分层 Dirichlet 过程及其应用综述

周建英 [1] 王飞跃 [1] 曾大军 [1]

摘 要 Dirichlet 过程是一种应用于非参数贝叶斯模型中的随机过程, 特别是作为先验分布应用在概率图模型中. 与传统的参数模型相比, Dirichlet 过程的应用更加广泛且模型更加灵活, 特别是应用于聚类问题时, 该过程能够自动确定聚类数目和生成聚类中心的分布参数. 因此, 近年来, 在理论和应用上均得到了迅速的发展, 引起越来越多的关注. 本文首先介绍 Dirichlet 过程, 而后描述了以 Dirichlet 过程为先验分布的 Dirichlet 过程混合模型及其应用, 接着概述分层 Dirichlet 过程及其在相关算法构造中的应用, 最后对分层 Dirichlet 过程的理论和应用进行了总结, 并对未来的发展方向作了探讨.

关键词 Dirichlet 过程, 概率图模型, 聚类, 分层 Dirichlet 过程

DOI 10.3724/SP.J.1004.2011.00389

Hierarchical Dirichlet Processes and Their Applications: A Survey

ZHOU Jian-Ying[1] WANG Fei-Yue[1] ZENG Da-Jun[1]

Abstract Dirichlet processes are a type of stochastic processes widely used in nonparametric Bayesian models, especially in research that involves probabilistic graphical models. Over the past few years, significant effort has been made in the study of such processes, mainly due to their modeling flexibility and wide applicability. For instance, Dirichlet processes are capable of learning the number of clusters as well as the corresponding parameters of each cluster whereas other clustering or classification models usually are not able to. In this survey, we first introduce the definitions of Dirichlet processes. We then present Dirichlet process mixture models and their applications, and discuss in detail hierarchical Dirichlet processes (HDP), their roles in constructing other models, and examples of related applications in many important fields. Finally, we summarize recent developments in the study and applications of hierarchical Dirichlet processes and offer our remarks on future research.

Key words Dirichlet processes, probabilistic graphical models, clustering, hierarchical Dirichlet processes (HDP)

机器学习是人工智能的核心研究领域之一, 其主要目的是通过经验来提高某任务处理的性能[1]. 在实际应用中, "经验" 通常是以某些格式的数据存在的, 因此机器学习要完成的工作是从 (观测) 数据 (样本) 出发挖掘数据的潜在规律, 利用这些规律对未来数据或无法观测的数据进行预测判断. 目前机器学习在搜索引擎、网络安全监控、国防安全、交通安全和电子商务系统等等很多领域中均已得到了广泛的应用.

机器学习的这些应用主要是基于统计学习理论的, 即其假设训练样本将数据的所有信息实现统计描述, 基于训练样本得到的模型可以对未来新的数据进行描述和预测判断等. 但是随着各种应用需求的增加, 人们获取数据的手段和工具也越来越强大, 数据的数量、维数和类型也在增多, 进行数据的聚类或分类学习时, 利用有限参数对数据建模将变得不可靠. 同时人为的标注和分类很难将各种类别都考虑到, 而新的数据中也可能会有未知类型出现, 依靠训练样本得到的模型将无法继续应用在新的数据建模中. 因此, 传统的统计学习算法将不再适用, 我们需要一种模型, 这种模型不依赖于训练样本, 而且随着数据的变化, 模型能够实现自适应的变化, 实现模型的参数学习和分类数目自动更新等任务.

在这种需求下, 近年来, Dirichlet 过程得到了广泛的关注和发展[2−5], 以 Dirichlet 过程为先验分布的各种非参数贝叶斯模型在文本处理、视频监控数据处理和静态图像内容理解以及统计认知的研究应用等领域均得到了应用[6−13].

文献 [14−15] 综述性地介绍了 Dirichlet 过程作为先验分布的各种模型在自然语言处理中的应用情况, 与之不同, 本文主要以 Dirichlet 过程的发展和应用为主线, 第 1 节给出了 Dirichlet 过程的定义和其概率图表示, 并简要叙述了 Dirichlet 过程的应用. 第 2 节概述了基于 Dirichlet 过程发展起来的分层 Dirichlet 过程 (Hierarchical Dirichlet process, HDP), 并给出关于 HDP 模型的实验分析. 第 3 节引出了 HDP 作为先验分布的其他非参数贝叶斯模型, 以目前受到广泛关注和发展的分层 Dirichlet 过程

收稿日期 2010-07-05 录用日期 2010-12-02
Manuscript received July 5, 2010; accepted December 2, 2010
国家自然科学基金 (70890084, 60921061, 71025001) 资助
Supported by National Natural Science Foundation of China (70890084, 60921061, 71025001)
1. 中国科学院自动化研究所复杂系统智能控制与管理国家重点实验室 (筹) 北京 100190
1. State Key Laboratory for Intelligent Control and Management of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

— 隐马尔科夫模型 (HDP-Hidden Markov models, HDP-HMM) 为例进行重点介绍. 第 4 节概述了以 HDP 模型为先验分布的各种算法以及 HDP-HMM 等各种算法在视频图像数据处理和文本处理等领域的相关应用. 第 5 节简要概括了目前比较受关注的与 HDP 功能类似的算法 LDA (Latent Dirichlet allocation) 模型, 并与 HDP 比较, 阐述两种模型之间的区别和联系. 最后第 6 节对基于 Dirichlet 过程的非参数贝叶斯模型提出了作者认为在算法和应用中有待突破的几点, 给出本文的结论.

1 Dirichlet 过程

Dirichlet 过程 (Dirichlet process) 是一种应用于非参数贝叶斯模型中的随机过程. 在 1973 年, Ferguson 给出 Dirichlet 过程的定义, 并证明 Dirichlet 过程以概率 1 离散[16].

Dirichlet 过程的定义如下[5, 16]: 假设 G_0 是测度空间 Θ 上的随机概率分布, 参数 α_0 是正实数, 空间 Θ 上的概率分布 G 如果满足以下条件:

对测度空间 Θ 的任意一个有限划分 $A_1, \cdots, A_r$, 均有以下关系存在:

$$(G(A_1), \cdots, G(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r)) \quad (1)$$

则 G 服从由基分布 G_0 和 Concentration 参数 α_0 组成的 Dirichlet 过程, 即

$$G \sim \mathrm{DP}(\alpha_0, G_0) \quad (2)$$

反之, 如果满足 $G \sim \mathrm{DP}(\alpha_0, G_0)$, 则有式 (1) 成立.

因此, Dirichlet 过程是关于分布的分布, 即 Dirichlet 过程的每个采样本身即是一个随机分布. 该随机过程之所以被称为 Dirichlet 过程, 是因为其任意有限维边缘分布均是 Dirichlet 分布. 这和另外一种常见的随机过程 — 高斯过程相似, 在高斯过程中, 其任意有限维边缘分布均是高斯分布. 从 Dirichlet 过程中采样得到的分布是可数无限个离散概率, 无法用有限数量的参数描述, 因此 Dirichlet 过程是非参数模型[3].

Dirichlet 过程和 Dirichlet 分布一样具有共轭性. 若 $G \sim \mathrm{DP}(\alpha_0, G_0)$, 样本 $\theta \sim G$, 则对测度空间的有限划分的后验分布也为 Dirichlet 过程, 观测数据只影响其所在划分区域的分布参数[17], 即

$$(G(A_1), \cdots, G(A_r)|\theta \in A_k) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_k) + 1, \cdots, \alpha_0 G_0(A_r))$$

仅仅根据 Dirichlet 过程的定义, 无法实现 Dirichlet 过程的采样, 但是 Dirichlet 过程的三种不同形式的构造使得 Dirichlet 过程的应用成为可能, 下面我们将分别对这三种构造给予介绍.

1.1 Stick-breaking 构造

基于相互独立的变量序列 $(\beta_k)_{k=1}^{\infty}$ 和 $(\phi_k)_{k=1}^{\infty}$ 的 Stick-breaking 构造:

$$\beta_k|\alpha_0, G_0 \sim \mathrm{Beta}(1, \alpha_0), \;\; \phi_k|\alpha_0, G_0 \sim G_0$$

定义一随机概率分布 G 如下:

$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l), \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \quad (3)$$

此处, $\mathrm{Beta}(1, \alpha_0)$ 是常见的 Beta 分布, δ_ϕ 是 ϕ 点的概率测度. Sethuraman 证明如此构造的分布函数 G 是服从 Dirichlet 过程 $\mathrm{DP}(\alpha_0, G_0)$ 分布的一个随机分布[18].

式 (3) 中, $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ 以概率 1 满足 $\sum_{k=1}^{\infty} \pi_k = 1$. 因此, 可以将 $\boldsymbol{\pi}$ 视为关于正整数的随机概率分布[5]. 通常, 用 $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0)$ 表示权重系数的构造关系 (GEM 分别是 Griffiths, Engen 和 McCloskey 的首字母[19]).

Stick-breaking 构造可以形象地解释为[20]: 对单位长度的棒在比例 β_1 处切割, 并将切掉的这部分长度赋值给 π_1, 而后对剩余长度为 $(1-\beta_1)$ 的棒在其比例 β_2 处切割, 并将切掉的棒的长度赋值给 π_2, 而后按照相同的方式对剩余的棒在比例 β_k 处切割, 并将切掉的棒的长度赋值给 π_k. 图 1 和图 2 分别是 Stick-breaking 的构造示意图和构造中前 30 个权重系数一次采样结果.

Dirichlet 过程的 Stick-breaking 构造体现了其离散的性质, 正是由于 Stick-breaking 构造, 与 Dirchlet 过程的相关的很多算法才得以采样或者变分实现, 这在后面的介绍中将陆续体现出来.

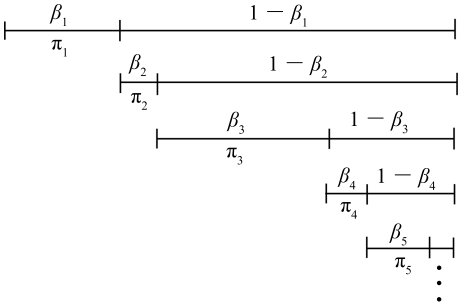
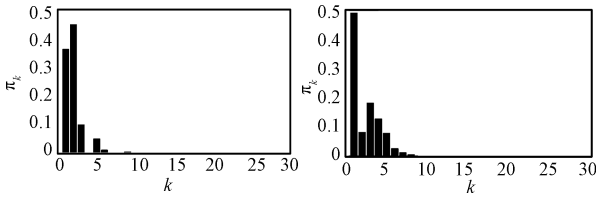


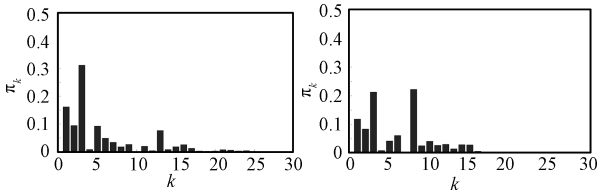
图 1 Stick-breaking 构造中权重系数的生成示意图[20]

Fig. 1 Sequential weights generated in the stick-breaking construction[20]



(a) $\alpha_0 = 1$ 时权重系数 1　(b) $\alpha_0 = 1$ 时权重系数 2

(a) The first sampling of weights with $\alpha_0 = 1$　(b) The second sampling of weights with $\alpha_0 = 1$



(c) $\alpha_0 = 5$ 时权重系数 1　(d) $\alpha_0 = 5$ 时权重系数 2

(c) The first sampling of weights with $\alpha_0 = 5$　(d) The second sampling of weights with $\alpha_0 = 5$

图 2　Dirichlet 过程的 Stick-breaking 构造中, 前 30 个权重四次随机构造结果

Fig. 2　The first 30 mixture weights generated by four random stick-breaking construction of Dirichlet process

1.2　Polya urn scheme 构造

Dirichlet 过程的另外一种构造形式是 Polya urn 模型[21]. 该模型显示, Dirichlet 过程的采样不仅是离散的, 而且具有很好的聚类性质.

Polya urn 模型并不直接研究 G, 而是研究采样于 G 的样本. 令 $\theta_1, \theta_2, \cdots$ 是服从分布 G 的独立同分布的随机变量序列, 即 $\theta_1, \theta_2, \cdots$ 关于 G 条件独立, 因此, 变量序列是可交换的[22−23]. θ_i 关于其他变量 $\theta_1, \cdots, \theta_{i-1}$ 的条件分布具有以下形式[5, 21]:

$$\theta_i|\theta_1, \cdots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{l}^{i-1} \frac{1}{i-1+\alpha_0}\delta_{\theta_l} + \frac{\alpha_0}{i-1+\alpha_0}G_0 \quad (4)$$

我们可以用 Polya urn 来解释上述条件分布, 假设有一个罐子, 罐中装有很多球, 每个球对应一个不同的颜色, 最初球的颜色是等概率选择的. 随机从罐中取出一个球, 而后将该球和另外一个相同颜色的球放进罐中. 同时, 以正比例于 α_0 的概率取一个新球, 新球的颜色服从 G_0 分布[5].

