混合逆狄利克雷分布的变分学习及应用

赖裕平[1],周亚建[1],丁洪伟[2],郭玉翠[3],郭 春[1],杨义先[1]
(1、北京邮电大学计算机学院,北京 100876;2、云南大学信息学院,云南昆明 650091;3、北京邮电大学理学院,北京 100876)

摘 要: 混合逆狄利克雷分布是正的非高斯数据分析中一个重要的统计模型.但是利用常用的统计方法比如极大近似然估计、矩估计等往往很难得到模型参数估计的显性解析式.本文提出一个变分贝叶斯学习算法,它能够在估计参数的同时自动确定混合分量数.在合成数据集及实测数据集上的实验结果表明利用变分贝叶斯推理来估计混合逆狄利克雷分布是一种非常有效的方法.

关键词: 逆狄利克雷分布;贝叶斯估计;变分推理;拓展分解变分近似;模型选择
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2014)07-1435-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.07.030

Variational Learning for Finite Inverted Dirichlet Mixture Models and Applications

LAI Yu-ping[1], ZHOU Ya-jian[1], DING Hong-wei[2], GUO Yu-cui[3], GUO Chun[1], YANG Yi-xian[1]
(1. *Information Security Center, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *School of Information Science and Engineering, Yunnan University, Kunming, Yunnan* 650091, *China*;
3. *School of Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: Finite inverted Dirichlet mixture models play an important part in positive non-Gaussian data analysis. However, it is always different to obtain the analytical solutions to model parameters by using conventional approaches such as maximization likelihood estimation and moment estimation. In this paper, we have proposed a variational inference framework. Within this framework, parameter estimation and automatic model selection can be carried out simulta-neously. Experimental results on synthetic and real-world data sets demonstrate the effectiveness and the merits of the proposed approach.

Key words: inverted Dirichlet distribution; Bayesian estimation; variational inference; extended factorized variational approximation; model selection

1 引言

许多科学和工程领域均会涉及到大量数据的统计分析与数据建模.有限混合模型(Finite Mixture Models, FMM)就是一种用于统计数据分析非常有效的数学方法之一.FMM 属于一种半参数的模型,用于描述具有不同成分的混合数据,具有很强的灵活性.不管数据分布的结构如何复杂,FMM 总能通过增加混合分量数的方式来描述数据分布的局部特性.这使得它成为了最有效的密度估计工具以及最常用的无监督分类工具之一,在模式识别、数据挖掘和机器学习等领域中占有极其重要的地位.

对 FMM 的研究最早可以追溯到 100 年以前,1894 年皮尔逊用两个混合分量的高斯混合模型拟合一组观测数据,利用矩估法估计该模型的参数.这应该是关于 FMM 问题最早的研究.自此以后,FMM 研究逐渐受到人们的重视.目前,FMM 已广泛应用于语音识别[1]、语音增强[2]和生物序列膜体识别[3]等应用领域.其中,高斯混合模型(Gaussian Mixture Models, GMM)凭借其灵活性好、计算模型参数方便、能以很高的精度逼近任意概率密度等优点,已成为目前应用最为广泛的 FMM.高斯分布是无界的,而有些实际数据却是有界或半有界的,如直方图、归一化特征矢量(介于 0 到 1 之间且和等于 1,亦称为比例矢量)和线谱对频率参数(介于 0 和 π 之间).大量研究表明 GMM 对非高斯分布的数据的描述并不理想.近年来,一些研究者提出了若干其它分布模型,如文献[4]提出了基于 inverted Dirichlet 分布的混合模型并将其成功应用于正的非高斯数据聚类;文献[5]使用

收稿日期:2013-10-16;修回日期:2014-02-18;责任编辑:郭游
基金项目:国家自然科学基金(No.61121061, No.60972077, No.61072079, No.61303232, No.61363085);教育部博士点基金(No.20120005110017);国家 863 高技术研究发展计划(No.2009AA01AZ430);

了基于 Dirichlet 分布的混合模型并将其成功应用于图像分割.目前,非高斯统计模型研究成为了一个新研究领域[4~8].

