A STATISTICAL FRAMEWORK FOR POSITIVE DATA CLUSTERING WITH FEATURE SELECTION: APPLICATION TO OBJECT DETECTION

Mohamed Al Mashrgy$^{(1)}$, *Nizar Bouguila*$^{(1)}$, *Khalid Daoudi*$^{(2)}$

$^{(1)}$Concordia University, Montreal, Canada
m_almash@encs.concordia.ca, nizar.bouguila@concordia.ca
$^{(2)}$ INRIA Bordeaux Sud Ouest, France
khalid.daoudi@inria.fr

ABSTRACT

In this paper, we concern ourselves with the problem of simultaneous positive data clustering and feature selection. We propose a statistical framework based on finite mixture models of generalized inverted Dirichlet (GID) distributions. The GID offers a more practical and flexible alternative to the inverted Dirichlet which has a very restrictive covariance structure. For learning the parameters of the resulting mixture, we propose an approach based on minimum message length (MML) criterion. We use synthetic data and real data generated from a challenging application that concerns objects detection to demonstrate the feasibility and advantages of the proposed method.

Index Terms— Positive data, feature selection, clustering, mixture models, GID, MML, object detection.

1. INTRODUCTION

The clustering of real data is a challenging problem. This is especially true in most signal processing, computer vision and pattern recognition applications where the extracted feature spaces are known to be high-dimensional, complex, and noisy [1]. This kind of data poses different challenges for clustering algorithms since not all features are important. Indeed, some of the features may be irrelevant and then misguide the clustering process. Over the years, many feature selection techniques have bee proposed to handle high-dimensional vectors. In this paper, we are interested in feature selection when finite mixture models are used for clustering.
Finite mixture modeling has been the subject of much research in recent times and the reader is referred to [2]. Several research efforts have been devoted to the integration of feature selection into finite Gaussian mixtures (see, for instance, [3, 4]). It is well-known, however, that the Gaussian choice is not realistic in the majority of signal and image processing problems. For instance, it is well-known that the statistics of natural images are not Gaussian. Moreover, a crucial assumption in the majority of these previous works is that the features are supposed to be independent (e.g. treating the multivariate Gaussian as a product of univariate Gaussians [3]) which is not generally the case. Our work in this paper is closely related to these approaches, but considers positive data which are naturally generated by several tasks and for which the Gaussian assumption has been shown to be inappropriate [5]. In particular, we propose a feature selection model that builds on finite GID mixtures. The consideration of the GID has two main important advantages. First, the GID has a more general covariance structure than the inverted Dirichlet [6, 5] which has a very restrictive positive covariance matrix. Second, the mathematical properties of the GID distribution allow the representation of GID samples in a transformed space in which features are independent and follow inverted Beta distributions. Thus, as opposed to earlier works, the conditional independence assumption among features commonly used by researchers becomes a fact in our case. The resulting model is learned via an expectation maximization (EM) algorithm to minimize a message length objective for simultaneous parameter estimation, model order selection, and feature weighting.
The rest of this papers is organized as follows. In Section 2, we present our model and details the developed learning approach. Section 3 presents the experimental results to show the strengths of the proposed approach.

2. THE PROPOSED MODEL

2.1. The Generalized Inverted Dirichlet Finite Mixture

Let us consider a data set $\mathcal{Y} = (\vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_N)$ of N D-dimensional positive vectors, where $\vec{Y}_i = (Y_{i1}, \ldots, Y_{iD}), i = 1, \ldots, N$. We assume that $\vec{Y}_i$ follows a mixture of M GID distributions: $p(\vec{Y}_i|\Theta) = \sum_{j=1}^{M} \pi_j p(\vec{Y}_i|\vec{\Theta}_j)$, where $p(\vec{Y}_i|\vec{\Theta}_j)$ is a GID distribution [7]:

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

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

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 $\vec{\Theta}_j = (\alpha_{j1}, \beta_{j1}, \alpha_{j2}, \beta_{j2}, \ldots, \alpha_{jD}, \beta_{jD})$ is the set of parameters defining the jth component, and π_j is the mixing weight of the jth cluster.

