeling complex data in an unsupervised manner. In this paper, we introduce a finite generalized inverted Dirichlet mixture model for semi-bounded data clustering, where we also developed a variational entropy-based method in order to flexibly estimate the parameters and select the number of components. Experiments on real-world applications including breast cancer detection and image categorization demonstrate the superior performance of our proposed model.

Keywords: Unsupervised learning · Clustering · Generalized inverted dirichlet distribution · Entropy-based variational learning

1 Introduction

Nowadays, large amounts of complex data in various formats (e.g., image, text, speech) are generated increasingly at a bottleneck speed. This increase motivated data scientists to develop tactical models in order to automatically analyze and infer useful knowledge from these data [1]. In this context, statistical modeling plays a significant role in helping machines interpret data with statistics. An essential approach in statistical modeling is finite mixture models that are effectively used for clustering purposes, separating heterogeneous data into homogeneous groups [2]. The usefulness of mixture models has been widely demonstrated in many application areas including pattern recognition, text and image analysis [3]. However, there exist several challenges to address when working with mixture models: (1) Standard finite mixture models assume that the

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 130–143, 2021.
https://doi.org/10.1007/978-3-030-73280-6_11

observed data are normally distributed [4]. This is not always the case, in several applications. Lately, multiple studies have shown that other non-Gaussian statistical models (e.g., Dirichlet, inverted Dirichlet and Gamma) are effective in modeling data [5–15]. Thus, choosing a suitable probability distribution that better describes the nature and the properties of the observed data is crucial to the assessment of the validity of the model. For instance, the inverted Dirichlet mixture, has good flexibility in accepting different symmetric and asymmetric forms that results in better generalization capabilities. But, the model usually supposes that the features of the vectors are positively correlated, and that is not always applicable for real-life applications. (2) In most cases, the mixture model fitting is not straightforward and analytically intractable. Methods like Expectation-Maximization (EM) and Maximum likelihood [1] are widely used in this context, but they remain impractical as they are sensitive to initialization and usually lead to over-fitting [16]. An alternative approach to solve these problems is Bayesian learning, particularly, variational inference has made the parameter estimation process more computationally efficient. (3) The selection of the number of components is an important issue to consider in the design of mixture models, because a high number of components may lead to learning the data too much, whereas inference under a model with a small number of components can be biased. To this end, multiple effective methods have been proposed, like minimum message length criterion [17,18]. To overcome the aforementioned challenges, we introduce a novel finite variational Generalized Inverted Dirichlet Mixture Model for data clustering, which learns the latent parameters based on the variational inference algorithm. Our work is motivated by the success of the Generalized Inverted Dirichlet (GID) distribution [19]. The GID has great efficiency in comparison to Gaussian distribution when dealing with positive vectors and has been shown to be more practical due to its higher general covariance structure. Also the GID samples can be represented in a transformed space where features are independent and follow the inverted Beta distribution [20–22]. Moreover, the use of the variational inference algorithm allows us to minimize the Kullback–Leibler divergence between the true posterior and the approximated variational distribution, leading to accurate and computationally efficient parameter estimation of our proposed mixture model [22]. The main challenge here, is to design a good mixture model that better fits the observed semi-bounded data with the right number of components. We propose to apply an entropy-based variational inference combined with our GID Mixture Model. We started with one component and proceed incrementally to find the best number of components and we will explain the model complexity and approximate the perfect number of components by a compression between the estimated and theoretical entropy [23] similar to researches that have been successful on distributions like the Dirichlet mixture model [24]. To demonstrate the effectiveness of the proposed approach, we evaluate the Entropy-Based Variational Learning of Finite Generalized Inverted Dirichlet Mixture Model (EV-GIDMM) on real-world applications including breast cancer detection and image categorization.

The remainder of this paper is organized as follows. We provide an overview of the statistical background of our GID mixture model in Sect. 2. Section 3 is assigned to the variational inference process of our model. We explain the entropy-based variational inference of EV-GIDMM in Sect. 4. The results of our experiments on real data are provided in Sect. 5. Finally, we conclude the paper in Sect. 6.