从条件分布中可以看到, 某个颜色的球被取出的概率和该颜色的球的总数成正比, 某个颜色的球被取出的越多, 下一次取球时, 取到该颜色的球的几率就越大. 定义相互不同的变量序列 $\phi_1, \cdots, \phi_K$ 表示 $\theta_1, \cdots, \theta_{i-1}$ 中取的相互不同的值, 即球的不同颜色, 令 m_k 表示 $\theta_1, \cdots, \theta_{i-1}$ 序列中其值等于 ϕ_k 的 θ_l, $1 \le l \le i-1$ 的个数. 那么式 (4) 可以表示为[5]

$$\theta_i|\theta_1, \cdots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k}^{K} \frac{m_k}{i-1+\alpha_0}\delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0}G_0$$

从 Polya urn 模型中可以看到, Dirichlet 过程呈现了很好的聚类性质, 其将具有相同值的随机变量聚为一类.

1.3　Chinese restaurant process 构造

和 Polya urn 模型类似, CRP (Chinese restaurant process)[24] 为 Dirichlet 过程的构造提供了另外一种途径. 构造如下[5]: 考虑一中国菜餐厅, 可以容纳无限多张桌子. θ_i 被比作进入餐厅的顾客, 而不同的值 ϕ_k 对应顾客就座的桌子, 第一个顾客就座于第一张桌子, 第 i 个顾客以正比于已经就座于第 k 张桌子 ϕ_k 的顾客数 m_k 的概率就座于第 k 张桌子, 以正比于 α_0 的概率就座一张新桌子, 即 K 增加 1, 而 $\phi_K \sim G_0$, $\theta_i = \phi_K$.

图 3 为 CRP 的构造, 其中大圆表示餐桌, 其唯一代号为 ϕ_k, 周围的 θ_i 是就座的顾客.

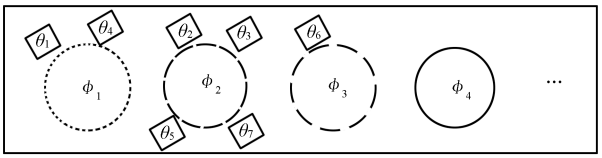


图 3　CRP 构造

Fig. 3　Chinese restaurant process

令第 i 个参数变量 θ_i 的指示因子为 z_i, 即 $\theta_i = \phi_{z_i}$, 可以得到[20]

$$z_i|z_1, \cdots, z_{i-1}, \alpha_0,\ G_0 \sim \sum_{k}^{K} \frac{m_k}{i-1+\alpha_0}\delta(z_i, k) + \frac{\alpha_0}{i-1+\alpha_0}\delta(z_i, \bar{k})$$

此处, $\bar{k}$ 表示空的新类.

从 Dirichlet 过程的三种构造可以看到, 无论是哪种方式, 均体现了其良好的聚类性质.

1.4　Dirichlet 过程混合模型

Dirichlet 过程表现了良好的聚类性质, 其将具有相同值的数据聚为一类, 但是如果两组数据不相等, 不管它们是多么具有相似性, 利用 Dirichlet 过程均无法实现聚类, 这大大限制了其应用. 为此, 人们引入 Dirichlet 过程混合模型[25−26].

在 Dirichlet 过程混合模型中, Dirichlet 过程作为数据的先验分布存在, 假设观测数据是 x_i, 其分布服从

$$\theta_i|G \sim G,\ \ x_i|\theta_i \sim F(\theta_i)$$

此处, 函数 $F(\theta_i)$ 表示在给定参数 θ_i 时, 观测量 x_i 的分布. 参数 θ_i 条件独立服从分布 G, 而观测变量 x_i 条件独立服从分布 $F(\theta_i)$. 当 G 服从 Dirichlet 过程分布时, 此时模型称为 Dirichlet 过程混合模型. Dirichlet 过程混合模型的有向图模型[27−31] 表示如图 4 所示[5, 20], 左图是模型的 Stick-breaking 构造, 其中 $G_0 = g(\lambda)$, λ 为分布参数. 在本文所有的有向图中, 空心圆表示变量, 阴影圆表示可观测量, 圆角矩形表示参数或者基本分布, 而矩形框表示迭代循环, 矩形框右下角的数字表示循环的次数.

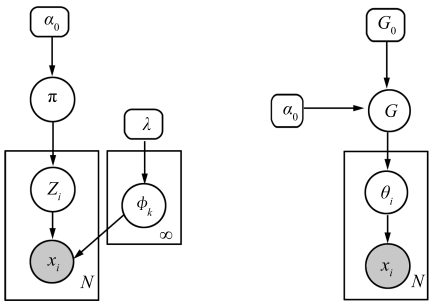


图 4 Dirichlet 过程混合模型 (根据文献 [5, 20] 重绘)

Fig. 4 Directed graphical representations of Dirichlet process mixture model (redraw based on [5, 20])

令 z_i 为 x_i 的指示因子, 则 Dirichlet 过程混合模型的 Stick-breaking 构造为[5]

$$\begin{aligned}&\boldsymbol{\pi}|\alpha_0 \sim \mathrm{GEM}(\alpha_0),\ \ z_i|\boldsymbol{\pi} \sim \boldsymbol{\pi}\\&\phi_k|G_0 \sim G_0,\ \ x_i|z_i,\ \ (\phi_k)_{k=1}^{\infty} \sim F(\phi_{z_i})\\&G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}\end{aligned}$$

由此可以看到, 利用 Dirichlet 过程混合模型能够实现数据聚类和分布参数估计. 在 Dirichlet 过程混合模型中, 目前实现数据的聚类分析有两种途径: 一种是近年来发展的, 利用变分推断近似计算数据的概率分布, 从而实现聚类分析或分布参数估计[6, 32−35]; 另外一种方式是通过 Gibbs 采样算法[36], 循环采样估计数据的聚类结果.

Dirichlet 过程的 Gibbs 采样得以实现[25, 37], 使得 Dirichlet 过程自 20 世纪 90 年代以来逐渐得到关注、发展和应用[3, 9, 20, 25, 38−40]. 目前在利用 Dirichlet 过程作为先验分布的非参数贝叶斯模型中, 主要是利用 Gibbs 采样算法, 这种方式比变分推断可行性强, 一般不需要作近似处理, 仅仅需要对一系列条件概率分布进行循环采样. 变分推断计算速度快, 但是得到有效变分推断算法比较困难. 因此, 本文主要以 Gibbs 采样算法为实现方式, 分析介绍基于 Dirichlet 过程为先验分布的各个算法.

1.4.1 Dirichlet 过程混合模型的采样

为更好地理解 Dirichlet 过程混合模型的重要作用和原理, 本小节对其采样算法进行介绍, 更加详细的内容和实验结果可参考文献 [20, 38].

假定有服从 Dirichlet 过程混合模型的观测数据集 $\mathcal{X} = \{x_1, \cdots, x_N\}$, 由于观测数据是可交换的, 即条件独立的, 在对观测数据进行聚类分析时, 不考虑观测数据的出现顺序. 实现聚类分析的目的是得到每个数据的指示因子 z_i. 约定: 当文中的某一变量的上角标或下角标有符号 "\" 时, 比如 $\mathcal{Z}_{\backslash i}$ 表示对应的变量集中移出下角标对应的变量, 即 $\mathcal{Z}_{\backslash i}$ 是将 z_i 从 $\mathcal{Z} = \{z_1, \cdots, z_N\}$ 中移出后由剩余的数据组成的数据集. 在给定其他数据的指示因子 $\mathcal{Z}_{\backslash i}$ 的情况下, 根据贝叶斯公式[22] 知关于 z_i 的条件分布为

$$\begin{aligned}&p(z_i|x_1, \cdots, x_N, \mathcal{Z}_{\backslash i}, \lambda, \alpha_0) \propto\\&\qquad p(z_i|\mathcal{Z}_{\backslash i}, \alpha_0) p(x_i|z_1, \cdots, z_N, \mathcal{X}_{\backslash i}, \lambda)\end{aligned} \tag{5}$$

式 (5) 中, 等号右边第 1 项可以用 Dirichlet 过程中的 CRP 表示, 由于各个观测量之间是可交换的, 可以把第 i 个观测数据视为最后一个观测量, 如果 $\mathcal{Z}_{\backslash i}$ 已有 K 个类别, 每一类中观测数据的个数为 $N_k^{\backslash i}$, 第一项为

$$\begin{aligned}z_i|\mathcal{Z}_{\backslash i}, \alpha_0 \sim &\sum_k^K \frac{N_k^{\backslash i}}{N-1+\alpha_0}\delta(z_i, k) +\\&\frac{\alpha_0}{N-1+\alpha_0}\delta(z_i, \bar{k})\end{aligned}$$

若第 i 个观测数据的指示因子为 $z_i = k$, 则有

$$\begin{aligned}&p(x_i|z_i = k, \mathcal{X}_{\backslash i}, \lambda) =\\&\qquad p(x_i|\{x_j|z_j = k, j \neq i\}, \lambda) =\\&\qquad \frac{\int_\theta f(x_i|\theta) \prod\limits_{z_j=k, j\neq i} f(x_j|\theta) g(\theta|\gamma)\mathrm{d}\theta}{\int_\theta \prod\limits_{z_j=k, j\neq i} f(x_j|\theta) g(\theta|\lambda)\mathrm{d}\theta}\end{aligned} \tag{6}$$

若 $z_i = \bar{k}$ 为一新类别, 则有

$$p(x_i|z_i = \bar{k}, \mathcal{X}_{\backslash i}, \lambda) = p(x_i|\lambda) =$$

$$\int_{\Theta} p(x_i|\theta)g(\theta|\lambda)\mathrm{d}\theta \tag{7}$$

因此,

$$p(z_i|x_1,\cdots,x_N,\mathcal{Z}_{\backslash i},\lambda,\alpha_0) \propto \sum_k^K \frac{N_k^{\backslash i}}{N-1+\alpha_0} \times p(x_i|\{x_j|z_j=k,j\neq i\},\lambda)\delta(z_i,k) + \frac{\alpha_0}{N-1+\alpha_0}\int_{\Theta} p(x_i|\theta)g(\theta|\lambda)\mathrm{d}\theta\delta(z_i,\bar{k}) \tag{8}$$

结合式 (6) ∼ (8), 可以得到 Dirichlet 过程混合模型的 Gibbs 采样算法. 采样算法中, 用 $\mathcal{Z}^{(t)}$ 描述第 t 次循环采样时观测数据的分类结果, $\mathcal{K}^{(t)}$ 表示此时的聚类个数, 输入第 $(t-1)$ 时的采样结果 $\mathcal{Z}^{(t-1)}$, $\mathcal{K}^{(t-1)}$, $\alpha_0^{(t-1)}$, 按照以下过程采样[20]:

1) 将 N 个观测数据随机排序, $\tau(i)$, $i=1,\cdots,N$.

2) 令 $\alpha_0=\alpha_0^{(t-1)}$, $\mathcal{Z}=\mathcal{Z}^{(t-1)}$, 每一个数据 $i\in\{\tau(1),\cdots,\tau(N)\}$, 对 z_i 进行采样.

a) 现有的 K 个聚类, 对每一个聚类计算该观测数据的似然估计 $f_k(x_i)=p(x_i|z_i=k,\mathcal{X}_{\backslash i},\lambda)$ 和 $f_{\bar{k}}(x_i)=p(x_i|z_i=\bar{k},\mathcal{X}_{\backslash i},\lambda)$;

b) 对 z_i 依据以下分布进行采样:

$$p(z_i|x_1,\cdots,x_N,\mathcal{Z}_{\backslash i},\lambda,\alpha_0)\sim \frac{1}{Z_i}\left(\sum_k^K N_k^{\backslash i}f_k(x_i)\delta(z_i,k)+\alpha_0 f_{\bar{k}}(x_i)\delta(z_i,\bar{k})\right)$$

其中,

$$Z_i=\sum_k^K N_k^{\backslash i}f_k(x_i)+\alpha_0 f_{\bar{k}}(x_i)$$

$N_k^{\backslash i}$ 是第 k 类内已有的数据量. 如果 $z_i=\bar{k}$, 则 K 增 1.

3) 检查各个类内的观测数据量, 如果某一类的观测数据总数为 0, 则将该类移除, 同时将聚类总数 K 减 1.

4) 若初始时参数采样于 $\alpha_0\sim\Gamma(a,b)$, 则按照文献 [25] 的方法更新参数, 采样关系如下:

$$\alpha_0^{(t)}\sim p(\alpha_0|K,N,a,b)$$

上述算法过程就是常用的 Collapsed Gibbs 采样算法[38], 其将不需要的变量积分掉, 只对我们关心的变量进行采样. 用 CRP 来描述上述算法如下: 随机选取一位顾客, 按照式 (8) 关系, 为其分配餐桌, 如果顾客选择新的餐桌, 则为餐厅新增一张桌子, 并将桌子个数增 1. 为所有的顾客分配餐桌后, 检查是否有餐桌没有顾客就座, 如果有, 则将该餐桌先从餐厅中移出, 并将就座的桌子总数减 1.

在采样过程中, 通常为了计算方便, 选择 $\theta_i\sim G_0$ 和 $x_i\sim F(\theta_i)$ 是共轭分布. 常用的有两种分布形式: Dirichlet 分布和其共轭多项式分布, Gaussian-Wishart 分布和其共轭 Gaussian 分布[23]. 共轭分布使得采样过程中式 (6) 和 (7) 中的积分等计算更易执行, 从而采样可行. 非共轭分布情况下, 通过引入辅助参数, 采样过程也会变得可行. 关于共轭分布和非共轭分布的 Dirichlet 过程混合模型的采样, 文献 [38] 给出了详细的介绍.