In mixture-based clustering [2], each vector $\vec{Y}_i$ is assigned to all classes with different posterior probabilities $p(j|\vec{Y}_i) \propto \pi_j p(\vec{Y}_i|\vec{\Theta}_j)$. It is possible to show that the properties of the GID distribution allows the factorization of the posterior probabilities as: $p(j|\vec{Y}_i) \propto \pi_j \prod_{l=1}^{D} p_{ib}(X_{il}|\theta_{jl})$, 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 $\mathcal{X} = (\vec{X}_1, \ldots, \vec{X}_N)$ described by the following mixture model with conditionally independent features:

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} p_{ib}(X_{il}|\theta_{jl}) \qquad (2)$$

Let us introduce latent variables $\vec{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iM}), i = 1, \ldots, N$ where $Z_{ij} \in 0, 1, \sum_{j=1}^{M} Z_{ij} = 1$ and $Z_{ij} = 1$ implies that $\vec{X}_i$ is in the jth component. Therefore, the likelihood of $\vec{X}_i$ given the class label $\vec{Z}_i$ is given by:

$$p(\vec{X}_i|\vec{Z}_i, \Theta) = \prod_{j=1}^{M} \Big(\prod_{l=1}^{D} p_{ib}(X_{il}|\theta_{jl}) \Big)^{Z_{ij}} \qquad (3)$$

According to Eq. 2, all features X_{il} are considered equally useful for the learning process which is naturally not true since some features might be irrelevant and then deteriorate the clustering process. An important challenge is then the selection of relevant features to improve modeling capabilities. In learning systems, the contribution of a given feature is normally not obvious. Indeed, some features may be relevant and then useful for the learning process and others may be irrelevant and then compromise the final model. In order to integrate feature selection in our mixture model, we consider the approach proposed in [3] in the case of finite Gaussian mixtures and extended in [8] for proportional data. The main idea is to suppose that each given feature is generated from a mixture of two univariate distributions. The first distribution is assumed to be generate relevant features and is different for each cluster and the second one is assumed to generate irrelevant features and is common to all clusters. This idea can be formulated as following in our case:

$$\begin{aligned} p(\vec{X}_i|\Theta^*) &= \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} [\rho_l p_{ib}(X_{il}|\theta_{jl}) \\ &+ (1 - \rho_l) p_{ib}(X_{il}|\xi_l)] \qquad (4) \end{aligned}$$

where $\Theta^* = \{\{\theta_{jl}\}, \{\rho_l\}, \{\xi_l\}\}$ is the set of all parameters representing our unsupervised feature selection model, ρ_l represents the probability that feature X_{il} is relevant for clustering, and $p_{ib}(X_{il}|\xi_l)$ is an inverted Beta distribution, with parameters $\xi_l = (\alpha_l, \beta_l)$ common to all clusters and supposed to generate irrelevant features.

2.2. MML-Based Learning

Our learning procedure will be based on the optimization of a message length objective which shall allow simultaneous parameters estimation and model selection (i.e. determination of the number of components M). The objective function to minimize is given by [9]:

$$\begin{aligned} MessLen(M) &\simeq -\log p(\Theta^*) + \frac{1}{2}\log|I(\Theta^*)| \\ &+ \frac{c}{2}(1 + \log\frac{1}{12}) - \log p(\mathcal{X}|\Theta^*) \qquad (5) \end{aligned}$$

where $p(\Theta^*)$ is a prior distribution over model's parameters, $|I(\Theta^*)|$ is the determinant of the expected Fisher information matrix, $p(\mathcal{X}|\Theta^*)$ is the model's likelihood function, c is the number of free parameters being estimated which is in our case $c = M - 1 + 3D + 2DM$.