2 Model Specification

2.1 Finite Generalized Inverted Dirichlet Mixture

Lets us assume $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ is a set of N independent identically distributed vectors, where every single $\boldsymbol{Y}_i$ is defined as $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iD})$, where D is the dimensionality of the vector. We are assuming that each $\boldsymbol{Y}_i$ follows a mixture of GIDs, where the probability density function of the GID is given by [19]:

$$p(\boldsymbol{Y_i} \mid \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_{jd} + \beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})} \frac{Y_{id}^{\alpha_{jd}-1}}{(1 + \sum_{l=1}^{d} Y_{il})^{\gamma_{jd}}} \tag{1}$$

where $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are the parameters of the GID, and they are defined as $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jd})$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})$ with constraints $\alpha_{jd} > 0$ and $\beta_{jd} > 0$. We can find γ_{id} according to $\gamma_{id} = \beta_{jd} + \alpha_{jd} - \beta_{j(d+1)}$. Supposing that the model consists of M different components [1], we are able to define the GID mixture model as follows:

$$p(\boldsymbol{Y}_i \mid \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{M} \pi_j p(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) \tag{2}$$

where $\boldsymbol{\pi}$ represents its mixing coefficients correlated with the components, where, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ with constrains $\pi_j \geq 0$ and $\sum_{j=1}^{M} \pi_j = 1$, and the shape parameters of the distribution are denoted as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ and $j = 1, \ldots, M$. According to [21], we can replace the GID distribution with a product of D Inverted Beta distributions, considering that it does not change the model, therefore, Eq. (2) can be rewritten as:

$$p(\mathcal{X} \mid \pi, \alpha, \beta) = \prod_{i=1}^{N} \left(\sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} P_{iBeta}(X_{il} \mid \alpha_{jl}, \beta_{jl}) \right) \tag{3}$$

By considering that $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iD})$, we have $X_{il} = Y_{il}$ and $X_{il} = \frac{Y_{il}}{1+\sum_{k=1}^{l-1} Y_{ik}}$ for $l > 1$. The inverted Beta distribution is defined by $P_{iBeta}(X_{il}|\alpha_{jl}, \beta_{jl})$ with the parameters α_{jl} and β_{jl} and given by:

$$P_{iBeta}(X_{il} \mid \alpha_{jl}, \beta_{jl}) = \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \frac{{X_{il}}^{\alpha_{jl}-1}}{(1 + X_{il})^{\alpha_{jl}+\beta_{jl}}} \tag{4}$$

In proportion to this design, we are able to estimate the parameters from Eq. (3) instead of the Eq. (2). We define the latent variables as $\mathcal{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ with the conditions $Z_{ij} \in \{0, 1\}$ that Z_{ij} is equal to 1 if $\boldsymbol{X}_i$ is assigned to cluster j and zero otherwise, and $\sum_{j=1}^{M} Z_{ij} = 1$. The conditional probability for the latent variables $\mathcal{Z}$ given $\boldsymbol{\pi}$ can be written as:

$$p(\mathcal{Z} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{j=1}^{M} {\pi_j}^{Z_{ij}} \tag{5}$$

We write the probability of the observed data vectors $\mathcal{X}$ given the latent variable and component parameters as:

$$p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \Big(\prod_{l=1}^{D} p_{iBeta}(X_{il} \mid \alpha_{jl}, \beta_{jl}) \Big)^{Z_{ij}} \tag{6}$$

By assuming that the parameters are independent and positive, we can suppose that the priors of these parameters are Gamma distributions $\mathcal{G}(.)$. According to [25], we can describe them as:

$$p(\alpha_{jl}) = \mathcal{G}(\alpha_{jl} \mid u_{jl}, \nu_{jl}) = \frac{{\nu_{jl}}^{u_{jl}}}{\Gamma(u_{jl})} {\alpha_{jl}}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \tag{7}$$

$$p(\beta_{jl}) = \mathcal{G}(\beta_{jl} \mid g_{jl}, h_{jl}) = \frac{{h_{jl}}^{g_{jl}}}{\Gamma(g_{jl})} {\beta_{jl}}^{g_{jl}-1} e^{-h_{jl}\beta_{jl}} \tag{8}$$

We define the joint distribution including all random variables , as follows:

$$p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{\pi}) = p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\mathcal{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) \tag{9}$$

$$\begin{aligned} p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{\pi}) = & \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\prod_{l=1}^{D} \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \frac{X_{il}^{\alpha_{jl}-1}}{(1 + X_{il})^{\alpha_{jl}+\beta_{jl}}} \right)^{Z_{ij}} \left(\prod_{i=1}^{N} \prod_{j=1}^{M} {\pi_j}^{Z_{ij}} \right) \\ & \prod_{j=1}^{M} \prod_{l=1}^{D} \Big(\frac{{\nu_{jl}}^{u_{jl}}}{\Gamma(u_{jl})} {\alpha_{jl}}^{u_{jl}-1} e^{-\nu_{jl}\alpha_{jl}} \times \frac{{h_{jl}}^{g_{jl}}}{\Gamma(g_{jl})} {\beta_{jl}}^{g_{jl}-1} e^{-h_{jl}\beta_{jl}} \Big) \end{aligned} \tag{10}$$

