

一种基于非参数贝叶斯模型的聚类算法

张媛媛

(宁波大学 信息科学与工程学院, 浙江 宁波 315211)

摘要: 鉴于聚类分析是机器学习和数据挖掘领域的一项重要技术, 并且与监督学习不同的是聚类分析中没有类别或标签的指导信息, 所以如何选择合适的聚类个数(即模型选择)一直是聚类分析中的难点. 由此提出了一种基于 Dirichlet 过程混合模型的聚类算法, 并用 collapsed Gibbs 采样算法对混合模型的参数进行估计. 新算法基于非参数贝叶斯模型的框架, 能够在不断的采样过程中优化模型参数并形成合适的聚类个数. 在人工合成数据集和真实数据集上的聚类实验结果表明: 基于 Dirichlet 过程混合模型的聚类算法不但能够自动确定聚类个数, 而且具有较强灵活性和鲁棒性.

关键词: 非参数贝叶斯模型; Dirichlet 过程混合模型; 聚类分析; Gibbs 采样

中图分类号: TP391

文献标志码: A

文章编号: 1001-5132 (2013) 04-0024-05

聚类分析是指在没有类别或标签信息的情况下, 将一组数据(或模式)进行分类的统计分析方法. 经过聚类分析后, 属于同类的数据比在不同类中的数据更具相似性. 经典的聚类算法有 K 均值聚类算法(K-means Algorithm)、基于 EM(Expectation Maximization)优化的有限混合模型(Finite Mixture Models)算法以及基于密度的方法(Density-based Methods)等^[1-2]. 但是这些算法都需要预先设定聚类或混合部件的个数, 若个数选择不当会导致模型的过拟合或欠拟合情况的发生. 因此, 如何选择合适的聚类个数是聚类分析的关键, 也是模式分类中的模型选择问题. 针对此问题, 一种直接的解决方式是尝试所有可能聚类个数的模型结构, 然后根据一定准则折中选取一个最优的、不易发生过拟合或欠拟合的模型. 常见评价方法有基于编码理论和最小信息长度的准则(Minimum Message Length, MML)、Akaike 信息准则(Akaike's Information Criterion, AIC)、贝叶斯信息准则(Bayesian Information Criterion, BIC)和最小描述长度准则(Minimum Description Length, MDL)^[3]. 但是现有的这些模型选择方法不仅计算量较大, 而且泛化

能力较弱, 尽管对训练样本数据拟合得较好, 但对未知数据可能拟合不好.

针对以上问题, 笔者设计实现了一种基于 Dirichlet 过程混合模型(Dirichlet Process Mixture Model, DPMM)的非参数聚类方法. 它为模型选择问题提供了一个比较灵活的解决之道, 是一种非参数贝叶斯模型(Nonparametric Bayesian Models), 也是近年来统计学习理论的研究热点^[4-6]. 与传统的参数模型相比, DPMM 采用单一复杂的混合模型, 并根据观测数据自动优化模型的结构, 使模型的分布参数随着观测数据进行调整^[7-8]. 实验结果表明, 当 DPMM 用于聚类分析时, 无需预先指定类别数, 而是根据观测数据自动计算目前所需的类别数, 并以概率方式允许将来的新数据出现时产生新的类别. 在计算 DPMM 后验概率时, 由于后验概率不能被直接计算, 所以我们采用 Gibbs 采样法反复从条件分布中采样并更新模型参数. Gibbs 采样法属马尔可夫链蒙特卡罗方法(Markov Chain Monte Carlo, MCMC), 是替代精确推理的近似推理方法. 和其它 MCMC 算法一样, Gibbs 采样法产生一条马尔可夫链, 经过足够次数的采样后,

收稿日期: 2013-06-09.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家自然科学基金(61175026); 浙江省新一代移动互联网用户端软件科技创新团队项目(2010R50009); 宁波市自然科学基金(2011A610193); 宁波大学学科项目(XKL09154).

作者简介: 张媛媛(1981-), 女, 浙江宁海人, 博士研究生/实验师, 主要研究方向: 机器学习和计算机视觉. E-mail: zhangyuanyuan@nbu.edu.cn

该马尔可夫链可以达到稳态.

