An Infinite Mixture Model of Generalized Inverted Dirichlet Distributions for High-Dimensional Positive Data Modeling

Nizar Bouguila, Mohamed Mashrgy

An Infinite Mixture Model of Generalized Inverted Dirichlet Distributions for High-Dimensional Positive Data Modeling

Nizar Bouguila[1] and Mohamed Al Mashrgy[2]

[1] Concordia Institute for Information Systems Engineering
Concordia University, Montreal, QC, Canada
nizar.bouguila@concordia.ca
[2] Department of Electrical Engineering
Concordia University, Montreal, QC, Canada
m_almash@encs.concordia.ca

Abstract. We propose an infinite mixture model for the clustering of positive data. The proposed model is based on the generalized inverted Dirichlet distribution which has a more general covariance structure than the inverted Dirichlet that has been widely used recently in several machine learning and data mining applications. The proposed mixture is developed in an elegant way that allows simultaneous clustering and feature selection, and is learned using a fully Bayesian approach via Gibbs sampling. The merits of the proposed approach are demonstrated using a challenging application namely images categorization.

Keywords: Clustering, feature selection, generalized inverted Dirichlet, mixture models, Bayesian inference, image databases.

1 Introduction

The important proliferation of digital content requires the development of powerful approaches for knowledge extraction, analysis, and organization. Clustering, in particular, has been widely adopted for knowledge discovery and data engineering. The main goal of any clustering algorithm is to partition a given data set into groups so that objects within a cluster are more similar than those in different clusters [26]. Many clustering techniques have been developed in the past and have been applied successfully on different data types (e.g. binary, discrete, continuous) extracted within various applications [6, 11, 15]. Among these techniques, mixture models have played important roles in many areas including (but not confined to) image processing, computer vision, data mining, and pattern recognition, thanks to their flexibility and strong statistical foundations which offer a formal principled way to clustering. In particular, Gaussian mixture model has drawn considerable attention in the machine learning community and has achieved good results [21]. However, recent concentrated research efforts have shown that this mixture model may fail to provide good generalization capabilities when the per-cluster data distributions are clearly non-Gaussian, which is

the case of positive data as deeply discussed in [3, 2, 1].
The main contribution of [3, 2] was the introduction of the finite inverted Dirichlet mixture model for the clustering of positive data which are naturally generated by many real-world applications. The authors have proposed a detailed approach for the learning of the parameters of this finite mixture, also. In order to handle huge number of classes and avoid over- or under-fitting problems (a.k.a. controlling variance, model selection), which is a central issue in learning-based techniques, the finite inverted Dirichlet mixture was extended to the infinite case in [2]. This extension was based on the consideration of Dirichlet processes which have been widely used in the case of nonparametric Bayesian approaches [9, 8]. Despite its advantages and flexibility, the inverted Dirichlet has a very restrictive covariance structure that is generally violated by data generated from real-life applications. Thus, we propose an alternative to the inverted Dirichlet namely the generalized inverted Dirichlet (GID) that has a more general covariance structure. Our work can be viewed as a principled and natural extension to the framework developed in [2], since we consider the GID within an infinite mixture model by taking feature selection into account. The feature selection process is formalized by introducing a background distribution, common to all mixture components, into the infinite model to represent irrelevant features. Moreover, we develop an algorithm for the learning of the resulting model using Markov chain Monte Carlo (MCMC) sampling techniques namely Gibbs sampling and Metropolis-Hastings [20].
The paper is organized as follows: Section 2 presents our infinite mixture model. Section 3 provides empirical evaluation based on the challenging problem of images categorization. Finally, Section 4 concludes the paper.

2 The Model

In this section, we start by presenting the finite GID mixture model, then its infinite counterpart is developed. A feature selection approach is proposed, also.

