A Hierarchical Infinite Generalized Dirichlet Mixture Model with Feature Selection

Wentao Fan[1], Hassen Sallay[2], Nizar Bouguila[3], and Sami Bourouis[4]

[1] Department of Computer Science and Technology, Huaqiao University, China
fwt@hqu.edu.cn
[2] Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
hmsallay@imamu.edu.sa
[3] Concordia University, Montreal, QC, Canada
nizar.bouguila@concordia.ca
[4] Taif University, Taif, Saudi Arabia
s.bourouis@tu.edu.sa

Abstract. We propose a nonparametric Bayesian approach, based on hierarchical Dirichlet processes and generalized Dirichlet distributions, for simultaneous clustering and feature selection. The resulting statistical model is learned within a variational framework that we have developed. The merits of the developed model are shown via extensive simulations and experiments when applied to the challenging problem of images categorization.

Keywords: Clustering, mixture model, feature selection, generalized Dirichlet, variational inference, image databases.

1 Introduction

Dirichlet process (DP) has became as one of the most popular nonparametric Bayesian techniques [12,10]. It can be viewed as a stochastic process and can be considered as distribution over distributions, also. Recently, hierarchical DP [20,19] has been developed as a hierarchical nonparametric Bayesian model and has shown promising results to the problem of model-based clustering of grouped data with sharing clusters. It is built on the DP and involves Bayesian hierarchy where the base measure for a DP is itself distributed according to a DP. The hierarchical DP framework is particularly useful in problems involving the modeling of grouped data where observations are organized into groups, and allow these groups to remain statistically linked by sharing mixture components [20,19]. As several other statistical approaches, existing hierarchical DP-based models considers the Gaussian assumption. Unlike, these existing works, in this paper, we focus on a specific form of hierarchical DP mixture model where each observation within a group is drawn from a mixture of generalized Dirichlet (GD) distributions. We are mainly motivated by the fact that the generalized Dirichlet distribution has been shown to be efficient in modeling high-dimensional proportional data (i.e. normalized histograms) in a variety of applications from different disciplines (e.g. computer vision, data mining, pattern recognition) [5,6,8,9].

A. Bouchachia (Ed.): ICAIS 2014, LNAI 8779, pp. 1–10, 2014.

Having the hierarchical infinite generalized Dirichlet mixture model in hand, a principled variational approach is developed to learn its parameters. To validate the overall statistical framework a challenging application, which has attracted the attention of the computer vision community recently [2,16], namely the classification of images containing categories of animals is considered. The rest of this paper is organized as follows. In the next section, we propose our model. Section 3 develops our variational learning procedure. Section 4 is devoted to the experimental results. Finally, we conclude the paper in Section 5.

2 The Model

2.1 The Hierarchical Dirichlet Process Mixture Model

Let H be a distribution over some probability space Θ and γ be a positive real number, then a random distribution G follows a DP with a base distribution H and concentration parameter γ, denoted as $G \sim \mathrm{DP}(\gamma, H)$, if

$$(G(A_1), \ldots, G(A_t)) \sim \mathrm{Dir}(\gamma H(A_1), \ldots, \gamma H(A_t)) \tag{1}$$

where $(A_1, \ldots, A_t)$ is the set of finite partitions of Θ, and Dir is a finite-dimensional Dirichlet. Now, let us introduce the general setting of a two-level hierarchical Dirichlet process: Assume that we have a grouped data set, in which each group is associated with an infinite mixture model (a DP G_j). This indexed set of DPs $\{G_j\}$ shares a base distribution G_0, which is itself distributed as a DP:

$$G_0 \sim \mathrm{DP}(\gamma, H) \quad G_j \sim \mathrm{DP}(\lambda, G_0) \qquad \text{for each } j,\ j \in \{1, \ldots, M\} \tag{2}$$

where j is an index for each group of data. In this work, we represent a hierarchical DP in a more intuitive and straightforward form through two stick-breaking constructions which involves a global-level and a group-level construction [18,11]. In the global-level construction, the global measure G_0 is distributed according to a $\mathrm{DP}(\gamma, H)$ and can be described using a stick-breaking representation

