CDP Mixture Models for Data Clustering

Yangfeng Ji, Tong Lin, Hongbin Zha
Key Laboratory of Machine Perception (Ministry of Eduction)
School of EECS, Peking University,
Beijing 100871, China
jiyf, lintong, zha@cis.pku.edu.cn

***Abstract*—In Dirichlet process (DP) mixture models, the number of components is implicitly determined by the sampling parameters of Dirichlet process. However, this kind of models usually produces lots of small mixture components when modeling real-world data, especially high-dimensional data. In this paper, we propose a new class of Dirichlet process mixture models with some constrained principles, named *constrained Dirichlet process (CDP) mixture models*. Based on general DP mixture models, we add a resampling step to obtain latent parameters. In this way, CDP mixture models can suppress noise and generate the compact patterns of the data. Experimental results on data clustering show the remarkable performance of the CDP mixture models.**

***Keywords*-Clustering, Dirichlet process, Dirichlet process mixture models, Gaussian mixture models**

I. Introduction

In data clustering and modeling, various mixture models are basic tools to model data and discover the patterns of data. In recent years, Dirichlet process (DP) mixture models [1]–[3], have received much attention in the machine learning community. The DP mixture model is one kind of mixture models with a Dirichlet process prior. With the nonparametric nature of the Dirichlet process, DP mixture models can generate countable components in modeling data. In contrast to finite mixture models such as Gaussian mixture models, DP mixture models do not need the number of components as a given parameter.

The DP mixture models have witnessed several successful applications [4]–[6], due to the flexibility of DP mixture models. Some problems of DP mixture models have been brought about when apply them to practical problems. For example, in clustering on high-dimensional data space, the DP mixture models always produce more components than that the real data should have. These small mixture components are mainly caused by noise or sparsity of high-dimensional data distribution.

Some approaches have been proposed for solving this problem. In [7], an upper bound of the number of components is fixed in advance to limit the number in modeling with mixture models. In [1], components with little data points are simply discarded, and these data points are reassigned to other existing components. However, these two methods based on simple thresholds can not be directly used for real world data, since choosing the best thresholds is usually difficult. Note that, another constrained Dirichlet process mixture model is also proposed in [8] for verb clustering, which is different with our model, because it considers some particular properties in natural language processing.

In this paper, we focus on the problem that how to obtain a reasonable mixture models in modeling data, especially high-dimensional and noisy data. We propose *constrained Dirichlet process* (CDP) mixture models, which impose some constraints to the DP mixture models. By introducing a resampling step, the CDP mixture models can handle the problem and provide more reasonable results in practice. In experiments, data clustering and motion segmentation show the better performance of the CDP mixture models than those of DP mixture models and finite mixture models.

II. Constrained Dirichlet Process Mixture Models

The Dirichlet process (DP) is a stochastic process whose sample paths are probability measures with probability one. For a random distribution G of a DP, its marginal distributions have to be Dirichlet. Specifically, let G_0 be a distribution over Θ, and α be a positive real number. Then for any finite and measurable partition $A_1, \cdots, A_r$ of Θ, if

$$(G(A_1), \cdots, G(A_r)) \sim \mathrm{Dir}(\alpha G_0(A_1), \cdots, \alpha G_0(A_r)),$$

then G is a Dirichlet process with base distribution G_0 and concentration parameter α, written as $G \sim \mathrm{DP}(\alpha, G_0)$.

With the Dirichlet process prior, we model a set of observations $X = \{x_1, \ldots, x_n\}$ using a set of latent parameters $\theta_1, \ldots, \theta_n$. Each θ_i is drawn independently and identically from G, while each x_i has distribution $F(\theta_i)$ parametrized by θ_i:

$$\begin{aligned} G|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0), \\ \theta_i|G &\sim G, \\ x_i|\theta_i &\sim F(\theta_i). \end{aligned} \tag{1}$$

Because G is discrete, some of $\{\theta_i\}_{i=1}^n$ could take on the same value simultaneously and the above model can be viewed as a mixture model, since x_i with the same value of θ_i belong to the same mixture component.