FMM 研究的问题包括两个方面:混合模型参数估计和混合分量数估计(模型选择).用于拟合 FMM 的经典算法有极大近似然估计的 EM 算法和 Bayes 估计法.前者存在迭代初始值敏感、过拟合等问题;后者通过对模型参数赋予先验分布能够避免过拟合.求解贝叶斯层次模型真正的后验分布涉及到高维积分困难的问题,一般需要用近似推理算法.常见的近似推理算法有马尔可夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)、拉普拉斯近似.MCMC 对大数据集计算量太大,收敛速度慢;拉普拉斯推理中假设模型似然函数通是单峰的,而大多数 FMM 的似然函数却是多峰的,这导致推理结果一般不准确.故对于 FMM 的一些快速逼近参数推断方法相继提出.变分推理方法(Variational Bayesian Approach)用易处理的一簇分布来逼近隐变量的后验分布,加快了算法参数推断速度.

FMM 研究的一个重要的工作是进行模型选择.在利用 FMM 拟合数据时,过多的混合分量容易引起模型过学习,导致没有理想的泛化性能;而过少的混合分量会使得模型缺乏柔韧性,不能较准确地逼近真实后验分布.EM 算法需要结合某种模型选择准,例如 Bayesian 信息准则(Bayesian Information Criterion, BIC)[9], Akaike 信息准则(Akaike Information Criterion, AIC), MML 准则(Minimum Information Criterion, MML),通过对候选的模型逐个进行参数估计和准则值计算,最后选择合适的模型.这种方法通常比较耗时.而变分贝叶斯推理将模型参数估计与模型选择纳入一个统一的算法框架中,能自动确定分量数和同时得到模型参数估计值.

本文的主要工作是研究有限逆狄利克雷混合模型(Finite Inverted Dirirhlet Mixture Models, IDM)的变分学习算法.文献[5]提出 IDM 的 EM 学习算法(E-IDM),并通过优化 MML 信息准则获得最佳分量数.其中,M 步采用了牛顿莱布尼茨迭代算法,因此它很耗时.故本文提出一种拓展分解变分近似算法,在合成数据和真实数据上的实验结果验证了本算法的有效性:既能得到较为准确的混合分量参数,也能得到较为准确的混合分量数.

2 逆狄利克雷混合模型

假设观测数据 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\mid \boldsymbol{x}_n\in \boldsymbol{R}_+^d\}$ 由 IDM 产生,样本矢量之间相互独立.IDM 定义为[4]

$$p(\boldsymbol{x}_n\mid\boldsymbol{\alpha},\boldsymbol{\pi})=\sum_{i=1}^{M}\pi_i\mathrm{IDir}(\boldsymbol{x}_n\mid\boldsymbol{\alpha}_i) \tag{1}$$

其中,$\boldsymbol{\pi}=(\pi_1,\cdots,\pi_M)$是混合权重矢量,它满足 $\pi_i>0, i=1,\cdots,M$ 且 $\sum_{i=1}^{M}\pi_i=1$;$\boldsymbol{\alpha}_i=(\alpha_{i1},\cdots,\alpha_{iD+1})$是第 i 个混合分量的参数;$\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_M)$代表所有的模型参数,$\mathrm{IDir}(\boldsymbol{x}_n\mid\boldsymbol{\alpha}_i)$是代表 D 维逆狄利克雷分布的密度函数,其形式为

$$\mathrm{IDir}(\boldsymbol{x}_n\mid\boldsymbol{\alpha}_i)=\frac{\Gamma(\mid\boldsymbol{\alpha}_i\mid)}{\prod_{d=1}^{D+1}\Gamma(\alpha_{id})}\prod_{d=1}^{D}x_{nd}^{\alpha_{id}-1}\Big(1+\sum_{d=1}^{D}x_{nd}\Big)^{-\mid\boldsymbol{\alpha}_i\mid} \tag{2}$$

其中,$\mid\boldsymbol{\alpha}_i\mid=\sum_{d=1}^{D+1}\alpha_{id}$,$\Gamma(\cdot)$是伽马函数.对观测数据 $\boldsymbol{X}$ 引入指示向量 $\boldsymbol{S}=(\boldsymbol{s}_1,\cdots,\boldsymbol{s}_N)$, $\boldsymbol{s}_n=(s_{n1},\cdots,s_{nM})$,它代表每个样本属于的哪个混合分量.其中,$s_{ni}$取值为 0 或 1 且 $\sum_{i=1}^{M}s_{ni}=1$.若 $\boldsymbol{x}_n$来自于第 i 个分量,则 $s_{ni}=1$,而 $s_{nj}=0, j\neq i$.在给定混合权重矢量 $\boldsymbol{\pi}$ 的条件下,隐变量 $\boldsymbol{S}$ 的条件分布为

$$p(\boldsymbol{S}\mid\boldsymbol{\pi})=\prod_{n=1}^{N}\prod_{i=1}^{M}\pi_i^{s_{ni}} \tag{3}$$

