第48卷第6期 中国科学技术大学学报 Vol.48,No.6
2018年6月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jun. 2018

文章编号:0253-2778(2018)06-0477-09

基于KS检验的高斯混合模型分裂与合并算法

蒋硕然,陈亚瑞*,秦智飞,杨巨成

(天津科技大学计算机科学与信息工程学院,天津 300457)

摘要:高斯混合模型是有限个独立高斯模型的线性组合,估计其子模型个数是一个重要的研究问题.一类典型的算法是以最短描述长度为目标函数,在迭代过程中通过对子模型进行分裂与合并操作确定子模型个数.这类方法一般采用熵比、KL散度和模型相似度作为分裂和合并的判别准则.但是熵比或KL散度准则对稀疏子模型和凹形子模型过于敏感导致过度分裂,模型相似度准则不能反映合并后模型的高斯拟合优度导致过度合并.在算法的迭代过程中,这些过度分裂与合并操作产生振荡现象.针对估计子模型个数时出现的过度分裂与合并问题,基于KS检验的高斯混合模型的分裂与合并算法选择熵比与KS检验作为分裂的判别准则,模型相似度和KS检验作为合并的判别准则.最后在六个数据集上进行了实验证明算法的有效性.

关键词:高斯混合模型;最短描述长度;熵比;KS检验

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.06.006

引用格式:蒋硕然,陈亚瑞,秦智飞,等. 基于KS检验的高斯混合模型分裂与合并算法[J]. 中国科学技术大学学报,2018,48(6):477-485.

JIANG Shuoran, CHEN Yarui, QIN Zhifei, et al. Split and merge algorithm for Gaussian mixture model based on KS test[J]. Journal of University of Science and Technology of China, 2018,48(6):477-485.

Split and merge algorithm for Gaussian mixture model based on KS test

JIANG Shuoran, CHEN Yarui, QIN Zhifei, YANG Jucheng

(*Tianjin University of Science & Technology, College of Computer Science and Information Engineering, Tianjin* 300457)

Abstract: Gaussian mixture model is a linear combination of finite numbers of independent Gaussian models. Estimating the number of components is an important research area. One class of algorithms based on the minimum description length determine the number of components by splitting and merging components during the iterations. Traditional algorithms use entropy ratio, KL divergence, model similarity as split and merge criteria. However, entropy ratio and KL divergence might result in excessive split because of their excessive sensitivity to sparse or concave models, and model similarity might result in excessive merge because of its inability to assess the merged models' goodness of fitting Gaussian. In the iterations of algorithm, these excessive splitting and merging operations may cause oscillations. For these problems, a split and merge algorithm for Gaussian mixture model based on KS test is proposed, with entropy ratio and KS test used as split criteria and model similarity and KS test as merge criteria. This

收稿日期:2017-09-20;**修回日期**:2018-04-10

基金项目:国家自然科学基金(61402332, 61502338, 61402331);天津市科学技术委员会项目(17JCQNJC00400, 15ZCZDGX00200, 15JCQNJC00700)资助.

作者简介:蒋硕然,男,1992年生,硕士。研究方向:机器学习。E-mail: srjiang@mail.tust.edu.cn

通讯作者:陈亚瑞,博士/副教授. E-mail:yrchen@tust.edu.cn

algorithm is capable of preventing excessive split and merge, as validated by experiments conducted on seven datasets.

Key words: Gaussian mixture model; minimum description length; entropy ratio; KS test

0 引言

高斯混合模型[1-2](Gaussian mixture model, GMM)是有限个独立高斯模型的线性组合,每个样本点由其中的一个高斯模型产生.GMM 广泛应用于模式识别、计算机视觉、机器学习和生物信息等领域,主要完成图像分割、聚类、概率密度函数构建等任务. EM 算法[1-2](expectation maximization algorithm)是求解 GMM 参数的一种重要方法,是对带隐变量模型进行参数估计的常用方法.EM 以似然函数为目标函数,通过迭代优化参数,每步迭代包含期望步(E 步)和极大步(M 步).E 步利用高斯模型参数计算隐变量的期望,M 步根据隐变量的期望对高斯模型参数最大似然估计.由于 EM 算法的目标是最大化似然函数,只考虑了数据拟合度,当子模型个数的初始设置与实际值相差过大时,算法难以估计正确的子模型个数.

关于 GMM 子模型个数选择的研究,已经提出两类很重要的方法.一类是无参贝叶斯方法,如 Ferguson 和 Antoniak 提出的狄利克雷过程高斯混合模型(Dirichlet process Gaussian mixture models, DPGMM)[1-2],该方法通过狄利克雷过程分布中的 scalar 参数控制基分布,即子模型个数,然后从离散化的基分布中随机抽样高斯混合模型的参数.DPGMM 的概率模型关系比较复杂,为得到子模型个数引入了狄利克雷分布,并且需要对狄利克雷过程进行抽样.Blei[3]等利用变分推理对 DPGMM 模型改进,提出了 VDPGMM 算法.基于狄利克雷过程混合模型算法在高维度以及大规模数据集上的效果较好,但是在样本空间中子模型不均衡的样本中表现较差.另一类重要方法是以最短描述长度[4-5](minimum description length,MDL)为目标函数的分裂与合并算法[6-9].MDL 是由香农提出的一种综合考虑数据拟合度与模型复杂度建立可以精确描述对象的数学模型,最小化 MDL 可以避免选择过适应的模型.但该准则中模型复杂度的形式过于简单,不适用于子模型间样本数量不均衡的 GMM.文献[6]改进后的 MDL 更适应于不均衡的 GMM,采用分裂和合并操作的 GMM 参数估计算法就以此版本的 MDL 为目标函数.Li 等[9]提出以 MDL 作为目标函数的分裂与合并的 EM 算法(a noval split and merge EM algorithm for Gaussian mixture model, SMEM),其中以熵比作为子模型分裂与合并的判别准则,分裂与合并操作都要满足熵比增加且 MDL 减小.SMEM 算法迭代执行分裂与合并操作,当无子模型分裂与合并时算法达到收敛,估计出 GMM 参数.SMEM 算法通过分裂与合并操作提高了对子模型个数估计的精度,但熵比对稀疏或凹形子模型过于敏感.实际问题中这两类子模型不应被分裂,并且合并判别只能以遍历所有子模型对的方式确定,算法的计算量较大.针对合并搜索时计算量较大问题,文献[10]提出以 MDL 为目标函数的改进分裂与合并 EM 算法(evolutionary split & merge for expectation maximization, ESM-EM),其中以 KL 散度为子模型分裂的判别准则,以模型相似度为子模型对合并的判别准则,对 KL 散度最大且分裂后 MDL 减小的子模型进行分裂,对模型相似度最大并且合并后 MDL 值减小的子模型对进行合并.算法的分裂与合并迭代过程与 SMEM 算法相似.合并操作以模型相似度为判别准则不需要遍历所有子模型对,相比于 SMEM 算法减少了计算量,但模型相似度判别的合并操作会对空间邻近的子模型对过度合并.SMEM 和 ESM-EM 算法中不适当的分裂或合并操作导致部分子模型需经过多次重复的分裂与合并,出现振荡现象,增大了迭代次数.