From stick-breaking construction, we can get another viewpoint of the DP mixture model. Let c_i be a variable

for component assignment (also known as latent variable in mixture models), which takes value k with probability π_k. DP mixture models can be equivalently expressed as following:

$$\begin{aligned} \pi|\alpha &\sim \mathrm{GEM}(\alpha), \\ c_i|\pi &\sim \mathrm{Multional}(\pi), \\ \phi_k|G_0 &\sim G_0, \\ x_i|c_i,\phi_k &\sim F(\phi_{c_i}), \end{aligned} \tag{2}$$

with $G = \sum_{k=1}^{\infty} \pi_k \delta(\phi_k)$ and $\theta_i = \phi_{c_i}$. In Eq.(2), π is the mixing proportion, ϕ_k are the mixture component parameter, $F(\phi_k)$ is the distribution over observations in mixture components k, and G_0 is the prior distribution over ϕ_k. Because the value of π_k decrease exponentially with increasing k, only a finite number of components will be used to model the data. In the DP mixture model, the actual number of mixture components used to model data is not fixed.

When modeling real-world data, especially high-dimensional data, using DP mixture models, there are always more mixture components than the mixture models where the data come from. Furthermore, some of these mixture components are employed to model noise. This is often unreasonable for many practical applications in data clustering. For example, when we segment one sequence of motion data using DP mixture models, noise in the data leads to more mixture components as illustrated in Figure 3. To cope with this kind of problems, we propose the CDP mixture model.

The CDP mixture model can be expressed as following:

$$\begin{aligned} G|\alpha,G_0 &\sim DP(\alpha,G_0), \\ \theta_i|G &\sim G, \\ \xi_i|\theta_i &\sim H(\theta_i), \\ x_i|\xi_i,\theta_i &\sim F(\xi_i). \end{aligned} \tag{3}$$

In CDP mixture model (3), latent variables $\{\xi_i\}$ is introduced with $\{\theta_i\}$ to indicate the distribution of $\{x_i\}$: $x_i|\xi_i,\theta_i \sim F(\xi_i), i = 1,\ldots,n$. H is a finite discrete distribution over distinct values of θ_i, where the probability of ξ_i are proportional to the number of data related to θ_i, and to the possibility of data belong to a mixture component. Using the probability distribution H, we can sample the parameter ξ_i which can be viewed as resampling from θ_i. This resampling procedure make the value of ξ_i concentrate on a small number of θ_i. Consequently, the data concentrate on a small number of mixture models.

A. Thrifty Chinese Restaurant Process with Finite Customers

Using DP mixture model for cluster analysis, the number of mixture components can be determined automatically. The clustering effect can be viewed as a partition of integers, and the distribution over partitions is called Chinese restaurant process. To get better understanding, suppose that there is a Chinese restaurant with an infinite number of tables. Each table can allow an infinite number of customers. The first customer enters the restaurant and chooses any one table to sit. The second customer enters and decides either to sit with the first customer, or to sit at a new table. In general, the nst customer either joins an already occupied table k with probability proportional to the number n_k of customers already sitting there, or sits at a new table with the probability proportional to α without further restricts.

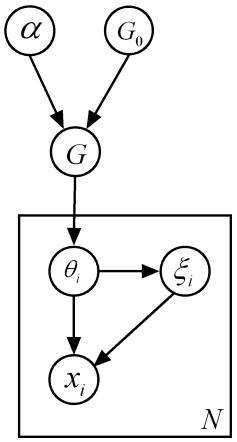

Figure 1. Graphical illustration of the CDP mixture model.

However, in an thrifty restaurant, the manager of this restaurant always wants to vacate some tables for new customers and utilize only a small number of tables. Suppose that N customers have occupied K tables currently. The manager may specify only $K'(< K)$ tables to be served, and then each customer should re-choose his/her seat among these K' tables.

III. Inference

Inference algorithms on DP mixture models include some sampling methods (mainly the Markov chain Monte Carlo algorithm) [3], [9], and also some variational inference algorithms [10]. For our CDP mixture models, we proposed a sampling method based on the algorithms on DP mixture models to infer the latent variables. According to model (3), the main objective of inference is to sample latent variables θ_is and ξ_is from posterior distribution with Dirichlet process prior.