结合上面的采样过程和 Dirichlet 过程混合模型的有向图结构, 以文档数据为例, 对 Dirichlet 过程混合模型的生成过程进行解构: 由 $\boldsymbol{\pi}\sim\mathrm{GEM}(\alpha_0)$ 生成 $\boldsymbol{\pi}$, 而后按照 $z_i\sim\boldsymbol{\pi}$ 选取主题号, 从基分布 G_0 中生成主题参数 ϕ_{z_i}, 而后按照该主题中单词的分布 $F(x_i|\,\phi_{z_i})$ 生成单词 x_i. 如此操作重复进行 N 次, 即可生成一个含有 N 个单词的文档. 这就是 Dirichlet 过程混合模型的生成过程, 和聚类过程恰好相反. 因此, 从一定程度上我们可以这么解释 Dirichlet 过程中的 Concentration 参数 α_0 和基分布 G_0 的角色: 基分布 G_0 决定了模型中基本组元的分布, 而参数 α_0 决定了各个基本组元在随机过程中的分布权重.

1.5 有限混合模型的无限极限近似

Dirichlet 过程混合模型还可以由有限混合模型取极限得到, 即令有限混合模型中的有限个单元趋近于无限近似得到[5, 20, 41−42].

如图 5 所示, 假设有限混合模型中有 K 个组成单元, 令各个组成单元的权重系数为 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_K)$, 其先验分布为具有等同参数 $(\alpha_0/K,\cdots,\alpha_0/K)$ 的 Dirichlet 分布, 参数 ϕ_k 是第 k 个单元的分布参数, 其先验分布为 G_0. 观测量 x_i 的指示因子为 $z_i=k$, x_i 的分布参数为 ϕ_k, 有以下关系存在[5, 20]:

$$\begin{aligned}&\boldsymbol{\pi}|\alpha_0\sim\mathrm{Dir}(\frac{\alpha_0}{K},\cdots,\frac{\alpha_0}{K}),\ \ z_i|\boldsymbol{\pi}\sim\boldsymbol{\pi}\\&\phi_k|G_0\sim G_0,\ \ x_i|z_i,(\phi_k)_{k=1}^K\sim F(\phi_{z_i})\\&G^K=\sum_{k=1}^K\pi_k\delta_{\phi_k}\end{aligned}$$

根据 Dirichlet 分布的性质[20], 给定除了第 i 个观测数据外的指示因子分布 $\mathcal{Z}_{\backslash i}$, 第 i 个观测数据的指示因子预测分布为

$$p(z_i=k|\mathcal{Z}_{\backslash i},\alpha_0)=\frac{N_k^{\backslash i}+\frac{\alpha_0}{K}}{N-1+\alpha_0} \tag{9}$$

其中, $N_k^{\backslash i}=\sum_{j\neq i}\delta(z_j,k)$.

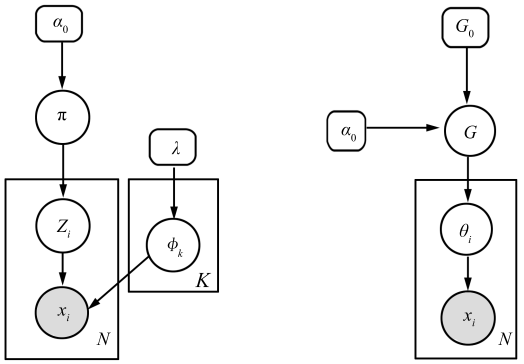


图 5 有限单元混合模型的有向图表示
(根据文献 [5, 20] 重绘)

Fig. 5 Directed graphical representations of finite mixture model (redraw based on [5, 20])

由式 (9) 可以看到, 当 $K \to \infty$ 时, 存在

$$p(z_i = k|\mathcal{Z}_{\setminus i}, \alpha_0) = \frac{N_k^{\setminus i}}{N - 1 + \alpha_0}$$

因此, 当 x_i 属于一个空的新类时,

$$p(z_i = \bar{k}|\mathcal{Z}_{\setminus i}, \alpha_0) = 1 - \sum_{k, N_k^{\setminus i} \geq 0} p(z_i = k|\mathcal{Z}_{\setminus i}, \alpha_0) = 1 - \sum_k \frac{N_k^{\setminus i}}{N - 1 + \alpha_0} = \frac{\alpha_0}{N - 1 + \alpha_0}$$

从上面初步的推导过程可以看到, 当混合模型中的单元数趋于无穷大时, 指示因子的预测分布和 Dirichlet 过程混合模型的性质一致[20]. Ishwaran 和 Zarepour 证明, 当 $K \to \infty$ 时, 有限组成单元混合模型等价于 Dirichlet 过程混合模型, $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0)$, $G \sim \mathrm{DP}(\alpha_0, G_0)$[41].

有限混合模型的近似过程为我们提供了近似求解 Dirichlet 过程的一种途径, 有些近似算法[33, 43] 和聚类模型[44] 正是基于这种近似思想发展而来的.

1.6 Dirichlet 过程的应用

Dirichlet 过程混合模型为 Dirichlet 过程的应用提供了思路. 近年来, Dirichlet 过程作为先验分布在很多领域的数据处理中得到了应用.

Bouguila 等充分利用 Dirichlet 过程混合模型可以由有限模型取无限近似得到的性质, 建立一种延伸 Dirichlet 过程混合模型[44], 采样实现图像数据的分类和建模, 如图 6 所示.

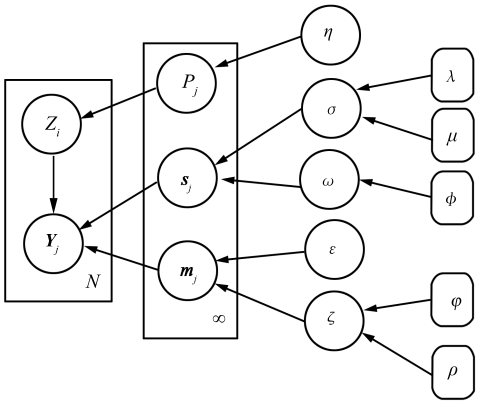


图 6 延伸 Dirichlet 过程混合模型的有向图模型
(根据文献 [44] 重绘)

Fig. 6 Directed graphical representations of generalized Dirichlet process mixture model (redraw based on [44])

Fox 等采用 Dirichlet 过程混合模型的聚类性质, 利用数据关联信息, 在机动目标跟踪中实现了跟踪目标数量的确定[45].

在图像分割中, 马尔科夫随机场是一个重要的工具, 马尔科夫随机场能将像素的空间关系结合在一起, 将像素间的相互作用加以传播. 但是在对图像分割时, 分割区域个数始终是一个难以解决的问题. Orbanz 等将 Dirichlet 过程引入到马尔科夫随机场模型中, 对模型 Gibbs 采样实现图像分割区域个数的自动生成[46].

Zhang 等将 Dirichlet 过程混合模型应用在多元有监督学习的回归问题上, 实现 MCMC 采样, 得到了良好的计算和统计学效果[47].

Qi 等将 Dirichlet 过程混合模型用在压缩传感稀疏变换系数的学习中, 并利用变分推断, 实现模拟数据和真实图像数据的分析和理解[48].

以 Dirichlet 过程为基本模型, 考虑到时间、空间等因素的可变性, MacEachern 提出一种可变的 Dirichlet 过程: DDP 过程 (Dependent Dirichlet process)[49−53]. 在 DDP 中, 基分布 G_0 是关于测度空间的一个随机过程, 参数 $\phi_k(\omega) \sim G_0$. DDP 的 Stick-breaking 构造可以表示如下[20]:

$$\pi_k(\omega) = \beta_k(\omega) \prod_{l=1}^{k-1} (1 - \beta_l(\omega))$$

$$G = \sum_{k=1}^{\infty} \pi_k(\omega) \delta(\theta(\omega), \phi_k(\omega))$$

DDP 使得数据具有了时间或空间等关联性, 目前基于 DDP 的算法在连续时间序列[54] 和离散时间的数据[55] 以及预测问题 (Predictor dependent)[56] 中得到了应用. Caron 等利用 Dirichlet 过程作为线

性动态系统中噪声的先验分布进行密度估计, 实现线性动态系统的优化和时间序列估计[57−58].

2　分层 Dirichlet 过程

Dirichlet 过程可以实现一组数据的聚类和分析, 但在研究多组数据的聚类问题时, 单纯利用 Dirichlet 过程混合模型是无法实现建模分析的. 非参数贝叶斯模型分层 Dirichlet 过程 (HDP)[3−5] 的提出, 为实现多文档之间共享无限多个聚类提供了解决途径.

HDP 的有向图表示如图 7 所示[5, 20], 因而 HDP 是 Dirichlet 过程混合模型的多层形式, 其中左图是 HDP 的 Stick-breaking 构造结构. 从图中可以看到, 各个文档的主题均是服从基分布 H 分布, 这保证了各个文档之间的主题共享. 首先, 以基分布 H 和 Concentration 参数 γ 构成了 Dirichlet 过程, $G_0 \sim \mathrm{DP}(\gamma, H)$. 而后以 G_0 为基分布, 以 α_0 为 Concentration 参数, 对每一组数据构造 Dirichlet 过程混合模型, $G_j \sim \mathrm{DP}(\alpha_0, G_0)$. 依据该层 Dirichlet 过程为先验分布, 构造 Dirichlet 过程混合模型

$$\theta_{ji}|G_j \sim G_j,\ \ x_{ji}|\theta_{ji} \sim F(\theta_{ji})$$

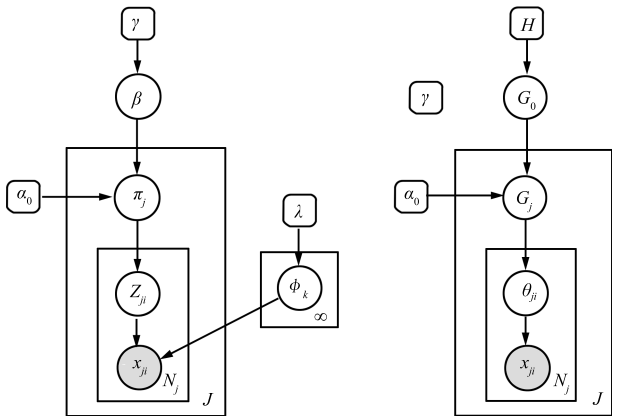


图 7　HDP 的有向图模型表示 (根据文献 [5, 20] 重绘)

Fig. 7　Directed graphical representations of hierarchical Dirichlet process (redraw based on [5, 20])

在 HDP 模型中, 参数 γ 和 α_0 的作用和 Dirichlet 过程混合模型中的参数 α_0 类似, 只是构造关系稍微复杂一些. 下面我们将以类似于 Dirichlet 过程的构造来解析 HDP 模型.

2.1　Stick-breaking 构造

根据文献 [5], HDP 的 Stick-breaking 构造可以分为两层进行解析. 首先, 第一层 Dirichlet 过程

$$\begin{aligned} G_0(\phi) &= \sum_{k=1}^{\infty} \beta_k \delta(\phi, \phi_k) \\ \boldsymbol{\beta} &\sim \mathrm{GEM}(\gamma) \\ \phi_k &\sim H(\lambda),\ k = 1, 2, \cdots \end{aligned}$$

对每一组数据来说, 以 G_0 为基分布, 每组的 G_j 在 $(\phi_k)_{k=1}^{\infty}$ 各个点也有意义, 因此第二层将每组 G_j 用 Stick-breaking 构造表示为

$$G_j(\phi) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\phi, \phi_{jk}),\ \boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty}$$

由于 G_j 关于 G_0 条件独立, 因此 $\boldsymbol{\pi}_j$ 关于 $\boldsymbol{\beta}$ 条件独立, 下面分析两者之间的构造关系.

对测度空间 Θ 的一个任意划分 $A_1, \cdots, A_r$, 令 $K_l = \{k : \phi_k \in A_l\}$, $l = 1, \cdots, r$. 因此, K_l 是对正整数的有限划分, 结合 Dirichlet 过程的定义知

$$\begin{aligned} &(G_j(A_1), \cdots, G_j(A_r)) \sim \\ &\quad \mathrm{Dir}(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r)) \end{aligned}$$

因而有以下关系存在:

$$\begin{aligned} &\left(\sum_{k \in K_1} \pi_{jk}, \cdots, \sum_{k \in K_r} \pi_{jk} \right) \sim \\ &\quad \mathrm{Dir}(\alpha_0 \sum_{k \in K_1} \beta_k, \cdots, \alpha_0 \sum_{k \in K_r} \beta_k) \end{aligned} \tag{10}$$

从式 (10) 看到, $\boldsymbol{\pi}_j$ 是关于 $\mathrm{DP}(\alpha_0, \boldsymbol{\beta})$ 的 Dirichlet 过程条件独立分布.