为了阻止熵比判别准则对 GMM 样本的稀疏或凹形子模型的过度分裂以及模型相似度判别准则对子模型的过度合并,本文提出基于 KS(Kolmogorov-Smirnov)检验[11]的高斯混合模型分裂与合并算法(KSGMM 算法),该算法以 MDL 为目标函数.在分裂操作中,首先用 KS 检验判断子模型的高斯分布形态,对不满足 KS 检验的子模型中熵比最大的进行分裂.在合并操作中,若模型相似度最大子模型对合并后满足 KS 检验且 MDL 值减小,则执行 EM 算法更新参数,否则放弃合并.对分裂或合并后的子模型执行 EM 算法更新参数.迭代执行分裂与合并操作,当无子模型可分裂或合并时,算法结束得到模型参数.最后设计实验对比分析了 KSGMM 算法与其

他 5 种 GMM 子模型选择算法，验证 KSGMM 算法估计 GMM 子模型个数及参数的有效性.

1　高斯混合模型及 MDL 准则

1.1　高斯混合模型与 EM 算法

GMM 是由 K 个独立高斯模型加权线性组合而成，假设 x 表示 d 维随机变量，其概率分布为

$$p(\boldsymbol{x})=\sum_{k=1}^{K}\boldsymbol{\pi}_k N(\boldsymbol{x}\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{1}$$

式中，$N(\boldsymbol{x}\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ 表示第 k 个高斯分布的概率密度，$\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k$ 是该分布的均值向量和协方差矩阵，K 是独立高斯模型个数，π_k 表示随机变量 $\boldsymbol{x}$ 取自第 k 个高斯模型的概率，满足 $0\leqslant\pi_k\leqslant 1$ 且 $\sum_{k=1}^{K}\boldsymbol{\pi}_k=1$.

令 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 为样本集，$\boldsymbol{Z}\in\mathbb{R}^{N\times K}$ 为模型隐变量矩阵，$Z_{nk}=1$ 代表第 n 个样本由第 k 个模型产生，$\boldsymbol{\gamma}\in\mathbb{R}^{N\times K}$ 为权值矩阵，$\gamma_{nk}=p(Z_{nk}=1)$ 表示第 n 个样本由第 k 个模型产生的概率.EM 算法首先初始化子模型个数 K 、子模型均值 $\boldsymbol{\mu}=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_K\}$、协方差矩阵 $\boldsymbol{\Sigma}=\{\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_K\}$ 及混合权重值 $\boldsymbol{\pi}=\{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_K\}$.

E(expectation)步，根据高斯模型参数 $\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}$，计算混合权重值：

$$\boldsymbol{\gamma}(Z_{nk})=\frac{\boldsymbol{\pi}_k N(\boldsymbol{x}_n\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\boldsymbol{\pi}_j N(\boldsymbol{x}_n\mid\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \tag{2}$$

$$\boldsymbol{\pi}_k=\frac{N_k}{N}=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}(Z_{nk}) \tag{3}$$

式中，N_k 是由第 k 个模型产生的样本总数.

M(maximum)步，根据 E 步得到的权值矩阵 $\boldsymbol{\gamma}$ 更新每个独立高斯模型的参数：

$$\boldsymbol{\mu}_{\mathrm{k}}=\frac{1}{N_k}\sum_{n=1}^{N}\boldsymbol{\gamma}(Z_{nk})\,\boldsymbol{x}_n \tag{4}$$

$$\boldsymbol{\Sigma}_k=\frac{1}{N_k}\sum_{n=1}^{N}\boldsymbol{\gamma}(Z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}} \tag{5}$$

根据 E、M 步得到参数 $\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\gamma}$，计算样本的似然函数值：

$$P(\boldsymbol{X}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\gamma})=\sum_{n=1}^{N}\sum_{k=1}^{K}\boldsymbol{\gamma}(Z_{nk})N(\boldsymbol{x}_n\mid\boldsymbol{\mu}_{\mathrm{k}},\boldsymbol{\Sigma}_k) \tag{6}$$

如果参数 $\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}$ 与 $P(\boldsymbol{X}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\gamma})$ 都不发生较大变化，则停止迭代，否则返回 E 步.

EM 算法以最大化似然函数为目标，仅考虑了数据拟合度.MDL[4-5] 利用数据拟合度与模型复杂度描述数学对象，以 MDL 为目标函数能够在数据拟合度与模型复杂度之间做出平衡.