$$\psi'_k \sim \mathrm{Beta}(1, \gamma) \quad \Omega_k \sim H \quad \psi_k = \psi'_k \prod_{s=1}^{k-1}(1 - \psi'_s) \quad G_0 = \sum_{k=1}^{\infty} \psi_k \delta_{\Omega_k} \tag{3}$$

where $\{\Omega_k\}$ is a set of independent random variables distributed according to H, δ_{Ω_k} is an atom at Ω_k. The random variables ψ_k are the stick-breaking weights that satisfy $\sum_{k=1}^{\infty} \psi_k = 1$, and are obtained by recursively breaking a unit length stick into an infinite number of pieces. In this work, we apply the conventional stick-breaking representation [21] to construct each group-level DP G_j

$$\pi'_{jt} \sim \mathrm{Beta}(1, \lambda) \quad \varpi_{jt} \sim G_0 \quad \pi_{jt} = \pi'_{jt} \prod_{s=1}^{t-1}(1 - \pi'_{js}) \quad G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\varpi_{jt}}$$

where $\delta_{\varpi_{jt}}$ are group-level atoms at ϖ_{jt}, $\{\pi_{jt}\}$ is a set of stick-breaking weights which satisfies $\sum_{t=1}^{\infty}\pi_{jt}=1$. Since ϖ_{jt} is distributed according to the base distribution G_0, it takes on the value Ω_k with probability ψ_k. This can also represented using a binary latent variable W_{jtk} as an indicator variable, such that $W_{jtk}\in\{0,1\}$, $W_{jtk}=1$ if ϖ_{jt} maps to the base-level atom Ω_k which is indexed by k; otherwise, $W_{jtk}=0$. Then, we can have $\varpi_{jt}=\Omega_k^{W_{jtk}}$. As a result, group-level atoms ϖ_{jt} do not need to be explicitly represented which further simplifies the inference process as it shall be clearer later. The indicator variable $\boldsymbol{W}=(W_{jt1},W_{jt2},\ldots)$ is distributed according to $\boldsymbol{\psi}$ in the form

$$p(\boldsymbol{W}|\boldsymbol{\psi})=\prod_{j=1}^{M}\prod_{t=1}^{\infty}\prod_{k=1}^{\infty}\psi_k^{W_{jtk}} \tag{4}$$

Since $\boldsymbol{\psi}$ is a function of $\boldsymbol{\psi}'$ according to a stick-breaking construction, then

$$p(\boldsymbol{W}|\boldsymbol{\psi}')=\prod_{j=1}^{M}\prod_{t=1}^{\infty}\prod_{k=1}^{\infty}\left[\psi_k'\prod_{s=1}^{k-1}(1-\psi_s')\right]^{W_{jtk}} \tag{5}$$

The prior distribution of $\boldsymbol{\psi}'$ is a specific Beta distribution

$$p(\boldsymbol{\psi}')=\prod_{k=1}^{\infty}\mathrm{Beta}(1,\gamma_k)=\prod_{k=1}^{\infty}\gamma_k(1-\psi_k')^{\gamma_k-1} \tag{6}$$

Let i index the observations within each group j, we assume that each θ_{ji} is a factor corresponding to an observation X_{ji}, and the factors $\boldsymbol{\theta}_j=(\theta_{j1},\theta_{j2},\ldots)$ are distributed according to DP G_j, one for each j: $\theta_{ji}|G_j\sim G_j, X_{ji}|\theta_{ji}\sim F(\theta_{ji})$, where $F(\theta_{ji})$ represents the distribution of the observation X_{ji} given θ_{ji}. According to Eq.(2), the base distribution H of G_0 provides the prior for θ_{ji}. This setting forms the definition of a *hierarchical DP mixture model*, where each group is associated with a mixture model, and the components are shared among these mixtures due to the sharing of atoms Ω_k among all G_j. Furthermore, since each θ_{ji} is distributed according to G_j, it takes the value ϖ_{jt} with probability π_{jt}. Next, we introduce a binary latent variable $Z_{jit}\in\{0,1\}$ as an indicator variable. That is, $Z_{jit}=1$ if θ_{ji} is associated with component t and maps to the group-level atom ϖ_{jt}; otherwise, $Z_{jit}=0$. Therefore, we have $\theta_{ji}=\varpi_{jt}^{Z_{jit}}$. Since ϖ_{jt} also maps to the global-level atom Ω_k, we then have $\theta_{ji}=\varpi_{jt}^{Z_{jit}}=\Omega_k^{W_{jtk}Z_{jit}}$. The indicator variable $\boldsymbol{Z}=(Z_{ji1},Z_{ji2},\ldots)$ is distributed according to $\boldsymbol{\pi}$ as