和 Dirichlet 过程混合模型中一样, 参数 θ_{ji} 服从分布 G_j, 以概率 π_{jk} 取值 ϕ_k, 设其指示因子为 z_{ji}, 则 HDP 模型可表示为

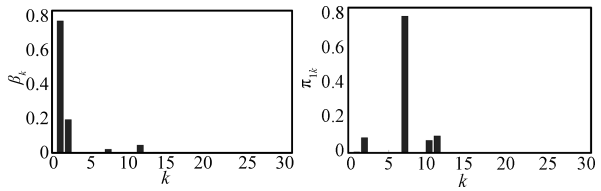
$$\begin{aligned} &\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma),\ \ \boldsymbol{\pi}_j|\alpha_0,\ \ \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}) \\ &z_{ji}|\boldsymbol{\pi}_j \sim \boldsymbol{\pi}_j,\ \ \phi_k| \sim H \\ &x_{ji}|z_{ji},\ \ (\phi_k)_{k=1}^{\infty} \sim F(\phi_{z_{ji}}) \end{aligned}$$

令基分布 H 的概率密度函数为 $h(.|\lambda)$, HDP 的 Stick-breaking 构造如图 7 中左图所示[20]. 图 8 是对 HDP 的 Stick-breaking 构造中各层权重系数的前 30 个采样结果.

2.2　Chinese restaurant franchise 构造

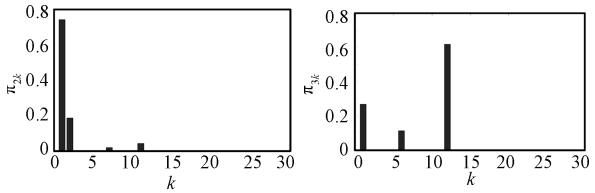
和 Dirichlet 过程的构造类似, HDP 也有类似 CRP 的构造, 只是其模型更加复杂, 称为 CRF (Chinese restaurant franchise) 构造[5].

模型中, 一个 CRF 中所有的餐厅共用一份相同



(a) 权重系数 β_k (b) 权重系数 π_{1k}

(a) Weights of β_k (b) Weights of π_{1k}



(c) 权重系数 π_{2k} (d) 权重系数 π_{3k}

(c) Weights of π_{2k} (d) Weights of π_{3k}

图 8 HDP 模型的 Stick-breaking 构造前 30 个权重系数采样, 其中 $\alpha_0 = 5, \gamma = 5$

Fig. 8 The first 30 mixture weights of stick-breaking construction in hierarchical Dirichlet process, where $\alpha_0 = 5, \gamma = 5$

的菜单, 菜单中菜的种类可以无穷多项. 和 CRP 中一样, 每个餐厅可容纳无穷多张餐桌, 每张餐桌可容纳无穷多位顾客, 第一位顾客就座第一张餐桌, 每一张餐桌上的第一位客人负责点菜, 一张餐桌只有一道菜, 其他后来就座于该餐桌的客人共同享用该道菜. 不同餐厅的不同餐桌可以点用同一道菜, 同一餐厅的不同餐桌也可点用同一道菜[5]. HDP 的 CRF 构造如图 9 所示[20].

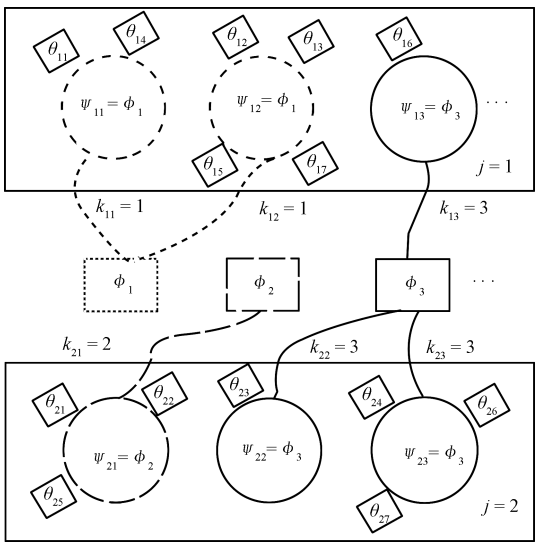


图 9 HDP 的 CRF 图模型表示 (根据文献 [20] 重绘)

Fig. 9 CRF interpretations of hierarchical Dirichlet process model (redraw based on [20])

在 HDP 中, 将要用到的各个符号意义说明如下[5]: 观测变量的分布参数 θ_{ji} 视为客人, 从基分布 H 中采样得到的参数序列 $(\phi_k)_{k=1}^{\infty}$ 是菜单中不同的菜, 变量 ψ_{jt} 表示第 j 个餐厅里第 t 张餐桌上的客人点的菜. 因此, θ_{ji} 只对应一个 ψ_{jt}, 而 ψ_{jt} 对应一个 ϕ_k. 引入指示因子, 令 t_{ji} 表示 θ_{ji} 和 ψ_{jt} 之间的关联, k_{jt} 表示 ψ_{jt} 和 ϕ_k 之间的关联, 因此在 CRF 中, 第 j 个餐厅的第 i 位顾客就座于第 t_{ji} 张餐桌上, 第 t 张餐桌上点了第 k_{jt} 道菜.

同时还定义了几个变量表示餐桌的个数和顾客的个数, n_{jtk} 表示第 j 个餐厅里第 t 张餐桌上在享用第 k 道菜的顾客数, 因此 $n_{jt\cdot}$ 表示第 j 个餐厅里第 t 张餐桌上的顾客数, $n_{j\cdot k}$ 表示第 j 个餐厅里所有餐桌上在享用第 k 道菜的顾客数. m_{jk} 表示第 j 个餐厅里所有在享用第 k 道菜的餐桌数, $m_{j\cdot}$ 表示第 j 个餐厅里所有的餐桌数, 而 $m_{\cdot k}$ 表示所有餐厅里点了第 k 道菜的桌子数, $m_{\cdot\cdot}$ 表示所有餐厅里所有已有顾客就座的餐桌数.

而后, 根据 CRP 中式 (1) 和 (2), 将 G_j 和 G_0 积分消去后, 分别得到式 (11) 和 (12)[5].

$$\theta_{ji}|\theta_{j1},\cdots,\theta_{j,i-1},\alpha_0,\ G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha_0}\delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0}G_0 \tag{11}$$

$$\psi_{jt}|\psi_{11},\psi_{12},\cdots,\psi_{21},\cdots,\psi_{j,i-1},\gamma,\ H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma}H \tag{12}$$

根据 CRF 解释式 (11) 和 (12), 如果顾客就座于一张已有顾客的餐桌, 则 $\theta_{ji} = \psi_{jt}, t_{ji} = t$. 如果就座于一张新的餐桌, 则 $m_{j\cdot}$ 增 1, 同时由分布 G 中采样得到 $\psi_{jm_{j\cdot}} \sim G_0$, $\theta_{ji} = \psi_{jm_{j\cdot}}$, $t_{ji} = m_{j\cdot}$.

如果餐桌上选用一道已经有顾客点过的菜, 那么 $\psi_{jt} = \phi_k, k_{jt} = k$. 如果选用一道新的菜, 则 K 增 1, 同时从 H 中采样得到 $\phi_K, \psi_{jt} = \phi_K, k_{jt} = K$.

在采样时, 对每一个 j 和 i, 首先利用式 (11) 对 θ_{ji} 进行采样, 如果需要从 G_0 中取新的样本, 利用式 (12), 得到新的 ψ_{jt}, 并令 $\theta_{ji} = \psi_{jt}$.

因此, HDP 的 CRF 构造即是为顾客分配餐桌和菜的过程, 首先为每一位顾客分配餐桌, 顾客选择就座于哪张餐桌和这张餐桌已有的顾客数成正比, 同时顾客也可以选择一张新的餐桌就座. 在分配完餐桌后, 为每张餐桌分配菜, 每道菜都有可能被点到, 某道菜被点到的概率与已点用这道菜的餐桌个数成正比, 同时也可能有新的菜被点到.

HDP 和 Dirichlet 过程一样, 也可以由有限混合模型取极限得到, 具体细节可参考文献 [5].

2.3 HDP 的采样

本节中, 主要分析基于 MCMC (Markov chain Monte Carlo) 的采样算法对 HDP 实现采样, 目前采样算法主要采用 Collapsed Gibbs 采样[5, 20, 39]. 分三个小节, 第 2.3.1 节和第 2.3.2 节介绍两种基于 CRF 的采样算法, 在第 2.3.3 节介绍了一种直接分配的采样算法. 为了简化推导, 也不失一般应用性, 假设基分布 H 和数据分布函数 F 是共轭分布. 为了更方便查找文献和理解算法, 在采样过程中要用到的符号意义和文献 [5] 中类似.

假设模型中的观测数据 x_{ji} 服从分布 $F(\theta_{ji})$, 关于 $\phi_{k_{jt}}$ 的先验分布为 H, 引入观测数据的 x_{ji} 的指示因子 z_{ji}, 使得 $z_{ji}=k_{jt_{ji}}$.

变量之间有以下关系存在: $\mathcal{X}=\{x_{ji}:j,i\}$, $\mathcal{X}_{jt}=\{x_{ji}:i,t_{ji}=t\}$, $\mathcal{T}=\{t_{ji}:j,i\}$, $\mathcal{K}=\{k_{jt}:j,t\}$, $\mathcal{Z}=\{z_{ji}:j,i\}$, $\mathcal{M}=\{m_{jk}:j,k\}$.

假设观测数据的分布函数 F 的分布密度函数为 $f(\cdot|\theta)$, 基分布函数 H 有密度分布函数 $h(\cdot|\lambda)$. 由于选用的是函数 F 和 H 共轭分布, 因此我们可以尝试将分布参数积分消去, 利用贝叶斯公式, 关于观测数据的条件分布为[5]

$$f_k^{\backslash x_{ji}}(x_{ji})=\frac{\int f(x_{ji}|\phi_k)\prod\limits_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}|\phi_k)h(\phi_k)\mathrm{d}\phi_k}{\int\prod\limits_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}|\phi_k)h(\phi_k)\mathrm{d}\phi_k} \tag{13}$$

同样, 一组数据的条件分布 $f_k^{\backslash\mathcal{X}_{jt}}(\mathcal{X}_{jt})$ 也可以采用式 (13) 的关系计算得到. 本节的条件分布中, 将 Concentration 参数等条件省略, 即

$$f_k^{\backslash x_{ji}}(x_{ji})=f_k^{\backslash x_{ji}}(x_{ji}|\mathcal{X}^{\backslash ji},\alpha_0,\lambda,\gamma)$$

2.3.1 基于 CRF 后验采样算法

在 HDP 中, 利用 Collapsed Gibbs 采样算法, 分两个层次对数据实现聚类. 和 Dirichlet 过程混合模型中类似, 在 CRF 中, 首先为每一位顾客分配餐桌, 而后再为餐桌分配菜. 因此, 有以下关系存在:

$$p(t_{ji}=t|,\mathcal{T}^{\backslash ji},\mathcal{K})\propto\begin{cases}n_{jt\cdot}^{\backslash ji}f_{k_{jt}}^{\backslash x_{ji}}(x_{ji}), & t\text{ 为已有顾客就座的餐桌}\\ \alpha_0 p(x_{ji}|\mathcal{T}^{\backslash ji},t_{ji}=t^{\mathrm{new}},\mathcal{K}), & t=t^{\mathrm{new}}\end{cases} \tag{14}$$

在上式中, $f_{k_{ji}}^{\backslash x_{ji}}(x_{ji})$ 可按照式 (13) 得到. 当顾客就座于一张新的餐桌时, 观测数据的条件分布为

$$p(x_{ji}|\mathcal{T}^{\backslash ji},\ t_{ji}=t^{\mathrm{new}},\ \mathcal{K})=\sum_{k=1}^{K}\frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}f_k^{\backslash x_{ji}}(x_{ji})+\frac{\gamma}{m_{\cdot\cdot}+\gamma}f_{k^{\mathrm{new}}}^{\backslash x_{ji}}(x_{ji}) \tag{15}$$

式 (15) 中, 等号右边第 1 项是新的餐桌点用已有顾客点过的菜之概率和, 第 2 项是该新的餐桌点一道新的菜的概率.

而后, 对每个餐厅中的餐桌分配菜, 因此, 此时的分析对象是各个餐厅的不同餐桌, 而由于在整个 CRF 中共用一个菜谱, 则有

$$p(k_{jt}=k|\mathcal{K}^{\backslash jt},\mathcal{T})\propto\begin{cases}m_{\cdot k}^{\backslash jt}f_k^{\backslash\mathcal{X}_{jt}}(\mathcal{X}_{jt}), & k\text{ 为已有顾客点用的菜}\\ \gamma f_{k^{\mathrm{new}}}^{\backslash\mathcal{X}_{jt}}(\mathcal{X}_{jt}), & k=k^{\mathrm{new}}\end{cases} \tag{16}$$

按照上面的分配方式实现数据的聚类分析是完全基于 CRF 的采样方式, 称为基于 CRF 的后验采样算法[5], 其中将 G_0 和 H 均积分消去, 只留下数据的聚类特征.