1 Dirichlet 过程混合模型

1.1 Dirichlet 过程

Dirichlet 过程是 Ferguson 在 1973 年提出的一种随机过程, 它是一种描述概率分布的分布^[9-11]. 假设 H_0 是测度空间 Ω 上的一个概率分布, α_0 是 1 个正实数, 那么对空间 Ω 任意一种有限剖分 $(T_1, T_2, \dots, T_k)$, 如果随机概率测度 G 都满足下式:

$$G(T_1), \dots, G(T_k) \sim \text{Dir}(\alpha_0 H_0(T_1), \dots, \alpha_0 H_0(T_k)), \quad (1)$$

则称 G 服从 Dirichlet 过程, 记作 $G \sim DP(\alpha_0, H_0)$.

由于按照以上的数学定义无法对 Dirichlet 过程实现采样和对未知数据的预测, 所以在实际应用中往往采用 Dirichlet 过程的其它构造形式, 比如 stick-breaking 构造过程就适合于采样算法, 在计算新观测数据的预测分布时, 采用的是 Polya 罐子构造模型(Polya Urn Model, PUM)和中国餐馆过程(Chinese Restaurant Process, CRP).

假设混合系数 $\pi = \{\pi_k\}_{k=1}^\infty$ 是从以下的 stick-breaking 过程中得到的 1 个无穷序列^[6]:

$$\begin{aligned} v_k &\sim \text{Beta}(1, \alpha_0), \\ \pi_k &= v_k \sum_{l=1}^{k-1} (1 - v_l), k = 1, \dots, \infty. \end{aligned} \quad (2)$$

H_0 是测度空间上 Ω 的一个基分布, 若概率测度 G 满足下列条件:

$$\begin{aligned} G(\theta) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \\ \theta_k &\sim H_0, \end{aligned} \quad (3)$$

则该测度 G 服从 Dirichlet 过程分布, 混合系数序列 π 的产生过程记作 $\pi \sim GEM(\alpha_0)$, 而 GEM 为 3 个人名 Griffiths, Engen 和 McCloskey 的缩写.

假设 G 服从 Dirichlet 过程: $G \sim DP(H_0, \alpha_0)$, $\eta_1, \dots, \eta_n$ 是服从分布 G 的独立同分布的随机观测变量, $\eta_1^*, \dots, \eta_K^*$ 是其观测到的 K 个不同取值, 那么新的随机观测变量 η_{n+1} 的预测条件分布具有以下形式:

$$\begin{aligned} p(\eta_{n+1} = \eta^* | \eta_1, \dots, \eta_n, \alpha_0, H_0) &= \\ \frac{1}{n + \alpha_0} \sum_{k=1}^K N_k \delta(\eta^*, \eta_k^*) + \frac{\alpha_0}{n + \alpha_0} H_0(\eta^*), \end{aligned} \quad (4)$$

其中, N_k 是指序列 $\eta_1, \dots, \eta_n$ 中其值等于 η_k^* 的个数.

从(4)式可见, Dirichlet 过程具有很好的聚类性

质, 它可以具有相同值的观测变量聚为一类. 假设 $c_1, \dots, c_n$ 表示 $\eta_1, \dots, \eta_n$ 所属的类标签变量, 那么下个观测变量的类标签 c_{n+1} 服从以下分布:

$$\begin{aligned} p(c_{n+1} = j | c_1, \dots, c_n) &= \\ \frac{1}{n + \alpha_0} \sum_{k=1}^K N_k \delta(j, k) + \frac{\alpha_0}{n + \alpha_0} \delta(j, \bar{c}), \end{aligned} \quad (5)$$

其中, $\bar{c}$ 表示产生新的一个类别.

(5)式表示的概率分布可以用中国餐馆过程来形容: 假设有 1 个中国餐馆可以容纳无限张餐桌, 每张餐桌可以容纳无限数量的顾客. 在这里, 桌子对应某个类别 c_i , 而顾客对应观测变量 η_i . 根据(5)式, 第 η_{n+1} 个顾客将以正比于已经就坐于第 k 张桌子的顾客数 N_k 的概率坐于第 k 张餐桌, 而且以正比于 α_0 的概率就坐于 1 张新的餐桌. 由此可知, 某个餐桌上就座的顾客越多, 就越有可能加入新的顾客, 这也是一种 “rich-gets-richer” 的性质.