在给定隐变量 $\boldsymbol{S}$ 的条件下,观测数据 $\boldsymbol{X}$ 的条件分布为

$$p(\boldsymbol{X}\mid\boldsymbol{\alpha},\boldsymbol{S})=\prod_{n=1}^{N}\prod_{i=1}^{M}\mathrm{IDir}(\boldsymbol{x}_n\mid\boldsymbol{\alpha}_i)^{s_{ni}} \tag{4}$$

为了便于变分贝叶斯推理,应先给模型参数指定适当的先验分布.为了便于推理,本文选择 gamma 分布作为先验分布.假设模型参数是相互独立的,则先验分布为

$$p(\boldsymbol{\alpha}\mid u,\upsilon)=\prod_{i=1}^{M}\prod_{d=1}^{D+1}\frac{\upsilon_{id}^{u_{id}}}{\Gamma(u_{id})}\alpha_{id}^{u_{id}-1}\mathrm{e}^{-\upsilon_{id}\alpha_{id}} \tag{5}$$

其中,$\{u_{id},\upsilon_{id}\}$是超参数且满足 $u_{id}>0,\upsilon_{id}>0$.

3 变分学习

3.1 标准变分后验

为了估计 IDM 的参数及最佳的混合分量数,我们采用文献[10]中提出的变分推理框架.用 $\boldsymbol{\Theta}=(\boldsymbol{S},\boldsymbol{\alpha})$表示其所有的隐变量和未知参数.联合式(3),(4)和(5)及利用 Bayes 定理,得到观测数据 $\boldsymbol{X}$ 和随机变量 $\boldsymbol{\Theta}$ 的联合分布

$$\begin{aligned}\ln p(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})=&\sum_{n=1}^{N}\sum_{i=1}^{M}s_{ni}\Big\{\ln\pi_i+\sum_{d=1}^{D}(\alpha_{id}-1)\ln x_{nd}\\&+\boldsymbol{H}_i-\ln\Big(1+\sum_{d=1}^{D}x_{nd}\Big)\sum_{d=1}^{D+1}\alpha_{idi}\Big\}+\sum_{i=1}^{M}\sum_{d=1}^{D+1}\big[u_{id}\ln\upsilon_{id}\\&-\ln\Gamma(u_{id})+(u_{id}-1)\ln\alpha_{id}-\upsilon_{id}\alpha_{id}\big]\end{aligned} \tag{6}$$

其中,$\boldsymbol{H}_i$ 是多元 LIB[7]函数(MLIB),其表达式为

$$\boldsymbol{H}_i=\ln\frac{\Gamma\big(\sum_{d=1}^{D+1}\alpha_{id}\big)}{\prod_{d=1}^{D+1}\Gamma(\alpha_{id})} \tag{7}$$

假设变分分布族分解为

$$q(\boldsymbol{\Theta})=q(\boldsymbol{S})q(\boldsymbol{\alpha})=\prod_{i=1}^{M}\prod_{d=1}^{D+1}q(\alpha_{id})\prod_{i=1}^{N}\prod_{i=1}^{M}q(s_{ni}) \tag{8}$$

利用标准分解变分近似[11],得到各因子的最佳解为:

$$\ln q^*(s_{ni}) = \ln\pi_i + \sum_{d=1}^{D}(\bar{\alpha}_{id}-1)\ln x_{nd} - \ln\Big(1+\sum_{d=1}^{D}x_{nd}\Big)\times\sum_{d=1}^{D+1}\bar{\alpha}_{id} + \langle \boldsymbol{H}_i\rangle_{\alpha_{i1},\cdots,\alpha_{iD+1}} + \text{const} \tag{9}$$

$$\ln q^*(\alpha_{id}) = \sum_{n=1}^{N}\langle s_{ni}\rangle\Big[\langle \boldsymbol{H}_i\rangle_{\boldsymbol{\Theta}\neq\alpha_{id}} - \alpha_{id}\ln\Big(1+\sum_{d=1}^{D}x_{nd}\Big) - \alpha_{id}\ln x_{nd}\Big] + (u_{id}-1)\ln\alpha_{id} - \upsilon_{id}\alpha_{id} + \text{const}, 1\leqslant d\leqslant D \tag{10}$$