3 Model Learning with Variational Inference

The GID mixture model contains hidden variables that can not be estimated directly. In order to estimate them, we apply the variational inference method, in which we aim to find an approximation of the posterior probability distribution of $p(\Theta|\mathcal{X}, \boldsymbol{\pi})$ by having $\Theta = \{\mathcal{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$. Inspired by [24], we introduce $Q(\Theta)$ as an approximation of the true posterior distribution $p(\Theta|\mathcal{X}, \boldsymbol{\pi})$. We make use of the Kullback-Leibler (KL) divergence in order to minimize the difference between the

true posterior distribution and the approximated one, which can be expressed as follows:

$$KL(Q \mid\mid P) = -\int Q(\Theta) \ln \left(\frac{p(\Theta \mid \mathcal{X}, \boldsymbol{\pi})}{Q(\Theta)} \right) \mathrm{d}\Theta = \ln\ p(\mathcal{X} \mid \boldsymbol{\pi}) - \mathcal{L}(Q) \quad (11)$$

where $\mathcal{L}(Q)$ is defined as:

$$\mathcal{L}(Q) = \int Q(\Theta) \ln \left(\frac{p(\mathcal{X}, \Theta \mid \boldsymbol{\pi})}{Q(\Theta)} \right) \mathrm{d}\Theta \quad (12)$$

Starting from the fact that $\mathcal{L}(Q) \leq \ln p(\mathcal{X}|\boldsymbol{\pi})$, we can see that $\mathcal{L}(Q)$ is the lower bound of the log likelihood. Thus, we have to maximize $\mathcal{L}(Q)$ in order to minimize the KL divergence. We assume a factorization assumption around $Q(\Theta)$ to apply it in variational inference. This assumption is called the Mean Field Approximation. We can factorize the posterior distribution $Q(\Theta)$ as $Q(\Theta) = Q(\mathcal{Z})Q(\boldsymbol{\alpha})Q(\boldsymbol{\beta})Q(\boldsymbol{\pi})$ [26,27]. In order to obtain a variational solution for the lower bound with respect to all the model parameters, we consider an optimal solution for a fix parameter s that is defined as $\ln Q_s^*(\Theta_s) = \langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s}$ where $\langle \cdot \rangle_{i \neq s}$ refers to the expectation with respect to all the parameters apart from Θ_s, if an exponential is taken from both sides, the normalized equation is as follows.

$$Q_s(\Theta_s) = \frac{exp\big\langle \ln\ p(\mathcal{X}, \Theta) \big\rangle_{i \neq s}}{\int exp\big\langle \ln\ p(\mathcal{X}, \Theta) \big\rangle_{i \neq s} \mathrm{d}\Theta} \quad (13)$$

We obtain the optimal variational posteriors solution that are formulated as:

$$Q(\mathcal{Z}) = \prod_{i=1}^{N} \prod_{j=1}^{M} r_{ij}^{Z_{ij}} \quad (14)$$

$$Q(\boldsymbol{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\alpha_{jl} \mid u_{jl}^*, \nu_{jl}^*)\ ,\ Q(\boldsymbol{\beta}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\beta_{jl} \mid g_{jl}^*, h_{jl}^*) \quad (15)$$

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M} \tilde{r}_{ij}} \quad (16)$$

$$\ln \tilde{r}_{ij} = \ln \pi_j + \sum_{l=1}^{D} \tilde{R}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} - (\bar{\alpha}_{jl} + \bar{\beta}_{jl}) \ln(1 + X_{il}) \quad (17)$$

$$\begin{aligned} \tilde{R} = {} & \ln \frac{\Gamma(\bar{\alpha} + \bar{\beta})}{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})} + \bar{\alpha}[\psi(\bar{\alpha} + \bar{\beta}) - \psi(\bar{\alpha})](\langle \ln \beta \rangle - \ln \bar{\beta}) + 0.5\alpha^2[\psi'(\bar{\alpha} + \bar{\beta}) \\ & - \psi'(\bar{\alpha}))]\langle (\ln \alpha - \ln \bar{\alpha})^2 \rangle + 0.5\beta^2[\psi'(\bar{\alpha} + \bar{\beta}) - \psi'(\bar{\beta}))]\langle (\ln \beta - \ln \bar{\beta})^2 \rangle \\ & + \bar{\alpha}\bar{\beta}\psi'(\bar{\alpha} + \bar{\beta})(\langle \ln \alpha \rangle - \ln \bar{\alpha})(\langle \ln \beta \rangle - \ln \bar{\beta}) \end{aligned} \quad (18)$$

