Pattern Analysis and Applications (2019) 22:63–74
https://doi.org/10.1007/s10044-018-00767-y

THEORETICAL ADVANCES

A nonparametric Bayesian learning model using accelerated variational inference and feature selection

Wentao Fan[1] · Nizar Bouguila[2] · Xin Liu[1]

Received: 5 July 2016 / Accepted: 26 December 2018 / Published online: 8 January 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract

Developing effective machine learning methods for multimedia data modeling continues to challenge computer vision scientists. The capability of providing effective learning models can have significant impact on various applications. In this work, we propose a nonparametric Bayesian approach to address simultaneously two fundamental problems, namely clustering and feature selection. The approach is based on infinite generalized Dirichlet (GD) mixture models constructed through the framework of Dirichlet process and learned using an accelerated variational algorithm that we have developed. Furthermore, we extend the proposed approach using another nonparametric Bayesian prior, namely Pitman–Yor process, to construct the infinite generalized Dirichlet mixture model. Our experiments, which were conducted through synthetic data sets, the clustering analysis of real-world data sets and a challenging application, namely automatic human action recognition, indicate that the proposed framework provides good modeling and generalization capabilities.

Keywords Infinite mixtures · Variational Bayes · Generalized Dirichlet · Feature selection · Human action recognition

1 Introduction

With the rapid development of digital technologies, providing approaches that can model visual data is more and more pressing. Several data mining and machine learning techniques have been proposed [4, 5]. Mixture models have received a particular attention among these techniques during the past decade [19]. Mixture models suppose that data are generated by a collection of populations where each population can be modeled using a probability density function. These models are widely used in many pattern recognition, image processing and computer vision applications [6]. The Gaussian mixture model has received particular attention in the computer vision and pattern recognition literature [1]. However, we have shown recently that other models such as the infinite generalized Dirichlet (GD) mixture [7], that we will consider in this paper, may provide better modeling capabilities thanks to its flexibility.

The main advantage of the Dirichlet process mixture of GD model (also known as the infinite GD mixture model) is that it allows explicit use of prior information, thereby giving insights into problems where classic frequentist techniques fail. This prior information is introduced via Dirichlet process that has gained a lot of spotlights recently [11, 21, 28, 29, 31, 32]. A crucial problem when using these kinds of models is feature selection. Feature selection helps to prevent overfitting and provides meaningful interpretation of the data. Many approaches in use are mainly stepwise selection techniques which ignore data uncertainty and the strong dependency between model and feature selection problems. An approach that handles simultaneously parameters estimation, model and feature selection problems, in the case of infinite GD mixtures, has been proposed in [10]. This approach can be considered as an infinite extension of the unsupervised feature selection framework previously proposed in [8]. A related model has been proposed in [10] in which a variational Bayes method is adopted for model learning. One major limitation of the approach in [10] is that it cannot extend to large-scale problems directly. To tackle this problem, we may adopt an

✉ Wentao Fan
fwt@hqu.edu.cn

Nizar Bouguila
nizar.bouguila@concordia.ca

Xin Liu
xliu@hqu.edu.cn

[1] Department of Computer Science and Technology, Huaqiao University, Xiamen, China

[2] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada

accelerated variational Bayes algorithm [15] which is originally designed for learning Gaussian mixture models . However, the original accelerated variational Bayes algorithm is not able to discriminate the significance of each feature. Thus, it is crucial to design a model that can perform fast clustering and feature selection simultaneously for high-dimensional large-scale data sets.

The major contributions of this work can be summarized as follows: Firstly, we develop an efficient nonparametric model for both clustering and feature selection based on the Dirichlet process mixture model with GD distributions, and is learnt by an accelerated variational Bayes algorithm; secondly, we extend the Dirichlet process mixture model of GD distributions with feature selection using another nonparametric Bayesian prior, namely the Pitman–Yor process mixture model. Similar to the Dirichlet process mixture model, the Pitman–Yor process mixture can also be considered as an infinite mixture model but with a power-law behavior, which makes it more appropriate for modeling data describing natural phenomena than the Dirichlet process mixture model does; lastly, the resulting statistical framework performance is assessed based on synthetic data sets, the clustering analysis of real-world data sets and a challenging application that concerns human action recognition which has been drawing growing interest because of its importance in surveillance and content-based research tasks [14, 25, 33].

The remaining parts of this paper are organized as follows. In Sect. 2, we briefly introduce the infinite GD mixture model together with an unsupervised feature selection scheme. In Sect. 3, an accelerated variational framework is developed to learn the parameters of the corresponding model. In Sect. 4, we propose an extension to the proposed approach by introducing the Pitman–Yor process mixture model. Section 5 demonstrates the experimental results. Lastly, the conclusion and future work are presented in Sect. 6.

2 Dirichlet process mixture model of GD distributions with feature selection

In this part, we briefly present the infinite generalized Dirichlet (GD) mixture model with unsupervised feature selection, which is based on a Bayesian nonparametric framework, namely Dirichlet process [13, 20]. In our case, a stick-breaking representation [23] of the Dirichlet process framework is adopted due to its intuitive representation and intrinsic clustering property. Assume that a random distribution G follows a Dirichlet process $G \sim DP(\xi, H)$, with concentration parameter ξ and base distribution H. Its stick-breaking representation is then given by

$$G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}, \qquad \pi_j = \lambda_j \prod_{s=1}^{j-1} (1 - \lambda_s), \qquad \lambda_j \sim \mathrm{Beta}(1, \xi), \qquad \theta_j \sim H, \tag{1}$$

where δ_{θ_j} represents the Dirac delta measure centered at θ_j, and ξ is a positive real number. The variables $\{\pi_j\}$ denote the mixing weights where $\sum_{j=1}^{\infty} \pi_j = 1$.

By exploiting the framework of Dirichlet process, if a D-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_D)$ is distributed according to a GD mixture model which contains an infinite number of mixture components, we have

$$p(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{j=1}^{\infty} \pi_j \mathrm{GD}(\mathbf{Y}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j), \tag{2}$$

where $\boldsymbol{\pi}$ denotes the mixing weights. $\mathrm{GD}(\mathbf{Y}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ is the GD distribution that belongs to class j with positive parameters $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jD})$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jD})$ and is given as

$$\mathrm{GD}(\mathbf{Y}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) = \prod_{l=1}^{D} \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} Y_l^{\alpha_{jl}-1} \left(1 - \sum_{k=1}^{l} Y_k\right)^{\tau_{jl}}, \tag{3}$$

where $0 < Y_l < 1$ and $\sum_{l=1}^{D} Y_l < 1$ for $l = 1, \ldots, D$; $\tau_{jl} = \beta_{jl} - \beta_{jl+1} - \alpha_{jl+1}$ for $l = 1, \ldots, D-1$; $\tau_{jD} = \beta_{jD} - 1$. $\Gamma(\cdot)$ is the gamma function.

