基于 BIC 准则和 Gibbs 采样的有限混合模型无监督学习算法

刘伟峰, 杨爱兰
(杭州电子科技大学自动化学院, 浙江杭州 310018)

摘 要: 针对有限混合模型无监督学习算法分布元个数未知, 本文提出了一种基于 BIC 准则和 Gibbs 采样的无监督学习算法, 通过 Gibbs 采样算法对混合模型的参数进行估计, 进一步计算观测数据的 Bayes 信息准则(BIC)指标, 确定混合分布元个数. 仿真实验以高斯混合分布为例, 分别利用具有不同个数的分布模型拟合观测数据, 分析表明, 该算法能够很好地学习混合高斯分布参数及个数.

关键词: BIC 准则; Gibbs 采样; 无监督学习; 有限混合模型
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2011) 3A-134-06

Unsupervised Learning for Finite Mixture Models Based on BIC Criterion and Gibbs Sampling

LIU Wei-feng, YANG Ai-lan
(*School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018 China*)

Abstract: An unsupervised learning algorithm for finite mixture models (FMM) by using the BIC criterion and the Gibbs sampling is proposed. The FMM parameters are estimated by the Gibbs sampling algorithm. The number of the models is further given by calculating the BIC criterion. In the final simulations, we propose two examples of Gaussian mixture models, where the FMM with different number of components is adopted to fit observation data. The final result shows that the proposed algorithm can effectively estimate the parameters and the number of the components.

Key words: BIC criterion; Gibbs sampling; unsupervised learning; finite mixture models

1 引言

有限混合模型属于一种半参数模型, 用于描述具有不同成分的混合数据, 其研究始于 20 世纪初, 生物学家沃尔登获得了某种蝎类头部和身体长度比值的 1000 组样本值, 他想知道这些样本值包含了几类物种, 数学家皮尔逊把这个问题描述为混合模型, 这应该是有混合模型问题研究的开始[1]. 自此以后, 混合模型研究逐渐受到人们重视. 目前, 它已广泛应用于生物[2], 医学, 图像, 天文和航空等领域.

皮尔逊用两个高斯混合分布模型拟合这组数据, 利用矩法估计模型的 5 个未知参数, 但是矩法对高阶的混合分布却无能为力. 此后, 蒙特卡洛采样方法的提出, 特别是 Markov 链方法, 为高阶混合分布的研究提供了新的手段[3]. 上世纪 70 年代前后, Dempster 提出了基于极大似然方法的 EM 优化算法, 并把 EM 算法用于解决混合模型估计问题[4], 这两种方法都可看作是一种优化学习算法, 时至今日, 混合分布的研究仍然是一个很活跃的领域[5~9].

混合模型研究的问题可以归类为两点: 混合成份个数估计, 混合模型参数估计. 如果混合模型的个数已知, 分布模型已知, 分布参数未知, 这类学习过程就属于有监督的学习. 如果模型个数未知, 则属于无监督的学习算法. 如果分布未知, 则属于非参数学习问题. 从目前主要的学习算法来看, 可以分为非确定的学习算法, 以 Markov 链为代表的 Bayes 学习算法为主. 另一个是确定性学习算法, 以 EM 算法为代表的极大似然学习方法 (EM 也可扩展用于 Bayes 估计). 混合模型无监督的学习问题一直是一个比较棘手的问题. 对此, 非确定算法中主要代表是格林的可逆跳 Markov 链蒙特卡洛(RJMCMC)方法[10], 通过在不同维参数空间进行采样, 获得参数的估计, 混合分布的个数可以通过参数维数确定. 确定性算法通过结合某种信息准则, 例如 AIC, BIC 信息准则等方法[7], 通过优化该信息准则获得混合分布个数. RJMCMC 属于一种采样方法, 混合分布的个数也属于一种概率方法. EM 算法容易受初值影响, 使混合估

收稿日期: 2010-09-08; 修回日期: 2010-12-15
基金项目: 国家自然科学基金(No. 91016020, No. 60934009); 浙江省自然科学基金(No. Y1101218); 杭州电子科技大学科研启动基金(No. KYS065609051)

计陷入局部极值. 正是基于这两类算法各自的不足以及各自的优点, 在前期研究成果中, 作者提出了一种基于修正 Gibbs 采样方法[11], 其中结合了分布元的管理技术, 可以有效提高学习效率, 但是分布元个数获得是一种随机学习过程, 没有一种量值刻化学习效果.