$$u_{jl}^{*} = u_{jl} + \sum_{i=1}^{N} \langle Z_{ij} \rangle \bar{\alpha}_{jl} \left[\psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl} \psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl})(\langle \ln \beta_{jl} \rangle - \ln \bar{\beta}_{jl}) \right] \tag{19}$$

$$\nu_{jl}^{*} = \nu_{jl} - \sum_{i=1}^{N} \langle Z_{ij} \rangle \ln \frac{X_{il}}{1 + X_{il}} \tag{20}$$

$$g_{jl}^{*} = g_{jl} + \sum_{i=1}^{N} \langle Z_{ij} \rangle \bar{\beta}_{jl} \left[\psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \psi(\bar{\beta}_{jl}) + \bar{\alpha}_{jl} \psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl})(\langle \ln \alpha_{jl} \rangle - \ln \bar{\alpha}_{jl}) \right] \tag{21}$$

$$h_{jl}^{*} = h_{jl} - \sum_{i=1}^{N} \langle Z_{ij} \rangle \ln \frac{1}{1 + X_{il}} \tag{22}$$

Furthermore $\psi(\cdot)$ and $\psi^{'}(\cdot)$ are representing the Digamma and Trigamma functions, respectively. As $R = \langle \ln \frac{\Gamma(\bar{\alpha}+\bar{\beta})}{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})} \rangle$ is intractable, we have used the second order Taylor expansion for its approximation. The expected values of the above equations are as follows:

$$\langle Z_{ij} \rangle = r_{ij} \tag{23}$$

$$\bar{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{u_{jl}^{*}}{\nu_{jl}^{*}} \ , \ \langle \ln \alpha_{jl} \rangle = \psi(u_{jl}^{*}) - \ln \nu_{jl}^{*} \tag{24}$$

$$\bar{\beta}_{jl} = \langle \beta_{jl} \rangle = \frac{g_{jl}^{*}}{h_{jl}^{*}} \ , \ \langle \ln \beta_{jl} \rangle = \psi(g_{jl}^{*}) - \ln h_{jl}^{*} \tag{25}$$

$$\pi_{j} = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \tag{26}$$

4 Entropy-Based Variational Model Learning

In this section, we develop an entropy-based variational inference to learn the generalized inverted Dirichlet mixture model, that is mainly motivated by [23]. The core idea is to evaluate the quality of fitting of a component of our mixture model. Hence, we do a comparison between the theoretical maximum entropy and the MeanNN entropy [28]. In case of a significant difference, we proceed with a splitting process to fit the component, which consists in splitting the component into two new clusters.

4.1 Differential Entropy Estimation

The probability density function of an observation $\boldsymbol{X}_i = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_D)$ is defined as $p(\boldsymbol{X}_i)$, with a set of N samples $\{\boldsymbol{X}_i, \ldots, \boldsymbol{X}_N\}$, the differential entropy can be defined as:

$$H(\boldsymbol{X}_i) = -\int p(\boldsymbol{X}_i) log_2 P(\boldsymbol{X}_i) d\boldsymbol{X}_i \tag{27}$$

We introduce the maximum differential entropy of the GID as follows:

$$H_{GID}(\boldsymbol{X}_i \mid \alpha_j, \beta_j) = \sum_{l=1}^{D} \Big[-\ln \Gamma(\alpha_{jl} + \beta_{jl}) + \ln \Gamma(\alpha_{jl}) + \ln \Gamma(\beta_{jl}) \qquad (28)$$
$$- (\alpha_{jl} - 1)\big[-\psi(\alpha_{jl} + \beta_{jl}) + \psi(\alpha_{jl})\big] + (\alpha_{jl} + \beta_{jl})[-\psi(\alpha_{jl} + \beta_{jl})]\Big]$$