Feature selection is a common technique in high-dimensional data modeling to improve the learning performance by selecting a subset of most relevant features. In this work, we incorporate a feature selection technique with the infinite GD mixture model by following a mathematical transformation of the GD distribution which is described in [8]. Specifically, the original data points are transformed into a new D-dimensional space with independent features. Then, the infinite GD mixture model as shown in Eq. (2) can be rewritten in terms of Beta distributions as

$$p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{\infty} \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_l|\alpha_{jl}, \beta_{jl}), \tag{4}$$

where $X_1 = Y_1$ and $X_l = \frac{Y_l}{(1-\sum_{k=1}^{l-1} Y_k)}$, for $l > 1$. $\mathrm{Beta}(X_l|\alpha_{jl}, \beta_{jl})$ is a Beta distribution associated with parameters α_{jl} and β_{jl}. As a result, in contrast to previous unsupervised feature selection approaches with Gaussian mixture models [9, 18], the independency between all features in the new data space becomes a fact instead of an assumption.

Here, we integrate the method of unsupervised feature selection introduced in [18] with the infinite GD mixture model. The main idea is that if the lth feature is independent of the hidden label Z (i.e., it is distributed according to a common density), it is considered as an irrelevant feature. Consequently, we can define the infinite GD mixture model with unsupervised feature selection method as

$$p(\mathbf{X}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\alpha}',\boldsymbol{\beta}',\boldsymbol{\pi},\boldsymbol{\phi}) = \sum_{j=1}^{\infty}\pi_j\prod_{l=1}^{D}\left[\text{Beta}(X_l|\alpha_{jl},\beta_{jl})\right]^{\phi_l}\left[\text{Beta}(X_l|\alpha'_l,\beta'_l)\right]^{1-\phi_l}, \tag{5}$$

where $\boldsymbol{\phi} = (\phi_1,\dots,\phi_D)$ is the feature relevance indicator and $\phi_l = \{0,1\}$. When $\phi_l = 0$, it denotes that the lth feature is irrelevant and has the probability distribution $\text{Beta}(X_l|\alpha'_l,\beta'_l)$. By contrast, when $\phi_l = 1$, it indicates that feature l is relevant and is distributed according to $\text{Beta}(X_l|\alpha_{jl},\beta_{jl})$. The prior distribution of $\boldsymbol{\phi}$ has the following form

$$p(\boldsymbol{\phi}|\epsilon) = \prod_{l=1}^{D}\epsilon_{l_1}^{\phi_l}\epsilon_{l_2}^{1-\phi_l}, \tag{6}$$

where $p(\phi_l = 1) = \epsilon_{l_1}$ and $p(\phi_l = 0) = \epsilon_{l_2}$. Vector $\boldsymbol{\epsilon}$ denotes the probabilities that the data features are relevant and we have $\boldsymbol{\epsilon}_l = (\epsilon_{l_1},\epsilon_{l_2})$ and $\epsilon_{l_1} + \epsilon_{l_2} = 1$. Thus, $\boldsymbol{\epsilon}$ can also be considered as the 'feature saliency.'

Next, for the observed data set $(\mathbf{X}_1,\dots,\mathbf{X}_N)$, we introduce a variable $\mathbf{Z} = (Z_1,\dots,Z_N)$, where Z_i is an integer. When $Z_i = j$, it means that $\mathbf{X}_i$ is drawn from component j. The conditional probability of Z given π_j is given by

$$p(\mathbf{Z}|\pi) = \prod_{i=1}^{N}\prod_{j=1}^{\infty}\pi_j^{\mathbf{1}[Z_i=j]}, \tag{7}$$

where $\mathbf{1}[\cdot]$ equals 1 when $Z_i = j$; otherwise, $\mathbf{1}[\cdot]$ takes the value of 0. Moreover, according to the stick-breaking representation shown in Eq. (1), $p(\mathbf{Z})$ can also be defined by

$$p(\mathbf{Z}) = \prod_{j=1}^{\infty}\prod_{i=1}^{N}\left[\lambda_j\prod_{s=1}^{j-1}(1-\lambda_s)\right]^{\mathbf{1}[Z_i=j]}. \tag{8}$$

The prior distribution of λ is a specific Beta distribution given in Eq. (1) and is explicitly defined by

$$p(\lambda|\xi) = \prod_{j=1}^{\infty}\text{Beta}(\lambda_j,\xi_j) = \prod_{j=1}^{\infty}\xi_j(1-\lambda_j)^{\xi_j-1}. \tag{9}$$

Furthermore, in our Bayesian framework, we need to place priors over the Beta parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$. Since these parameters have to be positive, Gamma distributions are adopted as their priors

$$p(\boldsymbol{\alpha}) = \text{Gamma}(\boldsymbol{\alpha}|\mathbf{u},\mathbf{v}), \qquad p(\boldsymbol{\beta}) = \text{Gamma}(\boldsymbol{\beta}|\mathbf{g},\mathbf{h}),$$
$$p(\boldsymbol{\alpha}') = \text{Gamma}(\boldsymbol{\alpha}'|\mathbf{u}',\mathbf{v}'), \qquad p(\boldsymbol{\beta}') = \text{Gamma}(\boldsymbol{\beta}'|\mathbf{g}',\mathbf{h}'). \tag{10}$$

where the associated hyperparameters $\mathbf{u}$, $\mathbf{v}$, $\mathbf{g}$, $\mathbf{h}$, $\mathbf{u}'$, $\mathbf{v}'$, $\mathbf{g}'$ and $\mathbf{h}'$ are positive.

3 Accelerated variational Bayes model learning

According to recent studies, infinite GD mixture models can be learned using either Markov chain Monte Carlo (MCMC) techniques or variational Bayes methods [7, 10]. However, the performance of these approaches is significantly limited when dealing with large amount of data (e.g., millions of data instances). In order to efficiently handle large-scale data set, we may follow the idea of an accelerated version of variational Bayes inference method as proposed in [15]. In this part, we develop an accelerated variational Bayes method based on kd-tree structure in order to learn infinite GD mixture models with unsupervised feature selection.

3.1 Conventional variational Bayes model learning

In conventional variational Bayes model learning, the main goal is to discover a suitable approximation $q(\Omega)$ to the original posterior distribution $p(\Omega|\mathcal{X})$, where $\Omega = \{\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\alpha}',\boldsymbol{\beta}',\boldsymbol{\phi},\lambda\}$ represents the unknown parameters in our model. In our case, we factorize $q(\Omega)$ into the product of disjoint factors based on the mean-field assumption

$$q(\Omega) = q(\boldsymbol{\alpha})q(\boldsymbol{\beta})q(\boldsymbol{\alpha}')q(\boldsymbol{\beta}')q(\mathbf{Z})q(\boldsymbol{\phi})q(\lambda) \tag{11}$$