Since $\vec{\pi} = (\pi_1, \ldots, \pi_M)$ and $\vec{\rho_l} = (\rho_{l1}, \rho_{12}), \rho_{l2} = 1 - \rho_{l1}$ are actually proportional vectors, we choose as priors for them Dirichlet distributions with hyperparameters set to 0.5:

$$p(\vec{\pi}) \propto \prod_{j=1}^{M} \pi^{-\frac{1}{2}} \qquad p(\vec{\rho_l}) \propto \rho_{l1}^{-\frac{1}{2}} \rho_{l2}^{-\frac{1}{2}} \qquad (6)$$

As for $\alpha_{jl}^{\theta}, \beta_{jl}^{\theta}, \alpha_l^{\xi}$, and β_l^{ξ}, we consider:

$$\begin{aligned} p\big(\alpha_{jl}^{\theta}, \beta_{jl}^{\theta}\big) &= \frac{\hat{A}^{\theta}}{\alpha_{jl}^{\theta}\beta_{jl}^{\theta}(\hat{A}^{\theta} - \alpha_{jl}^{\theta} - \beta_{jl}^{\theta})} \\ p\big(\alpha_l^{\xi}, \beta_l^{\xi}\big) &= \frac{\hat{A}^{\xi}}{\alpha_l^{\xi}\beta_l^{\xi}(\hat{A}^{\xi} - \alpha_l^{\xi} - \beta_l^{\xi})} \end{aligned} \qquad (7)$$

where the superscript θ and ξ denote the relevant and the irrelevant components, respectively, and $\hat{A} = e^{6\frac{(\hat{\alpha}+\hat{\beta})^2}{\hat{\alpha}\hat{\beta}}}$. The expected Fisher information can be approximated by taking the determinant of the second derivative of the negative log likelihood of the data with respect to all model's parameters:

$$|I(\vec{\pi})| = N^{M-1} \prod_{j=1}^{M} \pi_j^{-1}, |I(\vec{\rho})| = N\rho_{l1}^{-1}\rho_{l2}^{-1} \qquad (8)$$

$$\begin{aligned} |I(\theta_{jl})| = N_{jl}^2 \big|\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \big(\Psi'(\alpha_{jl} + \beta_{jl}) \\ (\Psi'(\alpha_{jl}) + \Psi'(\beta_{jl}))\big)\big| \qquad (9) \end{aligned}$$

$$\begin{aligned} |I(\xi_l)| = N_{jl}^2 \big|\Psi'(\alpha_l)\Psi'(\beta_l) - \big(\Psi'(\alpha_l + \beta_l) \\ (\Psi'(\alpha) + \Psi'(\beta_l))\big)\big| \qquad (10) \end{aligned}$$

where $\Psi'(.)$ is the second derivative of the logarithm of the Gamma function. By substituting Eqs. 4, 6, 7, 8, 9 and 10 into Eq. 5, we obtain the message length of our model.

2.3. Estimation And Selection Algorithm

In this section, we summarize our EM algorithm for minimizing the message length of our model. The minimization is done under the constraints $0 < \pi_j < 1$, $\sum_{j=1}^{M} \pi_j = 1$ and $\rho_{l1} + \rho_{l2} = 1$. To satisfy these constraints, we introduce Lagrange multipliers Λ and $\lambda_l^{\rho}, l = 1, \dots, D$:

$$\begin{aligned} S(\Theta, \mathcal{X}) &= -MessLen(M) + \Lambda(1 - \sum_{j=1}^{M} \pi_j) \\ &+ \sum_{l=1}^{D} \lambda_l^{\rho}(1 - \rho_{l1} - \rho_{l2}) \end{aligned} \tag{11}$$

The minimization of the previous function gives us the following updating equations in the M-step:

$$\pi_j \propto max\big(\sum_{i=1}^{N} p(j|\vec{X}_i) - D, 0\big) \tag{12}$$

$$\frac{1}{\rho_{l1}} = 1 + \frac{\max\Big(\sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{X}_i)\frac{\rho_{l2}p_{ib}(X_{il}|\xi_l)}{\rho_{l1}p_{ib}(X_{il}|\theta_{jl})+\rho_{l2}p_{ib}(X_{il}|\xi_l)} - 1, 0\Big)}{\max\Big(\sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{X}_i)\frac{\rho_{l1}p_{ib}(X_{il}|\theta_{jl})}{\rho_{l1}p_{ib}(X_{il}|\theta_{jl})+\rho_{l2}p_{ib}(X_{il}|\xi_l)} - M, 0\Big)} \tag{13}$$

where $p(j|\vec{X}_i)$ is the posterior:

$$p(j|\vec{X}_i) \propto \pi_j \prod_{l=1}^{D} [\rho_l p_{ib}(X_{il}|\theta_{jl}) + (1 - \rho_l) p_{ib}(X_{il}|\xi_l)] \tag{14}$$