4.2 MeanNN Entropy Estimator

In order to make sure that the specified component is indeed distributed according to a generalized inverted Dirichlet distribution, we choose the MeanNN entropy estimator [23], to estimate $H(\boldsymbol{X}_i)$ for random variable $\boldsymbol{X}_i$ with D dimensions, that has an unknown density function $P(\boldsymbol{X}_i)$ [29]. By considering the fact that the Shannon entropy estimator in (27) can be considered equal to the average of $-\log P(\boldsymbol{X}_i)$, we can exploit an unbiased estimator by estimating $\log P(\boldsymbol{X}_i)$ [28,29]. We assume that $\boldsymbol{X}_i$ is the center of a ball with diameter ϵ, and that there is a point within the distance $[\epsilon, \epsilon + d_\epsilon]$ from $\boldsymbol{X}_i$. We have $\hat{k} - 1$ points in a smaller distance, and the other $N - \hat{k} - 1$ points are within a large distance from $\boldsymbol{X}_i$. Consequently, we can define the probability of the distances and the k-th nearest neighbor as follows:

$$p_{i\hat{k}}(\epsilon) = \frac{(N-1)!}{\left(\hat{k}-1\right)!\left(N-\hat{k}-1\right)!} \frac{dp_i(\epsilon)}{d\epsilon} p_i^{\hat{k}-1} (1-p_i)^{N-\hat{k}-1} \qquad (29)$$

where $p_i(\epsilon)$ denotes the mass of the ϵ-ball centered on $\boldsymbol{X}_i$:

$$p_i(\epsilon) = \int_{||\boldsymbol{X}-\boldsymbol{X}_i||<\epsilon} p(\boldsymbol{X}_i) d\boldsymbol{X}_i \qquad (30)$$

We can easily define the expectation of $\log p_i(\epsilon)$ with respect to $p_i(\epsilon)$ as mentioned in Eq. (31):

$$\mathbb{E}\big(\log p_i(\epsilon)\big) = \int_0^\infty p_{i\hat{k}} \log p_i(\epsilon) d\epsilon = \psi(\hat{k}) - \psi(N) \qquad (31)$$

Imagine $P(\boldsymbol{X}_i)$ is unchanging in the center of the ϵ-ball, we have $p_i(\epsilon) \simeq V_d \epsilon^d p(\boldsymbol{X}_i)$, where d corresponds to the dimension of $\boldsymbol{X}_i$, and V_d is the unit ball volume calculated by $V_d = \pi^{\frac{d}{2}} \Gamma(1 + d/2)$. Now, we are able to approximate $-\log P(\boldsymbol{X}_i)$ by substituting (30) into (31) we can get the Eq. (32). Hence, we get the unbiased K-NN estimator of the differential entropy, expressed in (33):

$$-\log p(\boldsymbol{X}_i) \simeq \psi(N) - \psi(\hat{k}) + dE(\log \epsilon) + \log V_d \qquad (32)$$

$$H_{\hat{k}}(\boldsymbol{X}) = \psi(N) - \psi\left(\hat{k}\right) + \frac{d}{N} \sum_{i=1}^{N} log\epsilon_i + \log V_d \qquad (33)$$

To reduce the high computational expenses of the K-NN estimator, we use an extension of the K-NN estimator called MeanNN, proposed in [30]. The main idea behind the MeanNN entropy estimator is to average the $\hat{k}$ nearest neighbor statics for all feasible values of order k in the range of $[1, N-1]$. The MeanNN estimator for the differential entropy is calculated according to (34).

$$H_M(\boldsymbol{X}) = \frac{1}{N-1}\sum_{\hat{k}=1}^{N-1} H_{\hat{k}}(\boldsymbol{X}) = \log V_d + \psi(N) + \frac{1}{N-1}\sum_{\hat{k}=1}^{N-1}\left[\frac{d}{N}\sum_{i=1}^{N}\log \epsilon_{i,\hat{k}} - \psi\left(\hat{k}\right)\right] \tag{34}$$

where $\epsilon_{i,\hat{k}}$ determines the $\hat{k}$-th nearest neighbor of $\boldsymbol{X}_i$. To find the maximum differential entropy of each individual cluster, we use:

$$H_{GID} = \sum_{j=1}^{M} \pi_j H_{GID}(j) \tag{35}$$

At this point, we are able to give an accurate evaluation of the model fitting, by evaluating and comparing the MeanNN and the theoretical maximum differential entropy [30]. Afterwards, we define Ω_{GID}, which is the normalized weighted sum of the difference between the theoretical and the estimated entropy of every component correlated with the generalized inverted Dirichlet mixture model, as expressed bellow:

$$\Omega_{GID} = \sum_{j=1}^{M} \pi_j \left[\frac{H_{GID}(j) - H_M(j)}{H_{GID}(j)}\right] = \sum_{j=1}^{M} \pi_j \left[1 - \frac{H_M(j)}{H_{GID}(j)}\right] \tag{36}$$

