Visual Scenes Clustering Using Variational Incremental Learning of Infinite Generalized Dirichlet Mixture Models

Wentao Fan
Electrical and Computer Engineering
Concordia University, Canada
wenta_fa@encs.concordia.ca

Nizar Bouguila
Institute for Information Systems Engineering
Concordia University, Canada
nizar.bouguila@concordia.ca

Abstract

In this paper, we develop a clustering approach based on variational incremental learning of a Dirichlet process of generalized Dirichlet (GD) distributions. Our approach is built on nonparametric Bayesian analysis where the determination of the complexity of the mixture model (i.e. the number of components) is sidestepped by assuming an infinite number of mixture components. By leveraging an incremental variational inference algorithm, the model complexity and all the involved model's parameters are estimated simultaneously and effectively in a single optimization framework. Moreover, thanks to its incremental nature and Bayesian roots, the proposed framework allows to avoid over- and under-fitting problems, and to offer good generalization capabilities. The effectiveness of the proposed approach is tested on a challenging application involving visual scenes clustering.

1 Introduction

Incremental clustering plays a crucial role in many data mining and computer vision applications [Opelt *et al.*, 2006; Sheikh *et al.*, 2007; Li *et al.*, 2007]. Incremental clustering is particularly efficient in the following scenarios: when data points are obtained sequentially, when the available memory is limited, or when we have large-scale data sets to deal with. Bayesian approaches have been widely used to develop powerful clustering techniques. Bayesian approaches applied for incremental clustering fall basically into two categories: parametric and non-parametric, and allow to mimic the human learning process which is based on iterative accumulation of knowledge. As opposed to parametric approaches in which a fixed number of parameters is considered, Bayesian nonparametric approaches use an infinite-dimensional parameter space and allow the complexity of models to grow with data size. The consideration of an infinite-dimensional parameter space allows to determine appropriate model complexity, which is normally referred to as the problem of model selection or model adaptation. This is a crucial issue in clustering since it permits to capture the underlying data structure more precisely, and also to avoid over- and under-fitting problems. This paper focuses on the latter one since it is more adapted to modern data mining applications (i.e. modern applications involve generally dynamic data sets).

Nowadays, the most popular Bayesian nonparametric formalism is the Dirichlet process (DP) [Neal, 2000; Teh *et al.*, 2004] generally translated to a mixture model with a countably infinite number of components in which the difficulty of selecting the appropriate number of clusters, that usually occurs in the finite case, is avoided. A common way to learn Dirichlet process model is through Markov chain Monte Carlo (MCMC) techniques. Nevertheless, MCMC approaches have several drawbacks such as the high computational cost and the difficulty of monitoring convergence. These shortcomings of MCMC approaches can be solved by adopting an alternative namely variational inference (or variational Bayes) [Attias, 1999], which is a deterministic approximation technique that requires a modest amount of computational power. Variational inference has provided promising performance in many applications involving mixture models [Corduneanu and Bishop, 2001; Constantinopoulos *et al.*, 2006; Fan *et al.*, 2012; 2013]. In our work, we employ an incremental version of variational inference proposed by [Gomes *et al.*, 2008] to learn infinite generalized Dirichlet (GD) mixtures in the context where data points are supposed to arrive sequentially. The consideration of the GD distribution is motivated by its promising performance when handling non-Gaussian data, and in particular proportional data (which are subject to two restrictions: nonnegativity and unit-sum) which are naturally generated in several data mining, machine learning, computer vision, and bioinformatics applications [Bouguila and Ziou, 2006; 2007; Boutemedjet *et al.*, 2009]. Examples of applications include textual documents (or images) clustering where a given document (or image) is described as a normalized histogram of words (or visual words) frequencies.

The main contributions of this paper are listed as the following: 1) we develop an incremental variational learning algorithm for the infinite GD mixture model, which is much more efficient when dealing with massive and sequential data as opposed to the corresponding batch approach; 2) we apply the proposed approach to tackle a challenging real-world problem namely visual scenes clustering. The effectiveness and merits of our approach are illustrated through extensive simulations. The rest of this paper is organized as follows. Section 2 presents the infinite GD mixture model. The incremental variational inference framework for model learning is described in Section 3. Section 4 is devoted to the experimental results. Finally, conclusion follows in Section 5.