$$p(\boldsymbol{Z}|\boldsymbol{\pi})=\prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{t=1}^{\infty}\pi_{jt}^{Z_{jit}} \tag{7}$$

According to the stick-breaking construction, $\boldsymbol{\pi}$ is a function of $\boldsymbol{\pi}'$. We then have

$$p(\boldsymbol{Z}|\boldsymbol{\pi}')=\prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{t=1}^{\infty}[\pi_{jt}'\prod_{s=1}^{t-1}(1-\pi_{js}')]^{Z_{jit}} \tag{8}$$

$$p(\boldsymbol{\pi}') = \prod_{j=1}^{M}\prod_{t=1}^{\infty}\text{Beta}(1,\lambda_{jt}) = \prod_{j=1}^{M}\prod_{t=1}^{\infty}\lambda_{jt}(1-\pi'_{jt})^{\lambda_{jt}-1} \tag{9}$$

2.2 Hierarchical DP Mixture Model of GD Distributions with Feature Selection

Assume that we have a D-dimensional vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ drawn from a GD distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$ and $\boldsymbol{\beta}_j = (\beta_1, \ldots, \beta_D)$

$$\text{GD}(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{l=1}^{D}\frac{\Gamma(\alpha_l+\beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)}Y_l^{\alpha_l-1}\big(1-\sum_{f=1}^{l}Y_f\big)^{\gamma_l} \tag{10}$$

where $\sum_{l=1}^{D}Y_l < 1$ and $0 < Y_l < 1$ for $l = 1, \ldots, D$, $\alpha_l > 0$, $\beta_l > 0$, $\gamma_l = \beta_l - \alpha_{l+1} - \beta_{l+1}$ for $l = 1, \ldots, D-1$, $\gamma_D = \beta_D - 1$, and $\Gamma(\cdot)$ is the gamma function. Based on an interesting mathematical property of the GD distribution which is thoroughly discussed in [7], we can transform the original data point $\boldsymbol{Y}$ using a geometric transformation into another D-dimensional data point $\boldsymbol{X}$ with independent features in the form of $\text{GD}(\boldsymbol{X}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{l=1}^{D}\text{Beta}(X_l|\alpha_l,\beta_l)$, where $\boldsymbol{X} = (X_1, \ldots, X_D)$, $X_1 = Y_1$ and $X_l = Y_l/(1-\sum_{f=1}^{l-1}Y_f)$ for $l > 1$, and $\text{Beta}(X_l|\alpha_l,\beta_l)$ is a Beta distribution defined with parameters $\{\alpha_l,\beta_l\}$. Now let us consider a data set $\mathcal{X}$ that contains N random vectors separated into M groups, then each vector $\boldsymbol{X}_{ji} = (X_{ji1}, \ldots, X_{jiD})$ is represented in a D-dimensional space and is drawn from a hierarchical infinite GD mixture model.