The normalized weight Ω_{GID} operates in the range of $[0, 1]$ and it is equal to zero, only if the data was genuinely distributed. The splitting process is performed by choosing the cluster j^* with the highest Ω_{GID} according to Eq. (37), and split the chosen component j^* into two new components.

$$j^* = \arg\max_j \left[\Omega_{GID}(j)\right] = \arg\max_j \left[\pi_j \frac{H_{GID}(j) - H_M(j)}{H_{GID}(j)}\right] \tag{37}$$

The overall entropy-based variational learning algorithm of the GID mixture model is illustrated in Algorithm 1.

Algorithm 1. Entropy-based variational learning of GID mixture models

1. Initialization: Set $M = 1$, $j^* = M$, $\pi_1 = 1$. and initialize hyperparameters u_{jl},ν_{jl},g_{jl},h_{jl}.
2. The splitting process.
   - Split j^* into two new components j_1 and j_2 with equal proportion $\pi^*/2$.
   - Set $M = M + 1$.
   - Initialize the parameters of j_1 and j_2 using the same parameters of j^*.
3. Apply standard variational Bayes until convergence.
4. Determine the number of components through the evaluation of the mixing coefficients π_j according to 26.
5. If $\pi_j \approx 0$. where $j \in 1, \ldots, M$ then set $M = M - 1$ and terminate the program.
6. Else evaluate Ω_{MD}, choose j^* according to 37 and go back to the splitting process in step 2.

5 Experimental Results

In order to demonstrate the effectiveness of the proposed model, Entropy-Based Variational Learning of Finite Generalized Inverted Dirichlet Mixture Model (EV-GIDMM), we conduct several experiments on two real-world challenging applications, including breast cancer detection and image categorization. In the first one, we used the standard breast cancer (Wisconsin Prognostic) dataset with numerical features, whereas in the second one, we run our experiments on two other popular datasets, namely, Caltech101 and Describable Texture Dataset (DTD). To validate the performance of our model, we compared our proposed EV-GIDMM against three unsupervised state-of-the-art mixture models, including the Entropy-based variational inference on Multivariate Beta Mixture Model (EV-MBMM) [23], variational Dirichlet Mixture Model (varDMM) [25] and Entropy-based variational on Dirichlet Mixture Model (E-DMM) [24].

5.1 Breast Cancer

The first application that we considered to evaluate the performance of our proposed model is breast cancer detection. According to the WHO (World Health Organization), breast cancer has been declared as the most frequent cancer among women that affects about 2.1 million women every year. Machine learning techniques can be of great help in this context, in early detection of women breast cancer, thus, they can have a great impact on the breast cancer treatment. To this end, we applied our proposed model on the breast cancer Wisconsin dataset that is publicly available[1]. This dataset includes 569 data samples of patients seen by Dr. Wolberg, that have been diagnosed with either malignant or benign cancer. The number of patients having a benign tumor is 357, whereas 212 cases with malignant tumor cancer. This dataset was obtained by applying the Fine Needle Aspiration (FNA) method [31,32], and it contains cases showing invasive

[1] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

breast cancer and no sign of distant metastases. The first 30 features describe the characteristics of each nuclei cell in the images of the tissue. Table 1 shows the experimental results of our model as well as the baseline methods for the breast cancer detection task. We can see that our proposed EV-GIDMM successfully achieved the best accuracy on this task.

Table 1. Accuracy performance of our model and the baselines on the breast cancer dataset

Method	Accuracy(%)
EV-GIDMM	**92.6**
EV-MBMM	90.8
E-DMM	89.7
varDMM	63.5

5.2 Image Analysis

We are now ready to evaluate the performance of the proposed approach on the image categorization task, which is a significant research topic and aims at classifying images into their corresponding category. To do so, we used two popular image datasets, namely, Caltech101 and Describable Texture Dataset (DTD). In this experiment, we first considered the Caltech101 image dataset[2] [33], which originally contains a set of images depicting objects belonging to 101 classes, from which we selected three main object categories: Airplane, Sea Horse and Brain. Some sample images from this dataset are illustrated in Fig. 1.

Fig. 1. Sample images of each cluster from the Caltech101 dataset.