针对这个问题, 本文提出一种基于 Gibbs 采样和 Bayes 信息准则(BIC: (Bayesian information criterion))[12], 作为有限混合分布无监督学习算法, 它结合了不确定方法和确定性方法的优点, 可以有效度量学习效果. 基本思想是利用 Gibbs 采样学习混合模型参数, 然后计算该混合分布的信息准则值, 最终确定在最优准则下的混合分布的个数.

2　Gibbs 采样与混合模型 Bayes 分布

2.1　背景与问题描述

针对 RJMCMC 方法的不足, 文献[11] 提出了基于 Gibbs 采样算法的修正 Gibbs 采样无监督学习算法. 其中, 采样参数维数的变化是利用分布元管理技术来完成的: 管理包括剔除权重较小的分布元, 合并近似的分布元, 该方法的优点是学习效率比较高. 但是存在一个问题, 这种管理也是随机的, 在剔除权重较小的分布元时, 如果采样点处于一个局部平缓区域, 还没有进入稳态, 可能会剔除正确的分布元. 为了解决这个问题, 本文提出一种结合优化指标和采样算法的无监督学习算法, 也是考虑结合确定性算法和不确定性算法的优点, 尽可能避开各自的缺点.

2.2　有限混合模型

具有 m 分布元的 FMM 可以表示为如下形式[11]:

$$f(y_i|\theta)=\pi_1 f_1(y_i|\theta_1)+\cdots+\pi_j f_j(y_i|\theta_j)+\cdots+\pi_m f_m(y_i|\theta_m) \tag{1}$$

其中, $y_i\in R^d$ 是观测数据, θ_j 是第 j 分布元的参数. π_j 是混合权重并且满足 $\pi_j\geqslant 0$, $\sum_{j=1}^{m}\pi_j=1$, 参数 $\theta=\{\pi_1,\cdots,\pi_m,\theta_1,\cdots,\theta_m\}$. 对于 n 个独立观测数据 $y=\{y_1,\cdots y_i,\cdots,y_n\}$, 那么参数似然函数可以表示成:

$$L_y(\theta)=L(\theta|y)=\prod_{i=1}^{n}\sum_{j=1}^{m}\pi_j f_j(y_i|\theta_j) \tag{2}$$

根据式(2), FMM 的参数后验分布可描述为:

$$p(\theta|y)=\frac{1}{C}L_y(\theta)p(\theta) \tag{3}$$

其中, $C=\int L_y(\theta)\mathrm{d}\theta$ 是正则常量. 对每一个获得的观测数据 y_i, 观察者并不知道该观测是由哪个分布产生的, 因此, 定义一个 m 维的指示变量, 称为缺失变量, 每一维指示混合分布中的一个分布元, 显然, 一个观测只能由一个分布产生, 因此, 该 m 维向量只能有一维为 1, 其它各维为 0, 即表示为: $e_i=\{e_{i,1},\cdots,e_{i,m}\}$, 并且满足 $e_{i,j}\in\{0,1\}$, $\sum_{j=1}^{m}e_{i,j}=1$, 其中 $e_{i,j}=1$ 表示数据 y_i 由分布元 $f_j(\cdot)$ 产生. 由于高斯混合分布良好的拟合性能, 本文以高斯混合分布为例, 相应参数集为 $\theta_j=\{\mu_j,\Sigma_j\}$, 其中 μ_j 是均值, Σ_j 是协方差阵.

2.3　Gibbs 采样算法

如何从一个分布中获得采样样本, 最常用的方法是采用蒙特卡洛采样方法, 但是直接从分布中采样往往非常困难, 因此, 一种方法是采用重要性采样的方法, 另一种方法就是马尔科夫链方法; 本文采用第二种方法. 有两类常用的方法产生马尔科夫链, 一类是 Metropolis-Hasting(MH) 算法[13], 另一种是 Gibbs 采样方法[13, 14]. MH 采样的不足是采样函数(重要性采样函数)应该尽可能精确, 否则采样效率会降低很多. 并且 MH 一般是在高维空间采样, 采样难度增加. 其优点是算法在采样函数比较精确的情况下, 采样效率比较高. 由于 MH 需要在高维空间采样, 为降低采样空间的维数, Gibbs 采样是从高维空间中的每一维分别采样, 逐步逼近高维采样点. 其优点是采样难度降低, 但是采样次数会增加. 正是基于此, 本文采用基于 Gibbs 采样的方法. 它可以简单描述如下: 假设 $\varphi(x)=\varphi(x_1,\cdots,x_m)$ 指的是 m 维联合分布, $\varphi(x_i|x_{-i})$ 是全条件分布, 其中, $x_{-i}\triangleq\{x_j, j\neq i\}$, 传统的 Gibbs 采样算法[15] 如下:

(1) 首先, 假设初始值是 $x^{(0)}=\{x_1^{(0)},\cdots,x_m^{(0)}\}$, 然后从条件密度 $\varphi(x_i|x_{-i})$ 采样.