In practice, not all the features are important, some of them may be irrelevant and may even degrade the clustering performance. Therefore, feature selection technique is important and is adopted here as a tool to chooses the "best" feature subset. The most common feature selection technique, in the context of unsupervised learning, defines an irrelevant feature as the one having a distribution independent from class labels [13]. Therefore, in our work the distribution of each feature X_l can be defined by

$$p(X_{jil}) = \text{Beta}(X_{jil}|\alpha_{kl},\beta_{kl})^{\phi_{jil}}\text{Beta}(X_{jil}|\alpha'_l,\beta'_l)^{1-\phi_{jil}} \tag{11}$$

where ϕ_{jil} is a binary latent variable represents the feature relevance indicator, such that $\phi_{jil} = 0$ denotes the feature l of group j is irrelevant (i.e. noise) and follows a Beta distribution: $\text{Beta}(X_{jil}|\alpha'_l,\beta'_l)$; otherwise, the feature X_{jil} is relevant. The prior distribution of $\boldsymbol{\phi}$ is defined as

$$p(\boldsymbol{\phi}|\boldsymbol{\epsilon}) = \prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{l=1}^{D}\epsilon_{l_1}^{\phi_{jil}}\epsilon_{l_2}^{1-\phi_{jil}} \tag{12}$$

where each ϕ_{jil} is a Bernoulli variable such that $p(\phi_{jil}=1) = \epsilon_{l_1}$ and $p(\phi_{jil} = 0) = \epsilon_{l_2}$. The vector $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_D)$ represents the features saliencies such that

$\boldsymbol{\epsilon}_l = (\epsilon_{l_1}, \epsilon_{l_2})$ and $\epsilon_{l_1} + \epsilon_{l_2} = 1$. Furthermore, a Dirichlet distribution is chosen over $\boldsymbol{\epsilon}$ as

$$p(\boldsymbol{\epsilon}) = \prod_{l=1}^{D} \mathrm{Dir}(\boldsymbol{\epsilon}_l|\boldsymbol{\xi}) = \prod_{l=1}^{D} \frac{\Gamma(\xi_1 + \xi_2)}{\Gamma(\xi_1)\Gamma(\xi_2)} \epsilon_{l_1}^{\xi_1 - 1} \epsilon_{l_2}^{\xi_2 - 1} \tag{13}$$

In the next step, we need to introduce Gamma prior distributions for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$.

3 Variational Model Learning

In order to simplify notations, we define $\Theta = (\Xi, \Lambda)$ as the set of latent and unknown random variables, where $\Xi = \{\boldsymbol{Z}, \boldsymbol{\phi}\}$ and $\Lambda = \{\boldsymbol{W}, \boldsymbol{\epsilon}, \boldsymbol{\psi}, \boldsymbol{\pi}', \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}', \boldsymbol{\beta}'\}$. Variational inference [1,3] is a deterministic approximation technique that is used to find tractable approximations for posteriors of a variety of statistical models. The goal is to find an approximation $Q(\Theta)$ to the true posterior distribution $p(\Theta|\mathcal{X})$ by maximizing the lower bound of $\ln p(\mathcal{X})$:

$$\mathcal{L}(Q) = \int Q(\Theta) \ln[p(\mathcal{X}, \Theta)/Q(\Theta)] d\Theta \tag{14}$$

In this work, we adopt factorial approximation which is commonly used in variational inference [3] to factorize $Q(\Theta)$ into disjoint tractable distributions. Moreover, we apply a truncation technique as in [4] to truncate the variational approximations of global- and group-level Dirichlet process at K and T. It is noteworthy that the truncation levels K and T are variational parameters which can be freely initialized and will be optimized automatically during the learning process. By using truncated stick-breaking and factorization, the approximated posterior distribution $q(\Theta)$ can be fully factorized into disjoint distributions as

$$q(\Theta) = q(\boldsymbol{Z})q(\boldsymbol{W})q(\boldsymbol{\phi})q(\boldsymbol{\pi}')q(\boldsymbol{\psi}')q(\boldsymbol{\alpha})q(\boldsymbol{\beta})q(\boldsymbol{\alpha}')q(\boldsymbol{\beta}')q(\boldsymbol{\epsilon}) \tag{15}$$