In order to use our model for the selected dataset, we need to form a bag of visual words model (BoVW) [34]. Before applying the BoVW, we first need to apply some descriptor extraction method, that, we choose SIFT [35]. Therefore we extract the features with the help of SIFT and then apply K-means clustering on the descriptors extracted with SIFT from the image. As a result a BOVW feature vector is formed for each image. Our experiments revealed that the SIFT

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101.html.

method is more suitable for our selected dataset, resulting in more discriminative descriptors. After applying SIFT to all images, we obtain a matrix that serves as an input for our model. We report the results of this experiment in Table 2, which shows that our proposed model outperformed all the baseline methods in image clustering, with a considerable accuracy margin of almost 6.6%.

Table 2. Accuracy comparison of our proposed model and the baseline methods on the Caltech101 dataset.

Method	Accuracy(%)
EV-GIDMM	**90.9**
EV-MBMM	84.3
E-DMM	74.9
varDMM	40.3

In the second part of our experiments, we focus on texture differentiation. This dataset will be a good challenge for our model as images are very similar. In order to show how machines are becoming more capable of detecting and recognizing fine-grained images, in this experiment, we chose to use the Describable Texture Dataset[3] that includes 120 images per class where each class consists of different types of textures. We have chosen Dotted, Frilly and Meshed image categories to evaluate our model as illustrated in Fig. 2.

Fig. 2. Sample images of each cluster from the DTD dataset.

Similarly, we performed the BoVW and used SIFT, to generate a discriminative input for our EV-GIDMM. The results of clustering evaluation on DTD are listed in Table 3. From this table it can be confirmed that our proposed mixture model achieves the best accuracy performance among all the other mixture models.

[3] https://www.robots.ox.ac.uk/~vgg/data/dtd/.

Table 3. Accuracy comparison of our EV-GIDMM approach and the baseline methods on the DTD dataset.

Method	Accuracy(%)
EV-GIDMM	**85.5**
EV-MBMM	65.3
E-DMM	65.8
varDMM	71.9

6 Conclusion

In this paper, we introduced an unsupervised entropy-based variational framework that effectively learns the finite generalized inverted Dirichlet mixture model. In our method, we used a splitting technique called Entropy, where we started by comparing the theoretical maximum entropy and the resulting entropy from MeanNN. Thereafter, we proceeded to split the component that has the highest difference into two smaller components, since it was concluded that the mixture model is not describing the component properly. Our experimental results have demonstrated that EV-GIDMM works very well and has outperformed other models on two real-world applications, namely, breast cancer detection and image categorization, across three different benchmark data sets. The results indicate that our proposed mixture model is able to produce high quality data clusters.

Acknowledgment. The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Natural Science Foundation of China (61876068).

References

1. Bishop., C.M : Pattern recognition and machine learning. Information science and statistics. Springer, New York, NY, (2006). Softcover published in 2016
2. McLachlan, G.J., Peel, D.: Finite Mixture Models. John Wiley & Sons, Hoboken (2004)
3. Ho, T.K., Baird, H.S.: Large-scale simulation studies in image pattern recognition. IEEE Trans. Pattern Analy. Mach. Intell. **19**(10), 1067–1079 (1997)
4. Fan, W., Bouguila, N.: Non-Gaussian data clustering via expectation propagation learning of finite Dirichlet mixture models and applications. Neural Process. Lett. **39**(2), 115–135 (2014)
5. Bdiri, T., Bouguila, N.: Positive vectors clustering using inverted Dirichlet finite mixture models. Expert Syst. Appl. **39**(2), 1869–1882 (2012)
6. Bouguila, N., Ziou, D.: A countably infinite mixture model for clustering and feature selection. Knowl. Inf. Syst. **33**(2), 351–370 (2012)
7. Bouguila, N., Amayri, O.: A discrete mixture-based kernel for SVMs: application to spam and image categorization. Inf. Process. Manag. **45**(6), 631–642 (2009)