Instead of dealing with an infinite number of classes directly, a common trick in learning infinite mixture model is to use the truncation technique [3, 10], to truncate the number of mixture components of variational posteriors into a finite value. Nevertheless, a disadvantage of the truncation technique is that it may cause undesirable consequence that the approximating variational families are not nested [15]. We may address this issue by considering the idea proposed in [15] where the number of mixture components of variational posteriors remains infinite, but the variational parameters of all models are tied after a particular level M. That is, if a component is associated with the index $j > M$, then $q(\Omega_j)$ is set to its prior. Using the parameter tying assumption for $j > M$, the approximated posterior $q(\Omega)$ can then be calculated through the minimization of the following free energy

$$F = \sum_{i=1}^{N}\left[\left\langle \ln\frac{q(Z_i)}{p(Z_i|\lambda)p(\mathbf{X}_i|\Lambda)}\right\rangle + \sum_{l=1}^{D}\left\langle \ln\frac{q(\phi_{il})}{p(\phi_{il})}\right\rangle\right] + \sum_{j=1}^{M}\left[\sum_{l=1}^{D}\left\langle \ln\frac{q(\alpha_{jl})}{p(\alpha_{jl})}\right\rangle + \left\langle \ln\frac{q(\beta_{jl})}{p(\beta_{jl})}\right\rangle + \left\langle \ln\frac{q(\lambda_j)}{p(\lambda_j)}\right\rangle\right] + \sum_{l=1}^{D}\left[\left\langle \ln\frac{q(\alpha'_l)}{p(\alpha'_l)}\right\rangle + \left\langle \ln\frac{q(\beta'_l)}{p(\beta'_l)}\right\rangle\right] \tag{12}$$

$$\rho_{ij} = \sum_{l=1}^{D}\langle\phi_{il}\rangle\left[(\bar{\alpha}_{jl}-1)\ln X_{il} + (\bar{\beta}_{jl}-1)\ln(1-X_{il}) + \left\langle \ln\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}\right\rangle\right] + \sum_{s=1}^{j-1}\langle\ln(1-\lambda_s)\rangle + \langle\ln\lambda_j\rangle \tag{22}$$

$$\widetilde{f}_{il} = \sum_{j=1}^{M}\langle Z_i = j\rangle\left[(\bar{\alpha}_{jl}-1)\ln X_{il} + (\bar{\beta}_{jl}-1)\ln(1-X_{il}) + \left\langle \ln\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}\right\rangle\right] + \ln\epsilon_{l_1} \tag{23}$$

where $\Lambda = \{\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\phi}\}$; $\langle\cdot\rangle$ represents the corresponding expected value. Then, the variational solutions to each factor can be calculated by

$$q(\mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{\mathbf{1}[Z_i=j]} \tag{13}$$

$$q(\lambda) = \prod_{j=1}^{M}\mathrm{Beta}(\lambda_j|a_j, b_j) \tag{14}$$

$$q(\boldsymbol{\phi}) = \prod_{i=1}^{N}\prod_{l=1}^{D} f_{il}^{\phi_{il}}(1-f_{il})^{(1-\phi_{il})} \tag{15}$$

$$q(\boldsymbol{\alpha}) = \prod_{j=1}^{M}\prod_{l=1}^{D}\mathrm{Gamma}(\alpha_{jl}|\tilde{u}_{jl}, \tilde{v}_{jl}) \tag{16}$$

$$q(\boldsymbol{\beta}) = \prod_{j=1}^{M}\prod_{l=1}^{D}\mathrm{Gamma}(\beta_{jl}|\tilde{g}_{jl}, \tilde{h}_{jl}) \tag{17}$$

$$q(\boldsymbol{\alpha}') = \prod_{l=1}^{D}\mathrm{Gamma}(\alpha'_l|\tilde{u}'_l, \tilde{v}'_l) \tag{18}$$

$$q(\boldsymbol{\beta}') = \prod_{l=1}^{D}\mathrm{Gamma}(\beta'_l|\tilde{g}'_l, \tilde{h}'_l) \tag{19}$$

where associated hyperparameters are calculated by

$$r_{ij} = \frac{\exp(\rho_{ij})}{\sum_{j=1}^{\infty}\exp(\rho_{ij})} \tag{20}$$

$$f_{il} = \frac{\exp(\widetilde{f}_{il})}{\exp(\widetilde{f}_{il}) + \exp(\widehat{f}_{il})} \tag{21}$$

$$\widehat{f}_{il} = \left\langle \ln\frac{\Gamma(\alpha'_l+\beta'_l)}{\Gamma(\alpha'_l)\Gamma(\beta'_l)}\right\rangle + (\bar{\beta}'_l-1)\ln(1-X_{il}) + (\bar{\alpha}'_l-1)\ln X_{il} + \ln\epsilon_{l_2} \tag{24}$$

$$\tilde{u}_{jl} = u_{jl} + \sum_{i=1}^{N}\langle Z_i = j\rangle\bar{\alpha}_{jl}\langle\phi_{il}\rangle\big[\psi(\bar{\alpha}_{jl}+\bar{\beta}_{jl}) - \psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl}\psi'(\bar{\alpha}_{jl}+\bar{\beta}_{jl}) \times (\langle\ln\beta_{jl}\rangle - \ln\bar{\beta}_{jl})\big] \tag{25}$$

$$\tilde{g}_{jl} = g_{jl} + \sum_{i=1}^{N}\langle Z_i = j\rangle\bar{\beta}_{jl}\langle\phi_{il}\rangle[\psi(\bar{\alpha}_{jl}+\bar{\beta}_{jl}) - \psi(\bar{\beta}_{jl}) + \bar{\alpha}_{jl}\psi'(\bar{\alpha}_{jl}+\bar{\beta}_{jl}) \times (\langle\ln\alpha_{jl}\rangle - \ln\bar{\alpha}_{jl})] \tag{26}$$

$$\tilde{v}_{jl} = v_{jl} - \sum_{i=1}^{N}\langle\phi_{il}\rangle\langle Z_i = j\rangle\ln X_{il} \tag{27}$$

$$\tilde{h}_{jl} = h_{jl} - \sum_{i=1}^{N}\langle\phi_{il}\rangle\langle Z_i = j\rangle\ln(1-X_{il}) \tag{28}$$

$$a_j = 1 + \sum_{i=1}^{N}\langle Z_i = j\rangle \tag{29}$$

$$b_j = \xi_j + \sum_{i=1}^{N} \sum_{s=j+1}^{\infty} \langle Z_i = s \rangle \tag{30}$$

$$\sum_{j=M+1}^{\infty} \exp(\rho_{ij}) = \frac{\exp(\rho_{i,M+1})}{1 - \exp[\psi(\xi_j) - \psi(1 + \xi_j)]} \tag{31}$$

where $\psi(\cdot)$ in the above equations represents the digamma function. We can calculate the updating equations for hyperparameters of $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$ similar to Eqs. (25)–(28). Please notice that $\sum_{s=j+1}^{\infty}\langle Z_i = s\rangle$ in Eq. (30) can be obtained using (31) based on the parameter tying assumption that we previously discussed for $j > M$. The expectations in the above equations can be calculated by

$$\langle Z_i = j \rangle = r_{ij}, \qquad \langle \phi_{il} \rangle = f_{il} \tag{32}$$

$$\bar{\alpha}_{jl} = \langle \alpha_{jl} \rangle = \frac{\tilde{u}_{jl}}{\tilde{v}_{jl}}, \qquad \bar{\beta}_{jl} = \langle \beta_{jl} \rangle = \frac{\tilde{g}_{jl}}{\tilde{h}_{jl}} \tag{33}$$

$$\langle \ln \alpha_{jl} \rangle = \psi(\tilde{u}_{jl}) - \ln \tilde{v}_{jl}, \qquad \langle \ln \beta_{jl} \rangle = \psi(\tilde{g}_{jl}) - \ln \tilde{h}_{jl} \tag{34}$$

$$\langle \ln(1 - \lambda_j) \rangle = \psi(b_j) - \psi(a_j + b_j), \qquad \langle \ln \lambda_j \rangle = \psi(a_j) - \psi(a_j + b_j) \tag{35}$$

Since there are no closed form solutions to $\left\langle \ln \frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \right\rangle$ and $\left\langle \ln \frac{\Gamma(\alpha'_l+\beta'_l)}{\Gamma(\alpha'_l)\Gamma(\beta'_l)} \right\rangle$ in Eqs. (22)–(24), their lower bound approximations were obtained by the second-order Taylor expansion.