1.2　高斯混合模型的 MDL 准则

MDL 准则[12-14]用数据拟合度与模型复杂度描述数学对象.其中，样本的似然表示数据拟合度，模型的描述长度表示模型复杂度.MDL 的公式如下：

$$L(\boldsymbol{X}\mid\boldsymbol{\theta})=\log\frac{1}{L(\boldsymbol{X}\mid\boldsymbol{\theta})}+\frac{D}{2}\sum_{k=1}^{K}\log(\frac{N\alpha_k}{12})-\frac{K}{2}\log(\frac{N}{12})-\frac{K(D+1)}{2} \tag{7}$$

式中，$\boldsymbol{X}$ 为样本集，$\boldsymbol{\theta}$ 为模型参数，D 是模型参数的元素个数，K 是模型个数，N 是样本总量，α_k 为第 k 个模型中样本比例.

结合式(1)GMM 的概率分布与式(7)MDL，GMM 的 MDL 公式为

$$L(\boldsymbol{X}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})=\sum_{n=1}^{N}\sum_{k=1}^{K}-\log(\boldsymbol{\gamma}(Z_{nk})N(\boldsymbol{x}_n\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k))+\frac{D}{2}\sum_{k=1}^{K}\log(\frac{N\pi_k}{12})-\frac{K}{2}\log(\frac{N}{12})-\frac{K(D+1)}{2} \tag{8}$$

式中，$D=(d+d(d+1)/2)$ 为 d 维高斯模型参数 $\boldsymbol{\mu}_k$ 、$\boldsymbol{\Sigma}_k$ 中元素个数.

MDL 准则综合考虑了数据拟合度与模型复杂度，基于分裂与合并的 GMM 算法可以通过最小化 MDL 估计子模型个数及 GMM 参数.

2　基于 KS 检验的高斯混合模型分裂与合并算法

基于 KS 检验的高斯混合模型分裂与合并算法(KSGMM 算法)以 MDL 为目标函数，包括分裂操作与合并操作两部分.KSGMM 算法首先执行 EM 算法初始化子模型个数和 GMM 参数，并计算 MDL 值.分裂操作以熵比[9]与 KS 检验[7]作为子模型分裂的判别准则，合并操作以模型相似度[6]与 KS 检验作为子模型对合并的判别准则.对分裂或合并后的子模型执行 EM 算法更新参数.算法迭代执行分裂与合并操作至没有子模型可分裂与合并为止，给出子模型个数和 GMM 参数.

2.1　KS 检验

KS 检验[13]用来计算有限数量样本的经验分布与假设分布的拟合优度.该方法通过比较经验累积分布函数与理论累积分布函数的最大距离判断样本是否来自假设分布.

对于高斯模型，令 $\boldsymbol{x}_n$ 为随机变量，F_N 表示由样本得到的经验累积分布函数，F_0 表示为理论累积分布函数.取 $|F_N(\boldsymbol{x}_n)-F_0(\boldsymbol{x}_n)|$ 在随机变量 $\boldsymbol{x}_n$ 上的最大值作为 KS 检验统计量，则

$$D=\max_{1\leqslant n\leqslant N}|F_N(\boldsymbol{x}_n)-F_0(\boldsymbol{x}_n)|\ ,\ n=1,2,\cdots,N \tag{9}$$

通过查找 KS 检验的临界值表得 $D(N,\alpha)$，其中 N 为样本总量，α 为置信度水平，一般取 0.05.如果 $0.05\leqslant\hat{D}(N,\alpha)$，就认为经验分布符合高斯分布形态；否则，不符合高斯分布形态.

2.2 分裂操作

信息熵[12]是随机变量取值的随机性度量，最大熵[15]是随机变量完全按照概率分布随机取值时的信息熵值，此时变量取值的随机性达到最大，熵比是信息熵与最大熵的比值.对 GMM 第 k 个子模型，令 $H^k(\boldsymbol{X})$ 表示信息熵，$H^k_{\max}(\boldsymbol{X})$ 表示最大熵，该模型熵比 Sp_{k} 为

$$Sp_k=\frac{H^k(\boldsymbol{X})}{H^k_{\max}(\boldsymbol{X})}=\frac{-\dfrac{1}{N_k}\displaystyle\sum_{t=1}^{N_{\mathrm{k}}}\log N(\boldsymbol{X}_t\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\pi}_k)}{-\dfrac{1}{2}\log[(2\pi e)^d|\boldsymbol{\Sigma}_k|]} \tag{10}$$

分裂操作以熵比作为判别准则，选择熵比最小的子模型进行分裂.GMM 中稀疏或凹形子模型的熵比虽然较小，但这类子模型满足聚类条件，不应被分裂为两个子模型.分裂操作通过 KS 检验判断这类子模型各分量上的高斯累积分布形态，阻止分裂.

基于熵比和 KS 检验的分裂操作对熵比最小子模型 k 进行 KS 检验，如果 $D_k>\hat{D}(N_k,\alpha)$ 且其分裂后 GMM 的 MDL 值减小，则完成分裂并更新参数.否则放弃分裂，令 $Sp_k=+\infty$.