2.1 Finite Model

Let us consider a data set $\mathcal{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N)$ of N D-dimensional positive vectors, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iD}), i = 1, \ldots, N$. We assume that $\boldsymbol{Y}_i$ follows a mixture of M GID distributions:

$$p(\boldsymbol{Y}_i|\Theta) = \sum_{j=1}^{M} \pi_j p(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j) \tag{1}$$

where $p(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j)$ is a GID distribution [18]:

$$p(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j) = \prod_{l=1}^{D} \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \frac{Y_{il}^{\alpha_{jl}-1}}{T_{il}^{\eta_{jl}}} \tag{2}$$

where $T_{il} = 1 + \sum_{k=1}^{l} Y_{ik}$ and $\eta_{jl} = \beta_{jl} + \alpha_{jl} - \beta_{j(l+1)}$ with $\beta_{j(D+1)} = 0$. Each $\boldsymbol{\Theta}_j = (\alpha_{j1}, \beta_{j1}, \alpha_{j2}, \beta_{j2}, \ldots, \alpha_{jD}, \beta_{jD})$ is the set of parameters defining the jth component, and π_j is the mixing weight of the jth cluster. Of course, being probabilities, the π_j must satisfy: $\pi_j > 0, j = 1, \ldots, M$, and $\sum_{j=1}^{M} \pi_j = 1$.
In mixture-based clustering [21], each vector $\boldsymbol{Y}_i$ is assigned to all classes with different posterior probabilities $p(j|\boldsymbol{Y}_i) \propto \pi_j p(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j)$. It is possible to show that the properties of the GID distribution allows the factorization of the posterior probabilities as: $p(j|\boldsymbol{Y}_i) \propto \pi_j \prod_{l=1}^{D} p_{ib}(X_{il}|\theta_{jl})$, where $X_{i1} = Y_{i1}$ and $X_{il} = \frac{Y_{il}}{1+\sum_{l=1}^{D} Y_{il}}$ for $l > 1$, $p_{ib}(X_{il}|\theta_{jl})$ is an inverted Beta distribution with $\theta_{jl} = (\alpha_{jl}, \beta_{jl})$, $l = 1, \ldots, D$:

$$p_{ib}(X_{il}|\alpha_{jl}, \beta_{jl}) = \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} X_{il}^{\alpha_{jl}-1}(1 + X_{il})^{-\alpha_{jl}-\beta_{jl}}$$

Thus, the clustering structure underlying $\mathcal{Y}$ is the same as that underlying $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ described by the following mixture model with conditionally independent features:

$$p(\boldsymbol{X}_i|\Theta) = \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} p_{ib}(X_{il}|\theta_{jl}) \tag{3}$$

This means that GID mixture model has the ability to reduce complex multidimensional clustering problems to a sequence of one-dimensional ones.

2.2 Infinite model

Let Z_i be a variable indicating from which cluster each vectors $\boldsymbol{X}_i$ arose (i.e $Z_i = j$ means that $\boldsymbol{X}_i$ comes from component j), thus $\pi_j = p(Z_i = j), j = 1, \ldots, M$ and

$$p(Z|P) = \prod_{j=1}^{M} \pi_j^{n_j} \tag{4}$$

where $P = (\pi_1, \ldots, \pi_M)$, $Z = (Z_1, \ldots, Z_N)$, $n_j = \sum_{i=1}^{N} \mathbb{I}_{Z_i=j}$ is the number of vector in cluster j. It is common to consider a Dirichlet distribution as a prior for P which is justified by the fact that the Dirichlet is conjugate to the multinomial [7]:

$$p(P|\eta_1, \ldots, \eta_M) = \frac{\Gamma(\sum_{j=1}^{M} \eta_j)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} p_j^{\eta_j - 1} \tag{5}$$

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

$$p(P|\eta) = \frac{\Gamma(\eta)}{\Gamma(\frac{\eta}{M})^M} \prod_{j=1}^{M} p_j^{\eta - 1} \tag{6}$$

Because the Dirichlet is a conjugate prior to the multinomial, we can marginalize out P:

$$p(Z|\eta) = \int_P p(Z|P)p(P|\eta)dP = \frac{\Gamma(\eta)}{\Gamma(\eta+N)} \prod_{j=1}^{M} \frac{\Gamma(\frac{\eta}{M}+n_j)}{\Gamma(\frac{\eta}{M})}$$