In order to estimate the θ_{jl} and ξ_l parameters, we will use Fisher's scoring methods which is given as following in the case of θ_{jl} (similar formula is used to update ξ_l):

$$\hat{\theta}_{jl}^{t+1} = \hat{\theta}_{jl}^{t} - \big(\frac{\partial^2}{\partial \theta_{jl}^2} S(\Theta, \mathcal{X})\big)^{-1}_{\hat{\theta}_{jl}^{t}} \times \big(\frac{\partial}{\partial \theta_{jl}} S(\Theta, \mathcal{X})\big)_{\theta_{jl} = \hat{\theta}_{jl}^{t}} \tag{15}$$

The complete learning process is summarized in Algorithm 1.

3. EXPERIMENTAL RESULTS

This section is dedicated to show the merits of our model and its ability to perform simultaneous clustering and feature selection. In particular, we perform extensive experiments involving synthetic data sets and a real-world challenging application namely object detection. For all conducted experiments, we set M_{max} and M_{min} to 15 and 2, respectively.

Algorithm 1

```
Input: $D$-dimensional data set $\mathcal{X}$, $M_{max}$, $M_{min}$.
Output: $M^*$ and $\Theta^*$.
initialization: $M = M_{max}$, $\rho_{l1} = \rho_{l2} = 0.5$ and apply Fuzzy C-Means to have an initial partition of the data.
While $M \geq M_{min}$ do
repeat
    E-step:
    Compute $p(j|\vec{X}_i)$ using Eq. 14.
    M-step:
    Update $\pi_j$ and $\rho_{l1}$ using Eqs. 12 and 13.
    Update $\xi_l$ and $\theta_{jl}$ using Fisher scoring.
    If $\pi_j = 0$ then prune the $jth$ component.
    If $\rho_l = 0$ then prune the $lth$ feature.
    Else if $\rho_{l2} = 0$ then prune the irrelevant feature.
until Convergence criterion is reached.
Record $\Theta$, $M$, and MML of the model.
Remove the $jth$ component $\theta_{jl}$ with the lowest mixing weight $\pi_j$.
Endwhile
Return $\Theta^*$, $M^*$ with the lowest MML.
```

3.1. Evaluation on Synthetic Data

In this experiment we evaluate the performance of the proposed model using 3-dimensional synthetic data sets which are generated from 2, 3, and 4 components GID mixtures. The parameters used to generate these datasets are given in Table 1. Moreover, eight "noisy" features, generated from one inverted Beta with parameters $\alpha = 3$ and $\beta = 15$, are appended to the datasets which increase the dimensionality of the data to 11 dimensions. For all three data sets, our algorithm selected the exact number of clusters with classification accuracies of 98.89%, 97.78% and 95.91%, respectively. The saliencies of all the 11 features, computed automatically by our algorithm, for each of the synthetic datasets are shown in figure 1. According to this figure, it is obvious that high relevancies have been assigned to features 1, 2, and 3 which is consistent with the ground-truth. We conclude that, for syn-

Table 1. Parameters used to generate the synthetic data sets. n_j represents the number of elements in cluster j.

	n_j	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	α_{j3}	β_{j3}
Dataset 1	300	1	40	30	33	46	18	40
	300	2	30	44	25	40	35	22
Dataset 2	300	1	30	44	25	40	35	22
	300	2	18	35	43	25	21	14
	300	3	40	28	33	46	18	40
Dataset 3	300	1	16	28	17	32	21	41
	300	2	18	35	43	25	21	14
	300	3	40	28	33	46	18	40
	300	4	30	44	25	40	35	22