(2) $x_1^{(1)}$ 从 $\varphi(x_1|x_{-1}^{(0)})$ 中采样;

$x_2^{(1)}$ 从 $\varphi(x_1|x_{-2}^{(0)})$ 中采样;

…

$x_m^{(1)}$ 从 $\varphi(x_1|x_{-m}^{(0)})$ 中采样.

(3) 这样我们完成一次从 $x^{(0)}$ 到 $x^{(1)}$ 的转移.

(4) 重复以上各步直到达到 $\{x^{(0)}\}$ 稳态. 基于以上的迭代过程, 我们可以获得一条 Markov 链.

2.4　FMM 参数先验分布

考虑公式(3), 首先分析如何获得先验分布 $P(\theta)$, 在高斯混合情况下, 先验参数为 π_j, μ_j, Σ_j, 那么

$$p(\theta_j)=p(\pi_j,\mu_j,\Sigma_j)=p(\pi_j|\mu_j,\Sigma_j)p(\mu_j|\Sigma_j)p(\Sigma_j) \tag{4}$$

由于混合权重反映的是观测数据所占量成分的多少, 如果各成分比例未知, 最简单的先验分布可以采用等成分的 Dirichelet 分布.

$$p(\pi_1,\cdots,\pi_m)=Dir(\frac{1}{m},\cdots,\frac{1}{m}) \tag{5}$$

均值 μ_j 先验分布采用高斯分布[10],

$$p(\mu_j|\Sigma_j)=N(\mu_j;\tau_j,\Sigma_j) \tag{6}$$

先验方差 $p(\Sigma_j)$ 一般符合 Wishart 分布[14]：

$$p(\Sigma_j^{-1}) = W(\beta_j, \boldsymbol{I}) \tag{7}$$

其中 β_j 是自由度，$\boldsymbol{I}$ 是单位矩阵.

2.5 FMM 参数后验分布

在先验分布分别为式(5)，(6)，(7)，观测数据似然函数为高斯混合条件下，根据 Bayes 更新公式(3)可得：

$$p(\theta|y) \propto L_y(\theta)p(\theta) = \prod_{i=1}^{n}\sum_{j=1}^{m}\pi_j N(y_i, \mu_j, \Sigma_j) \cdot \prod_{j=1}^{m} p(\pi_j|\mu_j, \Sigma_j)p(\mu_j|\Sigma_j)p(\Sigma_j) \tag{8}$$

各个参数的后验分布为：后验权重为 Dirichlet 分布，后验均值为正态分布，后验方差为 Wishart 分布[14]. 下面我们逐个分析这几个参数：

(1)混合权重$\{\pi_j\}$

混合权重满足下面的 Dirichelet 分布：

$$p(\pi_1, \cdots, \pi_m) = Dir(\alpha_1 + l_1, \cdots, \alpha_m + l_m) \tag{9}$$

其中 $\alpha_j > 0$ 是常数. l_j 是属于第 j 个分布元观测数据的个数.

(2)缺失变量$\{e_{i,j}\}$，缺失数据可以根据以下的 Bayes 公式估计

$$\hat{e}_{i,j} = \frac{\pi_j N(y_i; \mu_j, \Sigma_j)}{\sum_{j=1}^{m}\pi_j N(y_i; \mu_j, \Sigma_j)} \tag{10}$$

$$l_j = \sum_{i=1}^{n}\hat{e}_{i,j} \tag{11}$$

(3)方差 Σ_j，方差逆服从 Wishart 分布

$$p(\Sigma_j^{-1}) = W(\alpha_0 + l_j/M_0, \beta_0 + \kappa_j^2/N_0) \tag{12}$$

$$\kappa_j^2 = \frac{\sum_{i=1}^{n}(y_j - \mu_j)(y_j - \mu_j)^{\mathrm{T}}\hat{e}_{i,j}}{\sum_{i=1}^{n}\hat{e}_{i,j}} \tag{13}$$

其中是 α_0，β_0 正常数. M_0，$N_0 > 0$ 起调节作用.