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

$$p(P|Z,\eta) = \frac{p(Z|P)p(P|\eta)}{p(Z|\eta)} = \frac{\Gamma(\eta+N)}{\prod_{j=1}^{M}\Gamma(\frac{\eta}{M}+n_j)} \prod_{j=1}^{M} p_j^{n_j+\frac{\eta}{M}-1} \tag{7}$$

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

$$p(Z_i = j|\eta, Z_{-i}) = \frac{n_{-i,j}+\frac{\eta}{M}}{N-1+\eta} \tag{8}$$

where $Z_{-i} = \{Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_N\}$, $n_{-i,j}$ is the number of vectors, excluding $\boldsymbol{Y}_i$, in cluster j. Letting $M \to \infty$ in Eq. 8, the conditional prior gives the following limits [23]

$$p(Z_i = j|\eta, Z_{-i}) = \begin{cases} \frac{n_{-i,j}}{N-1+\eta} \text{ if } n_{-i,j} > 0 \text{ (cluster } j \in \mathcal{R}) \\ \frac{\eta}{N-1+\eta} \text{ if } n_{-i,j} = 0 \text{ (cluster } j \in \mathcal{U}) \end{cases} \tag{9}$$

where $\mathcal{R}$ and $\mathcal{U}$ are the sets of represented and unrepresented clusters, respectively. The previous equation describes actually a Dirichlet process of mixtures which learning is generally based on the MCMC technique of Gibbs sampling [1] by generating the assignments of vectors according to the posterior distribution

$$p(Z_i = j|Z_{-i},\mathcal{X}) \propto p(Z_i = j|Z_{-i}) \int p(\boldsymbol{X}_i|Z_i = j,\Theta_j)p(\Theta_j|Z_{-i},\mathcal{X}_{-i})d\Theta_j \tag{10}$$

where Z_{-i} represents all the vectors assignments except Z_i and $\mathcal{X}_{-i}$ represents all the vectors except $\boldsymbol{X}_i$.
In order to obtain the conditional posterior distributions of our infinite model's parameters given the data that we would like to cluster, we need to choose appropriate priors. Here, we consider the same priors previously proposed in [1] for the inverted Dirichlet which is actually the multivariate case of the inverted Beta in Eq. 3. Thus, we need to parametrize the inverted Beta as following:

$$p_{ib}(X_{il}||\alpha_{jl}|,\mu_{jl}) = \frac{\Gamma(|\alpha_{jl}|)}{\Gamma(\mu_{jl}|\alpha_{jl}|)\Gamma((1-\mu_{jl})|\alpha_{jl}|)} X_{il}^{\mu_{jl}|\alpha_{jl}|-1}(1+X_{il})^{-|\alpha_{jl}|} \tag{11}$$

where $|\alpha_{jl}| = \alpha_{jl}+\beta_{jl}$, $\mu_{jl} = \frac{\alpha_{jl}}{|\alpha_{jl}|}$, and for which we impose independent uniform and inverse Gamma priors, respectively:

$$p(\mu_{jl}) \sim \mathcal{U}_{[0,1]}^{jl} \qquad p(|\boldsymbol{\alpha}_{jl}||\sigma,\varpi) \sim \frac{\varpi^\sigma \exp(-\varpi/|\boldsymbol{\alpha}_{jl}|)}{\Gamma(\sigma)|\boldsymbol{\alpha}_{jl}|^{\sigma+1}} \tag{12}$$

where σ and ϖ are hyperparameters, common to all components, representing shape and scale of the distribution, respectively, and for which we consider the following priors to add more flexibility to the model:

$$p(\sigma|\lambda,\delta) \sim \frac{\delta^{\lambda}\exp(-\delta/\sigma)}{\Gamma(\lambda)\sigma^{\lambda+1}} \qquad p(\varpi|\phi) \sim \phi\exp(-\phi\varpi) \tag{13}$$