$$\ln q^*(\alpha_{id}) = \sum_{n=1}^{N}\langle s_{ni}\rangle\Big[\langle \boldsymbol{H}_i\rangle_{\boldsymbol{\Theta}\neq\alpha_{id}} - \alpha_{id}\ln\Big(1+\sum_{k=1}^{D}x_{nk}\Big)\Big] + (u_{id}-1)\ln\alpha_{id} - \upsilon_{id}\alpha_{id} + \text{const}, d = D+1 \tag{11}$$

其中,$\langle\cdot\rangle$表示相关变量关于变分分布的数学期望,$\bar{\alpha}_{id}=\langle\alpha_{id}\rangle$.因为上述最佳解与相应的先验分布形式都不同,所以我们无法得到闭合形式的变分后验.

3.2　拓展分解变分近似

根据标准分解变分近似[11],最大化的变分目标函数(变分下限)为

$$\boldsymbol{L}(q) = \int q(\boldsymbol{\Theta})\ln\Big\{\frac{p(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})}{q(\boldsymbol{\Theta})}\Big\}\mathrm{d}\boldsymbol{\Theta} = \langle\ln p(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})\rangle_{\boldsymbol{\Theta}} - \langle\ln q(\boldsymbol{\Theta})\rangle_{\boldsymbol{\Theta}} \tag{12}$$

在 3.1 节我们提到,利用这种方法有时无法直接得到闭合形式的变分后验.为了解决这个问题,我们提出拓展变分分解近似算法.

定理 1　拓展分解变分近似

假如我们能够找到一个似然函数 $g(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})$,满足

$$\int q(\boldsymbol{\Theta})\ln p(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})\mathrm{d}\boldsymbol{\Theta}\geqslant\int q(\boldsymbol{\Theta})\ln g(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})\mathrm{d}\boldsymbol{\Theta}, \tag{13}$$

则我们想要最大化的变分目标函数为

$$\boldsymbol{L}(q)\geqslant\int q(\boldsymbol{\Theta})\ln g(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})\mathrm{d}\boldsymbol{\Theta} - \int q(\boldsymbol{\Theta})\ln q(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} - KL(q(\boldsymbol{\Theta})\,\|\,\tilde{g}(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})) + \ln c = \tilde{\boldsymbol{L}}(q) \tag{14}$$

其中,c 是似然函数 $g(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})$的归一化常量,满足 $g(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})=c\tilde{g}(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})$.取 $q(\boldsymbol{\Theta})=\tilde{g}(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})$,可得到变分下限 $\boldsymbol{L}(q)$的最大值.尽管我们无法直接最大化 $\boldsymbol{L}(q)$,但是可以通过渐进最大化 $\tilde{\boldsymbol{L}}(q)$求解 $\boldsymbol{L}(q)$的最大值.我们把这种方法称为拓展分解变分近似,变分目标函数变为

$$\tilde{\boldsymbol{L}}(q)=\langle\ln g(\boldsymbol{X},\boldsymbol{\Theta}\mid\boldsymbol{\pi})\rangle_{\boldsymbol{\Theta}} - \langle\ln q(\boldsymbol{\Theta})\rangle_{\boldsymbol{\Theta}} \tag{15}$$

对于指定的分解因子 $q_j(\boldsymbol{\Theta}_j)$,其最佳解为

$$\ln q_j^*(\boldsymbol{\Theta}_j)=\langle\ln g(\boldsymbol{X},\boldsymbol{\Theta})\mid\boldsymbol{\pi}\rangle_{l\neq j}+\text{const} \tag{16}$$

定理 2　MLIB 函数的下界

$$\boldsymbol{H}_i\geqslant\ln\frac{\Gamma\big(\sum_{d=1}^{D+1}\bar{\alpha}_{id}\big)}{\prod_{d=1}^{D+1}\Gamma(\bar{\alpha}_{id})} + \sum_{d=1}^{D+1}\Big[\Psi\Big(\sum_{l=1}^{D+1}\bar{\alpha}_{il}\Big)-\Psi(\bar{\alpha}_{id})\Big]\times(\ln\alpha_{id}-\ln\bar{\alpha}_{id})\bar{\alpha}_{id} + \frac{1}{2}\sum_{l=1}^{D+1}\sum_{k\neq l}^{D+1}\Psi'\Big(\sum_{l=1}^{D+1}\bar{\alpha}_{il}\Big)\times(\ln\alpha_{il}-\ln\bar{\alpha}_{il})(\ln\alpha_{ik}-\ln\bar{\alpha}_{ik})\bar{\alpha}_{il}\bar{\alpha}_{ik} = \tilde{\boldsymbol{H}}_i \tag{17}$$