2.3.2 Augmented representation 后验采样算法

基于 CRF 的后验采样算法无法得到 Stick-breaking 构造中的权重系数 π_{jk}, 这将为 HDP 模型的应用带来问题, 比如在后续要提到的模型 HDP-HMM 中, 需要权重系数 π_{jk} 的定量值. 为此, 在 CRF 的基础上得到 $(\mathcal{T},\mathcal{K})$ 时, 由于 $G_0\sim\mathrm{DP}(\gamma,H)$, 每个 t 对应的 ψ_{jt} 采样于 G_0, 因此在所有 t 对应的 $\boldsymbol{\psi}_{jt}$ 条件下有

$$G_0|\boldsymbol{\psi}_{jt}\sim\mathrm{DP}\left(\gamma+m_{\cdot\cdot},\frac{\gamma H+\sum\limits_{k=1}^{K}m_{\cdot k}\delta_{\phi_k}}{\gamma+m_{\cdot\cdot}}\right)$$

因此,

$$\begin{aligned}&\boldsymbol{\beta}=(\beta_1,\cdots,\beta_K,\beta_u)\sim\mathrm{Dir}(m_{\cdot 1},\cdots,m_{\cdot K},\gamma)\\&G_u\sim\mathrm{DP}(\gamma,H)\\&p(\phi_k|\mathcal{T},\mathcal{K})\propto h(\phi_k)\prod_{ji:k_{jt_{ji}}=k}f(x_{ji}|\phi_k)\\&G_0=\sum_{k=1}^{\mathcal{K}}\beta_k\delta_{\phi_k}+\beta_u G_u\end{aligned} \tag{17}$$

在采样过程中, 我们关心的是 $\mathcal{T}$ 和 $\mathcal{K}$, 因此需要将式 (17) 中 ϕ_k 和 G_u 积分消去. $\mathcal{T}$ 和 $\mathcal{K}$ 的采样与

式 (14) 和式 (16) 一致. 需要注意的是, 此时上两式中需要用 β_k 代替 $m_{\cdot k}$, β_u 代替 γ 进行计算. 采样中, 新的 k^{new} 产生时, $\beta_{k^{\text{new}}} = b\beta_u$, $\beta_u^{\text{new}} = (1-b)\beta_u$, 其中, $b \sim \text{Beta}(1,\gamma)$, b 的采样充分利用了 Dirichlet 过程的 Stick-breaking 构造的性质.

对 $\boldsymbol{\beta}$ 采样

$$(\beta_1, \cdots, \beta_K, \beta_u)|\mathcal{T}, \mathcal{K} \sim \text{Dir}(m_{\cdot 1}, \cdots, m_{\cdot K}, \gamma) \tag{18}$$

上面采样过程是 Augmented representation 后验采样算法[5], 该采样算法实现了权重系数的跟踪和计算.

2.3.3 直接分配后验采样算法

前面两小节中的采样算法是基于 CRF 构造的采样, 在采样过程中, 首先分配餐桌 t_{ji}, 而后为餐桌分配菜 k_{jt}, 这两种方式都是间接对数据进行聚类分析的方式. 在 HDP 采样中, 还有直接分配 (Direct assignment) 后验采样算法[5], 这种算法的主要思路与 CRF 不同, 直接对观测数据的指示因子 z_{ji} 进行聚类分析, 其中 z_{ji} 和前两节中的 $k_{jt_{ji}}$ 对应, 而此时的餐桌通过 m_{jk} 起作用.

根据文献 [5] 首先对指示因子 $\mathcal{Z}$ 采样, 由式 (14) 和 (15) 得到以下分配关系:

$$p(z_{ji} = k|\mathcal{Z}^{\backslash ji}, \mathcal{M}, \boldsymbol{\beta}) \propto \begin{cases} (n_{j\cdot k}^{\backslash ji} + \alpha_0\beta_k) f_k^{\backslash x_{ji}}(x_{ji}), & k \text{ 为已有顾客点用的菜} \\ \alpha_0\beta_u f_{k^{new}}^{\backslash x_{ji}}(x_{ji}), & k = k^{\text{new}} \end{cases}$$

此时, 由 β_k 替换 $m_{\cdot k}$, β_u 替换 γ.

而后, 为每一个餐厅的顾客分配餐桌个数, Antoniak 证明 m_{jk} 的采样存在以下关系[26]:

$$p(m_{jk} = m|\mathcal{Z}, \mathcal{M}^{\backslash jk}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k + n_{j\cdot k})} s(n_{j\cdot k}, m)(\alpha_0\beta_k)^m$$

其中, $s(n,m)$ 是 Stirling 数.

然后根据式 (18), 计算权重系数 $\boldsymbol{\beta}$. 如此完成直接分配后验采样算法, 在算法中, 直接得到数据的指示因子, 实现聚类分析.

2.3.4 三种采样算法的比较分析

在三种采样算法中, 前两种的实现相对较繁琐, 但是这两种采样算法和 HDP 的 CRF 构造一致, 比较容易理解. 在采样过程中, 为餐桌分配菜时, 同时更新很多数据的聚类属性, 因此, 这两种采样方法中, 收敛速度快, 效果好. 第三种方法, 对指示因子直接进行聚类分析, 每次只更新一个数据的聚类属性, 算法的收敛速度相对会慢. 另一方面, 第一种方法将 G_0, H 积分消去, 无法得到权重系数的分布, 第二种和第三种采样算法能够得到权重系数的信息, 而 HDP 算法作为先验分布用在其他模型中, 如 HDP-HMM 模型, 一般需要这些权重信息[5].

值得说明的是, 在三种采样算法中, Concentration 参数 α_0 和 γ 的更新可参考文献 [25].

2.4 实验分析

为了验证 HDP 的聚类性质, 本文作者采用 C++ 语言在 Visual Studio 6.0 编译环境下, 对常用来做主题 (Topics) 提取算法测试[59−61] 的数据, 利用 HDP 的三种采样算法分别进行主题提取. 如图 10 所示, 文档的字典是从 5 像素 ×5 像素 的图像中选择的, 即用图中像素点的位置 $(1, \cdots, 25)$ 等 25 个数字作为字典数据, 而数据的分布主题有竖条和横条十种分布情况. 每一个文档均是由竖条和横条主题随机组合生成, 而后由这些主题生成文档中的单词. 因此, 利用 HDP 模型要实现的聚类任务即是将 10 个主题从多文档中的提取出来. 图 11 是作者采用 CRF 进行采样的前 1 000 次结果.

图 10 10 个主题分布图示

Fig. 10 Graphical representations of 10 topics

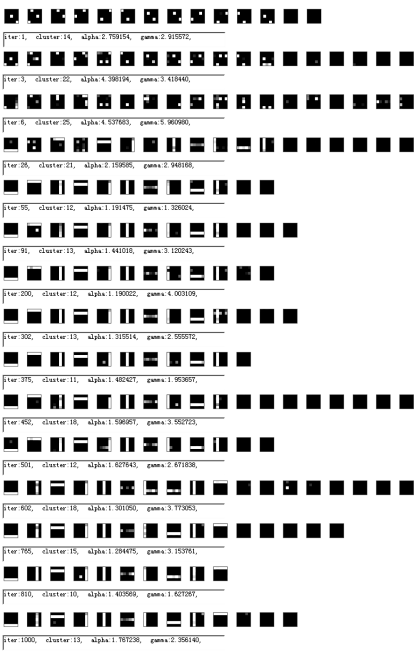


图 11 Gibbs 采样中前 1 000 次采样中若干次采样结果

Fig. 11 Some results in the first 1 000 iterations of Gibbs sampling

在该实验中共有 40 组文档, 每组文档有 50 个单词组成, 先验分布 H 是 Dirichlet 分布, 观测数

据服从 F 多项式分布. 图 11 的每一步循环结果中, 第 1 项 iter 对应采样循环次数, cluster 项是聚类个数, alpha 和 gamma 项分别是采样中 HDP 的 Concentration 参数 α_0 和 γ 的更新值. 从图中实验结果可看到, 通过随机初始化时主题分布很乱, 但在经过大约 20 步采样后, 主题的分布模式即开始呈现. HDP 的良好聚类效果在 Teh 公布的 Matlab 版本的测试程序[61] 中给出了很好的实验结果.

3 Hierarchical Dirichlet 过程 — 隐马尔科夫模型 (HDP-HMM)

3.1 HDP-HMM 介绍

隐马尔科夫模型 (Hidden Markov models, HMM)[62] 在空间、时间序列数据的分析、建模中得到了广泛的发展和应用. HMM 模型是一个两层的随机过程: 马尔科夫链和一般随机过程. 其中马尔科夫链描述了状态的转移, 一般用转移概率矩阵描述; 而一般随机过程描述状态和观测序列间的关系, 用观察概率矩阵描述.

图 12 是 T 时刻时 HMM 的模型示意图. 其中, v_0, $\cdots$, v_T 是在 T 时刻前出现的状态, 各时刻的状态均可以从 K 个不同的状态中取值, 不同状态之间的转换由转换概率矩阵 A 决定. 而 $y_1, \cdots, y_T$ 分别是不同状态下出现的观测值, 观测值 y_t 关于状态 v_t 条件独立. 假设在给定状态 v_t 下, 观测值 y_t 的分布为 $F_{\phi_{v_t}}(y_t)$, 此分布通常称为 "发散分布 (Emission distribution)", 各个状态下观测值的分布组成发散分布矩阵 B.

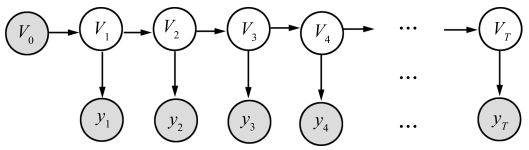


图 12　HMM 的有向图模型

Fig. 12　Directed graphical representations of HMM

在 HMM 中, 状态数 K 是需要提前确定的, 而且是有限的, 即状态 $v_0, \cdots, v_T$ 只能在此 K 个状态之间进行转移. HDP-HMM 模型解决了 HMM 中有限状态的限制[5, 63], 因此有时又称 iHMM (Infinite HMM) 模型, 如图 13 所示[5] 是 HDP-HMM 的 Stick-breaking 构造. HDP-HMM 充分利用 Dirichlet 过程的性质, 实现 HMM 中状态数自动生成.

根据文献 [5], HDP-HMM 的 Stick-breaking 构造如下:

$$\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma)$$

$$\boldsymbol{\pi}_k|\alpha_0, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$$
$$\phi_k|H \sim H \tag{19}$$

在时间 $t = 1, \cdots, T$ 时, 状态之间的转移和观测值的分布分别为

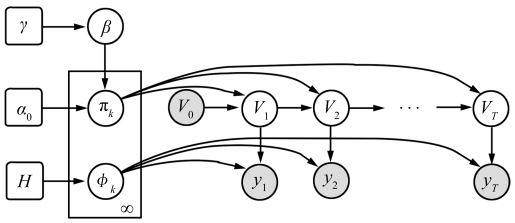


图 13　HDP-HMM 的有向图模型[5]

Fig. 13　Directed graphical representations of HDP-HMM[5]

$$v_t|v_{t-1}, (\boldsymbol{\pi}_k)_{k=1}^{\infty} \sim \boldsymbol{\pi}_{v_{t-1}}$$
$$y_t|v_t, (\phi_k)_{k=1}^{\infty} \sim F(y_t|\phi_{v_t}) \tag{20}$$

根据式 (19) 和 (20), 对 HDP-HMM 进行采样[5, 64−65], 首先对 $\boldsymbol{\beta}$ 采样,

$$p(\beta_1, \cdots, \beta_K, \beta_{\bar{k}}|\mathcal{T}, \mathcal{K}, y_1, \cdots, y_T, \gamma) \propto \mathrm{Dir}(m_{\cdot 1}, \cdots, m_{\cdot K}, \gamma)$$

对状态量 v_t 采样,

$$p(v_t|\mathcal{V}_{\backslash t}, y_1, \cdots, y_T, \boldsymbol{\beta}, \alpha_0, \lambda) \propto p(v_t|\mathcal{V}_{\backslash t}, \boldsymbol{\beta}, \alpha_0) p(y_t|\mathcal{Y}_{\backslash t}, v_t, \mathcal{V}_{\backslash t}, \lambda) \tag{21}$$

在式 (21) 中, 等号右边第 1 项根据马尔科夫链和 Dirichlet 过程的性质可以得到, 当状态 $v_t = k$, k 为已出现过的状态时,

$$p(v_t = k|\mathcal{V}_{\backslash t}, \boldsymbol{\beta}, \alpha_0) \propto (\alpha_0\beta_k + n_{v_{t-1}k}^{\backslash t}) \times \left(\frac{\alpha_0\beta_k + n_{kv_{t+1}}^{\backslash t} + \delta(v_{t-1}, k)\delta(v_{t+1}, k)}{\alpha_0 + n_{k\cdot}^{\backslash t} + \delta(v_{t-1}, k)}\right)$$

其中, n_{kj} 是状态 k 到状态 j 的转移次数, $n_{\cdot j}$ 是由其他状态转移到状态 j 的次数, $n_{k\cdot}$ 是由状态 k 转移到其他状态的次数.

状态 $v_t = \bar{k}$ 为一新出现的状态时,

$$p(v_t = \bar{k}|\mathcal{V}_{\backslash t}, \boldsymbol{\beta}, \alpha_0) \propto \alpha_0\beta_{\bar{k}}\beta_{v_{t+1}}$$

式 (21) 中等号右边第 2 项, $p(y_t|\mathcal{Y}_{\backslash t}, v_t, \mathcal{V}_{\backslash t}, \lambda)$ 的计算参考式 (6) 和 (13) 可得.

然后对 m_{jk} 采样,

$$p(m_{jk} = m|n_{jk}, \boldsymbol{\beta}, \alpha_0) =$$

$$\frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k+n_{jk})}s(n_{jk},m)(\alpha_0\beta_k)^m$$

其中, $s(n,m)$ 是 Stirling 数.

因此, HDP-HMM 既有 HMM 处理序列数据的特点, 又具有 HDP 自动生成聚类数目和实现聚类的功能. 因此, 近年来, HDP-HMM 模型受到了广泛的关注和研究. 除了 HDP-HMM 模型, HDP 的出现同时也为 HMM 模型的应用提供了更多的灵活性和适用性.