The feature saliencies can then be estimated by minimizing free energy F in Eq. (12) by making the derivative of F with respect to ϵ_l equal to zero

$$\epsilon_l = \frac{\sum_{i=1}^{N} \langle \phi_{il} \rangle}{N}. \tag{36}$$

3.2 Accelerated variational Bayes model learning

In this subsection, the conventional variational Bayes learning process is extended into an accelerate version based on the idea as proposed in [15] through a kd-tree structure [2]. Assume that the data set $\mathcal{X}$ is stored in a kd-tree with the assumption that all instances in the outer node T will have the same responsibility [i.e., $q(Z_i) \equiv q(Z_T)$]. Following this constraint, the variational solutions with kd-tree structure are calculated as

$$\langle Z_T = j \rangle = \frac{\exp(\rho_{Tj})}{\sum_{j=1}^{\infty} \exp(\rho_{Tj})} \tag{37}$$

$$f_{Tl} = \frac{\exp(\widetilde{f}_{Tl})}{\exp(\widetilde{f}_{Tl}) + \exp(\widehat{f}_{Tl})} \tag{38}$$

$$\begin{aligned}\rho_{Tj} = &\sum_{l=1}^{D} \langle \phi_{il} \rangle [\ln\langle X_l \rangle_T (\bar{\alpha}_{jl} - 1) + \ln(1 - \langle X_l \rangle_T)(\bar{\beta}_{jl} - 1) \\ &+ \left\langle \ln \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \right\rangle + \sum_{s=1}^{j-1} \langle \ln(1 - \lambda_s) \rangle + \langle \ln \lambda_j \rangle\end{aligned} \tag{39}$$

$$\begin{aligned}\widetilde{f}_{Tl} = &\ln \epsilon_{l_1} + \sum_{j=1}^{M} \langle Z_T = j \rangle \big[(\bar{\alpha}_{jl} - 1) \ln\langle X_l \rangle_T + (\bar{\beta}_{jl} - 1) \ln(1 - \langle X_l \rangle_T) \\ &+ \left\langle \ln \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \right\rangle \big]\end{aligned} \tag{40}$$

$$\begin{aligned}\widehat{f}_{Tl} = &\left\langle \ln \frac{\Gamma(\alpha'_l + \beta'_l)}{\Gamma(\alpha'_l)\Gamma(\beta'_l)} \right\rangle + \ln \epsilon_{l_2} + \ln(1 - \langle X_l \rangle_T)(\bar{\beta}'_l - 1) \\ &+ \ln\langle X_l \rangle_T (\bar{\alpha}'_l - 1)\end{aligned} \tag{41}$$

$$\begin{aligned}\tilde{u}_{jl} = &u_{jl} + \sum_{T} |n_T| \langle \phi_{Tl} \rangle \bar{\alpha}_{jl} \langle Z_T = j \rangle [\psi(\bar{\beta}_{jl} + \bar{\alpha}_{jl}) - \psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl}(\langle \ln \beta_{jl} \rangle \\ &- \ln \bar{\beta}_{jl}) \psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl})]\end{aligned} \tag{42}$$

$$\begin{aligned}\tilde{g}_{jl} = &g_{jl} + \sum_{T} |n_T| \langle \phi_{Tl} \rangle \langle Z_T = j \rangle \bar{\beta}_{jl} [\psi(\bar{\beta}_{jl} + \bar{\alpha}_{jl}) - \psi(\bar{\beta}_{jl}) + \bar{\alpha}_{jl}(\langle \ln \alpha_{jl} \rangle \\ &- \ln \bar{\alpha}_{jl}) \psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl})]\end{aligned} \tag{43}$$

$$\tilde{v}_{jl} = v_{jl} - \sum_{T} |n_T| \langle Z_T = j \rangle \langle \phi_{Tl} \rangle \ln\langle X_l \rangle_T \tag{44}$$

$$\tilde{h}_{jl} = h_{jl} - \sum_{T} |n_T| \langle Z_T = j \rangle \langle \phi_{Tl} \rangle \ln(1 - \langle X_l \rangle_T) \tag{45}$$

$$a_j = 1 + \sum_{T} |n_T| \langle Z_T = j \rangle \tag{46}$$

$$b_j = \xi_j + \sum_{T} |n_T| \sum_{s=j+1}^{\infty} \langle Z_T = s \rangle \tag{47}$$

where $\langle X_l \rangle_T$ represents the average value of all instances contained in node T, $|n_T|$ represents the total number of instances stored in node T. The feature saliencies are obtained as $\epsilon_l = \frac{1}{|n_T|} \sum_T f_{Tl}$.

Inspired from [15], the setting of our learning algorithm is described as follows: We reorder the components in each update round and expand the kd-tree in every three rounds. It is also noteworthy that the free energy F

is always decreasing when refining the tree. The computational cost to learn infinite GD mixture together with unsupervised feature selection requires $O(MD|T|)$ for each update cycle using the variational inference with kd-tree structure, which is much more efficient than the one without using kd-tree structure (requires $O(MDN)$). The complete learning process with kd-tree structure is summarized in Algorithm 1.

Algorithm 1

```
1: Input: A data set $\mathcal{X}$ which is saved using the kd-tree data structure
2: Output: Hyperparameters, $\epsilon_l$, $M$
3: {Initialization}
4: Set $M = 1$. Initialize the expansion of the kd-tree to a initial level (such as five)
5: Based on the size $\sum_T |n_T|\langle Z_T = c\rangle$, draw several ‘candidate’ components $c$
6: for each potential component $c$ do
7:     Expand the outer nodes of the kd-tree by one level deeper which assign the highest
       responsibility $\langle Z_T = c\rangle$ to the component $c$ among all components.
8:     According to the bisector of the corresponding principal component, divide compo-
       nent $c$ into $c_1$ and $c_2$, to form two new components. The responsibilities $\langle Z_T = c_1\rangle$
       and $\langle Z_T = c_2\rangle$ are initialized
9:     Update variational factors for new components $c_1$ and $c_2$ using Eqs. (37)∼(47)
10: end for
11: Update variational factors for all $j \leq M + 1$, as well as expand the kd-tree and reorder
    the components
12: Evaluate feature saliencies $\epsilon_l$
13: if $F_M - F_{M+1}$ < threshold then
14:     Stop
15: else
16:     set $M = M + 1$ and go to step 5
17: end if
```

4 Pitman–Yor process mixture model of GD distributions with feature selection

Rather than using the Dirichlet process framework, the proposed infinite GD mixture model may be constructed based on another nonparametric Bayesian prior, namely Pitman–Yor process, which is a generalization of the Dirichlet processes, with heavier-tailed power-law prior distributions [27]. Recently, the Pitman–Yor process has shown its flexibility and better performance in modeling complex real-life data sets than the Dirichlet process does [24, 26, 27].