8. Sefidpour, A., Bouguila, N.: Spatial color image segmentation based on finite non-Gaussian mixture models. Expert Syst. Appl. **39**(10), 8993–9001 (2012)
9. Bouguila, N.: A model-based approach for discrete data clustering and feature weighting using MAP and stochastic complexity. IEEE Trans. Knowl. Data Eng. **21**(12), 1649–1664 (2009)
10. Bdiri, T., Bouguila, N.: Bayesian learning of inverted Dirichlet mixtures for SVM kernels generation. Neural Comput. Appl. **23**(5), 1443–1458 (2013)
11. Fan, W., Bouguila, N.: Online learning of a Dirichlet process mixture of beta-liouville distributions via variational inference. IEEE Trans. Neural Netw. Learn. Syst. **24**(11), 1850–1862 (2013)
12. Mashrgy, M.A.I., Bdiri, T., Bouguila, N.: Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted Dirichlet mixture models. Knowl. Based Syst. **59**, 182–195 (2014)
13. Bdiri, T., Bouguila, N.: Learning inverted Dirichlet mixtures for positive data clustering. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS (LNAI), vol. 6743, pp. 265–272. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21881-1_42
14. Bdiri, T., Bouguila, N.: An infinite mixture of inverted Dirichlet distributions. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011. LNCS, vol. 7063, pp. 71–78. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24958-7_9
15. Tirdad, P., Bouguila, N., Ziou, D.: Variational learning of finite inverted Dirichlet mixture models and applications. In: Laalaoui, Y., Bouguila, N. (eds.) Artificial Intelligence Applications in Information and Communication Technologies. SCI, vol. 607, pp. 119–145. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19833-0_6
16. Fukumizu, K., Amari, S.: Local minima and plateaus in hierarchical structures of multilayer perceptrons. Neural Netw. **13**(3), 317–327 (2000)
17. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized Drichlet mixture model based on minimum message length. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1716–1731 (2007)
18. Bouguila, N., Ziou, D.: MML-based approach for finite Dirichlet mixture estimation and selection. In: Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS (LNAI), vol. 3587, pp. 42–51. Springer, Heidelberg (2005). https://doi.org/10.1007/11510888_5
19. Maanicshah, K., Bouguila, N., Fan, W.: Variational learning for finite generalized inverted Dirichlet mixture models with a component splitting approach. In: 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), pp. 1453–1458. IEEE (2019)
20. Bourouis, S., Mashrgy, M.A.L., Bouguila, N.: Bayesian learning of finite generalized inverted Dirichlet mixtures: application to object classification and forgery detection. Expert Syst. Appl. **41**(5), 2329–2336 (2014)
21. Bdiri, T., Bouguila, N., Ziou, D.: Variational Bayesian inference for infinite generalized inverted Dirichlet mixtures with feature selection and its application to clustering. Appl. Intell. **44**(3), 507–525 (2016)
22. Mashrgy, M.A.L., Bdiri, T., Bouguila, N.: Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted Dirichlet mixture models. Knowl.-Based Syst. **59**, 182–195 (2014)
23. Manouchehri, N., Rahmanpour, M., Bouguila, N., Fan, W.: Learning of multivariate beta mixture models via entropy-based component splitting. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2825–2832. IEEE (2019)

24. Fan, W., Al-Osaimi, F.R., Bouguila, N., Du, J.: Proportional data modeling via entropy-based variational Bayes learning of mixture models. Appl. Intell. **47**(2), 473–487 (2017). https://doi.org/10.1007/s10489-017-0909-0
25. Fan, W., Bouguila, N., Ziou, D.: Variational learning for finite Dirichlet mixture models and applications. IEEE Trans. Neural Netw. Learn. Syst. **23**(5), 762–774 (2012)
26. Chandler, D.: Introduction to Modern Statistical. Mechanics. Oxford University Press, Oxford, UK (1987)
27. Celeux, G., Forbes, F., Peyrard, N.: Em procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recogn. **36**(1), 131–144 (2003)
28. Faivishevsky, L., Goldberger, J.: ICA based on a smooth estimation of the differential entropy. In: Advances in Neural Information Processing Systems, pp. 433–440 (2009)
29. Leonenko, N., Pronzato, L., Savani, V., et al.: A class of rényi information estimators for multidimensional densities. Ann. Stat. **36**(5), 2153–2182 (2008)
30. Penalver, A., Escolano, F.: Entropy-based incremental variational Bayes learning of Gaussian mixtures. IEEE Trans. Neural Netw. Learn. Syst. **23**(3), 534–540 (2012)
31. Dua, D., Graff, C.: UCI machine learning repository (2017)
32. Wolberg, W.H., Street, W.N., Mangasarian, O.L.: Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Lett. **77**(2–3), 163–171 (1994)
33. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 178–178. IEEE (2004)
34. Li, T., Mei, T., Kweon, I.-S., Hua, X.-S.: Contextual bag-of-words for visual categorization. IEEE Trans. Circ. Syst. Video Technol. **21**(4), 381–392 (2010)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)