(4)均值 μ_j. 均值满足参数为$\{\xi_j, \Sigma_j\}$的高斯分布，它可以从下式中采样

$$p(\mu_j) = N(\mu_j; \xi_j, \Sigma_j) \tag{14}$$

$$\xi_j = \frac{\sum_{i=1}^{n} y_i\hat{e}_{i,j}}{\sum_{i=1}^{n}\hat{e}_{i,j}} \tag{15}$$

基于以上的式(4)～(10)，我们可以利用 Gibbs 采样算法.

3 结合 BIC 信息准则的分布元估计算法

在各类雷达杂波特征分析中，例如杂波分布建模时[16]，通常需要分析杂波特征，如何估计分布元个数 m (或者模型阶次)有着潜在的应用价值. 本节主要解决混合分布个数的估计问题. 利用极大后验估计方法需要考虑模型阶次和参数联合分布 $p(m, \theta|y)$，可逆跳变的 MCMC(RJMCMC)方法采用模型阶次和参数同时采样的方法，利用极大后验准则的方法确定模型阶次，是一种完全的非确定性方法.

本文是利用 Gibbs 采样算法估计高斯分布的均值方差及各个高斯分布的权重，然后利用 BIC 准则来评价几个高斯分布拟合真实分布最正确，以此来达到无监督学习. 此方法比修正的 Gibbs 采样算法优势体现在估计上更加准确，但计算量相应增加. 为此，我们必须引入某种优化准则确定模型阶次. 最常用的决定模型阶次的方法有 Akaike 信息准则(AIC)[17]，Bayes 信息准则(BIC)[12]，信息编码长度(MCL)[7].

AIC 信息准则是上世纪七十年代日本学者赤池弘次(Akaike，1974)从极大似然法的信息论解释出发[17,18]，提出一个基本信息量的定阶准则. 该准则从随机建模观点出发，借助信息论提出确定模型阶次(分布元个数)的方法，按照使得该准则达到最小值来定阶，AIC 信息准则的定义为[18]：

$$AIC(m) = -2\log L(\theta|z) + 2M \tag{16}$$

上式中的 $\log L(\theta|y) = \sum_{i=1}^{n}\log L(\theta|y_i)$ 为模型参数极大似然估计的对数似然函数，它是各观测对数似然函数之和，M 为独立的模型参数个数. 在一组可供选择的模型类中，使 AIC 达到最小的那个模型是一个可取的模型. 但是 AIC 信息准则在大样本数据时通常会失效，由于似然函数值太大，淹没了模型参数 M 的影响. 因此，为弥补 AIC 准则不足，Schwarz 提出了 BIC 准则[11]：

$$BIC(m) = -2\log(\theta|z) + M\ln n \tag{17}$$

表 1 Gibbs 采样和 BIC 准则算法过程

第一步：Gibbs 采样

$t = 0$

for $t = 0$ to 100

$t = t + 1$

$(\pi_1^{(t)}, \cdots, \pi_m^{(t)}) \sim D(\alpha_1 + l_1^{(t-1)}, \cdots, \alpha_m + l_m^{(t-1)})$

for $j = 1$ to m

$u_{i,j}^{(t)} = N(y_i; u_j^{(t-1)}, \Sigma_j^{(t-1)})$，$e_{i,j}^{(t)} = \pi_j^{(t-1)}u_{i,j}^{(t)} / \sum_{j=1}^{m}\pi_j^{(t-1)}u_{i,j}^{(t)}$

$l_j^{(t)} = \sum_{i=1}^{n} e_{i,j}^{(t)}$，$\xi_j^{(t)} = \sum_{i=1}^{n} y_i e_{i,j}^{(t)} / l_j^{(t)}$

$\kappa_j^{(t)2} = \sum_{i=1}^{n}(y_i - \xi_j^{(t)})(y_i - \xi_j^{(t)})^{\mathrm{T}} e_{i,j}^{(t)} / l_j^{(t)}$

$\Sigma_j^{(t)-1} \sim W(\alpha_0 + l_j^{(t)}/2, \beta_0 + \kappa_j^{(t)2}/2)$

$u_j^{(t)} \sim N(\xi_j^{(t)}, \Sigma_j^{(t)})$

end for

end for

第二步：计算 BIC 准则

$BIC(m) = -2\log L(\theta|z) + M\ln n$

输出：$\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{m}$，$m$，$BIC(m)$

与AIC 信息准则相比, BIC 信息准则右边第二项用 $M\ln n$ 代替了 $2M$, 一般说来 $\ln n \gg 2$, 因此, 由极小化 BIC 定出的模型阶数估计值一般要比 AIC 判定的阶数估计值低. 更重要的是, BIC 信息准则考虑了样本数据的影响, 在区分模型阶数比 AIC 准则更明显. 因此本文采用 BIC 准则来评价用几个高斯分布拟合真实分布最正确. 详细算法见表 1.