4 HDP 和 HDP-HMM 模型的应用

4.1 HDP 模型的应用

由于具有很好的聚类性质, 近年来, HDP 算法在视频监控、图像理解和图像标注、文本分类、认知研究、信息检索等方面得到了广泛的关注和应用.

文档是由几个主题组成, HDP 能够将文档中的单词按照主题实现聚类, 实现文档间的主题共享和主题提取. 但是, HDP 无法实现文档层数据的聚类. 要实现这个目标, 单纯地在 HDP 上层添加一层 Dirichlet 过程是无法实现的, 因为这样只能实现不同文集之间的主题共享. 为了实现文档层的聚类, 必须对模型加以改进. Wang 等提出了几种基于 HDP 的改进算法[7−10, 39], 用于处理视频监控、雷达扫描数据和磁共振实验采集的大脑神经系统的活动信息等数据信息, 实现文档数据级别的聚类分析, 其中典型的模型是 Dual-HDP 模型[8−9], 如图 14 所示. 模型中, 在 HDP 的基础上, 嵌入 DDP 模型, 实现了不同文档之间主题的共享, 而且描述了同一类文档内具有的相同主题分布模式. Wang 等将视频中运动目标的运动轨迹视作文档, 将轨迹上的不同运动模式, 比如直行或右转弯等运动, 视为文档中的主题, 将轨迹上的每个点视作文档中的单词, 采用 Gibbs 采样, Dual-HDP 模型不仅实现了轨迹上不同运动模式的提取, 而且实现了轨迹文档的聚类和基于运动模式的视频检索等[8−9, 39].

根据文献 [8−9], Dual-HDP 模型的 Stick-breaking 构造如下:

$$Q=\sum_{c=1}^{\infty}\epsilon_c\delta_{\tilde{G}_c},\ \epsilon_c=\epsilon_c^{'}\prod_{l=1}^{c-1}(1-\epsilon_l^{'}),\ \epsilon_c^{'}\sim\text{Beta}(1,\mu)$$

$$\tilde{G}_c\sim\text{DP}(\rho,G_0),\ \ \tilde{G}_c=\sum_{k=1}^{\infty}\tilde{\boldsymbol{\pi}}_{ck}\delta_{\tilde{\phi}_{ck}}$$

$$G_j\sim\text{DP}(\alpha,\tilde{G}_{c_j})$$

由于 $G_0\sim\text{DP}(\gamma,H)$, $\phi_k\sim H$, 因此, 所有的 $(\tilde{G}_c)_{c=1}^{\infty}$ 有共同的主题, 即 $\tilde{\phi}_{ck}=\phi_k$, 只是主题分布的组合不同, 即权重系数 $\{\tilde{\boldsymbol{\pi}}_{ck}\}$ 的分布不同. 不同的文档 G_j 可以选择同一个先验分布 $\tilde{G}_c$, 这样不同的文档之间形成了聚类 c, 根据 $G_j\sim\text{DP}(\alpha,\tilde{G}_{c_j})$ 组合生成文档 G_j. 然后, 文档中主题和单词的生成原理和 HDP 模型中相同. 在 Dual-HDP 中, 参数 μ 控制文档在文集中的混合分布情况, 参数 ρ 控制主题在文档的分布情况. Dual-HDP 模型是对不同文档的聚类, 文档又随着主题的分布变化, 因此 $Q\sim\text{DDP}(\mu,\rho,G_0)$.

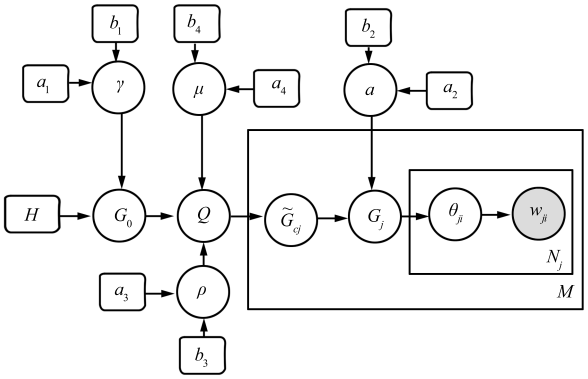


图 14 Dual-HDP 的有向图模型[8]

Fig. 14 Directed graphical representations of Dual-HDP[8]

在 Dual-HDP 模型中, 采样过程如下[8]:

1) 给定或根据上一步采样得到文档的分类属性 $\{c_j\}$ 进行采样, 此时 Dual-HDP 变成了 HDP 模型, 按照 HDP 模型的采样过程进行.

2) 根据 $\{z_{ji}\}$, $\{\boldsymbol{\pi}_{0k}\}$, $\{\tilde{\boldsymbol{\pi}}_{ck}\}$ 对每个文档的分类属性 $\{c_j\}$ 进行采样.

Dual-HDP 算法的详细采样过程可进一步参考文献 [8−9, 39].

在静态图像的处理中, Sudderth 等提出基于 Dirichlet 过程和 HDP 的 TDP (Transformed Dirichlet process) 等模型[11, 20, 66−67], 其中一个典型模型如图 15 所示. Sudderth 等将图像中像素点的位置和像素的颜色等信息作为分析数据, 利用 Gibbs 采样, 通过提取图像中的 "共享部分 (Sharing parts)", 得到图像中的共有主题, 实现视频图像场景的分类学习, 进而实现图像场景的理解和描述.

在静态图像的标注中, 图像的内容是由不同的区域 (Regions) 组成的, 而图像的题注信息则从文字上反映了图像的信息. 随着图像数据量的增大, 对图像实现人工标注的工作量变得越来越繁重. 基于 HDP, Yakhnenko 等提出一种称为 MoM-HDP (Multi-modal hierarchical Dirichlet process) 的模型[6, 32], 如图 16 所示, 根据该模型, 采用变分推断, 实现了对图像中不同区域的理解和图像题注的主题

提取分类学习, 为多模态的数据分类学习和自动进行图像标注提供了一种途径.

Xing 等利用 Dirichlet 过程和 HDP 模型作为先验分布应用在单体型 (Haplotye) 推断中, 描述了单体型的个数和统计信息[68−69].

在文本处理中, Cowans 即于 2004 年实现利用 HDP 模型作为先验分布的文本信息检索[70].

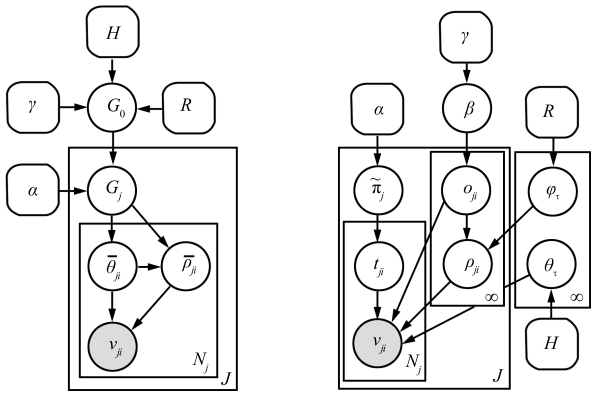


图 15　Transformed Dirichlet process 的有向图模型[11]

Fig. 15　Directed graphical representations of transformed Dirichlet process[11]

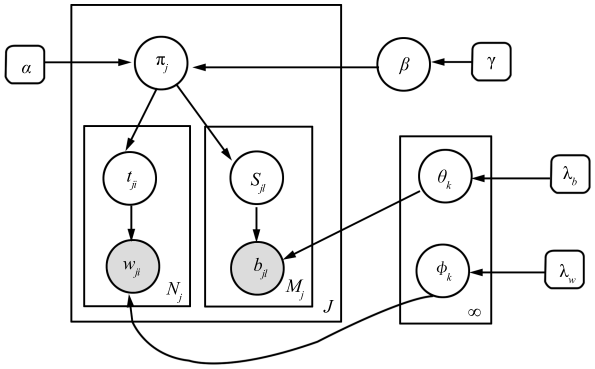


图 16　MoM-HDP 的有向图模型 (根据文献 [6] 重绘)

Fig. 16　Directed graphical representations of MoM-HDP (redraw based on [6])

Canini 等将 DP 和 HDP 模型分别用于对心理学研究中人种类别学习和迁移学习问题中主题共享的作用分析[71−72].

Li 等充分利用 Dirichlet 过程混合模型作为先验分布, 尝试实现视频监控中基于轨迹信息的视频检索[73].

4.2　HDP-HMM 以及其他类似模型的应用

HDP-HMM 模型是 HDP 模型的一种应用, 即将 HDP 模型作为基本的先验分布应用在其他概率图模型中. 本节以 HDP-HMM 为代表, 介绍 HDP 模型在其他概率图模型算法构造中的应用, 并将这些算法在语音数据、视频监控、静态图像处理、机动目标跟踪、生物信息数据、音乐数据等领域的应用作简要介绍.

在语音数据的处理中, Gael 等将 HDP-HMM 模型应用于语音处理的词性标注[74], 同时为了使得 HDP-HMM 取得良好的采样效果, 使用了一种 Beam sampling 的采样算法[64], 为 HDP-HMM 的应用提供了进一步的有利条件.

在视频监控数据处理中, Hu 和 Zhang 等利用 HDP-HMM 模型, 采用上述 Beam sampling 采样算法, 实现视频监控中的行为异常检测和运动模式识别[65, 75]. Pruteanu-Malinici 等利用 HDP 为先验分布的 iHMM 模型, 分别用 MCMC 采样和变分推断实现视频监控中运动模式的训练, 并依此检测异常行为事件[76−77].

参考 HDP-HMM 的结构, Xu 等提出一种 HDP-HTM (HDP with hierarchical transition matrix) 的算法, 利用此模型, 自动生成聚类数目, 尝试改善进化聚类中的聚类对应 (Cluster correspondence) 问题, 实验测试进化聚类效果得到改善[78−79].

Fox 等结合并改进了 HDP 和 HMM 模型, 实现了多机动目标的跟踪[12, 40, 80], 并提出一种 Sticky HDP-HMM 模型[12], 如图 17 所示. 该模型尝试改善 HDP-HMM 模型中标记偏置 (Label bias) 问题, 通过有效的采样算法, 分别利用模拟数据和实际语音数据对算法进行测试.

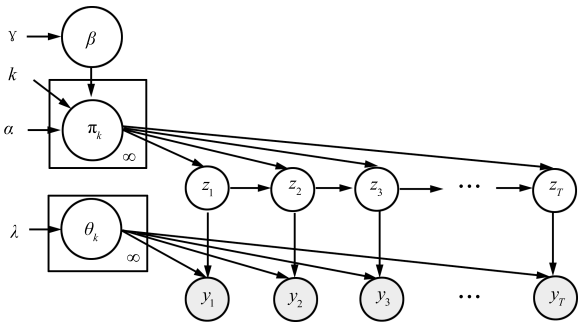


图 17　Sticky HDP-HMM 的有向图模型[12]

Fig. 17　Directed graphical representations of sticky HDP-HMM[12]

在静态图像处理中, Zhu 等将 HDP 模型成功应用在 2DHMM 模型中[81], 如图 18 所示. 利用 HDP-2DHMM 实现图像纹理学习和合成, 同时得到纹理基元个数和反应纹理基元之间空间关系的转移矩阵, 采样实现纹理图像的分割和合并等.

与 HDP 模型在 HMM 中的应用类似, Kivinen 等将 HDP 模型作为 Markov tree 中状态和观测值的先验分布, 如图 19 所示. 利用 HDP-HMT (HDP-

hidden Markov trees) 模型将图像中由 Sift 提取得到的特征之间的空间相关性进行建模, 实现图像场景的学习和描述[82−83].

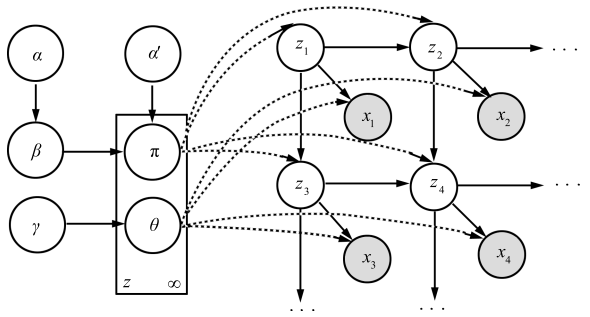


图 18 HDP-2DHMM 的有向图模型[81]

Fig. 18 Directed graphical representations of HDP-2DHMM[81]

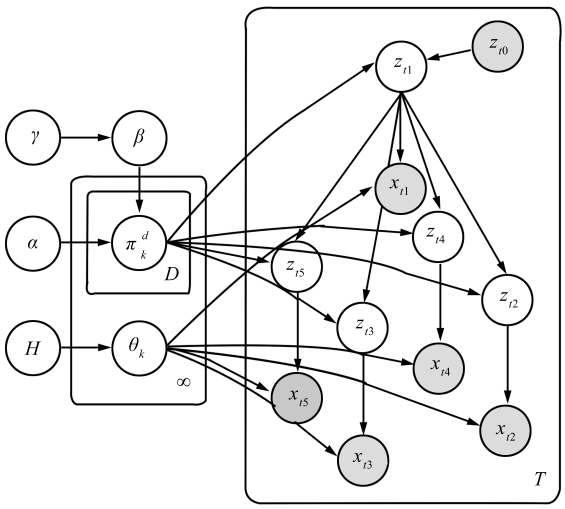


图 19 HDP-HMT 的有向图模型[82]

Fig. 19 Directed graphical representations of HDP-HMT[82]

在 HDP-HMM 模型中, 每一个初始状态对应一个 Dirichlet 过程混合模型来描述其状态转移矩阵, 而事实上所有的初始状态和目标状态都来自于一个无限状态空间. 基于这一事实, Sohn 和 Xing 等提出 HMDP (Hidden Markov Dirichlet process), 对每个初始状态对应的 Dirichlet 过程建立了统一的模型, 利用分层 Polya urn 构造实现 HMDP 采样过程, 并将 HMDP 算法分别应用于模拟和真实数据的单体型推断的 Ancestral 推断等问题中, 与利用传统 HMM 等参数模型的实验结果相比较, HMDP 模型显现了明显的优势[84−85].