The approach to sample θ_i is to repeatedly draw samples from its conditional distribution given the data $X = \{x_1,\cdots,x_n\}$ and $\theta_j (j \neq i)$ (written as θ_{-i} for short). The likelihood for θ_i is $F_{x_i}(\theta_i)$ since x_i is under distribution $F(\theta_i)$. The prior distribution of θ_i is

$$\theta_i|\theta_{-i} \sim \frac{1}{n-1+\alpha}\left\{\sum_{j\neq i}\delta(\theta_j) + \alpha G_0\right\}. \tag{4}$$

When integrating the likelihood with the prior, we have the following conditional distribution

$$\theta_i|\theta_{-i}, x_i \sim b\left\{\sum_{j\neq i} F_{x_i}(\theta_j)\delta(\theta_j) + H_i\alpha\int F_{x_i}(\theta)dG_0(\theta)\right\}. \quad (5)$$

where b a is normalizing constant, H_i is the posterior distribution of θ, which can be obtained from prior distribution G_0 and likelihood $F_{x_i}(\theta)$. In CDP mixture models, the integral in Eq.(5) is not analytically tractable. We adopt algorithms proposed in [3] to tackle this problem.

Here, $\Xi^* = \{\xi_1^*, \ldots, \xi_T^*\}$ is defined the set of distinct values of $\{\xi_i\}_{i=1}^n$. In the CDP mixture model, we need to resample parameters ξ_i from Ξ^*. The posterior distribution of ξ_i of data x_i is given as

$$\xi_i|\xi_{-i} \sim \sum_{j\neq i,\ \xi_j\in\Xi} F_{x_i}(\xi_j)\delta(\xi_j), \quad (6)$$

The resampling procedure is illustrated in Algorithm 1. The

Algorithm 1 Resampling process

Input: $X = \{x_i\}_{i=1}^N$, $\Xi^* = \{\xi_t^*\}_{t=1}^T$
for $i = 1$ **to** N **do**
  for $t = 1$ **to** T **do**
    Given ξ_t^*, $p_t = F_{x_i}(\xi_t^*)$
  end for
  $t = \text{argmax}_t\{p_1, \cdots, p_T\}$
  $c_i = t$
  $\xi_i = \xi_t^*$
end for

value of T in Algorithm 1 is computed through the following optimization problem:

$$T = \min_t \sum_{k=1}^K t_k, \quad s.t. \sum_{k=1}^K t_k N_k \geq \kappa N, \quad (7)$$

where $t = \{t_k\}_{k=1}^K \in \{0,1\}^K$ is a K-dimensional vector, N_k is the number of data belong to the kth mixture component, and $0 < \kappa < 1$. Note that if $\kappa = 1$, then $T = K$ and CDP mixture models is equivalent to DP mixture models.

IV. Experimental results

In experiments, we test our algorithm on real data with noise. In order to give an intuitive illustration, the CDP mixture models and DP mixture models were employed for clustering the old faithful geyser data. Then, the CDP mixture models were employed for motion segmentation, which can be viewed as clustering sequential data in high-dimensional space.

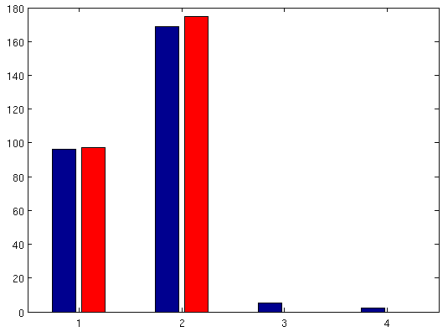

Figure 2. Clustering results using DP mixture models (blue bars) and CDP mixture models (red bars) respectively for Old Faithful Geyser data. The horizontal axis shows each component, and the vertical axis shows the number of data points in each component.

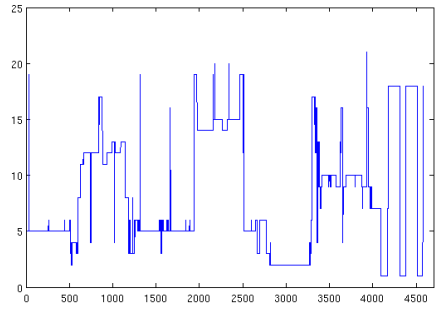

Figure 3. Motion segmentation on one sequence using a DP mixture model. Due to the noise in data, there are more than 20 mixture components in this DP mixture model.

A. Old Faithful Geyser data

The Old Faithful Geyser data is widely used in the task of clustering to show the performance of clustering algorithms. Here, we compare the results of clustering on these data using DP mixture models and CDP mixture models respectively.

Figure 2 shows the components obtained from the DP mixture model and the CDP mixture model. We can see that there are four mixture components in the DP mixture model and only two mixture models in the CDP mixture model. Therefore, we can see that the CDP mixture model successfully suppress small components, and generate better result than that of the DP mixture model.