子模型 k 分裂为两子模型 k^1 和 k^2，子模型个数为 $K=K+1$.对子模型 k 的协方差矩阵奇异值分解 $\boldsymbol{\Sigma}_k=\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}}$，获得最佳的子模型分裂方向 $\boldsymbol{A}_k=\boldsymbol{U}_k\sqrt{\boldsymbol{S}_k}$，以均分方式[9]在 $\boldsymbol{A}_k$ 方向上更新参数 $\{\pi_{k1},\pi_{k2},\boldsymbol{\mu}_{k1},\boldsymbol{\mu}_{k2},\boldsymbol{\Sigma}_{k1},\boldsymbol{\Sigma}_{k2}\}$，参数更新如下：

$$\boldsymbol{\pi}_{k1}=\boldsymbol{\pi}_{k2}=\frac{1}{2}\boldsymbol{\pi}_k \tag{11}$$

$$\boldsymbol{\mu}_{k1}=\boldsymbol{\mu}_k-\sqrt{\frac{\boldsymbol{\pi}_{k^2}}{\boldsymbol{\pi}_{k1}}}\boldsymbol{\mu}_k\boldsymbol{A}_k \tag{12}$$

$$\boldsymbol{\mu}_{k2}=\boldsymbol{\mu}_k-\sqrt{\frac{\boldsymbol{\pi}_{k1}}{\boldsymbol{\pi}_{k2}}}\boldsymbol{\mu}_k\boldsymbol{A}_k\sqrt{a^2+b^2} \tag{13}$$

$$\boldsymbol{\Sigma}_{k1}=\left(\frac{\boldsymbol{\pi}_{k2}}{\boldsymbol{\pi}_{k1}}\right)\boldsymbol{\Sigma}_k-0.25\,\boldsymbol{A}_k\boldsymbol{A}_k^{\mathrm{T}} \tag{14}$$

$$\boldsymbol{\Sigma}_{k2}=\left(\frac{\boldsymbol{\pi}_{k1}}{\boldsymbol{\pi}_{k2}}\right)\boldsymbol{\Sigma}_k-0.25\,\boldsymbol{A}_k\boldsymbol{A}_k^{\mathrm{T}} \tag{15}$$

以上更新的参数 $\{\boldsymbol{\pi}_{k1},\boldsymbol{\pi}_{k2},\boldsymbol{\mu}_{k1},\boldsymbol{\mu}_{k2},\boldsymbol{\Sigma}_{k1},\boldsymbol{\Sigma}_{k2}\}$ 作为 EM 算法的初始值，运行 EM 至收敛，调整 k^1 和 k^2 的参数.

2.3 合并操作

合并步骤以模型相似度[14]作为子模型对合并的判别准则.定义相似度矩阵为 $\boldsymbol{U}$，$\boldsymbol{U}\in\mathbb{R}^{K\times K}$ 为上三角矩阵.矩阵的元素 $U_{i,j}$ 为子模型对 $\{i,j\}$ 的相似度，该值计算公式如下：

$$U_{i,j}=\frac{\boldsymbol{\gamma}_{\cdot i}^{\mathrm{T}}\boldsymbol{\gamma}_{\cdot j}}{\|\boldsymbol{\gamma}_{\cdot i}\|\,\|\boldsymbol{\gamma}_{\cdot j}\|} \tag{16}$$

式中，$\boldsymbol{\gamma}\in\mathbb{R}^{N\times K}$ 为高斯混合模型的权值矩阵，$\boldsymbol{\gamma}_{\cdot i}$ 代表其中一列，即第 i 子模型的权值向量.相似度 $\boldsymbol{U}_{i,j}$ 等于两子模型权值向量 $(\boldsymbol{\gamma}_{\cdot i},\boldsymbol{\gamma}_{\cdot j})$ 的余弦相似度.

合并操作以模型相似度作为判别准则，合并模型相似度最大的子模型对.GMM 中空间近邻的两子模型的相似度较大，但这类子模型对合并后不符合高斯形态.合并操作后通过 KS 检验计算合并后的子模型高斯分布形态.如果满足高斯形态，执行 EM 算法更新参数，否则放弃合并操作.

基于 KS 检验和模型相似度的合并操作中，把相似度最大子模型对 $\{i,j\}$ 合并为 i^{new}，对子模型 i^{new} 进行 KS 检验，如果 $D_{i^{\mathrm{new}}}<\hat{D}(N_{i^{\mathrm{new}}},\alpha)$ 且合并后 GMM 的 MDL 值减小，则完成合并并更新参数.否则放弃合并，令 $U_{i,j}=-\infty$.

对需要合并的子模型对 $\{i,j\}$，合并后子模型为 i^{new}，子模型个数为 $K=K-1$.参数 $\{\boldsymbol{\pi}_{i^{\mathrm{new}}},\boldsymbol{\mu}_{i^{\mathrm{new}}},\boldsymbol{\Sigma}_{i^{\mathrm{new}}}\}$ 的更新公式如下：

$$\boldsymbol{\pi}_{i^{\mathrm{new}}}=\boldsymbol{\pi}_i+\boldsymbol{\pi}_j \tag{17}$$

$$\boldsymbol{\mu}_{i^{\mathrm{new}}}=(\boldsymbol{\pi}_i\boldsymbol{\mu}_i+\boldsymbol{\pi}_j\boldsymbol{\mu}_j)/\boldsymbol{\pi}_{i^{\mathrm{new}}} \tag{18}$$

$$\Sigma_{i^{\mathrm{new}}}=\frac{1}{\boldsymbol{\pi}_{i^{\mathrm{new}}}}\{\boldsymbol{\pi}_i(\Sigma_i+(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i^{\mathrm{new}}})\times(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i^{\mathrm{new}}})^{\mathrm{T}})+\boldsymbol{\pi}_j(\boldsymbol{\Sigma}_j+(\boldsymbol{\mu}_j-\boldsymbol{\mu}_{i^{\mathrm{new}}})\times(\boldsymbol{\mu}_j-\boldsymbol{\mu}_{i^{\mathrm{new}}})^{\mathrm{T}})\} \tag{19}$$

以上将 i^{new} 的均值设置为子模型 i 和 j 均值的中心点，并计算对应的协方差.以 $\{\boldsymbol{\pi}_{i^{\mathrm{new}}},\boldsymbol{\mu}_{i^{\mathrm{new}}},\boldsymbol{\Sigma}_{i^{\mathrm{new}}}\}$ 为初始值执行 EM 算法调整 i^{new} 的参数.