Having all our priors in hand, the calculation of the parameters posteriors given the rest of the variables becomes straightforward:

$$p(|\alpha_{jl}||\ldots) \propto \frac{\varpi^{\sigma}\exp(-\varpi/|\alpha_{jl}|)}{\Gamma(\sigma)|\alpha_{jl}|^{\sigma+1}} \prod_{Z_i=j} p(\boldsymbol{X}_i|\Theta) \tag{14}$$

$$p(\boldsymbol{\mu}_j|\ldots) \propto \prod_{Z_i=j} p(\boldsymbol{X}_i|\Theta) \tag{15}$$

$$p(\sigma|\ldots) \propto \frac{\varpi^{M\sigma}\delta^{\lambda}\exp(-\delta/\sigma)}{\Gamma(\sigma)^{M}\Gamma(\lambda)\sigma^{\lambda+1}} \prod_{j=1}^{M} \frac{\exp(-\varpi/|\alpha_{jl}|)}{|\alpha_{jl}|^{\sigma+1}} \tag{16}$$

$$p(\varpi|\ldots) \propto \frac{\varpi^{M\sigma}\phi\exp(-\phi\varpi)}{\Gamma(\sigma)^{M}} \prod_{j=1}^{M} \frac{\exp(-\varpi/|\alpha_{jl}|)}{|\alpha_{jl}|^{\sigma+1}} \tag{17}$$

With these posteriors, the learning algorithm can be summarized as follows:

- Initialization.
- Generate $\boldsymbol{Z}_i$ from Eq. 10, $i = 1,\ldots,N$ using the algorithm in [22].
- Update the number of represented components M.
- Update n_j and $\pi_j = \frac{n_j}{N+\eta}$, $j = 1,\ldots,M$.
- Update the mixing parameters of unrepresented components $\pi_U = \frac{\eta}{\eta+N}$.
- Generate $|\mu_{jl}|$ from Eq. 15 and $|\alpha_{jl}|$ from Eq. 14, $j = 1,\ldots,M$ using Metropolis-Hastings [20].
- Update the hyperparameters: Generate σ from Eq. 16 and ϖ from Eq. 17 using adaptive rejection sampling as proposed in [13].

Note that in the initialization step, the algorithm starts by assuming that all the vectors are in the same cluster and the initial parameters are generated as random samples from their prior distributions.

2.3 Feature Selection

It is noteworthy that the model proposed in the previous section does not take into account the fact that different features may have different weights in the clustering structure, and that some features may be noise and then compromise the generalization capabilities of the model [12]. In order to introduce feature selection in our model, it is possible to use the following formulation:

$$p(\boldsymbol{X}_i|\Xi) = \sum_{j=1}^{M}\pi_j \prod_{l=1}^{D}\left[\rho_l p_{ib}(X_{il}||\alpha_{jl}|,\mu_{jl}) + (1-\rho_l)p_{ib}(X_{il}||\alpha_{jl}^{irr}|,\mu_{jl}^{irr})\right] \tag{18}$$

where $\Xi = \{\Theta, \rho, \Theta^{irr}\}$ is the set of all the model parameters, $\rho = (\rho_1, \ldots, \rho_D)$, $\Theta^{irr} = \{|\alpha_{jl}^{irr}|, \mu_{jl}^{irr}\}$, and $p_{ib}(X_{il}||\alpha_{jl}^{irr}|, \mu_{jl}^{irr})$ is a background distribution, common to all mixture components, to represent irrelevant features. $\rho_l = p(z_{il} = 1)$ represents the probability that the l^{th} feature is relevant for clustering where z_{il} is a hidden variable equal to 1 if the l^{th} feature of $\boldsymbol{X}_i$ is relevant and 0, otherwise. By introducing feature selection, the learning algorithm proposed in the previous section has to be slightly modified by adding simulations from the posteriors of $|\alpha_{jl}^{irr}|$, μ_{jl}^{irr}, for which we choose the same priors considered for $|\alpha_{jl}|$, μ_{jl}, and ρ for which we consider a Beta prior with location δ_1 and scale δ_2 common to all dimensions:

$$p(\rho|\delta_1, \delta_2) = \left[\frac{\Gamma(\delta_2)}{\Gamma(\delta_1\delta_2)\Gamma(\delta_2(1-\delta_1))}\right]^D \prod_{d=1}^{D} \rho_d^{\delta_1\delta_2 - 1}(1-\rho_d)^{\delta_2(1-\delta_1)-1} \tag{19}$$