If a random distribution G is distributed according to a Pitman–Yor process, we have

$$G \sim \mathrm{PY}(\varphi, \gamma, H), \tag{48}$$

where $0 \leq \varphi < 1$ represents the discount parameter, $\gamma > -\varphi$ denotes the concentration parameter, and H is the corresponding base distribution. Please notice that when $\varphi = 0$, we obtain a Dirichlet process with concentration parameter γ. In contrast with the Dirichlet process, Pitman–Yor process contains a power-law behavior when $0 < \varphi < 1$ [22] which makes it more suitable for modeling data describing natural phenomena. Similar to the Dirichlet process mixture model, the Pitman–Yor process mixture can also be considered as an infinite mixture model via a stick-breaking construction:

$$\begin{aligned} G &= \sum_{j=1}^{\infty} \pi'_j \delta_{\theta_j}, & \pi'_j &= \lambda_j \prod_{s=1}^{j-1}(1 - \lambda_s), \\ \lambda_j &\sim \mathrm{Beta}(1 - \varphi, \gamma + j\varphi), & \theta_j &\sim H, \end{aligned} \tag{49}$$

Table 1 Parameters of the generated synthetic data sets

	j	N_j (million)	α_{j1}	β_{j1}	α_{j2}	β_{j2}	π_j
Data set 1	1	0.5	12	17	20	15	0.50
	2	0.5	22	20	36	43	0.50
Data set 2	1	1.0	12	17	20	15	0.25
	2	1.0	22	20	36	43	0.25
	3	2.0	15	30	15	20	0.50

N_j denotes the number of data points in cluster j

Table 2 Estimated parameters of the generated data sets by the proposed *InGD-Fs*

	j	N_j (million)	$\hat{\alpha}_{j1}$	$\hat{\beta}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\beta}_{j2}$	$\hat{\pi}_j$
Data set 1	1	0.5	11.59	17.83	19.27	15.51	0.492
	2	0.5	22.94	20.36	35.05	44.17	0.508
Data set 2	1	1.0	12.38	16.85	20.37	15.33	0.254
	2	1.0	22.63	20.62	34.91	42.18	0.252
	3	2.0	14.87	31.11	15.21	20.64	0.494

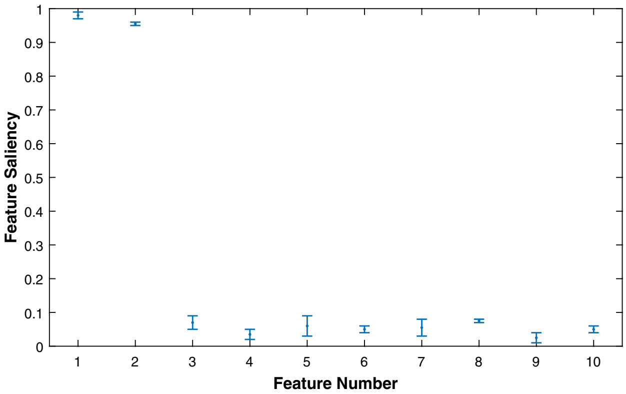

Fig. 1 Average feature saliences using *InGD-Fs* on the first synthetic data set over 10 runs

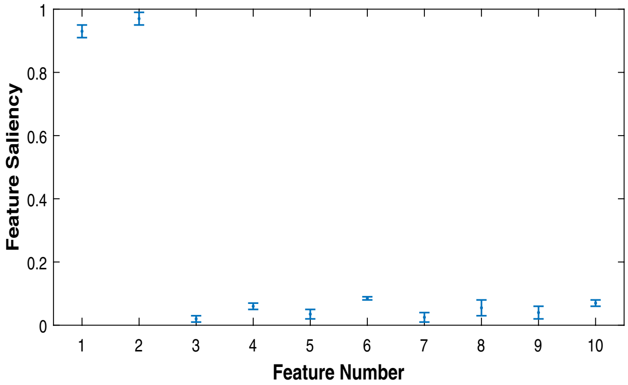

Fig. 2 Average feature saliences using *InGD-Fs* on the second synthetic data set over 10 runs

where δ_{θ_j} is an atom at θ_j. The variables π'_j represent the mixing weights where $\sum_{j=1}^{\infty}\pi'_j = 1$.

By adopting the Pitman–Yor process mixture model with stick-breaking construction, the infinite GD mixture model with unsupervised feature selection method can be defined as

$$p(\mathbf{X}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\alpha}',\boldsymbol{\beta}',\boldsymbol{\pi}',\boldsymbol{\phi}) = \sum_{j=1}^{\infty}\pi'_j\prod_{l=1}^{D}\left[\mathrm{Beta}(X_l|\alpha_{jl},\beta_{jl})\right]^{\phi_l}\left[\mathrm{Beta}(X_l|\alpha'_l,\beta'_l)\right]^{1-\phi_l}. \tag{50}$$

This model can also be learned using the accelerated variational Bayes approach as proposed in Sect. 3.2.

5 Experimental results

In this section, we illustrate the utility of the infinite GD mixture with unsupervised feature selection that is learned using accelerated variational Bayes (referred to as *InGD-Fs*) using both synthetic data sets and real-world applications with high-dimensional data. We start our experiments by validating the proposed *InGD-Fs* on two synthetic data sets with different settings of parameters. Then, we test our algorithm on several real data sets with different characteristics from the UCI Machine Learning Repository.[1] Lastly, we apply our algorithm on a challenging application concerns human action recognition using bag-of-visual words representation. For the real-world data sets and the application of human action recognition, we also apply the infinite GD mixture with unsupervised feature selection that is constructed using Pitman–Yor process mixture model (referred to as *InGD-FsPY*) as presented in Sect. 4. The initialization of the proposed *InGD-Fs* is summarized as follows: $(u_{jl}, v_{jl}, g_{jl}, h_{jl}, u'_l, v'_l, g'_l, h'_l, \xi_j) = (1, 0.05, 1, 0.05, 0.5, 0.01, 0.5, 0.01, 0.1)$.

5.1 Synthetic data sets

In this section, we validate the effectiveness of the proposed accelerated variational Bayes algorithm for learning *InGD-Fs* through synthetic data sets. Another target of this section is to demonstrate the advantages of accelerated variational Bayes for learning large-scale data sets by comparing it with the conventional variational Bayes learning algorithm.

[1] http://archive.ics.uci.edu/ml/.