1.2 Dirichlet 过程混合模型

Dirichlet 过程最重要的应用是作为概率模型中参数的先验分布. 假设有观测变量 $y_1, y_2, \dots$ 是服从下列分布的独立同分布的变量:

$$\begin{aligned} G &\sim DP(\alpha_0, H_0), \\ \eta_i | G &\sim G, \\ y_i | \eta_i &\sim F(\eta_i). \end{aligned} \quad (6)$$

观测变量 y_i 服从参数为 η_i 的分布 $F(\eta_i)$, G 是参数 η_i 的先验分布, 假如 G 服从基分布 H_0 和集中参数 α_0 的 Dirichlet 过程, 则称此模型为 Dirichlet 过程混合模型(DPMM), 其图模型结构如图 1 所示. 为计算方便, $F(\eta_i)$ 往往是属于指数族分布的, 而 H_0 是相应的共轭先验分布. Dirichlet 过程只能把相同值的数据聚为一类, 但是 DPMM 可以将服从相同参数分布的数据聚为一类, 所以特别适用于基于模型的聚类分析.

2 基于 DPMM 的聚类算法

基于 DPMM 聚类算法的基本思想是用混合部件(Components)的概率分布来拟合观测数据, 这里的每个观测数据来自其中某个部件, 并用一个隐变量指示数据的来源. 理论上 DPMM 有无限多个部件, 但是只用其中的一部分部件来解释现有的观测数据.

在计算图 1 模型的后验分布时, 用 Gibbs 采样

法从后验分布 $p(\eta_i | \eta_{-i}, y_i)$ 采样得到 η_i , 下标 $-i$ 表示除 i 以外的所有数据的集合. 由于这种直接的 Gibbs 采样法每次给定 1 个观测值后才更新 1 次, 所以收敛速度非常慢, 而且不适用于大规模的聚类分析. 因此笔者考虑在 DPMM 的另一个等价模型^[12-14]的基础上用 collapsed Gibbs 采样法对参数进行采样, 即不直接采样 η_i , 而是采样观测变量的类标签 c_i , 然后更新部件 c_i 的参数 ϕ_{c_i} ^[6,15](图 2).

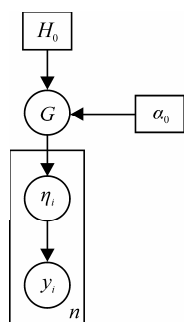


图 1 DPMM 的图模型结构

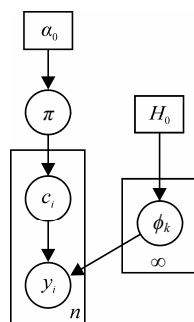


图 2 DPMM 的无限混合模型结构

在这种构造方式中, 混合系数的先验分布服从参数为 α_0 的 stick-breaking 过程, ϕ_k 表示部件 k 的分布参数, H_0 是 ϕ_k 的先验分布, 其构造形式是:

$$\begin{aligned} \pi | \alpha_0 &\sim \text{GEM}(\alpha_0), \\ c_i | \pi &\sim \pi, \\ \phi_k | H_0 &\sim H_0, \\ y_i | c_i &\sim F(\phi_{c_i}). \end{aligned} \quad (7)$$

给定独立同分布的观测数据 $y_1, \dots, y_n$ 和其它观测数据的类标签的条件下, 数据 y_i 的类标签 c_i 的后验分布服从以下条件分布形式:

$$p(c_i | y_1, \dots, y_n, c_{-i}, \alpha_0) \propto p(c_i | c_{-i}, \alpha_0) p(y_i | y_{-i}, c_i, c_{-i}). \quad (8)$$

(8)式右边第 1 项可以由中国餐馆过程构造, $\bar{c}$ 表示产生的新类别:

$$p(c_i = j | c_{-i}, \alpha_0) = \frac{1}{n-1+\alpha_0} \sum_{k=1}^K N_k \delta(j, k) + \frac{\alpha_0}{n-1+\alpha_0} \delta(j, \bar{c}). \quad (9)$$