这个定理的证明可参照 LIB 函数[6]的下界证明.将 $\tilde{\boldsymbol{H}}_i$ 代入式(6)并根据式(16)可以得到模型参数的后验分布

$$q^*(\boldsymbol{S}) = \prod_{n=1}^{N}\prod_{i=1}^{M}r_{ni}^{s_{ni}} \tag{18}$$

$$q^*(\boldsymbol{\alpha}) = \prod_{i=1}^{M}\prod_{d=1}^{D}\boldsymbol{G}(\alpha_{id}\mid u_{id}^*,\upsilon_{id}^*) \tag{19}$$

式中:

$$r_{ni} = \frac{\rho_{ni}}{\sum_{k=1}^{M}\rho_{nk}} \tag{20}$$

$$\rho_{ni} = \ln\pi_i + \langle\tilde{\boldsymbol{H}}_i\rangle + \sum_{d=1}^{D}(\bar{\alpha}_{id}-1)\ln x_{nd} - \ln\Big(1+\sum_{d=1}^{D}x_{nd}\Big)\sum_{d=1}^{D+1}\bar{\alpha}_{id} \tag{21}$$

$$u_{id}^* = u_{id} + \sum_{n=1}^{N}r_{ni}\Big[\Psi\Big(\sum_{l=1}^{D+1}\bar{\alpha}_{id}\Big)-\Psi(\bar{\alpha}_{id}) + \sum_{k\neq d}^{D+1}\Psi'\Big(\sum_{l=1}^{D+1}\bar{\alpha}_{il}\Big)(\langle\ln\alpha_{ik}\rangle-\ln\bar{\alpha}_{ik})\Big]\bar{\alpha}_{id} \tag{22}$$

$$\upsilon_{id}^* = \upsilon_{id} + \sum_{n=1}^{N}r_{ni}\Big[\ln\Big(1+\sum_{d=1}^{D}x_{nd}\Big)-\ln x_{nd}\Big] \tag{23}$$

$$\upsilon_{iD+1}^* = \upsilon_{iD+1} + \sum_{n=1}^{N}r_{ni}\Big(1+\sum_{d=1}^{D}x_{nd}\Big) \tag{24}$$

上述参数更新方程中期望值如下:

$$\langle s_{ni}\rangle = r_{ni},\bar{\alpha}_{id}=\frac{u_{id}^*}{\upsilon_{id}^*},\langle\ln\alpha_{id}\rangle=\Psi(u_{id}^*)-\ln\upsilon_{id}^* \tag{25}$$

其中,$\Psi(\cdot)$表示 digamma 函数,$\Psi'(\cdot)$表示双 digamma 函数.因为 $q^*(\boldsymbol{\alpha})$与 $q^*(\boldsymbol{S})$的解是相互耦合的,所以利用迭代算法可以求它们的解.在每次迭代中,首先是根据分解因子 q 最大化变分下限 $\tilde{\boldsymbol{L}}(q)$,然后根据混合加权 $\boldsymbol{\pi}$ 最大化目标函数 $\tilde{\boldsymbol{L}}(q)$.关于 $\boldsymbol{\pi}$ 对 $\tilde{\boldsymbol{L}}(q)$取偏导数并令导数为零,得到混合加权的估计值

$$\pi_i = \frac{1}{N}\sum_{n=1}^{N}r_{ni} \tag{26}$$

因为后验分布与变分下界的解都与加权系数有关,所以可以通过一种类似于 EM 算法的迭代算法来求其最优解.我们把这种算法称为变分 EM 算法(VE-EM),它是收敛的[12],收敛可以通过下界的值进行判断.迭代过程中,一些冗余的分量权值趋向 0,忽略将这些冗余分量,实现模型选择.算法流程总结如下:

初始化:

利用 K 均值算法初始化 r_{ni},设混合分量 $M=15$ 且混合权重等系数,超参数$\{u_{id},\upsilon_{id}\}=\{1,0.01\}$.