Table 3 The average computational runtime (in seconds) for each tested algorithm over 10 runs

Algorithm	Data set 1	Data set 2
InGD-Fs	463.32 (4.18)	1537.25 (3.22)
varInGD-Fs	2879.51 (3.52)	9943.65 (2.56)

The numbers in parentheses are the standard deviation of the corresponding quantities

Table 4 The four real-world data sets

Data set	*N*	*D*	*M*
Wine (WI)	178	13	3
Statlog (ST)	6435	36	6
Image segmentation (IS)	2310	19	7
Handwritten digits (HD)	5620	64	10

N, *D* and *M* denote the numbers of instances, features and classes, respectively

First, we evaluate the performance of the *InGD-Fs* on two ten-dimensional (two relevant features and eight irrelevant ones) synthetic data sets. We generate the relevant features in the transformed space from mixtures of Beta distributions with well-separated components as described in Sect. 2. The parameters for generating these two data sets with relevant features can be viewed in Table 1. Irrelevant features are generated from one common Beta distribution Beta$(1, 2)$. Table 2 shows the estimated parameters of the distributions representing the relevant features for each data set using the proposed *InGD-Fs*. Based on the results shown in this table, for each synthetic data set, the parameters representing relevant features of this model, and its mixing coefficients can be accurately estimated by *InGD-Fs*. Figures 1 and 2 demonstrate the results of the saliencies of all 10 features for the synthetic data sets over ten runs. Clearly, for each data set, features 1 and 2 are considered as relevant features in terms of high degree of relevance ($\epsilon_l > 0.9$), whereas features from 3 to 10 are recognized as irrelevant features due to low values of saliency ($\epsilon_l < 0.1$), which is consistent with the true setting.

Moreover, in order to demonstrate the advantages of the developed accelerated variational Bayes algorithm, we compare it with the conventional variational Bayes algorithm for learning the infinite GD mixture model with feature selection (referred to as *varInGD-Fs*), in terms of computational runtime. The corresponding results are shown in Table 3. It is clear that, for each synthetic data set, the proposed *InGD-Fs* requires less computational time than *varInGD-Fs*.

Table 5 The average error rates (%) using different algorithms over 30 random runs

Method	WI	ST	IS	HD
Kurihara et al. [15]	8.73 (0.97)	11.26 (1.05)	19.75 (1.23)	16.45 (1.18)
Fan et al. [10]	4.56 (0.58)	8.81 (0.85)	14.33 (1.75)	10.01 (1.21)
InGD-Fs	4.49 (0.82)	8.15 (1.12)	13.87 (1.46)	9.73 (1.63)
InGD-FsPY	4.41 (0.76)	8.04 (0.96)	10.09 (1.51)	9.69 (1.34)

The numbers in parentheses are the standard deviation of the corresponding quantities

Table 6 The average computational runtime (in seconds) for each tested algorithm over 30 runs

Method	WI	ST	IS	HD
Kurihara et al. [15]	2.46	19.52	10.19	18.36
Fan et al. [10]	7.39	61.41	28.67	56.71
InGD-Fs	2.68	22.34	11.95	20.15
InGD-FsPY	2.77	24.03	12.23	21.26

5.2 Real-world data sets

In this experiment, the proposed *InGD-Fs* and *InGD-FsPY* were tested on clustering four real-world data sets with different properties from the UCI Machine Learning Repository: **(1) The wine (WI) data set:** This data set is collected from a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. It contains 178 data instances, including 3 types of wines with 13 constituents found in each type; **(2) The statlog (ST) data set:** It consists of the multi-spectral values of pixels in 3×3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. In total, there are 6435 36-dimensional vectors from six classes: read soil, cotton crop, gray soil, damp gray soil, soil with vegetation stubble and very damp gray soil. The goal is to predict this classification, given the multi-spectral values; **(3) the image segmentation (IS) data set:** It contains 2310 data instances collected from a database of 7 outdoor image classes: brickface, sky, foliage, cement, window, path and grass. Each instance is a 3×3 region with 19 features; **(4) The handwritten digits (HD) data set:** This data set has 5620 data instances in total with 64 features from 10 classes: '0' to '9.' Features in this data set are integers in the range 0–16. The properties of aforementioned data sets are summarized in Table 4.

It is noteworthy that since we were performing clustering analysis, the class labels were not involved in this experiment. Moreover, all features in those data sets were normalized into the range of [0,1] as a preprocessing step.

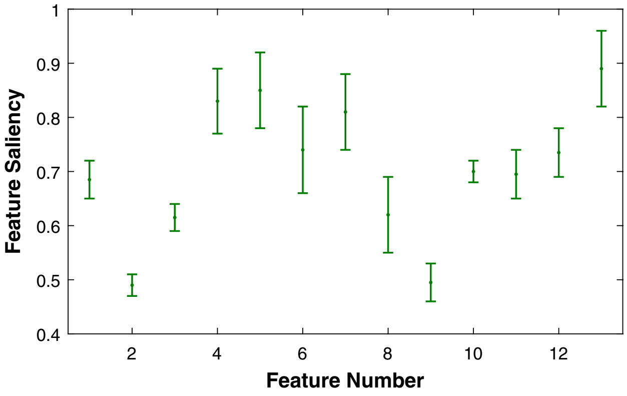

Fig. 3 Average feature saliences using *InGD-Fs* on the wine data set over 30 runs

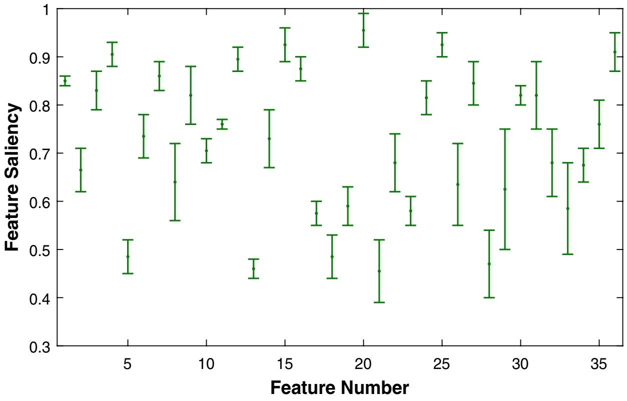

Fig. 4 Average feature saliences using *InGD-Fs* on the statlog data set over 30 runs

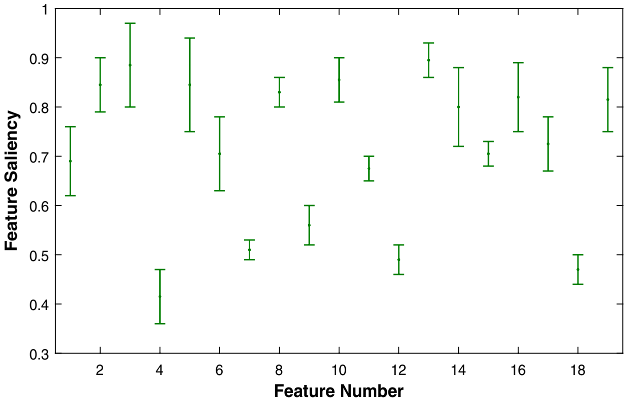

Fig. 5 Average feature saliences using *InGD-Fs* on the image segmentation data set over 30 runs