2 The Infinite GD Mixture Model

Let $\vec{Y} = (Y_1, \dots, Y_D)$ be a D-dimensional random vector drawn from an infinite mixture of GD distributions:

$$p(\vec{Y}|\vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^{\infty} \pi_j \mathrm{GD}(\vec{Y}|\vec{\alpha}_j, \vec{\beta}_j) \tag{1}$$

where $\vec{\pi}$ represents the mixing weights that are positive and sum to one. $\vec{\alpha}_j = (\alpha_{j1}, \dots, \alpha_{jD})$ and $\vec{\beta}_j = (\beta_{j1}, \dots, \beta_{jD})$ are the positive parameters of the GD distribution associated with component j, while $\mathrm{GD}(\vec{Y}|\vec{\alpha}_j, \vec{\beta}_j)$ is defined as

$$\mathrm{GD}(\vec{Y}|\vec{\alpha}_j, \vec{\beta}_j) = \prod_{l=1}^{D} \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} Y_l^{\alpha_{jl}-1} \left(1 - \sum_{k=1}^{l} Y_k\right)^{\gamma_{jl}} \tag{2}$$

where $\sum_{l=1}^{D} Y_l < 1$ and $0 < y_l < 1$ for $l = 1, \dots, D$, $\gamma_{jl} = \beta_{jl} - \alpha_{jl+1} - \beta_{jl+1}$ for $l = 1, \dots, D-1$, and $\gamma_{jD} = \beta_{jD} - 1$. $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du$. Furthermore, we exploit an interesting and convenient mathematical property of the GD distribution which is thoroughly discussed in [Boutemedjet *et al.*, 2009], to transform the original data points into another D-dimensional space where the features are conditionally independent and rewrite the infinite GD mixture model in the following form

$$p(\vec{X}|\vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^{\infty} \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_l|\alpha_{jl}, \beta_{jl}) \tag{3}$$

where $X_l = Y_l$ and $X_l = Y_l/(1 - \sum_{k=1}^{l-1} Y_k)$ for $l > 1$. $\mathrm{Beta}(X_l|\alpha_{jl}, \beta_{jl})$ is a Beta distribution parameterized with $(\alpha_{jl}, \beta_{jl})$.

In this work, we construct the Dirichlet process through a stick-breaking representation [Sethuraman, 1994]. Therefore, the mixing weights π_j are constructed by recursively breaking a unit length stick into an infinite number of pieces as $\pi_j = \lambda_j \prod_{k=1}^{j-1}(1 - \lambda_k)$. λ_j is known as the stick breaking variable and is distributed independently according to $\lambda_j \sim \mathrm{Beta}(1, \xi)$, where $\xi > 0$ is the concentration parameter of the Dirichlet process.

For an observed data set $(\vec{X}_1, \dots, \vec{X}_N)$, we introduce a set of mixture component assignment variables $\vec{Z} = (Z_1, \dots, Z_N)$, one for each data point. Each element Z_i of $\vec{Z}$ has an integer value j specifying the component from which $\vec{X}_i$ is drawn. The marginal distribution over $\vec{Z}$ is given by

$$p(\vec{Z}|\vec{\lambda}) = \prod_{i=1}^{N} \prod_{j=1}^{\infty} \left[\lambda_j \prod_{k=1}^{j-1} (1 - \lambda_k)\right]^{\mathbf{1}[Z_i=j]} \tag{4}$$

where $\mathbf{1}[\cdot]$ is an indicator function which equals to 1 when $Z_i = j$, and equals to 0 otherwise. Since our model framework is Bayesian, we need to place prior distributions over random variables $\vec{\alpha}$ and $\vec{\beta}$. Since the formal conjugate prior for Beta distribution is intractable, we adopt Gamma priors $\mathcal{G}(\cdot)$ to approximate the conjugate priors of $\vec{\alpha}$ and $\vec{\beta}$ as: $p(\vec{\alpha}) = \mathcal{G}(\vec{\alpha}|\vec{u}, \vec{v})$ and $p(\vec{\beta}) = \mathcal{G}(\vec{\beta}|\vec{s}, \vec{t})$, with the assumption that these parameters are statistically independent.