VB-E 步骤:

根据式(18)和(19)分别更新变分后验分布 $q^*(\boldsymbol{S})$ 和 $q^*(\boldsymbol{\alpha})$.

VB-M 步骤:

根据当前混合加权的估计值式(26)最大化变分下限 $\tilde{\boldsymbol{L}}(q)$.

迭代执行 E 步和 M 步直至算法收敛.忽略加权系数小于 10^{-5}的分量.根据式(18)和(19)重新估计模型参数.

4 实验结果及分析

4.1 合成数据

以下给出本文所提算法(varIDM)的性能.实验中用程序实现并产生了四个服从 IDM 分布的 2 维合成数据.算法收敛后每个分量对应的加权系数如图 1 所示.由图可知,算法最终收敛后通过模型选择.表 1 给出了对应于每个合成数据集的实际参数值及变分学习后的平均参数估计值.其中,N_i 代表第 i 簇中的样本数量,θ_{id}代表真实参数,$\hat{\theta}_{id},\hat{\pi}_i$ 代表由 varIDM 算法得到参数估计值.$\bar{\theta}_{id},\bar{\pi}_i$ 代表由 E-IDM 算法得到参数估计值.从表 1 可以看出,我们的算法能够比较准确的估计模型的参数.另外,与文献[5]中提出的 EM 算法相比,该算法具有着更高的精确度.此外,我们对比了 varIDM 与 E-IDM 两种算法的迭代次数和计算时间,实验结果如表 2 所示.很明显,varIDM 算法比 E-IDM 算法具有更快的收敛速度.

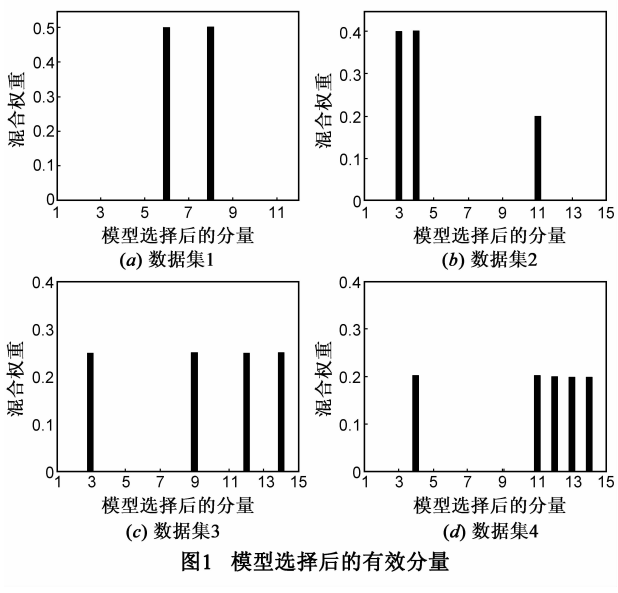

图1 模型选择后的有效分量

4.2 真实数据

首先的实验是一个关于目标检测的问题.目标检测是算机视觉研究的重要课题,在图像检索、智能视频监控等领域具有广泛的应用前景.尽管目标检测问题已经得到了大量的研究,但它仍旧具挑战性.特征提取是目标检测中的关键步骤之一.本部分实验数据来源 UIUC car、Caltech Motorbikes 和 Human faces 图片数据集.我们采用文献[7]提出的特征提取方法,用隐狄利克雷