与 HDP 在 HMM 中的应用思路稍微不同, Qi 等将 HMM 中三组参数: 转移矩阵 A、发散分布矩阵 B 和初始状态分布向量 $\boldsymbol{\pi}$, 作为一个参数组元 $\Theta_k = \{A_k, B_k, \boldsymbol{\pi}_k\}$, 利用 Dirichlet 过程混合模型作为此参数的先验分布同时实现 HMM 中状态个数和参数的学习, 如图 20 所示. 在 Dirichlet 过程混合模型中, 选用共轭分布 Gaussian-Wishart 分布和 Gaussian 分布进行建模, 分别采用 Gibbs 采样和变分推断将算法应用于音乐数据的分析[86].

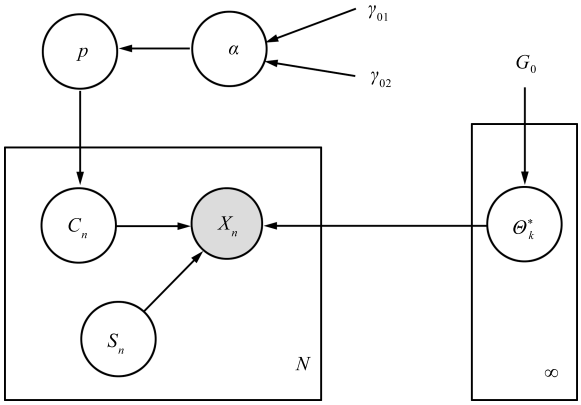


图 20 DP-HMM 的 Stick-breaking 构造有向图模型[86]

Fig. 20 Stick-breaking construction of DP-HMM[86]

到目前为止, HDP-HMM 中的 HMM 层和 DDP 模型考虑到时间序列性, 而 HDP 模型等不具有时间相关性, 然而在 HMM 层或 DDP 模型中, 其时间、空间等相关性上也很有限. Ren 等提出一种动态 HDP 模型 (Dynamic HDP, dHDP)[87−88], 即各个数据组 G_j 之间是存在动态顺序关系的, 如图 21 所示. Ren 等将 dHDP 模型作为 HMM 的先验分布进行建模, 并在音乐数据分析中进行了实验验证[88−89]. 因此, dHDP-HMM 模型同时考虑了数据之间和数据状态之间的时间序列性. Pruteanu-Malinici 等基于 dHDP 模型, 假设同一时间段的文档的主题分布服从同一混合模型, 分别利用 dHDP 和 dLDA (Dynamic LDA) 模型, 采用变分推断, 对带有时间标签的文档数据进行主题提取[90].

5 HDP 和 LDA 模型比较分析

由于 HDP 和 LDA[91] 同是近年来受到关注比较多的贝叶斯模型, 本节对两类算法进行简要比较分析.

5.1 LDA 模型介绍

和 HDP 模型一样, LDA 模型也是一种生成式分类器模型, 如图 22 所示. 首先按照模型的产生过程对模型进行描述.

假设已知文集中包含 K 个主题, 文档中的主题分布由参数 $\boldsymbol{\alpha}$ 决定, 主题中单词的分布由主题 z_n 和词典参数 $\boldsymbol{\beta}$ 共同决定, 按照以下关系生成文集中的每一个文档[91]:

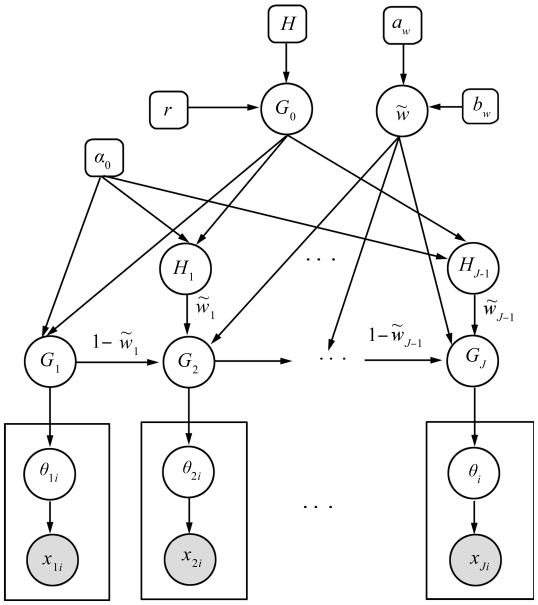


图 21　dHDP 的有向图模型 (根据文献 [88] 重绘)

Fig. 21　Directed graphical representations of dHDP (redraw based on [88])

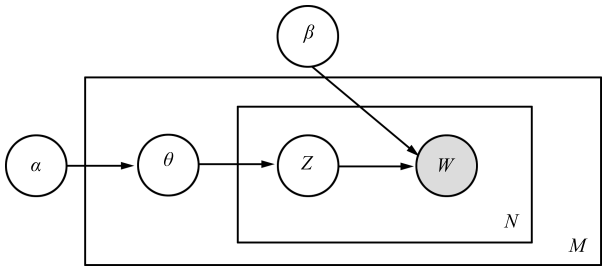


图 22　LDA 的有向图模型[91]

Fig. 22　Directed graphical representations of LDA[91]

1) 可以根据 $N \sim \text{Possion}(\xi)$ 随机选取文档中单词个数 N;

2) 根据参数向量选取主题分布概率, $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$, $\boldsymbol{\theta}$ 是 K 维向量;

3) 对每一个单词 w_n, $n = 1, \cdots, N$,

a) 根据单词 w_n 的指示因子 $z_n \sim \boldsymbol{\theta}$ 分配主题;

b) w_n 在主题 z_n 条件下, 在词典中服从多项式分布, 根据 $w_n \sim p(w_n|z_n, \boldsymbol{\beta})$ 选取单词 w_n.

文档生成过程和主题提取过程相反, 在进行主题提取时, 单词 w_n 是观测量, 而主题分布参数 θ 和单词的指示因子 z_n 是隐变量, 因此和 HDP 中一样, LDA 可以有两种解决方式: 变分推断和 Gibbs 采样. 最初 LDA 模型是通过 Mean-field 变分近似推断结合 EM 算法实现的[91−92], 目前 LDA 的 Gibbs 采样算法也已得到了推广和广泛应用[59, 93−94].

LDA 模型或者基于 LDA 模型改进的各种算法在文本主题提取、推荐系统、音乐视频内容、新闻报纸数据的处理等领域中得到了广泛的应用[59−60, 91−92, 94−96]. 以 LDA 为核心模型, 在 Blei 等的工作下, 形成了以挖掘文档数据中潜在的语义结构为主要目的的一类概率模型: 主题模型 (Topic models)[97]. 但是, LDA 模型中主题的个数需要提前输入, 主题提取效果受到主题个数的影响, 因此必须尽量精确地估算文档中的主题个数.

5.2　HDP 和 LDA 的异同

HDP 和 LDA 具有很多相同点, 归纳如下:

1) 模型的功能

a) 均实现文档数据内部的主题挖掘, 实现多文档之间的主题共享;

b) 同时对文档的主题和主题内部的单词建立概率描述.

2) 模型的机理

a) 在概率图模型中, 均是有向图表示;

b) 均是非监督学习模型;

c) 均假设文档是由服从某种概率分布的主题组成, 每个主题由服从某种概率分布的单词组成;

d) 均将文档数据看作是 Bag of word[98], 文档中的 Word 满足可交换性.

HDP 相比 LDA 具有更多优势, 两者具有下列不同点:

1) 模型中聚类数目的鲁棒性

a) HDP 模型不但能实现聚类和推断等功能, 而且能够自动生成聚类数目, 因此大大增强了算法的鲁棒性;

b) LDA 的聚类数目需人为提前给定.

2) 模型对其他算法的构造作用

a) Dirichlet 过程和 HDP 模型作为一种结构单元, 广泛应用于 HDP-HMM 模型等其他算法的构造中, 因此 HDP 模型不再是一种单纯的主题提取算法;

b) LDA 是参数贝叶斯模型, 且模型的概率分布受到限制, 因此, 对其他模型的构造受到限制.

3) 模型的复杂度

a) Dirichlet 过程和 HDP 均是非参数贝叶斯模型, 是随机过程的应用, 模型较为复杂;

b) LDA 是参数贝叶斯模型, 模型中主题和单词的分布均是简单的概率分布.

6　需要进一步研究的问题探讨和结论

机器学习和人工智能是一类迅速发展的学科, 即使是对该领域的某一类模型也很难形成既有充分信息支持又有充分根据的展望. 因此, 作者在此谨慎地对 Dirichlet 过程相关的一类模型需要进一步发展的几个方面作一探讨:

1) 目前基于 Dirichlet 过程的相关算法, 大多采用 MCMC 采样计算, 这一类算法有两个问题: 一是计算量大, 二是收敛问题没有得到解决. 因此当实验数据非常庞大, 尤其是文字处理中数据量非常大时, 实验很难进行下去. 研究快速、有效的采样算法是一种解决途径.

2) 虽然 Dirichlet 过程已经有变分近似解, 但是这种形式的解通常仅仅能完成简单模型的近似解, 对更加复杂的 HDP 或 HDP-HMM 等复杂的模型很难得到良好的近似解, 而且即使是变分推断的方式, 计算量仍然是一个问题, 因此, 在理论上需要进一步的研究和论证.

3) 随着数据量的迅速增多, 数据的内容越来越丰富, 比如视频监控数据和股票交易信息等很多数据具有时间性或 (和) 空间性. 未来进行数据分析和建模时, 必然要考虑到数据的时间、空间性. 因此, 充分利用 Dirichlet 过程良好聚类性质的基础上, 进一步建立和发展动态模型系统是很有必要的工作.

4) 在人工智能、机器学习等的模型和算法中, 实现数据的聚类学习是一项重要任务, Dirichlet 过程作为一种先验分布具备的良好的聚类性质. 因此, 和 HDP-HMM 类似, 将 Dirichlet 过程作为一种基本的算法框架应用于其他算法中, 也不失为一种好的应用前景, 比如可以将 Dirichlet 过程应用于条件随机场等对数据进行分类学习的模型中.

Dirichlet 过程作为先验分布的各种算法是一类新发展起来的聚类模型, 为机器学习、人工智能等领域提供了新的研究方法. 其良好的聚类性质和直观、简明的概率图模型表示, 使其受到了广泛的关注. 本文根据近年来 Dirichlet 过程和以其为先验分布的各种非参数贝叶斯模型的发展和应用为主线, 介绍 Dirichlet 过程、HDP 模型和 HDP-HMM 模型, 并介绍了以这些模型为代表的各种算法在静态图像内容处理、视频监控内容分析、文本分析、认知研究等领域的应用现状. 最后探讨了 Dirichlet 过程在理论和应用中有待研究和发展的几个问题.