B. Clustering for Motion Segmentation

Motion data from motion capture has frequently used in motion analysis in computer vision. Motion segmentation is one of the basic approaches to analyze motion behaviors in data from motion capture.

Currently, there are several method for motion segmentation based on statistical modeling and machine learning. In [11], Gaussian mixture models is employed for segmentation. As one kind of finite mixture models, the problem

Method	Precision	Recall
GMM [11]	0.77	0.71
CDPMM	0.86	0.82

Table I
THE PRECISION AND RECALL SCORES OF GAUSSIAN MIXTURE MODEL (GMM) AND CDP MIXTURE MODEL (CDPMM) FOR MOTION SEGMENTATION

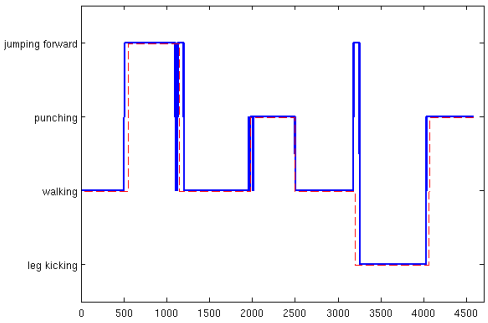

Figure 4. Motion segmentation on one sequence using CDP mixture models (Blue line: the result from CDP mixture model, Red line: hand-labeling ground truth). There are four kinds of motion in this sequence: jumping forward, punching, walking and lag kicking.

is the number of mixture components should be given at first as a input parameter. However, in dealing with practical problems, it is difficult to determine this parameter.

In this experiment, all the sequences are obtained from motion capture equipments at 120 frames per second. For comparison with [11], we use the same motion sequences. In preprocessing step, we use PCA to project the frames onto a lower dimensional subspace. The number of principal components is chosen so that 90% of the variance of the original data is preserved.

The results of motion segmentation on one sequence are illustrated in Figure 3 and 4, where the performance of DP mixture models and CDP mixture models are compared. In Figure 4, the performance of segmentation using CDP mixture models (blue line) is compared with hand-labeling ground-truth segmentation (red line). For this CDP mixture model, only a small number of frames are assigned to error clusters, mainly in the period of motion transitions.

The CDP mixture models were tested on 14 sequences, which consists of approximate $5000 \sim 8000$ frames respectively. Each sequence is a series of different motions, including walking, running, climbing, etc. The overall results in Table 1 show the better performance of CDP mixture models than that of Gaussian mixture models [11].

V. CONCLUSION

In this paper, we proposed CDP mixture models to model the high-dimensional data with noise. For this kind of data, the DP mixture models always trend to produce additional small components. Experiments on real-world data demonstrated that our CDP mixture models can achieve better performance than general DP mixture models and finite mixture models.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful suggestions and Dr. Wei Ma for the comments of this paper. This work was partially supported by the National Science Foundation of China(NSFC) Grants 60775006, the NHTRDP 863 Grant No. 2009AA01Z329 and the NHTRDP 863 Grant No. 2009AA012105.

REFERENCES

[1] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, June 1995.

[2] S. N. MacEachern and P. Muller, "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, pp. 223–238, 1998.

[3] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, June 2000.

[4] G. Boccignone, "Nonparametric Bayesian attentive video analysis," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2008.

[5] S. N. MacEachern and P. Muller, "Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models," in *Robust Bayesian Analysis*. Springer, 2000.

[6] P. Muller and F. A. Quintana, "Nonparametric Bayesian data analysis," *Statistical Science*, vol. 19, pp. 95–110, 2004.

[7] P. McCullagh and J. Yang, "How many clusters?" *Bayesian Analysis*, vol. 3, no. 1, pp. 101–120, 2008.

[8] A. Vlachos, A. Korhonen, and Z. Ghahramani, "Unsupervised and constrained Dirichlet process mixture models for verb clustering," in *EACL workshop on GEometrical Models of Natural Language Semantics*, 2009.

[9] M. D. Escobar, "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 268–277, Mar. 1994.

[10] D. Blei and M. Jordan, "Variational inference for Dirichlet process mixtures," *Journal of Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[11] J. Barbic, A. Safonova, J. Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *Graphics Interface*, May 2004. [Online]. Available: http://graphics.cs.cmu.edu/projects/segmentation/