Moreover, the z_i are generated from a D-variate Bernoulli distribution with parameters $(\hat{z}_{i1}, \ldots, \hat{z}_{iD})$, where $\hat{z}_{il} = \frac{\rho_l p_{ib}(X_{il}||\alpha_{jl}|, \mu_{jl})}{\rho_l p_{ib}(X_{il}||\alpha_{jl}|, \mu_{jl}) + (1-\rho_l) p_{ib}(X_{il}||\alpha_{jl}^{irr}|, \mu_{jl}^{irr})}$ denotes the expectation for z_{il}:

$$p(z|\rho) = \prod_{i=1}^{N}\prod_{d=1}^{D} \rho_d^{z_{id}}(1-\rho_d)^{1-z_{id}} = \prod_{d=1}^{D} \rho_d^{f_d}(1-\rho_d)^{N-f_d} \tag{20}$$

where $f_d = \sum_{i=1}^{N} \mathbb{I}_{z_{id}=1}$. Then, the posterior for ρ is

$$p(\rho|\ldots) \propto p(\rho|\delta_1, \delta_2)p(z|\rho) \propto \prod_{d=1}^{D} \rho_d^{\delta_1\delta_2 + f_d - 1}(1-\rho_d)^{\delta_2(1-\delta_1)+N-f_d-1} \tag{21}$$

Note that the feature selection process starts by assuming that all features have a probability of 0.5 to be relevant, then this relevancy value is updated during the learning iterations.

3 Experimental Results: Images Categorization

In this section we demonstrate the utility of our model by applying it on a challenging application namely visual scenes categorization. Moreover, we compare the proposed approach with the infinite inverted Dirichlet proposed in [2]. Comparing our results with many other generative and discriminative techniques is clearly out of the scope of this paper. In this application, the values of the hyperparameters have been set experimentally to one. This choice has been found reasonable according to our simulations.

The wealth of images generated everyday has spurred a tremendous interest in developing approaches to understand the visual content of these images. In this section, we shall focus on the challenging problem of images categorization, to validate our GID infinite mixture model, which is a crucial step in several applications such as annotation [5, 10], retrieval [14, 27], and object recognition [25].

A common recent approach widely used for images categorization, that we follow in this application, is the consideration of the so-called bag of visual words generated via quantization of local image descriptors such as SIFT [19].
We considered two challenging datasets in our experiments namely the 15 class scene recognition data set [16] and the 8 class sport events data set [17]. The 15 class scene recognition data set contains the following categories: 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), suburb residence (241 images), bedroom (174 images), kitchen (151 images), livingroom (289 images), and office (216 images), store (315 images), and industrial (311 images). Figure 1 displays examples of images from this data set. The 8 class sports event dataset contains the following cate-

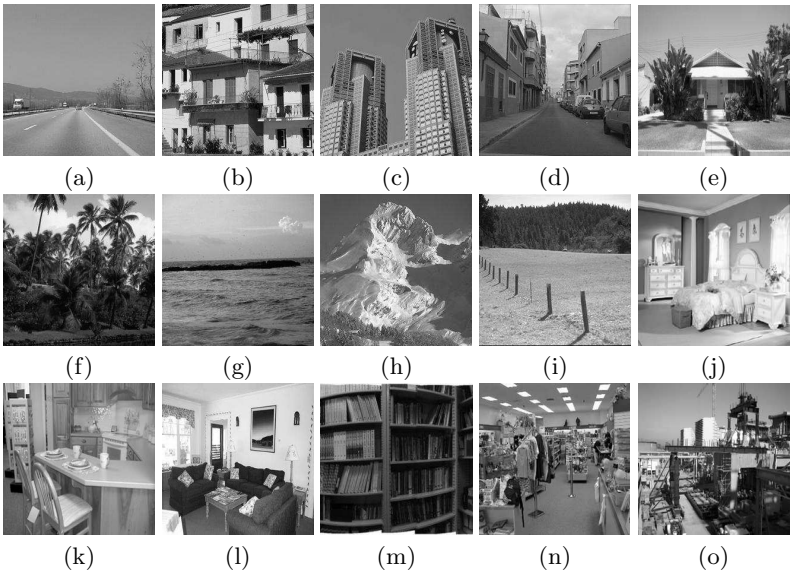

Fig. 1. Sample images from each group in the 15 class scene recognition data set. (a) Highway, (b) Inside of cities, (c) Tall buildings, (d) Streets, (e) Suburb residence, (f) Forest, (g) Coast, (h) Mountain, (i) Open country, (j) Bedroom, (k) Kitchen, (l) Livingroom, (m) Office, (n) Store, (O) Industrial.