4　仿真及实验结果

我们给出两个仿真算例, 一个四个高斯混合, 一个是六个高斯混合, 两个算例中混合分布都有重叠的分布元存在, 我们重点分析 BIC 信息准则和 Gibbs 采样算法的有效性.

4.1　四个高斯混合算例

该算例选自文献[7], 见图 1 所示, 考虑 4 个混合高斯分布, 观测数据 1000 个, 真实混合分布模型为:

$$f(y_i|\theta)=\pi_1 N(y_i;\mu_1,\Sigma_1)+\cdots+\pi_4 N(y_i;\mu_4,\Sigma_4);$$

$$\pi_1=\pi_2=\pi_3=0.3;\ \pi_4=0.1;\ \mu_1=\mu_2=[-4\quad -4]^{\mathrm{T}},$$

$$\mu_3=[2\quad 2]^{\mathrm{T}},\ \mu_4=[-1\quad -6]^{\mathrm{T}};$$

$$\Sigma_1=\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix},\ \Sigma_2=\begin{bmatrix}6 & -2\\ -2 & 6\end{bmatrix},$$

$$\Sigma_3=\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix},\ \Sigma_4=\begin{bmatrix}0.125 & 0\\ 0 & 0.125\end{bmatrix}$$

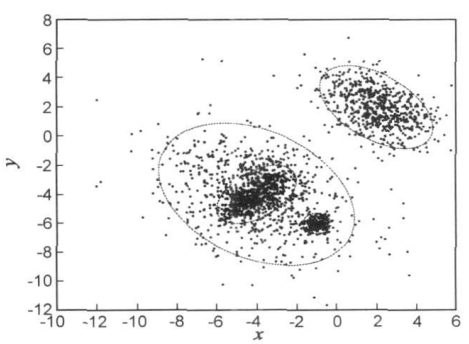

图1　4个高斯混合分布

初始参数取值为: 混合权重等权重 $\pi_1^{(0)}=\pi_2^{(0)}=\pi_3^{(0)}=\pi_4^{(0)}=0.25$, 均值 $\{\mu_j^{(0)}\}_{j=1}^4$ 在数据空间均匀分布, 协方差阵 $\{\Sigma_j^{(0)}\}_{j=1}^4$ 取单位阵 $\boldsymbol{I}$. 对于稳态的判断, 根据经验, 一般经过 50 ~ 60 步进入稳态, 我们采用 100 步蒙特卡洛, 第 100 步的稳态采样值作为 Gibbs 采样的参数估计值, 对不同的高斯分布元个数, 比较其稳态时混合模型的 AIC, BIC 优化信息准则. 如图 2, 可以看出当分布元为 4 时, 具有最小的 BIC 值 AIC 准则也对应具有最小的值, 不过, AIC 准则在分布元个数大于 4 时, 变化不明显, 主要是模型参数增加对 AIC 贡献不明显. 这时候, BIC 准则相对比较明显, 因此, 用 BIC 准则估计相对更准确些.

图 3 描述了 BIC 信息准则的优化过程. 我们取 100 步稳态迭代作为指标来判断, 高斯分布元个数从 2 变化到 7, 可以看出, 大部分采样在 50 步后进入稳态, 在 100 步稳态后, 分布元为 4 的高斯混合模型具有最小的 BIC 准则, 说明未知混合分布个数为 4. 分布元个数为 4 时估计的混合分布参数如下:

$$\hat{\pi}_1=0.3,\ \hat{\pi}_2=0.29,\ \hat{\pi}_3=0.31,\ \hat{\pi}_4=0.1;$$

$$\hat{\mu}_1=[-4.07\quad -3.99]^{\mathrm{T}},\ \hat{\mu}_2=[-3.94\quad -3.83]^{\mathrm{T}},$$

$$\hat{\mu}_3=[1.94\quad 2.07]^{\mathrm{T}},\ \hat{\mu}_4=[-1.03\quad -5.98]^{\mathrm{T}};$$

$$\hat{\Sigma}_1=\begin{bmatrix}0.85 & 0.32\\ 0.32 & 1.06\end{bmatrix},\ \hat{\Sigma}_2=\begin{bmatrix}7.02 & -1.76\\ -1.76 & 5.04\end{bmatrix},$$

$$\hat{\Sigma}_3=\begin{bmatrix}1.89 & -0.90\\ -0.90 & 1.99\end{bmatrix},\ \hat{\Sigma}_4=\begin{bmatrix}0.10 & 0.00\\ 0.00 & 0.11\end{bmatrix}.$$

可以看出, 估计的参数和实际的参数相对比较接近. 产生误差的原因主要有三点: (1) 有限样本数据; (2) 随机算法本身的随机误差; (3) 数值计算产生的计算误差.