表 1 实际参数与估计参数

	N_i	i	θ_{i1}	θ_{i2}	θ_{i3}	$\bar{\theta}_{i1}$	$\bar{\theta}_{i2}$	$\bar{\theta}_{i3}$	$\hat{\pi}_i$	$\hat{\theta}_{i1}$	$\hat{\theta}_{i2}$	$\hat{\theta}_{i3}$	$\tilde{\pi}_i$
数据集 1	200	1	12	31	44	11.63	29.89	42.15	0.5000	13.26	33.61	41.32	0.478
	200	2	24	16	90	24.17	16.12	90.19	0.5000	24.83	14.87	83.27	0.522
数据集 2	200	1	12	31	44	12.84	30.86	43.85	0.3998	11.02	29.34	46.32	0.431
	200	2	24	16	90	25.28	15.77	86.63	0.4001	22.27	16.24	94.22	0.356
	100	3	54	28	36	55.12	27.01	38.12	0.2001	51.28	29.33	37.21	0.213
数据集 3	200	1	12	31	44	11.45	29.67	41.78	0.2505	10.85	33.25	45.77	0.235
	200	2	24	16	90	22.63	15.19	85.31	0.2496	26.42	17.92	96.12	0.223
	200	3	54	28	36	51.65	26.23	33.68	0.2502	57.64	31.21	39.12	0.277
	200	4	30	52	18	29.56	53.29	17.92	0.2497	28.54	52.21	19.64	0.265
数据集 4	200	1	12	31	44	11.62	32.06	45.21	0.2000	14.01	31.78	47.28	0.176
	200	2	24	16	90	25.68	15.39	92.32	0.1998	22.06	14.68	85.42	0.221
	200	3	54	28	36	51.97	26.84	37.20	0.2002	56.72	31.27	33.19	0.216
	200	4	30	52	18	29.49	51.61	17.74	0.2000	32.06	55.84	20.32	0.183
	200	5	5	116	62	5.21	111.08	59.79	0.2000	4.32	108.23	64.32	0.204

表 2　varIDM 算法 E-IDM 算法的迭代次数和计算时间(s)

	varIDM		E-IDM	
数据集	计算时间	迭代次数	计算时间	迭代次数
数据集 1	2.19	513	4.32	612
数据集 2	1.63	439	3.16	539
数据集 3	1.78	406	2.72	468
数据集 4	1.32	298	2.45	401

分配(Latent Dirichlet Allocation, LDA)[13]降维.最佳主题数、词汇容量分别为 30 和 800.特征提取的结果为 30 维的比例矢量.假设特征集为 $\boldsymbol{Z}=(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N)$.利用下面的映射方案可以把比例矢量映射为正矢量集,标记为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$.

$$\boldsymbol{x}_n=\frac{\boldsymbol{z}_n}{1-\sum_{d=1}^{D-1}z_{nd}} \tag{28}$$

我们把基于特征集 $\boldsymbol{X}$ 分类的 IDM 分类器记为$\widehat{\text{var IDM}}$.

将每个特征集(正例样本和反例样本)随机分成两半,一半为训练集,另一半为测试集.平均检测率如表 3 所示.接下来的实验关于场景进行分类问题.实验数据来源于于文献[14]中提到场景图片数据集.其中四类为人工图片数据集,其余四类为自然图片数据集.另外一个图像数据集是 UIUC 运动场景图像.采用上述特征提取方法,其中最佳主题数和词汇容量分别为 35 和 800.平均分类准确率如表 4 所示.其中,E-GMM 表示基于 EM 与 MML 结合的 GMM 学习算法,B-GMM 表示文献[9]中提出的 GMM 学习算法,varGMM 表示文献[10]中提出的 GMM 学习算法,varDM 表示文献[7]中提出的狄利克雷混合模型的学习算法.

从表 3 和表 4 可以看出,本文提出的变分学习算法优于文献[4]中提出的 EM 算法;IDM 与狄利克雷混合模型的目标检测及图像分类的性能很接近接近;高斯混合模型不适合拟合比例数据.

表 3　各算法的平均检测率比较

图像集	E-GMM	B-GMM	varGMM	varDM	E-IDM	varIDM	$\widehat{\text{var IDM}}$
Car	79.11%	80.01%	80.31%	85.12%	81.14%	82.67%	84.91%
Face	80.11%	81..93%	82.41%	87.69%	84.32%	86.71%	88.02%
Motorbike	82.53%	84.71%	85.21%	89.13%	86.72%	88.18%	89.01%

表 4　各算法的平均分类准确率比较

图像集	E-GMM	B-GMM	varGMM	varDM	E-IDM	varIDM	$\widehat{\text{var IDM}}$
8 类图像	63.92%	70.61%	71.08%	74.11%	72.81%	73.75%	74.26%
自然图像	71.03%	71.56%	72.02%	75.72%	74.36%	75.24%	75.91%
人工图像	72.41%	72.71%	73.16%	77.41%	75.97%	76.79%	77.63%
运动场景	68.21%	68.63%	68.96%	73.46%	71.54%	72.68%	73.82%

5　总结

本文提出了一种变分推理的参数估计方法,并给出了详细的推导.仿真实验表明该算法可以有效、精确地估计 IDM 的参数.与此同时,通过 IDM 的参数估计方法还可以看出,应用拓展分解变分近似方法,可以对其它混合模型进行参数估计,比如狄利克雷混合、贝塔刘维尔混合及广义狄利克雷混合,该算法有较好的实用价值.今后研究工作将围绕分量数无限的混合逆狄利克雷模型及其特征选择问题展开.