gories: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). Figure 2 displays examples of images from this data set. We construct our visual vocabulary for each data set, from half of the available images in each data set, by detecting interest points from these images using the difference-of-Gaussians point detector, since it has shown excellent performance [19]. Then, we used SIFT descriptor [19], computed on

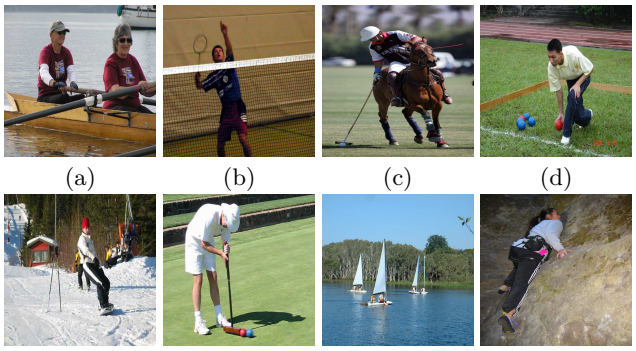

Fig. 2. Sample images from each group in the 8 class sports event dataset . (a) rowing, (b) badminton, (c) polo, (d) bocce, (e) snowboarding, (f) croquet, (g) sailing, (h) rock climbing.

detected keypoints of all images and giving 128-dimensional vector for each keypoint. Moreover, extracted vectors were clustered using the K-Means algorithm providing 250 visual-words. Each image in the data sets was then represented by a 250-dimensional positive vector describing the frequencies of visual words, provided from the constructed visual vocabulary. These vectors are separated into a test set of vectors and a training set of vectors. Then, we apply our learning algorithm to the training vectors in each class. After this stage, each class in the database is represented by a statistical model. Finally, in the classification stage each unknown image is assigned to the class increasing more its loglikelihood. 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 table 1. This table clearly shows that the GID infinite mixture outperforms the infinite inverted Dirichlet mixture. The results can be explained by the fact that the GID is more flexible than the inverted Dirichlet. We can clearly notice, also, that introducing feature selection improves further the results.

Table 1. Classification performance (%) obtained for the two tested data sets using three different approaches.

	GID	GID + feature selection	Inverted Dirichlet
Data set 1 (15 categories)	74.52	75.31	70.11
Data set 2 (8 events)	73.25	74.03	70.72

4 Conclusion

Clustering plays a crucial role in various data mining and knowledge discovery applications. The majority of existing clustering algorithms, however, either assume that clusters follow Gaussian distributions; or are very sensitive to the presence of irrelevant features. In this paper we have proposed a new clustering algorithm devoted to positive data that is robust to irrelevant features, and identifies automatically clusters having non-Gaussian distributions. Our approach achieves this by representing the data using an infinite mixture model of GID distributions in which a feature weighting component is introduced. Feature selection is introduced in order to remove irrelevant features that may compromise the clustering process. Our simulations based on the challenging problem of images categorization have shown the efficiency of the proposed model. A potential future work could be the development of a variational approach, like the one proposed in [12], to improve the learning of our model from a computational point of view. Several other directions present themselves for future efforts. Indeed, the developed approach could be applied to many real-world problems such as 3D object recognition [24] or to the generation of SVM kernels using the methodology recently proposed in [4].