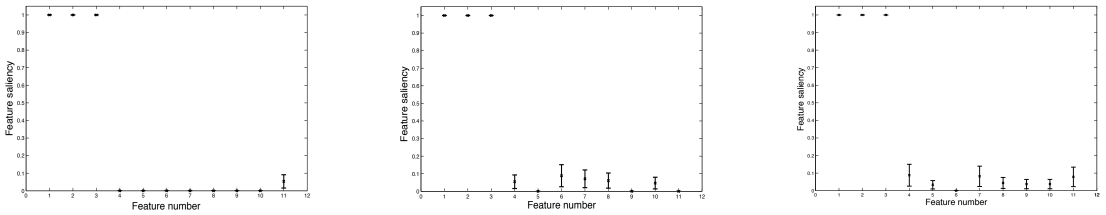

Fig. 1. Feature salience for all synthetic datasets.

thetic datasets, the proposed algorithm was able to successfully select the optimal number of components and to assigns features saliencies correctly.

3.2. Object Detection

With advances in multimedia technology, images are becoming available at an explosive rate. A crucial problem is then how to efficiently organize and index those multimedia data. A lot of approaches and techniques have been proposed in the past [10, 11]. Although different, all these approaches agree that an important step for efficient organization is object detection. Although object detection has been the subject of much research in the past, the problem is still challenging. In this section, we present the results of applying our statistical model on two common widely used tasks namely car and human detection. An important part of the object detection problem is feature extraction. Many visual descriptors have been proposed in the past (see, for instance, [12]). Here, we use local Histogram of Oriented Gradient (HOG) descriptor which generate positive features and which has been shown to be efficient and convenient for the object detection problem [13]. Experiments are conducted by considering three windows for the HOG descriptor which allows to represent each image by 81-dimensional vector of features. We conduct our experiments by considering the GID mixture with feature selection (GIDFS) and without feature selection (GIDnoFS). Moreover, we compare the obtained results with those obtained when considering finite Gaussian mixture model with (GMMFS) and without feature selection (GMMnoFS).

3.2.1. Car Detection

The dataset that we consider here contains images of cars side views which was collected at UIUC [1]. The dataset consists of 1050 images (550 car and 500 non-car images). Figures 2 and 3 show examples of images from this dataset. The first 100 images from both car and non-car images are used for training and the rest for testing. Table 2 shows the detection accuracies when both Gaussian and GID mixtures are considered with and without feature selectionm. According to this table, it is clear that the GID mixture outperforms the Gaussian mixture and that feature selection improves the results. Besides, Figure 4 shows the obtained features saliencies by both mixtures and displays clearly the fact that the different features have different saliencies.

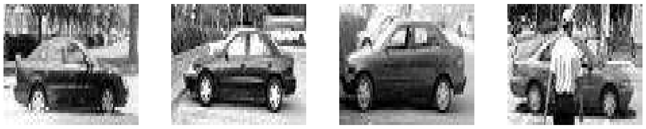

Fig. 2. Examples of car images.

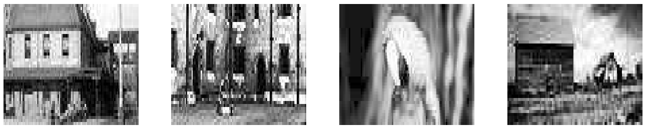

Fig. 3. Examples of non-car images.

Table 2. Car detection accuracies when different approaches are considered.

Model	GIDFS	GIDnoFS	GMMFS	GMMnoFS
Accuracy	83.69%	80.76%	74.00%	72.77%

3.2.2. Human detection

Another challenging task that we consider here is human detection. We consider the INRIA Static Person dataset [2] to evaluate the proposed model. The data consists of both positive (containing humans) and negative examples (images that do not contain humans). 400 images are used for training (200 positive examples and 200 negative ones). On the other hand, the testing set consists of 741 images, 288 of them are positive examples and the remaining 453 are negative examples. Figures 5 and 6 show samples of positive and negative examples, respectively. Table 3 shows the classification accuracy for the INRIA dataset. According to this table, its clear that the proposed model outperforms GMM and again feature selection improves the detection results.