(8)式右边的第 2 项是 y_i 的预测分布, 若 c_i 是已存在的类标签, 则可以由下列式子计算:

$$p(y_i | y_{-i}, c_i = j, c_{-i}) = p(y_i | \{y_k | c_k = j, k \neq i\}). \quad (10)$$

若 c_i 是新产生的类标签, 则 y_i 的预测分布是:

$$p(y_i | y_{-i}, c_i = j, c_{-i}) = \int F(y_i | \phi_j) dH_0(\phi_j). \quad (11)$$

Gibbs 采样的基本思想是在隐变量上定义 1 个马尔可夫链, 每个隐变量都是在给定其它隐变量和数据的条件分布中采样得到, 它的平稳分布是隐变量的后验分布. 若从图 2 所示的 DPMM 产生了 n 个观测序列 $y_1, \dots, y_n$, 则 Gibbs 采样为每个观测值分配类标签 $c_1, \dots, c_n$, 然后更新各部件的分布参数 ϕ_k 和类别数目 K .

综上所述, 当给定集中参数 α_0 及第 t 次采样后观测数据的分类结果 $\{c_i\}_{i=1}^t$ 和聚类个数 K 的情况下, collapsed Gibbs 采样法按照以下过程对类标签进行第 $t+1$ 次采样:

- (1) 将 n 个观测数据随机排序, 得到 $\sigma(1), \dots, \sigma(n)$.
- (2) 对每个数据 $\sigma(1), \dots, \sigma(n)$ 分别对其类标签 $\{c_i\}_{i=1}^{t+1}$ 采样: ①对现有的 K 个聚类按照(10)式计算 y_i 对每一个类的预测估计. 然后增加一个新类 $\bar{c}$, 并计算 y_i 属于该新类的概率估计. ②根据(8)式计算该 y_i 所属类别 c_i^{t+1} 的后验概率分布 $p(c_i^{t+1} | y_1, \dots, y_n, c_{-i}, \alpha_0)$, 并从该后验分布中采样得到 c_i^{t+1} .
- (3) 检查上述 $K+1$ 个类内的观测数据量, 如果某类为空, 则将该类移除, 并且将聚类总数减 1.

3 聚类分析试验

3.1 合成数据集

图 3 是 DPMM 用于人工合成数据集的实验. 该数据是由多个不同形状的高斯分布生成的二维观测值, 每个观测值是由其中某个高斯分布产生. 这 2 个数据集分别包括 $n=400$ 和 $n=600$ 个数据点, 真实的混合个数分别为 $k_1=5$ 和 $k_2=7$. 初始参数 α_0 的取值为: $\alpha_0=1$, 每个聚类用不同深浅的椭圆表示各类的密度轮廓. 图中每排图像显示在同个数据集上分别经过若干次 Gibbs 采样后得到的结果. 例如在对第 1 个数据集聚类时, 当 Gibbs 初始采样 7 次的时候, 算法只产生几个聚类; 经过 26 次采样后, 算法产生了很多冗余的聚类; Gibbs 采样经过 36 次循环后迅速找到了比较合适的聚类个数, 而且已经与真实模型非常地接近. 在此实验中, 聚类相当于高斯核, 因此 DPMM 也可以用于混合模型的核密度估计.