3 Model Learning

In our work, we adopt an incremental learning framework proposed in [Gomes *et al.*, 2008] to learn the proposed infinite GD mixture model through variational Bayes. In this algorithm, data points can be sequentially processed in small batches where each one may contain one or a group of data points. The model learning framework involves the following two phases: 1) model building phase: to inference the optimal mixture model with the currently observed data points; 2) compression phase: to estimate which mixture component that groups of data points should be assigned to.

3.1 Model Building Phase

For an observed data set $\mathcal{X} = (\vec{X}_1, \dots, \vec{X}_N)$, we define $\Theta = \{\vec{Z}, \vec{\alpha}, \vec{\beta}, \vec{\lambda}\}$ as the set of unknown random variables. The main target of variational Bayes is to estimate a proper approximation $q(\Theta)$ for the true posterior distribution $p(\Theta|\mathcal{X})$. This problem can be solved by maximizing the free energy $\mathcal{F}(\mathcal{X}, q)$, where $\mathcal{F}(\mathcal{X}, q) = \int q(\Theta) \ln[p(\mathcal{X}, \Theta)/q(\Theta)] d\Theta$. In our algorithm, inspired by [Blei and Jordan, 2005], we truncate the variational distribution $q(\Theta)$ at a value M, such that $\lambda_M = 1$, $\pi_j = 0$ when $j > M$, and $\sum_{j=1}^{M} \pi_j = 1$, where the truncation level M is a variational parameter which can be freely initialized and will be optimized automatically during the learning process [Blei and Jordan, 2005]. In order to achieve tractability, we also assume that the approximated posterior distribution $q(\Theta)$ can be factorized into disjoint tractable factors as: $q(\Theta) = [\prod_{i=1}^{N} q(Z_i)][\prod_{j=1}^{M} \prod_{l=1}^{D} q(\alpha_{jl}) q(\beta_{jl})][\prod_{j=1}^{M} q(\lambda_j)]$.

By maximizing the free energy $\mathcal{F}(\mathcal{X}, q)$ with respect to each variational factor, we can obtain the following update equations for these factors:

$$q(\vec{Z}) = \prod_{i=1}^{N} \prod_{j=1}^{M} r_{ij}^{\mathbf{1}[Z_i=j]}, \quad q(\vec{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}|u_{jl}^*, v_{jl}^*) \tag{5}$$

$$q(\vec{\beta}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\beta_{jl}|s_{jl}^*, t_{jl}^*), \quad q(\vec{\lambda}) = \prod_{j=1}^{M} \mathrm{Beta}(\lambda_j|a_j, b_j) \tag{6}$$

where we have defined

$$r_{ij} = \frac{\exp(\rho_{ij})}{\sum_{j=1}^{M} \exp(\rho_{ij})} \tag{7}$$

$$\rho_{ij} = \sum_{l=1}^{D} \left[\widetilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} + (\bar{\beta}_{jl} - 1) \ln(1 - X_{il})\right]$$

$$+ \langle \ln \lambda_j \rangle + \sum_{k=1}^{j-1} \langle \ln(1 - \lambda_k) \rangle$$

$$u_{jl}^* = u_{jl} + \sum_{i=1}^{N} \langle Z_i = j \rangle [\Psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \Psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl}$$

$$\times\Psi'(\bar{\alpha}_{jl}+\bar{\beta}_{jl})(\langle\ln\beta_{jl}\rangle-\ln\bar{\beta}_{jl})]\bar{\alpha}_{jl}$$

$$s^*_{jl}=s_{jl}+\sum_{i=1}^{N}\big\langle Z_i=j\big\rangle[\Psi(\bar{\alpha}_{jl}+\bar{\beta}_{jl})-\Psi(\bar{\beta}_{jl})+\bar{\alpha}_{jl}$$