2.4 算法描述

KSGMM 算法以 MDL 为目标函数.在分裂操

作中,以熵比最小原则和 KS 检验选择待分裂子模型,分裂后执行 EM 算法更新参数.在合并操作中,以模型相似度最大原则确定待合并子模型对,若合并后子模型满足 KS 检验,则执行 EM 算法更新参数,否则放弃合并,继续寻找需合并子模型对.迭代执行分裂与合并操作至无子模型可分裂与合并,得到估计的子模型个数和 GMM 参数.具体步骤如算法 2.1 所示.

(Ⅰ)首先执行 EM 算法初始化子模型个数 K 和 GMM 参数 $\{\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}\}$,计算 MDL 值 L.

(Ⅱ)分裂操作.计算所有子模型的熵比,对熵比最小子模型 k 进行 KS 检验.如果 $D_k > \hat{D}(N_k,\alpha)$,则根据式(11-15)把 k 分裂为 k^1 和 k^2,并执行 EM 算法更新 k^1 和 k^2 的参数,根据式(8)计算 L^{new},若满足 $L^{\text{new}} < L$ 则完成分裂,令 $L = L^{\text{new}}$,$K = K+1$.否则放弃分裂,令 $Sp_k = +\infty$.继续寻找待分裂子模型.

(Ⅲ)合并操作.计算子模型间相似度矩阵 $\boldsymbol{U}$,根据式(17-19)合并相似度最大子模型对 $\{i,j\}$ 为 i^{new},并执行 EM 算法更新参数.根据式(8)计算 GMM 的 MDL 值 L^{new}.如果 $D_{i^{\text{new}}} < \hat{D}(N_{i^{\text{new}}},\alpha)$ 且 $L^{\text{new}} < L$,则完成合并,令 $L = L^{\text{new}}$,$K = K-1$.否则放弃合并,令 $\boldsymbol{U}_{i,j} = -\infty$.继续寻找待合并子模型对.

(Ⅳ)KSGMM 算法迭代分裂与合并操作,如果不再有满足分裂与合并条件的子模型,结束算法.输出子模型个数与 GMM 参数.

算法 2.1　KSGMM 算法

算法描述

输入:样本集 X={x_1,x_2,…,x_n};初始模型个数 K;初始聚类中心

　　μ={μ_1,μ_2,…,μ_K};

```
do{
    根据式(8)计算 MDL 值 L;
    执行 EM 算法,计算 GMM 参数{μ,Σ,π};
    %分裂子模型
    根据式(10)计算所有子模型的熵比{Sp_i | i=1,2,…,K};
    sort({Sp_i | i=1,2,…,K}, ascend) ;
    for i in K{
      %KS 检验
      根据式(9)计算 KS 检验值 D_k;
      if(D_k>D̂(N_k,α)){
        根据式(11-15),分裂 k 并初始参数{μ_k1,Σ_k1,π_k1,μ_k2,Σ_k2,π_k2};
        执行 EM 算法,更新{μ_k1,Σ_k1,π_k1,μ_k2,Σ_k2,π_k2};
        %判断 MDL 是否减小
        根据式(8)计算分裂后 GMM 的 MDL 值 L^new;
        if(L^new<L){
          K=K++;
          L=L^new;
          break;
        }else{
          Sp_k=+∞;
          放弃分裂操作}
      }else{
        Sp_k=+∞;
      }
    }
    %合并子模型
    根据式(16)计算相似度矩阵 U;
    for i in size(U_i,j ~=0){
      合并相似度最大子模型对{i,j}为 i^new;
      根据式(17-19)初始参数{μ_inew,Σ_inew,π_inew};
      执行 EM 算法,更新参数;
      %KS 检验
      根据式(9),计算 KS 检验值 D_inew;
      if(D_inew<D̂(N_inew,α)){
        根据式(8),计算合并后 GMM 的 MDL 值 L^new;
        if(L^new<L){
          K=K--;
          L=L^new;
          break;
        }else{
          U_i,j=-∞;
          放弃合并;
        }
        }else{
          U_i,j=-∞;
        }
      }
    }while(无合并与分裂操作){
      终止算法;
    }
    输出:模型个数 K;模型参数{μ,Σ,π};
```

3　实验

本节在 UCI 数据集[3]、MNIST 手写体数据集[16]和生成数据集上设计了两组对比实验.

3.1 实验环境

数据集包括：5 个 UCI 数据集，分别是 Aggregation，k2far，Gaussian3，oldfaithful，sample1，全部为 2 维数据；MNIST 手写体数据集，是 0～9 的阿拉伯数字手写体，每个样本有 784 维特征；生成数据集 bigdata_40c，共有 40 个子模型，2 个属性，使用 Matlab 的 gmdistribution 函数生成.数据集具体信息如表 1 所示.

表 1 数据集信息

Tab.1 Information of dataset

数据集	标签	子模型个数	样本个数	特征个数
Aggregation	有	7	788	2
k2far	有	4	400	2
Gaussian3	无	4	5 000	2
oldfaithful	无	2	272	2
sample1	无	3	4 000	2
bigdata_40c	有	40	40 000	2
MNIST	有	10	60 000	784

实验对比算法包括 KSGMM 算法、EM 算法[14]、SMEM 算法[9]、ESM-EM 算法[6]、DPGMM 算法和 VDPGMM 算法. 其中，DPGMM 和 VDPGMM 是基于狄利克雷过程混合模型的 GMM 算法.算法的详细信息如表 2 所示.