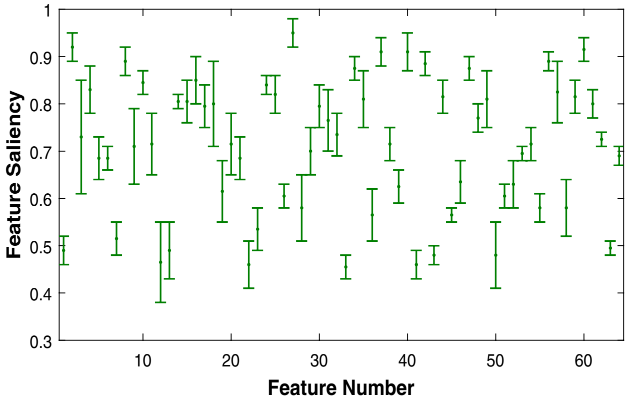

Fig. 6 Average feature saliences using *InGD-Fs* on the handwritten digits data set over 30 runs

We randomly partitioned each data set into two parts: one for training, another one for testing. The evaluation of the proposed *InGD-Fs* and *InGD-FsPY* was performed based on 30 runs. The advantages of the proposed algorithms were demonstrated by comparing them with two other state-of-the-art mixture modeling approaches: the infinite Gaussian mixture model with kd-tree structure as proposed in [15], the infinite GD mixture model with feature selection through conventional variational Bayes learning as proposed in [10].

The average results are summarized in Table 5 in terms of error rates. Based on the results shown in this table, we can observe that for all data sets the infinite Gaussian mixture model as proposed in [15] has obtained the worst performance in terms of the highest error rate among all tested algorithms. This fact illustrates the merits of using feature selection technique, as well as the advantages of using the GD mixture models over Gaussian ones in modeling proportional data. Furthermore, for data sets WI, ST and HD, the proposed *InGD-Fs* and *InGD-FsPY* can provide comparable performance to the one obtained by the infinite GD mixture model with feature selection as proposed in [10]. According to Student's *t*-tests, with 95 percent confidence, the differences in performance among those three approaches are not statistically significant (i.e., for the WI, ST and HD data sets, we have obtained *p*-values between 0.176 and 0.349 for different runs). In this case, *InGD-Fs* and *InGD-FsPY* are preferred since they are based on the accelerated variational Bayes and are significantly faster than the approach of [10] as shown in Table 6. Another important observation from Table 5 is that, for the IS data set, *InGD-FsPY* achieved the best performance in terms of the lowest error rate (10.09%), and the difference between the *InGD-FsPY* and the other

Fig. 7 Sample frames from the HMDB51 human action database

approaches is statistically significant (p-values between 0.025 and 0.043). This can be explained by the fact that the frequencies of observed objects in each image segmentation follow power-law distributions [26], and thus can be better modeled using the Pitman–Yor process framework than using the Dirichlet process.

We illustrate the features saliencies obtained by the proposed *InGD-Fs* for each tested data sets in Figs. 3, 4, 5 and 6. As shown in these figures, it is obvious that different features do not contribute equally in the clustering analysis due to different associated relevance degrees. More specifically, for the wine data set, two features (features number 4, 5, 7 and 13) have the highest relevance degrees where the feature saliencies are greater than 0.8. By contrast, there are two features (feature number 2 and 9) that have saliencies lower than 0.5, and therefore contribute less to clustering. For the statlog data set, we have obtained twelve features (features number 1, 3, 4, 7, 9, 12, 15, 16, 20, 24, 25, 27, 30, 31 and 36) that have higher degree of relevance (feature saliencies are greater than 0.8), while five features (feature number 5, 13, 18, 21 and 28) that have saliencies lower than 0.5. For the image segmentation and handwritten data sets, we have obtained three (feature number 4, 12 and 18) and nine irrelevant features (feature number 1, 12, 13, 22, 33, 41, 43, 50 and 63), respectively, which have feature saliencies lower than 0.5.

5.3 Experimental on human action recognition

In this experiment, we apply the proposed algorithms on a challenging task in the filed of computer vision, namely human action videos recognition. Our goal is to develop a statistical approach for recognizing human action videos using *InGD-Fs* or *InGD-FsPY* and local spatiotemporal features using bag-of-visual words representation.

5.3.1 Methodology and data set

We summarize our methodology for recognizing human actions in videos as follows. (1) By calculating space–time interest points, local spatiotemporal features were calculated from each video. In this work, the Harris3D detector [16] is applied to obtain the HOG/HOF feature descriptors [17]. (2) we apply the K-means algorithm to quantize the obtained HOG/HOF features in to visual words via the paradigm of bag-of-visual words. As a result, each video sequence can then be considered as a histogram over visual words. In our experiments, we obtained the optimal recognition performance when the size of the visual vocabulary is about 1000 by investigating different sizes (300–1200). (3) The probabilistic latent semantic analysis model is then applied [12] on

Table 7 The average recognition rate and runtime with the standard deviations by different algorithms in 30 runs

Method	Rec. rate (%)	Runtime (s)
Zhang et al. [33]	25.17 (1.28)	63.52
Song et al. [25]	23.79 (1.51)	59.31
Kuehne et al. [14]	22.13 (1.86)	49.65
Kurihara et al. [15]	24.49 (1.15)	27.16
Fan et al. [10]	29.77 (1.07)	70.34
InGD-Fs	29.83 (1.54)	29.51
InGD-FsPY	30.21 (1.13)	31.29

The numbers in parenthesis are the standard deviation of the corresponding quantities

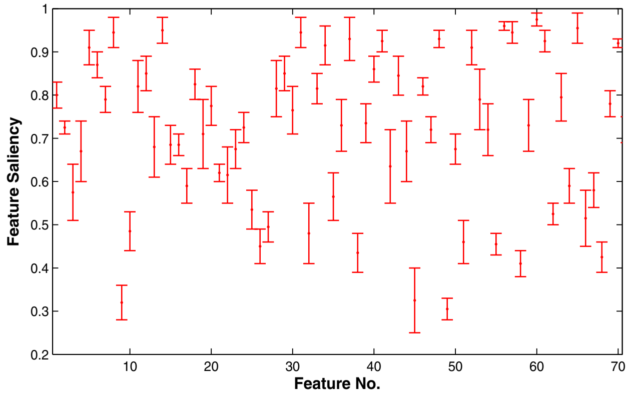

Fig. 8 Average feature saliences using *InGD-Fs* over 30 runs

the acquired histograms. Consequently, each video sequence is now represented as a proportional vector that the corresponding dimensionality may be considered as the number of latent aspects. In this work, the optimal performance was acquired when 70 aspects was considered.

Our recognition approach is developed based on a classifier. The inputs to the classifier are the 70-dimensional vectors extracted from the different action categories. These vectors are divided into two sets: the training set (50 vectors were randomly taken for training from each action category), whose class is known, and the testing set, whose class is unknown (i.e., unlabeled). The training set is used to adapt the classifier to each possible action category before the unknown set is submitted to the classifier. Then, we apply our algorithms, presented in Sects. 3 and 4, to the training vectors in each category. After this step, each action category in the data set is represented by an infinite GD mixture model. Finally, in the classification step, each testing vector (without labels) is assigned to the category that increases more its log-likelihood.

Our experiments were tested through one of the largest human action databases available nowadays known as the HMDB51 database [14],[2] which was collected from various sources (e.g., YouTube, movies, or Google videos). This database has 6849 video clips which can be classified into 51 action categories. Each category includes at least 101 video clips. Some examples of motion frames can be viewed in Fig. 7.