$$\times\Psi'(\bar{\alpha}_{jl}+\bar{\beta}_{jl})(\langle\ln\alpha_{jl}\rangle-\ln\bar{\alpha}_{jl})]\bar{\beta}_{jl}$$

$$v^*_{jl}=v_{jl}-\sum_{i=1}^{N}\big\langle Z_i=j\big\rangle\ln X_{il},\quad b_j=\xi_j+\sum_{i=1}^{N}\sum_{k=j+1}^{M}\langle Z_i=k\rangle$$

$$t^*_{jl}=t_{jl}-\sum_{i=1}^{N}\big\langle Z_i=j\big\rangle\ln(1-X_{il}),\quad a_j=1+\sum_{i=1}^{N}\langle Z_i=j\rangle$$

where $\Psi(\cdot)$ is the digamma function, and $\langle\cdot\rangle$ is the expectation evaluation. Note that, $\widetilde{\mathcal{R}}$ is the lower bound of $\mathcal{R}=\big\langle\ln\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\big\rangle$. Since this expectation is intractable, the second-order Taylor series expansion is applied to find its lower bound. The expected values in the above formulas are given by $\langle Z_i=j\rangle=r_{ij}$, $\bar{\alpha}_{jl}=\langle\alpha_{jl}\rangle=u^*_{jl}/v^*_{jl}$, $\bar{\beta}_{jl}=\langle\beta_{jl}\rangle=s^*_{jl}/t^*_{jl}$, $\langle\ln\lambda_j\rangle=\Psi(a_j)-\Psi(a_j+b_j)$, $\langle\ln(1-\lambda_j)\rangle=\Psi(b_j)-\Psi(a_j+b_j)$, $\langle\ln\alpha_{jl}\rangle=\Psi(u^*_{jl})-\ln v^*_{jl}$ and $\langle\ln\beta_{jl}\rangle=\Psi(s^*_{jl})-\ln t^*_{jl}$.

After convergence, the currently observed data points are clustered into M groups according to corresponding responsibilities r_{ij} through Eq. (7). According to [Gomes *et al.*, 2008], these newly formed groups of data points are also denoted as "clumps". Following [Gomes *et al.*, 2008], these clumps are subject to the constraint that all data points $\vec{X}_i$ in the clump c share the same $q(Z_i)\equiv q(Z_c)$ which is a key factor in the following compression phase.

Algorithm 1

```
Choose the initial truncation level $M$.
Initialize the values for hyper-parameters $u_{jl}$, $v_{jl}$, $s_{jl}$, $t_{jl}$ and $\xi_j$.
Initialize the values of $r_{ij}$ by $K$-Means algorithm.
while More data to be observed do
    Perform the model building phase through Eqs. (5) and (6).
    Initialize the compression phase using Eq. (10).
    while $\mathcal{MC}\geq\mathcal{C}$ do
        for $j=1$ to $M$ do
            if $evaluated(j)=$ false then
                Split component $j$ and refine this split using Eqs (9).
                $\Delta\mathcal{F}(j)$ = change in Eq. (8).
                $evaluated(j)=$ true.
            end if
        end for
        Split component $j$ with the largest value of $\Delta\mathcal{F}(j)$.
        $M=M+1$.
    end while
    Discard the current observed data points.
    Save resulting components into next learning round.
end while
```

3.2 Compression Phase

Within the compression phase, we need to estimate clumps that are possibly belong to the same mixture component while taking into consideration future arriving data. Now assume that we have already observed N data points, our aim is to make an inference at some target time T where $T\geq N$. we can tackle this problem by scaling the observed data to the target size T, which is equivalent to using the variational posterior distribution of the observed data N as a predictive model of the future data [Gomes *et al.*, 2008]. We then have a modified free energy for the compression phase in the following form