In our work, variational inference is performed based on a natural gradient method as introduced in [17]. The main idea is that, since our model has conjugate priors, the functional form of each factor in the variational posterior distribution is known. Therefore, the lower bound $\mathcal{L}(q)$ can be considered as a function of the parameters of these distributions by taking general parametric forms of these distributions. The optimization of variational factors is then obtained by maximizing the lower bound with respect to these parameters. In our case, the functional form of each variational factor is the same as its conjugate prior distribution, namely Discrete for $\boldsymbol{Z}$ and $\boldsymbol{W}$, Bernoulli for $\boldsymbol{\phi}$, Dirichlet for $\boldsymbol{\epsilon}$, Beta for $\boldsymbol{\psi}'$ and $\boldsymbol{\pi}'$, and Gamma for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$. Therefore, the parametric forms for these variational posterior distributions can be defined as the following

$$q(\boldsymbol{Z}) = \prod_{j=1}^{M} \prod_{i=1}^{N} \prod_{t=1}^{T} \rho_{jit}^{Z_{jit}} \quad q(\boldsymbol{W}) = \prod_{j=1}^{M} \prod_{t=1}^{T} \prod_{k=1}^{K} \vartheta_{jtk}^{W_{jtk}} \tag{16}$$

$$q(\boldsymbol{\phi}) = \prod_{j=1}^{M}\prod_{i=1}^{N}\prod_{l=1}^{D} \varphi_{jil}^{\phi_{jil}}(1-\varphi_{jil})^{1-\phi_{jil}} \quad q(\boldsymbol{\epsilon}) = \prod_{l=1}^{D} \mathrm{Dir}(\boldsymbol{\epsilon}_l|\boldsymbol{\xi}^*) \tag{17}$$

$$q(\boldsymbol{\pi}') = \prod_{j=1}^{M}\prod_{t=1}^{T} \mathrm{Beta}(\pi'_{jt}|a_{jt}, b_{jt}) \quad q(\boldsymbol{\psi}') = \prod_{k=1}^{K} \mathrm{Beta}(\psi'_k|c_k, d_k) \tag{18}$$

$$q(\boldsymbol{\alpha}) = \prod_{k=1}^{K}\prod_{l=1}^{D} \mathcal{G}(\alpha_{kl}|\tilde{u}_{kl}, \tilde{v}_{kl}) \quad q(\boldsymbol{\beta}) = \prod_{k=1}^{K}\prod_{l=1}^{D} \mathcal{G}(\beta_{kl}|\tilde{g}_{kl}, \tilde{h}_{kl}) \tag{19}$$

$$q(\boldsymbol{\alpha}') = \prod_{l=1}^{D} \mathcal{G}(\alpha'_l|\tilde{u}'_l, \tilde{v}'_l) \quad q(\boldsymbol{\beta}') = \prod_{l=1}^{D} \mathcal{G}(\beta'_l|\tilde{g}'_l, \tilde{h}'_l) \tag{20}$$

Consequently, the parameterized lower bound $\mathcal{L}(q)$ can be obtained by substituting Eqs.(16)$\sim$(20) into Eq.(14). Maximizing this bound with respect to these parameters then gives the required re-estimation equations. The variational inference for our model can be performed then as an EM-like algorithm.

4 Experimental Results

In this part, we validate the proposed hierarchical infinite GD mixture model with feature selection (referred to as *HInGDFs*) through a real-world application concerning images categorization. In our case, we concentrate on the problem of discriminating different images of breeds of cats and dogs, which is a specific type of object categorization. This problem is extremely challenging since cats and dogs are highly deformable and different breeds may differ only by a few subtle phenotypic details [15]. In our experiments, we initialize the global truncation level K to 800, and the group truncation level T to 100. The hyperparameters λ_{jt} and γ_k are initialized to 0.05. The initial values of hyperparameters u_{kl}, g_{kl}, u'_l, v'_l for the conjugate Beta priors are set to 0.1, and v_{kl}, h_{kl}, v'_l, h'_l are set to 0.01. The hyperparameters ξ_1 and ξ_2 of the feature saliency are both initialized to 0.5. These specific choices were found convenient according to our experiments.