[1] http://cogcomp.cs.illinois.edu/Data/Car/

[2] http://pascal.inrialpes.fr/data/human/

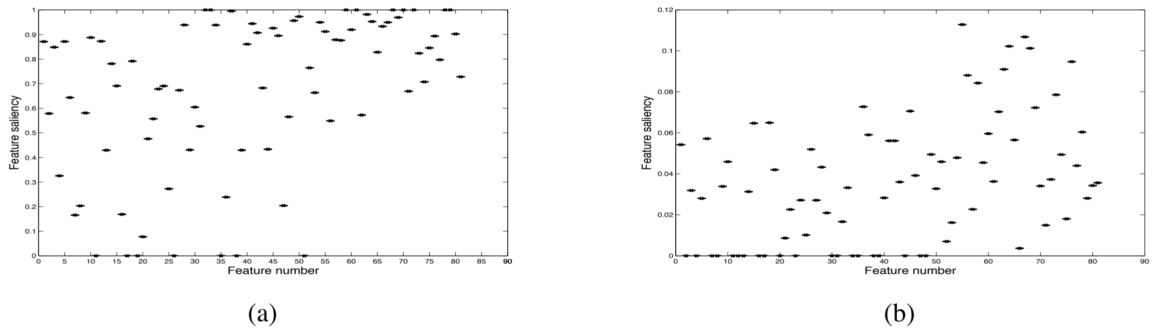

Fig. 4. Features saliencies in the case of the car detection application when using: (a) The GID mixture and (b) The Gaussian mixture.

Fig. 5. Examples of images containing humans.

Fig. 6. Examples of negative images used for human detection task.

Table 3. Human detection accuracies when different approaches are considered.

Model	GIDFS	GIDnoFS	GMMFS	GMMnoFS
Accuracy	68.55%	65.56%	57.35%	53.00%

4. CONCLUSION

We have introduced an approach for simultaneous clustering and feature selection in the case of positive data. Our approach is based on GID mixture models that have several interesting properties. We have developed a principled approach based on MML for the learning of the resulting model. The effectiveness and efficiency the proposed statistical framework was shown experimentally through quantitative evaluation in the case of object detection.

5. REFERENCES

[1] N.A. Gumerov C. Yang, R. Duraiswami and L. Davis, "Improved fast gauss transform and efficient kernel density estimation," in *Proc. of the IEEE 9th International Conference on Computer Vision (ICCV)*, 2003, pp. 664 –671.

[2] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley-Interscience, 2000.

[3] M. H. C. Law, M. A. T. Figueiredo and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.

[4] M. W. Graham and D. J. Miller, "Unsupervised Learning of Parsimonious Mixtures on Large Spaces with Integrated Feature and Component Selection," *IEEE Trans. on Signal Processing*, vol. 54, no. 4, pp. 1289–1303, 2006.

[5] T. Bdiri and N. Bouguila, "Positive vectors clustering using inverted dirichlet finite mixture models," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1869–1882, 2012.

[6] T. Bdiri and N. Bouguila, "Learning inverted dirichlet mixtures for positive data clustering," in *Proc. of the International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC)*. 2011, vol. 6743 of *Lecture Notes in Computer Science*, pp. 265–272, Springer.

[7] G. S. Lingappaiah, "On the generalised inverted dirichlet distribution," *Demostratio Mathematica*, vol. 9, no. 3, pp. 423–433, 1976.

[8] S. Boutemedjet, N. Bouguila and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429–1443, 2009.

[9] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Springer-Verlag New York, Inc., 2005.

[10] N. Bouguila, "Spatial color image databases summarization," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, pp. 953–956.

[11] N. Bouguila and K. Daoudi, "Learning concepts from visual scenes using a binary probabilistic model," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing, (MMSP)*, 2009, pp. 1–5.

[12] A. Rocha and S. Goldenstein, "PR: More than meets the eye," in *Proc. of the IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 2, pp. 886–893.