表 2 算法准则

Tab.2 Criterion of algorithm

算法	子模型选择准则	目标函数
KSGMM	分裂：KS 检验＋熵比 合并：KS 检验＋模型相似度	MDL
ESM-EM	分裂：KL 散度 合并：模型相似度	MDL
SMEM	分裂：熵比 合并：熵比	MDL
EM	后验概率最大模型	MLE
DPGMM	最大似然	MLE
VDPGMM	变分证据下界	ELBO

实验平台的 CPU 采用 Intel(R) Core(TM) i5-4590 CPU @ 3.30Hz，运行内存为 4 GB，操作系统 Windows 7 SP1，运行平台工具 Matlab 2016b.

本节设计了两组对比实验.实验一，在 7 个数据集上运行 KSGMM 和 EM，SMEM，ESM-EM，DPGMM，VDPGMM 算法，比较分析 6 种算法估计的子模型个数及其迭代变化.实验二，把含标签数据集 Aggregation，k2far，bigdata_40c 和 MNIST 分为训练集和测试集，对 6 种算法进行训练和测试，比较算法的训练和测试精度.

3.2 实验一

实验一在 Aggregation，Gaussian3，k2far，oldfaithful，sample1，MNIST，bigdata_40c 数据集上运行 KSGMM 和 EM，SMEM，ESM-EM，DPGMM，VDPGMM 算法，bigdata_40c 数据集的子模型个数初始值设置为 45，其他数据集的子模型个数的初始值都设置为 15.KSGMM，SMEM 和 ESM-EM 算法基于分裂与合并操作，子模型个数在迭代中会出现波动，这 3 种算法的迭代过程如图 1 所示.所有算法估计的子模型个数如表 3 所示.为了实验结果的可视化，选取 UCI 数据集的聚类结果进行绘图，如图 2 所示.

表 3 子模型个数估计值

Tab.3 Number of sub-models

数据集	EM	SMEM	ESM-EM	KSGMM	DPGMM	VDPGMM
Aggregation	4	9	9	6	4	2
k2far	6	5	5	4	5	4
Gaussian3	5	4	6	2	1	1
oldfaithful	4	6	4	2	2	1
sample1	5	4	4	3	10	9
bigdata_40c	2	6	6	34	14	6
MNIST	9	15	16	12	15	8

由表 3 可知，KSGMM 对 k2far，Gaussian3，oldfaithful，sample1 子模型个数的估计与实际值一致；在 Aggregation 数据集上子模型个数的估计值为 6，实际值为 7；在 bigdata_40c 数据集上子模型个数的估计值是 34，实际值是 40；在 MNIST 数据集上子模型个数的估计值是 12，实际值是 10.EM 算法在 Aggregation，bigdata_40c 数据集上的估计值小于实际值，在其他 5 个数据集上的估计值均大于实际值.ESM-EM、SMEM 算法在 bigdata_40c 数据集上的子模型估计值小于实际值，其他数据集子模型个数估计值小于实际值.DPGMM 算法在 oldfaithful