4.1 Experimental Setting

We perform the categorization of cats and dogs using the proposed *HInGDFs* model and the bag-of-visual words representation. Our methodology are summarized as follows: First, we extract the 128-dimensional scale-invariant feature transform (SIFT) [14] descriptors from each image using the Difference-of-Gaussians (DoG) interest point detectors and then normalized. Then, these extracted SIFT features are modeled using our *HInGDFs*. Specifically, each image $\mathcal{I}_j$ is considered as a "group" and is therefore associated with a Dirichlet process mixture (infinite mixture) model G_j. Thus, each extracted SIFT feature vector X_{ji} of the image $\mathcal{I}_j$ is supposed to be drawn from the infinite mixture

model G_j, where the mixture components of G_j can be considered as "visual words". A global vocabulary is constructed and is shared among all groups (images) through the common global infinite mixture model G_0 of our hierarchical model. This setting matches the desired design of a hierarchical Dirichlet process mixture model. It is noteworthy that an important step in image categorization approaches with bag-of-visual words representation is the construction of visual vocabulary. Nevertheless, most of the previously invented approaches have to apply a separate vector quantization algorithm (such as K-means) to build a visual vocabulary, where the vocabulary size is normally manually selected. In our approach, the construction of the visual vocabulary is part of our mixture framework, and therefore the size of the vocabulary (i.e., the number of mixture components in the global-level mixture model) can be automatically inferred from the data thanks to the property of nonparametric Bayesian model. Since the goal of our experiment is to determine which image category (breeds of cats and dogs) that a testing image $\mathcal{I}_j$ belongs to, we also need to introduce an indicator variable B_{jm} associated with each image (or group) in our hierarchical Dirichlet process mixture framework. B_{jm} means image $\mathcal{I}_j$ is from category m and is drawn from another infinite mixture model which is truncated at level J. This means that we need to add a new level of hierarchy to our hierarchical infinite mixture model with a sharing vocabulary among all image categories. In this experiment, we truncate J to 50 and initialize the hyperparameter of the mixing probability of B_{jm} to 0.05. Finally, we assign a testing image to a given category according to Bayes' decision rule.

4.2 Data Set

In this work, we evaluate the effectiveness of the proposed approach for categorizing cats and dogs using a publicly available database namely the Oxford-IIIT Pet database [15][1]. It contains 7,349 images of cats and dogs, and is composed of 12 different breeds of cats and 25 different breeds of dogs. Each of these breeds contains about 200 images. Some sample images of this database are displayed in Fig. 1 (cats) and Fig. 2 (dogs). This database is randomly divided into two partitions: one for training (to learn the model and build the visual vocabulary), the other one for testing.

Results. In our experiments, we demonstrate the advantages of the proposed *HInGDFs* approach by comparing its performance with two other mixture models including the hierarchical infinite GD mixture model without feature selection (*HInGD*) and the hierarchical infinite Gaussian mixture model with feature selection (*HInGFs*). To make a fair comparison, all of these models are learned using variational inference and we evaluated the categorization performance by running the approach 30 times. In the first part of our experiments, we measure the performance of our approach for categorizing cats (12 breeds) and dogs (25 breeds), respectively. The average categorization performances of our approach

[1] Database available at: http://www.robots.ox.ac.uk/∼vgg/data/pets/

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 1. Sample cat images from the Oxford-IIIT Pet database. (a) Abyssinian, (b) Bengal, (c) Birman, (d) Bombay, (e) British Shorthair, (f) Egyptian Mau, (g) Persian, (h) Ragdoll.

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Sample dog images from the Oxford-IIIT Pet database. (a) Basset hound, (b) Boxer, (c) Chihuahua, (d) Havanese, (e) Keeshond, (f) Pug, (g) Samyod, (h) Shiba inu.

and the two other tested approaches are shown in Table 1. As we can see from this table, our approach (*HInGDFs*) provided the highest categorization accuracy among all tested approaches. Moreover, *HInGDFs* outperformed *HInGD* and *HInGFs* which demonstrates the advantage of incorporating a feature selection scheme into our framework and verifies that the GD mixture model has better modeling capability for proportional data than Gaussian. Next, we have evaluated all approaches for the whole Oxford-IIIT Pet database (i.e., we do not separate cat and dog images). The corresponding results are shown in Table 1. Based on the obtained results, the proposed approach provides again the best categorization accuracy rate (42.73%) as compared toHInGD (39.58%) and

Table 1. The average categorization performance (%) and the standard deviation obtained over 30 runs using different approaches. The values in parenthesis are the standard deviations of the corresponding quantities.