$$\mathcal{F}=\sum_{j=1}^{M}\sum_{l=1}^{D}\bigg[\Big\langle\ln\frac{p(\alpha_{jl}|u_{jl},v_{jl})}{q(\alpha_{jl})}\Big\rangle+\Big\langle\ln\frac{p(\beta_{jl}|s_{jl},t_{jl})}{q(\beta_{jl})}\Big\rangle\bigg]$$

$$+\sum_{j=1}^{M}\Big\langle\ln\frac{p(\lambda_j|\xi_j)}{q(\lambda_j)}\Big\rangle+\frac{T}{N}\sum_{c}|n_c|\ln\sum_{j=1}^{M}\exp(\rho_{cj})\quad(8)$$

where $|n_c|$ represents the number of data points in clump c and $\frac{T}{N}$ is the data magnification factor. The corresponding update equations for maximizing this free energy function can be obtained as

$$r_{cj}=\frac{\exp(\rho_{cj})}{\sum_{j=1}^{M}\exp(\rho_{cj})}\quad(9)$$

$$\rho_{ij}=\sum_{l=1}^{D}\big[\widetilde{\mathcal{R}}_{jl}+(\bar{\alpha}_{jl}-1)\ln\langle X_{cl}\rangle+(\bar{\beta}_{jl}-1)\ln(1-\langle X_{cl}\rangle)\big]$$

$$+\big\langle\ln\lambda_j\big\rangle+\sum_{k=1}^{j-1}\big\langle\ln(1-\lambda_k)\big\rangle$$

$$u^*_{jl}=u_{jl}+\frac{T}{N}\sum_{c}|n_c|r_{cj}[\Psi(\bar{\alpha}_{jl}+\bar{\beta}_{jl})-\Psi(\bar{\alpha}_{jl})+\bar{\beta}_{jl}$$

$$\times\Psi'(\bar{\alpha}_{jl}+\bar{\beta}_{jl})(\langle\ln\beta_{jl}\rangle-\ln\bar{\beta}_{jl})]\bar{\alpha}_{jl}$$

$$s^*_{jl}=s_{jl}+\frac{T}{N}\sum_{c}|n_c|r_{cj}[\Psi(\bar{\alpha}_{jl}+\bar{\beta}_{jl})-\Psi(\bar{\beta}_{jl})+\bar{\alpha}_{jl}$$

$$\times\Psi'(\bar{\alpha}_{jl}+\bar{\beta}_{jl})(\langle\ln\alpha_{jl}\rangle-\ln\bar{\alpha}_{jl})]\bar{\beta}_{jl}$$

$$v^*_{jl}=v_{jl}-\frac{T}{N}\sum_{c}|n_c|r_{cj}\ln\langle X_{cl}\rangle$$

$$t^*_{jl}=t_{jl}-\frac{T}{N}\sum_{c}|n_c|r_{cj}\ln(1-\langle X_{cl}\rangle)$$

$$a_j=1+\frac{T}{N}\sum_{c}|n_c|\langle Z_c=j\rangle$$

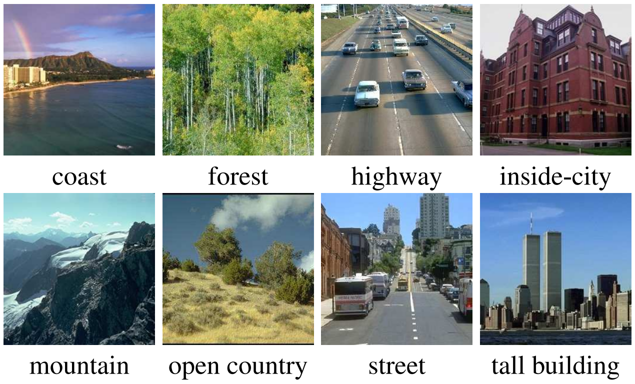
$$b_j=\xi_j+\frac{T}{N}\sum_{c}|n_c|\sum_{k=j+1}^{M}\langle Z_c=k\rangle$$

where $\langle X_{cl}\rangle$ denotes average over all data points contained in clump c.