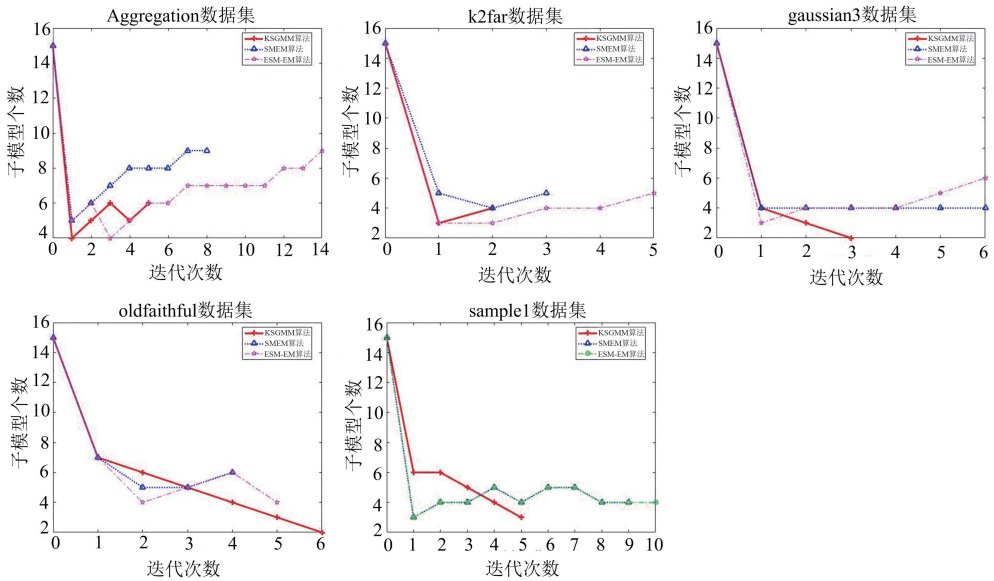

图 1　迭代过程

Fig.1　Iteration process

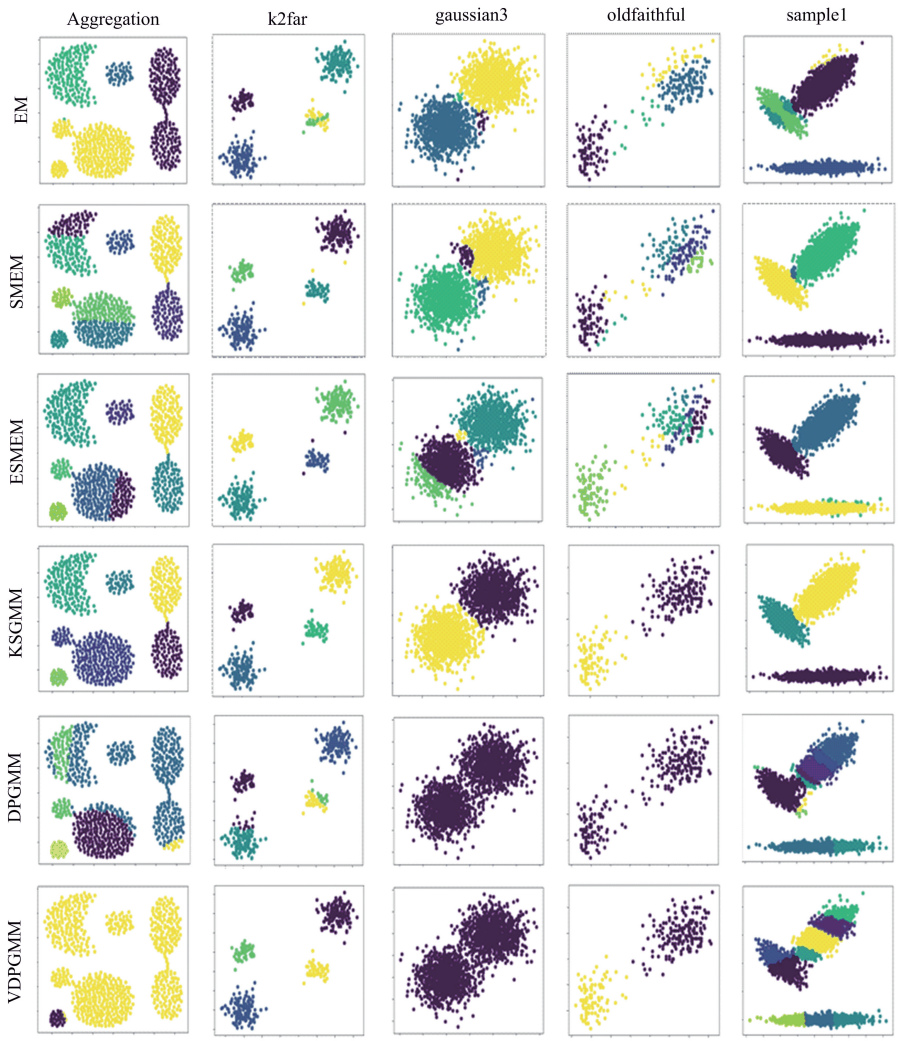

图 2　算法聚类结果

Fig.2　Results of Clustering

数据集上估计的子模型个数与实际值相等，在 Aggregation，Gaussian3 和 bigdata_40c 数据集上的子模型个数估计值小于实际值，在 k2far，sample1 和 MNIST 数据集上的子模型个数估计值均大于实际值.VDPGMM 算法在 k2far 数据集上得到正确的子模型个数估计值，在 Aggregation，Gaussian3，oldfaithful，bigdata_40c 和 MNIST 数据集上的子模型个数估计值小于实际值，其他数据集上子模型个数估计值均大于实际值.KSGMM 算法给出最优子模型个数选择结果.

由图 2 可知，KSGMM 算法把 Aggregation 中相互连接的两个子模型划分为 1 个子模型，在其他的数据集上均得到了正确的聚类结果.EM 算法把 Aggregation 中紧密连接的子模型划分为 1 个子模型，把 k2far，Gaussian3，oldfaithful 数据集中的稀疏样本划分为多个子模型.SMEM 和 ESM-EM 算法把 Aggregation 中凹形和圆形区域内样本划分为 2 个子模型，把 Gaussian3，oldfaithful，sample1 中稀疏样本划分为 2 个或多个子模型.基于狄利克雷过程的 DPGMM 算法和 VDPGMM 算法把 Aggregation，Gaussian3 和 oldfaithful 数据集中多个模型错误的估计为 1 个基分布，把 k2far，sample1 和数据集中的 1 个基分布错误估计为多个基分布.

由算法的迭代过程图 1 可知，KSGMM 算法在 Aggregation，k2far，Gaussian3，sample1 上的迭代次数都达到了最小，ESM-EM 算法在 Aggregation，Gaussian3，sample1 上存在振荡现象，SMEM 算法在 Gaussian3 上存在振荡现象.

综合以上分析，SMEM 算法在含有稀疏样本的数据中容易过度分裂，在学习过程中产生振荡现象，估计的子模型个数较实际值偏大，在大规模子模型数据中容易过度合并，子模型个数估计值较实际值偏小.ESM-EM 算法在样本分布不均衡与稀疏样本中都容易造成过度分裂，在大规模子模型数据中容易过度合并，这种过度操作导致算法迭代过程出现振荡现象.相对于以上 2 种算法，KSGMM 算法基于熵比和 KS 检验的分裂判别准则在分裂操作中能够保留 GMM 中的稀疏或凹形子模型，基于模型相似度和 KS 检验的合并判别准则在合并操作中能够阻止过度合并.这种准则能够消除振荡现象，减少迭代次数，并且估计的子模型个数接近实际值.基于狄利克雷过程的 DPGMM，VDPGMM 算法在不均衡和大规模子模型的样本中对基分布的离散化较小，低估子模型个数，在含有稀疏样本的样本中对基分布过于离散化，估计子模型偏大.

3.3 实验二

实验二把含有标签的数据集 Aggregation、k2far、bigdata_40c 和 MNIST 划分为训练集和测试集.其中，Aggregation 的训练集与测试集样本数量分别为 600 和 188，k2far 的训练集与测试集样本数量分别为 300 和 100，bigdata_40c 的训练集与测试集样本数量分别是 36 000 和 4 000，MNIST 的训练集与测试集样本数量分别是 60 000 和 10 000.在 Aggregation、k2far、bigdata_40c 和 MNIST 的训练集上运行 KSGMM 算法和 ESM-EM，SMEM，EM，DPGMM，VDPGMM 算法，并将 bigdata_40c 数据集对应的初始子模型个数设置为 45，其他数据集的初始子模型个数设置为 15.用训练结果对 Aggregation，k2far，bigdata_40c 和 MNIST 的测试集进行预测.算法训练和测试精度如表 4 所示.