Method	Cats	Dogs	Both
HInGDFs	56.38 (0.95)	44.23 (1.21)	42.94 (1.05)
HInGD	52.92 (1.46)	40.86 (1.07)	39.58 (1.19)
HInGFs	48.65 (1.39)	37.52 (1.13)	35.27 (0.93)

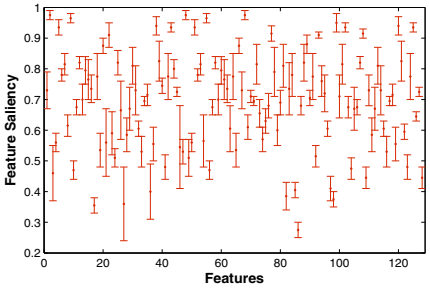

Fig. 3. Feature saliencies of different features calculated by *HInGDFs*

HInGFs (35.27%) approaches. Furthermore, the corresponding feature saliencies of the 128-dimensional SIFT vectors obtained by the proposed are illustrated in Fig. 3. According to this figure, it is clear that the different features do not contribute equally in the categorization, since they have different relevance degrees.

5 Conclusion

In this paper, we have developed a nonparametric Bayesian approach for data modeling, classification, and feature selection, using both hierarchical DP and generalized Dirichlet distributions which allows to model data without the need to define, a priori, the complexity of the entire model. In order to learn the parameters of the proposed model we have derived an inference procedure that relies on variational Bayes. The validation of the model has been based on a challenging application namely images categorization. We are currently working on the extension of our learning framework to online settings to handle the problem of dynamic data modeling.

Acknowledgments. The second author would like to thank King Abdulaziz City for Science and Technology (KACST), Kingdom of Saudi Arabia, for their funding support under grant number 11-INF1787-08.

References

1. Attias, H.: A variational Bayes framework for graphical models. In: Proc. of Advances in Neural Information Processing Systems (NIPS), pp. 209–215 (1999)
2. Berg, T., Forsyth, D.: Animals on the web. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1463–1470 (2006)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
4. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. Bayesian Analysis 1, 121–144 (2005)

5. Bouguila, N., Ziou, D.: A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture. IEEE Transactions on Image Processing 15(9), 2657–2668 (2006)
6. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(10), 1716–1731 (2007)
7. Boutemedjet, S., Bouguila, N., Ziou, D.: A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(8), 1429–1443 (2009)
8. Fan, W., Bouguila, N.: Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection. Pattern Recognition 46(10), 2754–2769 (2013)
9. Fan, W., Bouguila, N., Ziou, D.: Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference. IEEE Transactions on Knowlege and Data Engineering 25(7), 1670–1685 (2013)
10. Ferguson, T.S.: Bayesian Density Estimation by Mixtures of Normal Distributions. Recent Advances in Statistics 24, 287–302 (1983)
11. Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96, 161–173 (2001)
12. Korwar, R.M., Hollander, M.: Contributions to the theory of Dirichlet processes. The Annals of Probability 1, 705–711 (1973)
13. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9), 1154–1166 (2004)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
15. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3498–3505 (2012)
16. Ramanan, D., Forsyth, D.A.: Using temporal coherence to build models of animals. In: Proc. of the 9th IEEE International Conference on Computer Vision (ICCV), pp. 338–345. IEEE Computer Society (2003)
17. Sato, M.: Online model selection based on the variational Bayes. Neural Computation 13, 1649–1681 (2001)
18. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650 (1994)
19. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian Nonparametric Models with Applications. In: Hjort, N., Holmes, C., Müller, P., Walker, S. (eds.) Bayesian Nonparametrics: Principles and Practice. Cambridge University Press (2010)
20. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
21. Wang, C., Paisley, J.W., Blei, D.M.: Online variational inference for the hierarchical Dirichlet process. Journal of Machine Learning Research - Proceedings Track 15, 752–760 (2011)