该程序在一台 Intel core(TM) 2, 3.0G, 2G 内存, Win7 操作系统的台式电脑上运行, 迭代 100 步, 运行时间如表 2 所示, 可以看出随着混合模型个数增加, 运行时间也在增加, 并且呈现一种近似线性的增长过程, 运行时间都在 4s 以内, 相对比较快.

表 2　Gibbs 采样算法运行时间(100 步)　算例 1

分布元个数	2	3	4	5	6	7
运行时间(s)	2.12	2.25	2.6	2.76	3.12	3.26

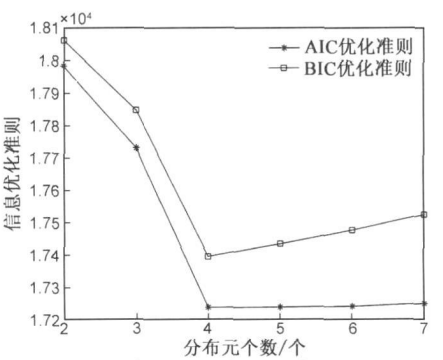

图2　稳态时分布元个数-优化准则关系(4个高斯混合模型)

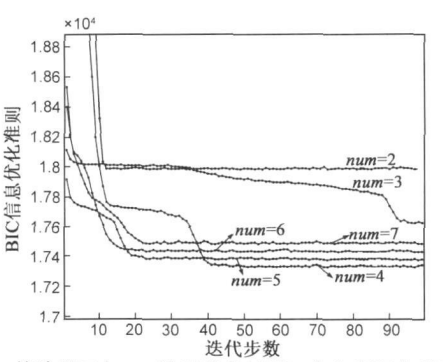

图3　算法学习与BIC准则优化过程(4个高斯混合模型)

4.2 算例 2 六个高斯混合算例

本算例中, 我们考虑 6 个混合高斯分布, 6 个分布两两相交, 有部分数据混叠在一起, 观测数据 3000 个, 真实混合分布模型如下式, 真实分布见图 4.

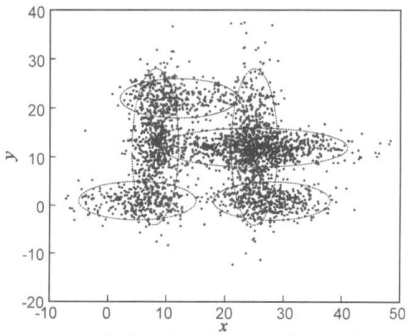

图4 6个高斯混合分布(6个高斯混合模型)

$$f(y_i|\theta)=\pi_1 N(y_i;\mu_1,\Sigma_1)+\cdots+\pi_6 N(y_i;\mu_6,\Sigma_6)$$

$$\pi=\{1/12,5/24,1/12,1/4,5/24,1/8\};$$

$$\mu_1=[5\quad 1]^T,\ \mu_2=[8\quad 12]^T,\ \mu_3=[12\quad 22]^T,$$

$$\mu_4=[25\quad 12]^T,\ \mu_5=[25\quad 12]^T,\ \mu_6=[28\quad 1]^T;$$

$$\Sigma_1=\begin{bmatrix}25&0\\0&4\end{bmatrix}^T,\ \Sigma_2=\begin{bmatrix}4&0.5\\0.5&4\end{bmatrix}^T,\ \Sigma_3=\begin{bmatrix}25&0\\0&4\end{bmatrix}^T,$$

$$\Sigma_4=\begin{bmatrix}64&0\\0&0\end{bmatrix},\ \Sigma_5=\begin{bmatrix}4&0.02\\0.02&64\end{bmatrix},\ \Sigma_6=\begin{bmatrix}25&0\\0&4\end{bmatrix};$$