表 4 算法精度比较

Tab.4 Accuracy comparison of algorithms

	Aggregation			k2far			bigdata_40c			MNIST		
	Components	Train	Test	Components	Train	Test	Components	Train	Test	Components	Train	Test
KSGMM	6	95%	97.87%	4	100%	100%	34	97.49%	85.00%	12	66.05%	47.82%
ESM-EM	9	70.17%	77.66%	5	98.3%	98.4%	6	15.0%	15.06%	16	48.08%	13.21%
SMEM	9	77.33%	77.13%	5	97.57%	98.1%	6	15.0%	15.0%	15	32.14%	9.63%
EM	4	66.45%	64.89%	6	90.67%	96%	2	7.50%	7.50%	9	59.61%	15.7%
DPGMM	4	41.17	40.43%	5	86.67%	90.0%	14	35.01%	12.5%	15	48.08%	59.5%
VDPGMM	2	34.5%	49.47%	4	100%	100%	6	15.0	10.0%	8	52.43%	68.05%

结合表 1 和表 4 分析.在 Aggregation 数据集上,KSGMM 算法的子模型个数估计值、训练精度和测试精度都得到最佳结果.在 k2far 数据集上,KSGMM 和 VDPGMM 算法的子模型个数估计值,训练精度和测试精度与真实值完全相同.在 bigdata_40c 数据集上,KSGMM 算法的子模型个数估计值与实际值最接近,算法的测试精度和训练精度分别为 97.49%和 85.0%.在 MNIST 数据集上,EM 算法的子模型个数估计值最接近真实值,KSGMM 算法的训练精度最好,为 66.05%;VDPGMM 算法的测试精度达到最高,为 68.05%.

综上可知,子模型个数的选择对算法预测精度影响很大,算法选择的子模型个数越接近实际值,训练和测试的精度越高.在稀疏、凹形以及子模型样本不均衡的样本中,以熵比为判别准则和以 KL 散度、模型相似度为判别标准的 GMM 算法以及基于狄利克雷过程的 GMM 估计算法表现都较差,尤其在大规模子模型样本中,这两类算法的子模型个数估计值与实际值差距较大.KSGMM 算法通过基于熵比和 KS 检验的分裂以及基于 KS 检验和模型相似度的合并估计的子模型个数最为接近实际值,提高了 GMM 参数的估计精度.

4　结论

本文提出 KSGMM 算法,以 MDL 为目标函数平衡数据拟合度与模型复杂度.基于熵比和 KS 检验的分裂判别准则在分裂操作中能够保留 GMM 中的稀疏或凹形子模型,基于模型相似度和 KS 检验的合并判别准则在合并操作中能够阻止过度合并.实验证明,KSGMM 算法能够精确估计子模型个数,提高了对 GMM 参数的估计精度.同时在算法的迭代中,每次运行 EM 之前通过 KS 检验阻止过度的分裂与合并,KSGMM 算法能够避免振荡现象,提高收敛速度.

参考文献(References)

[1] BISHOP C M. Pattern recognition [J]. Machine Learning, 2006, 128: 1-58.

[2] MURPHY K P. Machine Learning: A Probabilistic Perspective [M]. Cambridge, MA: The MIT Press, 2012.

[3] BLEI D M, JORDAN M I. Variational inference for Dirichlet process mixtures [J]. Bayesian Analysis, 2006, 1(1): 121-143.

[4] SHIFFRIN R M, CHANDRAMOULI S H, GRÜNWALD P D. Bayes factors, relations to minimum description length, and overlapping model classes[J]. Journal of Mathematical Psychology, 2016, 72: 56-77.

[5] RISSANEN J. Minimum description length principle [M]. John Wiley & Sons, Inc., 1985.

[6] COVÕES T F, HRUSCHKA E R. Splitting and merging Gaussian mixture model components: An evolutionary approach[C]// 2011 10th International Conference on Machine Learning and Applications and Workshops. Piscataway, NY, USA: IEEE Press, 2011, 1: 106-111.

[7] SCRUCCA L. Identifying connected components in Gaussian finite mixture models for clustering [J]. Computational Statistics & Data Analysis, 2016, 93: 5-17.

[8] LI Y, LI L. A greedy merge learning algorithm for Gaussian mixture model [C]// International Symposium on Intelligent Information Technology Application. Piscataway, NY, USA: IEEE Press, 2009:506-509.

[9] LI Y, LI L. A novel split and merge EM algorithm for Gaussian mixture model[C]// 2009 Fifth International Conference on Natural Computation. Piscataway, NY, USA: IEEE Press, 2009, 6: 479-483.

[10] COVÕES T F, HRUSCHKA E R, GHOSH J. Evolving Gaussian mixture models with splitting and merging mutation operators. [J]. Evolutionary Computation, 2016, 24(2):293.

[11] ZHU R, WANG L, ZHAI C, et al. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm [J]. Proceedings of Machine Learning Research, 2017, 70: 4180-4188.

[12] ASUNCION A, NEWMAN D H. UCI machine learning repository[EB/OL]. [2017-08-20]. http://archive.ics.uci.edu/ml/index.php.

[13] COVER T M, THOMAS J A. Elements of Information Theory[M]. John Wiley & Sons, Inc., 1991.

[14] MASSEY JR F J. The Kolmogorov-Smirnov test for goodness of fit[J]. Journal of the American Statistical Association, 1951, 46(253): 68-78.

[15] LI R, PERNECZKY R, DRZEZGA A, et al. Survival analysis, the infinite Gaussian mixture model, FDG-PET and non-imaging data in the prediction of progression from mild cognitive impairment[J]. arXiv preprint arXiv:1512.03955, 2015.

[16] LECUN Y, CORTES C, BURGES C. THE MNIST DATABASE of handwritten digits[DB/OL]. [2017-08-20]. http://yann.lecun.com/exdb/mnist/.