The first step of the compression phase is to assign each clump or data point to the component with the highest responsibility r_{cj} calculated from the model building phase as

$$I_c=\arg\max_{j} r_{cj}\quad(10)$$

where $\{I_c\}$ denote which component the clump (or data point) c belongs to in the compression phase. Next, we cycle through each component and split it along its principal component into two subcomponents. This split is refined by updating Eqs. (9). The clumps are then hard assigned to one of the two candidate components after convergence for refining the split. Among all the potential splits, we select the one that results in the largest change in the free energy (Eq. (8)). The splitting process repeats itself until a stopping criterion is met. According to [Gomes *et al.*, 2008], the stoping criterion for the splitting process can be expressed as a limit on the amount of memory required to store the components. In our case, the component memory cost for the mixture model is $\mathcal{MC} = 2DN_c$, where $2D$ is the number of parameters contained in a D-variate GD component, and N_c is the number of components. Accordingly, We can define an upper limit on the component memory cost $\mathcal{C}$, and the compression phase stops when $\mathcal{MC} \geq \mathcal{C}$. As a result, the computational time and the space requirement is bounded in each learning round. After the compression phase, the currently observed data points are discarded while the resulting components can be treated in the same way as data points in the next round of leaning. Our incremental variational inference algorithm for infinite GD mixture model is summarized in Algorithm 1.


coast forest highway inside-city

mountain open country street tall building

Figure 1: Sample images from the OT data set.

4 Visual Scenes Clustering

In this section, the effectiveness of the proposed incremental infinite GD mixture model (*InGDMM*) is tested on a challenging real-world application namely visual scenes clustering. The problem is important since images are being produced at exponential increasing rates and very challenging due to the difficulty of capturing the variability of appearance and shape of diverse objects belonging to the same scene, while avoiding confusing objects from different scenes. In our experiments, we initialize the truncation level M as 15. The initial values of the hyperparameters are set as: $(u_{jl}, v_{jl}, s_{jl}, t_{jl}, \xi_j) = (1, 0.01, 1, 0.01, 0.1)$, which have been found to be reasonable choices according to our experimental results.

4.1 Database and Experimental Design

In this paper, we test our approach on a challenging and publicly available database known as the OT database, which was introduced by Oliva and Torralba [Oliva and Torralba, 2001] [1]. This database contains 2,688 images with the size of 256 × 256 pixels, and is composed of eight urban and natural scene categories: coast (360 images), forest (328 images), highway (260 images), inside-city (308 images), mountain (374 images), open country (410 images), street (292 images), and tall building (356 images). Figure 1 shows some sample images from the different categories in the OT database.

Our methodology is based on the proposed incremental infinite GD mixture model in conjunction with a bag-of-visual words representation, and can be summarized as follows: Firstly, we use the Difference-of-Gaussians (DoG) interest point detector to extract Scale-invariant feature transform (SIFT) descriptors (128-dimensional) [Lowe, 2004] from each image. Secondly, K-Means algorithm is adopted to construct a visual vocabulary by quantizing these SIFT vectors into visual words. As a result, each image is represented as the frequency histogram over the visual words. We have tested different sizes of the visual vocabulary $|\mathcal{W}| = [100, 1000]$, and the optimal performance was obtained for $|\mathcal{W}| = 750$ according to our experimental results. Then, the Probabilistic Latent Semantic Analysis (pLSA) model [Hofmann, 2001] is applied to the obtained histograms to represent each image by a 55-dimensional proportional vector where 55 is the number of latent aspects. Finally, the proposed *InGDMM* is deployed to cluster the images supposed to arrive in a sequential way.

Table 1: Average rounded confusion matrix for the OT database calculated by *InGDMM*.