混合模型的初值设定方法和上例相同. 假设真实的分布元个数位于区间[2, 8], 我们搜索 2~8 共 7 组混合分布, 并计算其 BIC 指标, 从图 5 可以看出, 当分布元个数为 6 时, 模型 BIC 准则具有最小的优化值 BIC, 小于 6 或者大于 6, BIC 信息准则都会增加, 这也可已从图 6 中看出来, 用 2~8 个高斯混合模型进行学习, 大概 80 步后, 所有混合分布模型都进入稳态, 此时, 分布元个数为 6 对应的 BIC 准则具有最小的值, 这也说明未知混合分布的模型个数为 6. 当分布元个数为 6 时, 实际估计的参数值如下:

$$\hat{\pi}=\{0.08,0.22,0.09,0.27,0.21,0.13\}$$

$$\hat{\mu}_1=[4.47\quad 0.87]^T,\ \hat{\mu}_2=[8.15\quad 11.65]^T,$$

$$\hat{\mu}_3=[11.97\quad 21.98]^T,\ \hat{\mu}_4=[25.05\quad 11.97]^T,$$

$$\hat{\mu}_5=[24.89\quad 11.80]^T,\ \hat{\mu}_6=[27.70\quad 1.02]^T$$

$$\hat{\Sigma}_1=\begin{bmatrix}23.13&-1.1\\-1.1&4.55\end{bmatrix}^T,\ \hat{\Sigma}_2=\begin{bmatrix}3.85&1.87\\1.87&69.96\end{bmatrix}^T,$$

$$\hat{\Sigma}_3=\begin{bmatrix}25.14&-0.96\\-0.96&4.5\end{bmatrix}^T,\ \hat{\Sigma}_4=\begin{bmatrix}59.68&-1.25\\-1.25&4.31\end{bmatrix}^T,$$

$$\hat{\Sigma}_5=\begin{bmatrix}3.38&0.47\\0.47&66.19\end{bmatrix}^T,\ \hat{\Sigma}_6=\begin{bmatrix}25.94&-0.32\\-0.32&3.59\end{bmatrix}$$

可以看出估计值和实际基本吻合. 表 3 给出了该算例运行时间和分布元之间的关系, 同算例 1 一样, 迭代运行 100 步, 从表 3 中可以看出运行时间也近似线性增加, 最大运行时间基本上不超过 4s.

表 3 Gibbs 采样算法运行时间(100 步) 算例 2

分布元个数	2	3	4	5	6	7	8
运行时间(s)	2.26	2.41	2.64	2.89	3.66	3.8	4.00

需要指出的是, 无论是 EM 算法还是 MCMC 算法, 都会产生标签转置问题, 即不同的初始点, 参数具有不同的收敛路径, 这对整个混合分布是没有影响, 但参数收敛后分布元的次序可能会产生变换, 这对本研究对象没有影响, 不属于本文研究的重点, 这里不予考虑, 相关研究可以参阅文献[19].

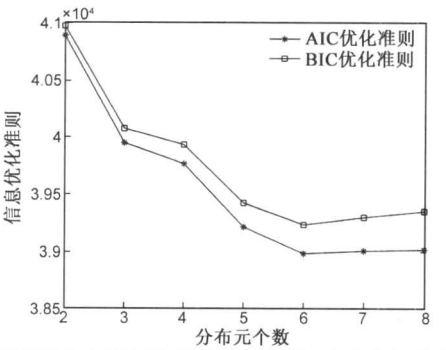

图5 稳态时分布元个数-优化准则关系(6个高斯混合模型)

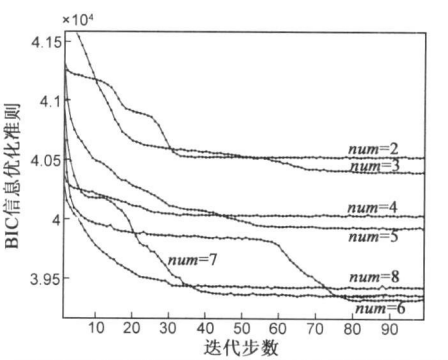

图6 算法学习与BIC准则优化过程(6个高斯混合模型)

5 总结

针对有限混合模型无监督学习过程中分布元个数未知问题, 本文给出了一种基于 BIC 信息准则和 Gibbs 采样的无监督学习算法. 算法利用 Gibbs 采样估计混合模型的参数和权重. 利用 BIC 准则确定分布元个数, 先验分布元的个数假设是在某个整数区间, 仿真实验表明该算法能够有效地估计高斯混合模型参数和高斯元个数, 算法可用于模式识别、分类, 图像识别, 跟踪估计等领域. 不过本文对分布元个数假设为在某个区间的整数, 这一点需要一定的先验信息, 后续的研究可以在这方面进行深入探讨.