3.2 真实数据集

实验中, 笔者采用 2 个从现实世界中收集的数

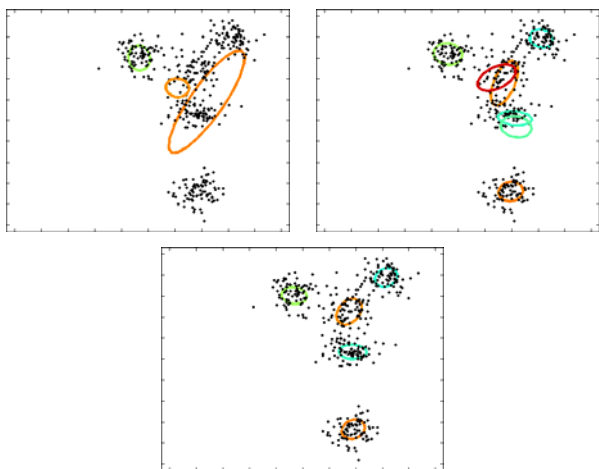
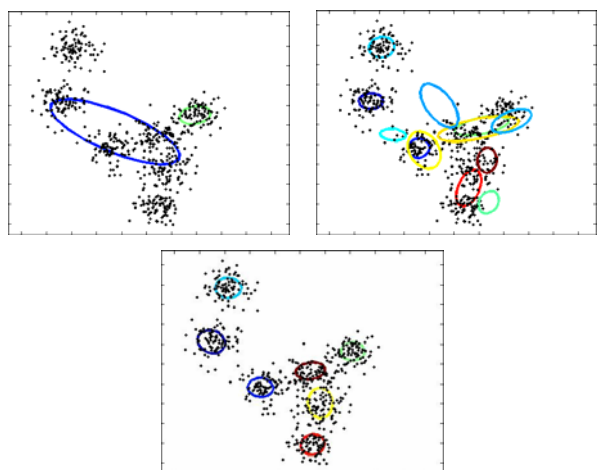
(a) $k_1=5, n=400$, 从左至右而下分别表示 7 次、26 次、36 次采样结果(b) $k_2=7, n=600$, 从左至右而下分别表示 10 次、42 次、80 次采样结果

图3 聚类实验结果

数据集来测试 DPMM 聚类算法性能. 第 1 个测试集来自 UCI 机器学习数据库(<http://archive.ics.uci.edu/ml/datasets/Iris>)的鸢尾科植物数据集(Iris dataset), 这是模式识别中经常用来测试聚类性能的数据集. 该数据集包含 3 种类别的鸢尾科植物, 每个类别有 50 个实例数据. 第一类和其它两类线性分离, 剩下的两类彼此非线性分离. 另 1 个测试集是来自 ML comp(<http://mlcomp.org/datasets/167>)的真实图像数据集, 由 7 个室外图像数据组成, 包括天空、路径、草地、天空、树叶等图像, 每个实例有 19 个连续属性.

从(8)式可见, 后验分布与集中参数 α_0 的选择相关, 所以为减少 α_0 对聚类结果的影响, 笔者选择一个弱信息的概率分布作为 α_0 的先验分布^[6]. 在实验中, 设置 α_0 服从伽玛分布: $\alpha_0 \sim \text{Gamma}(\kappa_1,$

$\kappa_2)$, 其中, κ_1, κ_2 是超参数. 由此, α_0 也是 DPMM 的 1 个隐变量, 因此在每次 Gibbs 采样中, 也要对 α_0 进行采样, 实验中超参数取值为 $\kappa_1 = 0.2, \kappa_2 = 0.1$.

由于 V-Measure 是基于同质性和完整性的一个聚类评价标准, 因此笔者选用 V-Measure 值来评估聚类的性能. 表 1 显示了该实验得到的结果, 由此可见, 图像数据集的 V-Measure 值低于植物数据集, 这是因为图像特征本身比较复杂, 尤其在无监督学习下更难以区分.

表1 真实数据集的聚类结果

Name	n	d	k	V-Measure
Iris	150	4	3	73%
Image	2310	19	7	65%

注: 每个数据集由 k 个类, n 个实例组成, 每个实例具有 d 维属性.

4 总结

提出了一种基于 DPMM 和 collapsed Gibbs 采样法的聚类算法, 实验结果表明该算法可以自动确定聚类个数和估计混合模型参数, 可用于聚类、模式识别、核密度估计等应用. 但是这种模型仍然有很多值得探讨和发展的地方^[16-18]. 首先, DPMM 适用于对一组数据的聚类分析, 但是不能用于成组数据的建模分析. 我们可以把 DPMM 扩展成分层 Dirichlet 混合模型, 这样就可以在组和组之间实现聚类的共享. 其次, 目前基于 Dirichlet 过程的模型主要采用 Gibbs 采样计算, 但是这种方法不仅计算量大而且往往收敛缓慢, 尤其需要处理的大规模数据量时, 需要用更加快速有效的近似方法来替代, 比如并行采样算法或者变分推断算法. 再次, DPMM 假设生成数据是跟次序无关(即可交换的), 但是在处理有时间或空间依赖数据时, 这种假设显然是不合适的, 所以研究有依赖关系的 Dirichlet 模型(Dependent Dirichlet Models)是很有必要的工作. 最后, 基于 DPMM 的聚类算法是一种无监督的学习, 但是在某些应用中监督信息是非常有用的, 所以如何把这两种学习方式有效地结合在一起也是今后的一个研究内容.