	C	F	H	I	M	O	S	T
Coast (C)	**127**	10	4	2	3	31	2	1
Forest (F)	2	**155**	1	2	1	3	0	0
Highway (H)	0	0	**122**	1	0	3	3	1
Inside-city (I)	2	4	2	**119**	3	2	15	7
Mountain (M)	6	21	4	5	**139**	9	1	2
Open country (O)	2	22	19	15	9	**131**	3	4
Street (S)	0	1	4	8	5	5	**122**	1
Tall building (T)	4	9	7	23	3	19	3	**110**

4.2 Experimental Results

In our experiments, we randomly divided the OT database into two halves: one for constructing the visual vocabulary, another for testing. Since our approach is unsupervised, the class labels are not involved in our experiments, except for evaluation of the clustering results. The entire methodology was repeated 30 times to evaluate the performance. For comparison, we have also applied three other mixture-modeling approaches: the finite GD mixture model (*FiGDMM*), the infinite Gaussian mixture model (*InGMM*) and the finite Gaussian mixture model (*FiGMM*). To make a fair comparison, all of the aforementioned approaches are learned through incremental variational inference. Table 1 shows the average confusion matrix of the OT database calculated by the proposed *InGDMM*. Table 2 illustrates the average categorization performance using different approaches for the OT database. As we can see from this table, it is obvious that our approach (*InGDMM*) provides the best performance in

[1] OT database is available at: http://cvcl.mit.edu/database.htm.

terms of the highest categorization rate (77.47%) among all the tested approaches. In addition, we can observe that better performances are obtained for approaches that adopt the infinite mixtures (*InGDMM* and *InGMM*) than the corresponding finite mixtures (*FiGDMM* and *FiGMM*), which demonstrate the advantage of using infinite mixture models over finite ones. Moreover, according to Table 2, GD mixture has higher performance than Gaussian mixture which verifies that the GD mixture model has better modeling capability than the Gaussian for proportional data clustering.

Table 2: The average classification accuracy rate (Acc) (%) obtained over 30 runs using different approaches.

Method	*InGDMM*	*FiGDMM*	*InGMM*	*FiGMM*
Acc(%)	77.47	74.25	72.54	70.19

5 Conclusion

In this work, we have presented an incremental nonparametric Bayesian approach for clustering. The proposed approach is based on infinite GD mixture models with a Dirichlet process framework, and is learned using an incremental variational inference framework. Within this framework, the model parameters and the number of mixture components are determined simultaneously. The effectiveness of the proposed approach has been evaluated on a challenging application namely visual scenes clustering. Future works could be devoted to the application of the proposed algorithm for other data mining tasks involving continually changing or growing volumes of proportional data.

References

[Attias, 1999] H. Attias. A variational Bayes framework for graphical models. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 209–215, 1999.

[Blei and Jordan, 2005] D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

[Bouguila and Ziou, 2006] N. Bouguila and D. Ziou. 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.

[Bouguila and Ziou, 2007] N. Bouguila and D. Ziou. 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*, 29:1716–1731, 2007.

[Boutemedjet *et al.*, 2009] S. Boutemedjet, N. Bouguila, and D. Ziou. 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.

[Constantinopoulos *et al.*, 2006] C. Constantinopoulos, M.K. Titsias, and A. Likas. Bayesian feature and model selection for Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):1013 –1018, 2006.

[Corduneanu and Bishop, 2001] A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. In *Proc. of the 8th International Conference on Artificial Intelligence and Statistics (AISTAT)*, pages 27–34, 2001.

[Fan *et al.*, 2012] W. Fan, N. Bouguila, and D. Ziou. Variational learning for finite Dirichlet mixture models and applications. *IEEE Transactions on Neural Netw. Learning Syst.*, 23(5):762–774, 2012.

[Fan *et al.*, 2013] Wentao Fan, Nizar Bouguila, and Djemel Ziou. 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.

[Gomes *et al.*, 2008] R. Gomes, M. Welling, and P. Perona. Incremental learning of nonparametric Bayesian mixture models. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[Hofmann, 2001] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.

[Li *et al.*, 2007] L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[Lowe, 2004] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Neal, 2000] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[Oliva and Torralba, 2001] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[Opelt *et al.*, 2006] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 3–10, 2006.

[Sethuraman, 1994] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[Sheikh *et al.*, 2007] Y.A. Sheikh, E.A. Khan, and T. Kanade. Mode-seeking by medoidshifts. In *Proc. of the IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[Teh *et al.*, 2004] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:705–711, 2004.