参考文献

[1] McLachlan G, Peel D. Finite Mixture Models[M]. New York: John Wiley Sons, 2000.

[2] 刘立芳, 霍红卫, 王宝树. 生物序列模体的混合 Gibbs 抽样识别算法[J]. 电子学报, 2008, 36(4): 750—755.
Liu Lifang, Huo Hongwei, Wang Baoshu. Multiple motif discovery in biological sequences by mixture Gibbs sampling[J]. Acta Electronica Sinica, 2008, 36(4): 750—755. (in Chinese)

[3] W K Hastings. Monto Carlo sampling methods using Markov chains and their Applications[J]. Biometrika, 1970, 57(1): 97—109.

[4] A P Dempster NML, D B Rubin. Maximum likelihood from Incomplete Data via the EM algorithm[J]. Journal of the Royal statistical Society, Series B, 1977, 39(1): 1—28.

[5] Constantinos Constantinopoulos, Michalis K. Titsias, and Aristidis Likas, Bayesian Feature and Model Selection for Gaussian Mixture Models[J]. IEEE Transactions of Pattern Analysis and Machine Intelligence, 2006, 6(28): 1013—1018.

[6] Nizar Bouguila, Djemel Ziou. A Hybrid Sem Algorithm for High-Dimensional Unsupervised Learning Using a Finite Generalized Dirichlet Mixture[J]. IEEE Transactions on Image Processing, 2006, 15(9): 2657—2668.

[7] Mario A T Figueiredo, Anil K. Jain. Unsupervised Learning of Finite Mixture Models [J]. IEEE Transactions of Pattern Analysis and Machine Intelligence, 2002, 3(24): 381—396.

[8] Bouguila N, Ziou D. High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(10): 1716—1731.

[9] Pemkopf F, Bouchaffra D. Genetic-based EM algorithm for learning Gaussian mixture models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (8): 1344—1348.

[10] Green P J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination[J]. Biometrika, 1995, 82(4): 711—732.

[11] 刘伟峰, 韩崇昭, 石勇. 修正 Gibbs 采样的有限混合模型无监督学习算法[J]. 西安交通大学学报, 2009, 43(2): 15—19.
Liu Weifeng, Hang Chongzhao, Shi Yong. Unsupervised learning for finite mixture models via modified Gibbs sampling[J]. Journal of Xi' an Jiaotong University, 43(2), 2009. 15—19. (in Chinese)

[12] Schwarz, Gideon E. Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6(2): 461—464.

[13] Geman S, Geman D. Stochastic relaxation, Gibbs Distributions and the Bayesian restoration of image[J]. IEEE Transactions of Pattern Analysis and Machine Intelligence, 1984, 6(6): 721—741.

[14] Jean Diebolt, Christian P. Robert. Estimation of finite mixture distributions through Bayesian sampling[J]. J. R. Statist. Soc. B, 1994, 56(2): 363—375.

[15] 韩崇昭, 朱洪艳, 段战胜. 多源信息融合[M]. 北京: 清华大学出版社, 2006. 25—27.
Han Chongzhao, Zhu Hongyan, Duan Zhansheng. Multi-source Information Fusion[M]. Beijing: Tsinghua University Press, 2006. 25—27. (in Chinese)

[16] 熊刚, 赵惠昌, 王李军. 海杂波背景下雷达引信的相关检测方法研究[J]. 电子学报, 2004, 32(12): 1937—1940.
Xiong Gang, Zhao Huichang, Wang Lijun. The correlation detection method of radar fuze in sea clutter[J]. Acta Electronica Sinica, 2004, 32(12): 1937—1940. (in Chinese)

[17] Akaike H. Information theory and an extension of the maximum likelihood principle[A]. In 2nd International Symposium on Inf ormation Theory[C]. 1973: 267—281, Budapest, Hungary.

[18] Akaike H. A new Look at the statistical model identification [J]. IEEE Transactions on Automatic Control, 1974, 19(6): 716—723.

[19] Sylvia Frhwirth-Schnatter, Markov chain Monte Carlo estimation of classical and the dynamic switching and mixture models[J]. Journal of the American Statistical Association, 2001, 96(453): 194—209.

作者简介

刘伟峰　男, 1973 年生, 陕西咸阳人, 博士, 讲师, 主要研究方向为多目标跟踪、随机集理论和模式识别.
E-mail: liuwf@hdu.edu.cn

杨爱兰　女, 1987 年出生于山东省潍坊市. 现为杭州电子科技大学在读硕士研究生, 主要研究方向为模式识别与图像处理.
E-mail: yangailanhao@163.com