参考文献:

- [1] Jain K A. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31:651-666.

- [2] Bishop M C. Pattern recognition and machine learning [M]. New York: Springer Press, 2007.
- [3] Claeskens G, Hjort N L. Model selection and model averaging[M]. New York: Cambridge University Press, 2008.
- [4] Shahbaba B, Neal R. Nonlinear models using Dirichlet process mixtures[J]. The Journal of Machine Learning Research, 2009, 10:1829-1850.
- [5] Fox E, Sudderth E, Jordan M, et al. Nonparametric Bayesian learning of switching linear dynamical systems [C]. In Advances in Neural Information Processing Systems 21, Vancouver and Whistler, BC, Canada: The MIT Press, 2008.
- [6] Sudderth E. Graphical models for visual object recognition and tracking[D]. MA, Cambridge: Massachusetts Institute of Technology, 2006.
- [7] Gershman S, Blei D. A tutorial on Bayesian nonparametric models[J]. Journal of Mathematical Psychology, 2012, 56:1-12.
- [8] Hjort L N, Holmes M P, Walker G S. Bayesian nonparametric[M]. New York: Cambridge University Press, 2010.
- [9] Ferguson S T. A Bayesian analysis of some nonparametric problems[J]. The Annals of Statistics, 1973, 12:209-230.
- [10] 周建英, 王飞跃, 曾大军. 分层 Dirichlet 过程及其应用综述[J]. 自动化学报, 2011, 37(4):390-407.
- [11] Teh W Y. Dirichlet processes[C]. Machine Learning Summer School Tutorial and Practical Course, 2007.
- [12] Teh W Y, Jordan I M, Beal J M, et al. Hierarchical Dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101:1566-1581.
- [13] Rasmussen E C. The infinite Gaussian mixture model[C]. In Advances in Neural Information Processing Systems 12, MIT Press, 2000:554-560.
- [14] Teh W Y. Bayesian nonparametric: DP mixtures[EB/OL]. [2009-10-11]. <http://www.gatsby.ucl.ac.uk/~ywtteh/teaching/npbayes.html>.
- [15] Neal R. Markov chain sampling methods for Dirichlet process mixture models[J]. Journal of Computational and Graphical Statistics, 2000, 9:249-265.
- [16] Blei M D, Jordan I M. Variational inference for Dirichlet process mixtures[J]. Bayesian Analysis, 2006(1):121-144.
- [17] Blei M D, Frazier P. Distance dependent Chinese restaurant processes[J]. The Journal of Machine Learning Research, 2011, 12:2461-2488.
- [18] Rasmussen, C, Williams C M. Gaussian processes for machine learning[M]. New York: Springer Press, 2006.

Data Clustering via Nonparametric Bayesian Models

ZHANG Yuan-yuan

(Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China)

Abstract: Clustering is one of the most useful techniques in machine learning and data mining. In cluster analysis, model selection concerning how to determine the number of clusters is an important issue. Unlike supervised learning, there are no class labels and criteria to guide the search, so the model for clustering is always difficult to select. To tackle this problem, we present the concept of nonparametric clustering approach based on Dirichlet process mixture model (DPMM), and apply a collapsed Gibbs sampling technique to sample the posterior distribution. The proposed clustering algorithm follows the Bayesian nonparametric framework and can optimize the number of components and the parameters of the model. The experimental result of clustering shows that this Bayes model has promising properties and robust performance.

Key words: nonparametric Bayesian models; Dirichlet process mixture model; clustering analysis; Gibbs sampling

(责任编辑 章践立)