References

1. Bdiri, T., Bouguila, N.: An infinite mixture of inverted dirichlet distributions. In: Lu, B.L., Zhang, L., Kwok, J.T. (eds.) ICONIP (2). Lecture Notes in Computer Science, vol. 7063, pp. 71–78. Springer (2011)
2. Bdiri, T., Bouguila, N.: Learning inverted dirichlet mixtures for positive data clustering. In: Kuznetsov, S.O., Slezak, D., Hepting, D.H., Mirkin, B. (eds.) RSFDGrC. Lecture Notes in Computer Science, vol. 6743, pp. 265–272. Springer (2011)
3. Bdiri, T., Bouguila, N.: Positive vectors clustering using inverted dirichlet finite mixture models. Expert Systems with Applications 39(2), 1869–1882 (2012)
4. Bdiri, T., Bouguila, N.: Bayesian learning of inverted dirichlet mixtures for svm kernels generation. Neural Computing and Applications 23(5), 1443–1458 (2013)
5. Benitez, A., Chang, S.F.: Semantic knowledge construction from annotated image collections. In: Proc. of the IEEE International Conference on Multimedia and Expo (ICME). pp. 205–208 vol.2 (2002)
6. Bezdek, J.C., Hathaway, R.J., Huband, J.M., Leckie, C., Ramamohanarao, K.: Approximate clustering in very large relational data. International Journal of Intelligent Systems 21(8), 817–841 (2006)
7. Bouguila, N., ElGuebaly, W.: On discrete data clustering. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD. Lecture Notes in Computer Science, vol. 5012, pp. 503–510. Springer (2008)
8. Bouguila, N., Ziou, D.: A nonparametric bayesian learning model: Application to text and image categorization. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.B. (eds.) PAKDD. Lecture Notes in Computer Science, vol. 5476, pp. 463–474. Springer (2009)
9. Bouguila, N., Ziou, D.: A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling. IEEE Transactions on Neural Networks 21(1), 107–122 (2010)

10. Chang, E.Y., Goh, K., Sychay, G., Wu, G.: Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuits Systems and Video Technology 13(1), 26–38 (2003)
11. Chen, W., Feng, G.: Spectral clustering with discriminant cuts. Knowledge-Based Systems 28, 27–37 (2012)
12. Fan, W., Bouguila, N., Ziou, D.: Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference. IEEE Transactions on Knowledge and Data Engineering 25(7), 1670–1685 (2013)
13. Gilks, W.R., Wild, P.: Algorithm as 287: Adaptive rejection sampling from log-concave density functions. Applied Statistics 42(4), 701–709 (1993)
14. He, J., Li, M., Zhang, H.J., Tong, H., Zhang, C.: Manifold-ranking based image retrieval. In: Proc. of the 12th Annual ACM International Conference on Multimedia (MM). pp. 9–16 (2004)
15. Huang, K.Y.: A hybrid particle swarm optimization approach for clustering and classification of datasets. Knowledge Based Systems 24(3), 420–426 (2011)
16. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of the EEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 2169–2178 (2006)
17. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: Proc. of the IEEE 11th International Conference on Computer Vision (ICCV). pp. 1–8 (2007)
18. Lingappaiah, G.S.: On the generalised inverted dirichlet distribution. Demostratio Mathematica 9(3), 423–433 (1976)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
20. Marin, J.M., Robert, C.P.: Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer (2007)
21. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley-Interscience (2000)
22. Neal, R.M.: Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics 9, 249–265 (2000)
23. Rasmussen, C.E.: The infinite gaussian mixture model. In: Advances in Neural Information Processing Systems (NIPS). pp. 554–560 (2000)
24. Selinger, A., Nelson, R.C.: A perceptual grouping hierarchy for appearance-based 3d object recognition. Computer Vision and Image Understanding 76(1), 83–92 (1999)
25. Spirkovska, L., Reid, M.B.: Higher-order neural networks applied to 2d and 3d object recognition. Machine Learning 15(2), 169–199 (1994)
26. Topchy, A., Law, M., Jain, A., Fred, A.: Analysis of consensus partition in cluster ensemble. In: Proc. of the IEEE International Conference on Data Mining (ICDM). pp. 225–232 (2004)
27. Wang, X.J., Ma, W.Y., Xue, G.R., Li, X.: Multi-model similarity propagation and its application for web image retrieval. In: Proc. of the 12th Annual ACM International Conference on Multimedia (MM). pp. 944–951 (2004)