参考文献

[1] 左国玉,刘文举,阮晓钢.声音转换技术的研究与进展[J].电子学报,2002,32(7):1165 – 1172.
Zuo G Y, Liu W J, Ruan X G. A voice conversion te-chnology and development [J]. Acta Electronica Sinia, 2002, 30(10): 1438 – 1440. (in Chinese)

[2] 梁岩,鲍长春,夏丙寅,等.基于高斯混合模型的压缩域语音增强方法[J].电子学报,2012,40(10):2031 – 2038.
Liang Y, Bao C C, Xia BY, et al. Compressed domain speech enhancement based on Gaussian mixture model [J]. Acta Electronica Sinica, 2012, 40(10): 2031 – 2038. (in Chinese)

[3] 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 L F, Huo H W, Wang B S. Multiple motif discovery in biological sequences by mixture Gibbs sampling [J]. Acta Electronica Sinica, 2008, 36(4): 750 – 755. (in Chinese)

[4] Bdiri T, Bouguila N. Positive vectors clustering using interted Dirichlet finite mixture Models [J]. Expert Systems with Applications, 2012, 39(2): 1869 – 1882.

[5] Bouguila N, ElGuebaly W. Discrete data clustering using finite

mixture models[J]. Pattern Recognition, 2009, 42(1): 33 – 42.
[6] Ma Z Y, Leijon A. Bayesian estimation of beta mixture models with variational inference [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(33): 2160 – 2173.
[7] Fan W T, Bouguila, Ziou N. Variaional learning for finite Dirichlet mixture models and applications [J]. IEEE Transaction on Neural Networks and Learning Systems, 2012, 23(5): 762 – 774.
[8] Bouguila N. Hybrid Generative/discriminative approaches for proportional data modeling and classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(12): 2184 – 2202.
[9] 刘伟峰,杨爱兰.基于 BIC 准则和 Gibbs 采样的有限混合模型无监督学习算法[J].电子学报,2011,39(3A):134 – 139.
Liu W F, Yang A L. Unsupervised learning for finite mixture models based on BIC criterion and gibbs sampling[J]. Acta Electronica Sinica, 2011, 39(3A): 134 – 1139. (in Chinese)
[10] Corduneanu A, Bishop C. Variational Bayesian model selection for mixture distributions[A]. Proceeding of the 8th International Conference on Artificial Intelligece and Statistics[C]. USA: IEEE, 2001. 27 – 34.
[11] Bishop C M. Pattern Recognition and Machine Learning [M]. New York: Springer-Verlag, 2006.
[12] Wang B, Titterington D M. Convergence and asymptotic normality of variational Bayesian approxima-tions for exponential family models with missing values[A]. Proc. UAI[C] USA: IEEE, 2004. 577 – 584.
[13] David MBlei, Andrew Y Ng, Michael I jordan. latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(33): 993 – 1022.
[14] JosefSivic, Andrew Zisserman. Video google: A text retrieval approach to object matching in videos [A]. Proceedings of Ninth IEEE International Conference on Computer Vision [C]. USA: IEEE, 2003, 1470 – 1477.

作者简介

赖裕平 男,1983 年 10 月生于江西瑞金,北京邮电大学博士研究生,主要研究方向为模式识别、机器学习、非高斯统计模型、贝叶斯网络、图像分类.
E-mail: Laiyp2009@126.com

丁洪伟 男,1964 年 5 月生于云南省景洪市,云南大学信息学院副教授,博士研究生,主要研究方向为轮询系统、随机多址通信系统、优化算法.

周亚建 男,1971 年生于陕西镇安.北京邮电大学副教授、硕士生导师,主要研究方向为移动通信、无线传感网络、信息安全等.

郭玉翠 女,1962 年 5 月生,北京邮电大学理学院教授.研究方向:微分方程边值问题的分析与求解.

郭 春 男,1986 年出生于贵州贵阳,北京邮电大学博士生.主要研究方向模式识别、机器学习、信息安全等

杨义先 男,1961 年 3 月生,北京邮电大学计算机学院教授,博士生导师,长江学者,主要研究方向为现代密码基础理论,信息隐匿技术与理论.