5.3.2 Results

We report the experimental results of the proposed *InGD-Fs* and *InGD-FsPY* based on 30 runs of our approach. For comparison, except the applications of the infinite Gaussian mixture model with kd-tree structure as proposed in [15] and the infinite GD mixture model proposed in [10], we also apply three other state-of-the-art approaches: the approach that is based on a boosted multi-class semi-supervised learning algorithm [33], the approach that is based on a localized, continuous and probabilistic video representation for human action recognition [25], and the action recognition approach as described in [14] where SVM with the RBF kernel is used for recognition. The average performance and the computational cost (in terms of computational runtime) are demonstrated in Table 7 for each approach. As shown in this table, both the proposed *InGD-Fs* and *InGD-FsPY* outperformed the other approaches with higher recognition rates. It is noteworthy that the approach of [10] performed slightly worse than the proposed two approaches, and its result was not significantly different from the proposed two approaches based on the Student's t-test. Specifically, with 95 percent confidence, we obtained p-values between 0.189 and 0.282 for different runs. Therefore, the *InGD-Fs* and *InGD-FsPY* were preferred in this case since they were significantly faster than the approach of [10] based on the results shown in Table 7, thanks to the accelerated variational learning with kd-tree structure.

We also evaluate the feature saliencies corresponding to the 70-dimensional aspects using *InGD-Fs* and present this result in Fig. 8. As shown in this figure, we have obtained different feature saliencies for different features. More specifically, 15 features were considered with high relevance degree since their feature saliencies were larger than 0.9. However, 12 features were considered having less contribution in recognition, since the resulted feature saliencies were smaller than 0.5.

6 Conclusion

In this paper, the Dirichlet process prior was used to provide nonparametric Bayesian estimates for generalized Dirichlet mixtures when used for simultaneous clustering and feature selection. This goal was achieved by providing an accelerated variational approach for model learning. Moreover, we have also proposed a construction of the infinite generalized Dirichlet mixture model using the framework of Pitman–Yor process, which can be considered as an extension to the infinite generalized Dirichlet mixture that is built through Dirichlet process mixture model. The experiments were based on the clustering analysis of several real-world data sets and the application of human activities recognition. The obtained results have shown the merits of our approach. Future works may include the extension of the proposed model to online settings. Another potential future work may be the inclusion of audio information, as done in [30], to improve distinguishing confusing human activities in videos.

[2] http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database.

Acknowledgements Funding was provided by National Natural Science Foundation of China (Grant No. 61502183), The Scientific Research Funds of Huaqiao University (Grant No. 600005-Z15Y0016) and Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (Grant No. ZQN-PY510).

References

1. Alfò M, Nieddu L, Vicari D (2008) A finite mixture model for image segmentation. Stat Comput 18(2):137–150
2. Bentley JL (1975) Multidimensional binary search trees used for associative searching. Commun ACM 18(9):509–517
3. Blei D, Jordan M (2005) Variational inference for Dirichlet process mixtures. Bayesian Anal 1:121–144
4. Bouguila N (2007) Spatial color image databases summarization. In: Proc. of the IEEE international conference on acoustics, speech and signal processing (ICASSP 2007), vol 1, pp I-953–I-956
5. Bouguila N, Ziou D (2004a) Improving content based image retrieval systems using finite multinomial Dirichlet mixture. In: Proc. of the 14th IEEE signal processing society workshop on machine learning for signal processing, pp 23–32
6. Bouguila N, Ziou D (2004b) A powerful finite mixture model based on the generalized Dirichlet distribution: unsupervised learning and applications. In: Proc. of the 17th international conference on pattern recognition (ICPR 2004), vol 1, pp 280–283 Vol 1
7. Bouguila N, Ziou D (2010) A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling. IEEE Trans Neural Netw 21(1):107–122
8. Boutemedjet S, Bouguila N, Ziou D (2009) A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering. IEEE Trans Pattern Anal Mach Intell 31(8):1429–1443
9. Constantinopoulos C, Titsias M, Likas A (2006) Bayesian feature and model selection for Gaussian mixture models. IEEE Trans Pattern Anal Mach Intell 28(6):1013–1018
10. Fan W, Bouguila N (2013) Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection. Pattern Recognit 46(10):2754–2769
11. Fan X, Cao L, Xu RYD (2015) Dynamic infinite mixed-membership stochastic blockmodel. IEEE Trans Neural Netw Learn Syst 26(9):2072–2085
12. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1/2):177–196
13. Korwar RM, Hollander M (1973) Contributions to the theory of Dirichlet processes. Ann Probab 1:705–711
14. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: Proc. of the international conference on computer vision (ICCV), pp 2556–2563
15. Kurihara K, Welling M, Vlassis N (2006) Accelerated variational Dirichlet process mixtures. In: Proc. of advances in neural information processing systems (NIPS)
16. Laptev I (2005) On space–time interest points. Int J Comput Vis 64(2/3):107–123
17. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proc. of IEEE conference on computer vision and pattern recognition (CVPR), pp 1–8
18. Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 26(9):1154–1166
19. McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York
20. Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9(2):249–265
21. Nguyen NT, Zheng G, Han Z, Zheng R (2011) Device fingerprinting to enhance wireless security using nonparametric Bayesian method. In: Proc. of the IEEE conference on INFOCOM, pp 1404–1412
22. Pitman J, Yor M (1997) The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. Ann Probab 25(2):855–900
23. Sethuraman J (1994) A constructive definition of Dirichlet priors. Stat Sin 4:639–650
24. Shyr A, Darrell T, Jordan M, Urtasun R (2011) Supervised hierarchical Pitman–Yor process for natural scene segmentation. In: Proc. of the 2011 IEEE conference on computer vision and pattern recognition (CVPR), pp 2281–2288
25. Song Y, Tang S, Zheng YT, Chua TS, Zhang Y, Lin S (2012) Exploring probabilistic localized video representation for human action recognition. Multimedia Tools and Applications 58(3):663–685
26. Sudderth EB, Jordan MI (2008) Shared segmentation of natural scenes using dependent Pitman-Yor processes. In: Proc. of Advances in Neural Information Processing Systems (NIPS), pp 1585–1592
27. Teh YW (2006) A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, pp 985–992
28. Walker SG (2007) Sampling the Dirichlet mixture model with slices. Communications in Statistics- Simulation and Computation 36:45–54
29. Walker SG, Gutierrez-Pena E (2007) Bayesian parametric inference in a nonparametric framework. Test 16:188–197
30. Wang T, Hammoud R, Zhu Z (2014) Ground-based activity recognition at distance and behind wall. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 231–236
31. Wei X, Li C (2012) The infinite student's t-mixture for robust modeling. Signal Processing 92(1):224–234
32. Wei X, Yang Z (2012) The infinite student's t-factor mixture analyzer for robust clustering and classification. Pattern Recognition 45(12):4346–4357
33. Zhang T, Liu S, Xu C, Lu H (2011) Boosted multi-class semi-supervised learning for human action recognition. Pattern Recognition 44(10):2334–2342

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.