References

1 Mitchell T M. *Machine Learning*. New York: McGraw-Hill, 1997

2 Teh Y W. Dirichlet processes. *Encyclopedia of Machine Learning*, Springer, 2010. Part 5, 280-287

3 Teh Y W, Jordan M I. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics Principles and Practice*. Cambridge University Press, 2009. 1−47

4 Teh Y W, Jordan M I, Beal M J, Blei D M. Sharing clusters among related groups: hierarchical Dirichlet processes. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: The MIT Press, 2004. 1385−1392

5 Teh Y W, Jordan M I, Beal M J, Blei D M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006, **101**(476): 1566−1581

6 Yakhnenko O, Honavar V. Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence. In: Proceedings of the SIAM International Conference on Data Mining. Sparks, USA: SIAM, 2009. 281−294

7 Wang X G, Ma X K, Grimson W E L. Unsupervised activity perception by hierarchical Bayesian models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE, 2007. 1−8

8 Wang X, Tieu K, Gee-Wah N, Grimson W E L. Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE, 2008. 1−8

9 Wang X G, Ma X X, Grimson W E L. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(3): 539−555

10 Wang X G, Grimson W E L, Westin C F. Tractography segmentation using a hierarchical Dirichlet processes mixture model. In: Proceedings of the 21st International Conference on Information Processing in Medical Imaging. Williamsburg, USA: Springer, 2009. 101−113

11 Sudderth E B, Torralba A, Freeman W T, Willsky A S. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 2008, **77**(1−3): 291−330

12 Fox E B, Sudderth E B, Jordan M I, Willsky A S. An HDP-HMM for systems with state persistence. In: Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008. 312−219

13 Goldwater S, Griffiths T L, Jordan M I. A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, 2009, **112**(1): 21−54

14 Sharif-Razavian N, Zollmann A. An overview of nonparametric Bayesian models and applications to natural language processing [Online], available: http://www.cs.cmu.edu/~zollmann/publications/nonparametric.pdf, August 16, 2009

15 Xu Qian, Zhou Jun-Sheng, Chen Jia-Jun. Dirichlet process and its applications in natural language processing. *Journal of Chinese Information Processing*, 2009, **23**(5): 25−46 (徐谦, 周俊生, 陈家骏. Dirichlet 过程及其在自然语言处理中的应用. 中文信息学报, 2009, **23**(5): 25−46)

16 Ferguson T S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1973, **1**(2): 209−230

17 Jordan M I. Dirichlet processes, Chinese restaurant processes, and all that. In: Proceedings of the Tutorial Presentation at the NIPS Conference. Whistler, Canada: The MIT Press, 2005

18 Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 1994, **4**: 639−650

19 Pitman J. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformation of an interval partition. *Combinatorics, Probability and Computing*, 2002, **11**(5): 501−514

20 Sudderth E B. Graphical Models for Visual Object Recognition and Tracking [Ph. D. dissertation], Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA, 2006

21 Blackwell D, MacQueen J B. Ferguson distributions via Polya Urn schemes. *The Annals of Statistics*, 1973, **1**(2): 353−355

22 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006

23 Gelman A, Carlin J B, Stern H S, Rubin D B. *Bayesian Data Analysis (Second Edition)*. Florida: Chapman and Hall /CRC, 2003

24 Pitman J. Combinatorial stochastic processes. *Lecture Notes in Mathematics 1875*. Berlin: Springer-Verlag, 2006

25 Escobar M D, West M, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 1995, **90**(430): 577−588

26 Antoniak C E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 1974, **2**(6): 1152−1174

27 Jordan M I. Bayesian nonparametric learning: expressive priors for intelligent systems. In: Proceedings of the Symposium on Heuristics, Probability and Causality: a Tribute to Judea Pearl. UCLA, USA: College Publications, 2010. 167−186

28 Hjort N L, Holmes C, Muller P, Walker S G. *Bayesian Nonparametrics*. Cambridge: Cambridge University Press, 2010

29 Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. Massachusetts: The MIT Press, 2009

30 Jordan M I. *Learning in Graphical Models*. Massachusetts: The MIT Press, 1998

31 Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008, **1**(1−2): 1−305

32 Yakhnenko O, Honavar V. Annotating images and image objects using a hierarchical Dirichlet process model. In: Proceedings of the 9th International Workshop on Multimedia Data Mining. New York, USA: ACM, 2008. 1−7

33 Blei D M, Jordan M I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 2006, **1**(1): 121−144

34 Wang C, Blei D. Variational inference for the nested Chinese restaurant process. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, 2009. 1−9

35 Kurihara K, Welling M, Teh Y W. Collapsed variational Dirichlet process mixture models. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2007. 2796−2801

36 Casella G, George E I. Explaining the Gibbs sampler. *The American Statistician*, 1992, **46**(3): 167−174

37 MacEachern S. Computational methods for mixture of Dirichlet process models. *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer-Verlag, 1998. 23−43

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

39 Wang X G. Learning Motion Patterns Using Hierarchical Bayesian Models [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 2009

40 Fox E B. Bayesian Nonparametric Learning of Complex Dynamical Phenomena [Ph. D. dissertation], Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA, 2009

41 Ishwaran H, Zarepour M. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 2002, **30**(2): 269−283

42 Rasmussen C E. The infinite Gaussian mixture model. In: Proceedings of the Advances in Neural Information Processing Systems. Denver, USA: The MIT Press, 2000. 554−560

43 Ishwaran H, James L F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 2001, **96**(453): 161−173

44 Bouguila N, Ziou D. A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling. *IEEE Transactions on Neural Networks*, 2010, **21**(1): 107−122

45 Fox E B, Choi D S, Willsky A S. Nonparametric Bayesian methods for large scale multi-target tracking. In: Proceedings of the 40th Asilomar Conference on Signals, Systems, and Computers. Pacific Grove, USA: IEEE, 2006. 2009−2013

46 Orbanz P, Buhmann J M. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 2008, **77**(1−3): 25−45

47 Zhang Z H, Dai G, Jordan M I. Matrix-variate Dirichlet process mixture models. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: The MIT Press, 2010. 980−987

48 Qi Y T, Liu D H, Dunson D, Carlin L. Multi-task compressive sensing with Dirichlet process priors. In: Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008. 768−775

49 MacEachern S N. Dependent nonparametric processes. In: Proceedings of the Section on Bayesian Statistical Science. Alexandria, USA: American Statistical Association, 1999. 50−55

50 MacEachern S N. Decision theoretic aspects of dependent nonparametric processes. In: Proceedings of Bayesian Methods with Applications to Science, Policy, and Official Statistics. Crete, Greece: International Society for Bayesian Analysis, 2001. 351−360

51 MacEachern S N, Kottas A, Gelfand A E. Spatial Nonparametric Bayesian Models, Technical Report 2001-10, Department of Statistical Science, Duke University, USA, 2001

52 Gelfand A E, Kottas A, MacEachern S N. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 2005, **100**(471): 1021−1035

53 Gelfand A E, Guindani M, Petrone S. Bayesian nonparametric modeling for spatial data using Dirichlet processes. *Bayesian Statistics*, **8**: 1−26

54 Griffin J E, Steel M F J. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 2006, **101**(473): 179−194

55 Rodriguez A, Horsty E T. Bayesian dynamic density estimation. *Bayesian Analysis*, 2008, **3**(2): 339−366

56 Chung Y S, Dunson D B. The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, 2009, **63**(1): 59−80

57 Caron F, Davy M, Doucet A, Duflos E, Vanheeghe P. Bayesian inference for dynamic models with Dirichlet process mixtures. In: Proceedings of the 9th International Conference on Information Fusion. Florence, Italy: IEEE, 2006. 1−8

58 Caron F, Davy M, Doucet A, Duflos E, Vanheeghe P. Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Transactions on Signal Processing*, 2008, **56**(1): 71−84

59 Griffithes T L, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences. California, USA: National Academy of Sciences, 2004. 5228−5235

60 Steyvers M, Smyth P, Rosen-Zvi M, Griffithes T L. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2004. 306−315

61 Teh Y W. Nonparametric Bayesian mixture models [Online], available: http://www.gatsby.ucl.ac.uk/~ywteh/research/software.html, June 20, 2009

62 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, **77**(2): 257−286

63 Beal M J, Ghahramani Z, Rasmussen C E. The infinite hidden Markov model. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: The MIT Press, 2001. 577−584

64 Gael J V, Saatci Y, Teh Y W, Ghahramanim Z. Beam sampling for the infinite hidden Markov model. In: Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008. 1088−1095

65 Hu D H, Zhang X X, Yin J, Zheng V W, Yang Q. Abnormal activity recognition based on HDP-HMM models. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2009. 1715−1720

66 Sudderth E B, Torralba A, Freeman W T, Willsky A S. Describing visual scenes using transformed Dirichlet processes. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: The MIT Press, 2005. 1297−1304

67 Sudderth E B, Torralba A, Freeman W T, Willsky A S. Learning hierarchical models of scenes, objects and parts. In: Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005. 1331−1338

68 Xing E, Sharan R, Jordan M I. Bayesian haplo-type inference via the Dirichlet process. In: Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM, 2004. 111−118

69 Xing E P, Sohn K A, Jordan M I, Teh T W. Bayesian multipopulation haplotype inference via a hierarchical Dirichlet process mixture. In: Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM, 2006. 1049−1056

70 Cowans P J. Information retrieval using hierarchical Dirichlet processes. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2004. 564−565

71 Canini K R, Griffithes T L. The hierarchical Dirichlet process as a model of human category learning. In: Proceedings of the Workshop on Machine Learning Meets Human Learning. Vancouver, Canada: The MIT Press, 2008

72 Canini K R, Shashkov M M, Griffiths T L. Modeling transfer learning in human categorization with the hierarchical Dirichlet process. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress, 2010. 151−158

73 Li X, Hu W, Zhang Z, Zhang X, Luo G. Trajectory-based video retrieval using Dirichlet process mixture models. In: Proceedings of British Machine Vision Conference. Leeds, UK: The British Machine Vision Association, 2008. 1−10

74 Gael J V, Vlachos A, Ghahramani Z. The infinite HMM for unsupervised PoS tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM, 2009. 678−687

75 Zhang X X, Liu H, Gao Y, Hu D H. Detecting abnormal events via hierarchical Dirichlet processes. In: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin, Germany: Springer-Verlag, 2009. 278−289

76 Pruteanu-Malinici I, Carin L. Infinite hidden Markov models and ISA features for unusual-event detection in video. In: Proceedings of IEEE International Conference on Image Proceessing. San Antonio, USA: IEEE, 2007. 137−140

77 Pruteanu-Malinici I, Carin L. Infinite hidden Markov models for unusual-event detection in video. *IEEE Transactions on Image Processing*, 2008, **17**(5): 811−822

78 Xu T B, Zhang Z F, Yu P S, Long B. Dirichlet process based evolutionary clustering. In: Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008. 648−657

79 Xu T B, Zhang Z F, Yu P S, Long B. Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state. In: Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008. 658−667

80 Fox E, Sudderth E B, Willsky A S. Hierarchical Dirichlet processes for tracking maneuvering targets. In: Proceedings of the 10th International Conference on Information Fusion. Quebec, Canada: IEEE, 2007. 1−8

81 Zhu L, Chen Y, Freeman W, Torralba A. Nonparametric Bayesian texture learning and synthesis. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: Curran Associates, 2009. 1−9

82 Kivinen J J, Sudderth E B, Jordan M I. Image denoising with nonparametric hidden Markov trees. In: Proceedings of the IEEE International Conference on Image Processing. San Antonio, USA: IEEE, 2007. 121−124

83 Kivinen J J, Sudderth E B, Jordan M I. Learning multiscale representations of natural scenes using Dirichlet processes. In: Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1−8

84 Sohn K A, Xing E P. Hidden Markov Dirichlet process: modeling genetic recombination in open ancestral space. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: The MIT Press, 2006. 1305−1312

85 Xing E P, Sohn K A. Hidden Markov Dirichlet process: modeling genetic recombination in open ancestral space. *Bayesian Analysis*, 2007, **2**(2): 501−528

86 Qi Y T, Paisley J W, Carin L. Music analysis using hidden Markov mixture models. *IEEE Transactions on Signal Processing*, 2007, **55**(11): 5209−5224

87 Ren L, Dunson D B, Carin L. The dynamic hierarchical Dirichlet process. In: Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008. 824−831

88 Ren L, Dunson D B, Lindroth S, Carin L. Music analysis with a Bayesian dynamic model. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, China: IEEE, 2009. 1681−1684

89 Ren L, Dunson D B, Lindroth S, Carin L. Dynamic nonparametric Bayesian models for analysis of music. *Journal of the American Statistical Association*, 2009, **105**(490): 458−472

90 Pruteanu-Malinici I, Ren L, Paisley J, Wang E, Carin L. Hierarchical Bayesian modeling of topics in time-stamped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(6): 996−1011

91 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**: 993−1022

92 Blei D M. Probabilistic Models of Text and Images [Ph. D. dissertation], University of California at Berkeley, USA, 2004

93 Heinrich G. Parameter Estimation for Text Analysis. Technical Report No. 09RP008-FIGD, Fraunhofer Institute for Computer Graphics, Germany, 2009

94 Steyvers M, Griffiths T L. Probabilistic topic models. *Handbook of Latent Semantic Analysis*. New Jersey: Lawrence Erlbaum Associates, 2007

95 Diane H, Lawrence S. Latent Dirichlet allocation for text, images, and music [Online], available: http://cseweb.ucsd.edu/~dhu/exam.html, December 20, 2009

96 Wei X, Croft W B. LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2006. 178−185

97 Blei D M, Lafferyt J D. Topic models. *Text Mining: Classification, Clustering, and Applications*. Florida: Chapman and Hall/CRC, 2009. 71−94

98 Lewis D D. Naive (Bayes) at forty: the independence assumption in information retrieval. *Lecture Notes in Computer Science*. Berlin: Springer, 1998. 4−15

周建英 中国科学院自动化研究所博士研究生. 主要研究方向为计算机视觉、模式识别与机器学习在智能交通中的应用. 本文通信作者.
E-mail: zhoujianyingbest@gmail.com
(**ZHOU Jian-Ying** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. Her research interest covers applications of computer vision, pattern recognition and machine learning in intelligent transportation systems (ITS). Corresponding author of this paper.)

王飞跃 中国科学院自动化研究所研究员. 主要研究方向为智能系统, 社会计算, 复杂系统的建模、分析和控制.
E-mail: feiyue@ieee.org
(**WANG Fei-Yue** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers intelligent control, social computing, and modeling, analysis, and control mechanism of complex systems.)

曾大军 中国科学院自动化研究所研究员. 主要研究方向为多智能体/代理系统及应用、随机图论与复杂系统建模与控制、情报与安全信息学、传染病信息学、时空数据分析、竞拍和谈判决策支持与自动推荐系统.
E-mail: dajun.zeng@ia.ac.cn
(**ZENG Da-Jun** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers software agents and multi-agent systems, random graphs and complex systems analysis, intelligence and security informatics, infectious disease informatics, spatio-temporal data analysis and online surveillance, automated negotiation and auction, and recommender systems.)