

医学图像的混合模型成份数估计

谢从华^{1,2} 宋余庆² 陈健美² 常晋义¹

(常熟理工学院计算机科学与工程学院 苏州 215500)¹ (江苏大学计算机科学与工程学院 镇江 212013)²

摘 要 混合模型成份数估计是医学图像聚类分析和密度估计的关键。针对基于信息准则的估计方法存在过拟合问题,提出了一种新的基于高斯混合模型特征函数的估计方法。首先定义医学图像高斯混合模型的特征函数,然后构造了一个基于特征函数的混合模型成份估计准则,最后设计了该准则的实现算法。新的估计方法通过选择合适的参数调控对数特征函数,让惩罚函数起到平衡作用。模拟数据和真实数据实验表明,此方法确定的混合模型的成份数 K 比其他经典的信息准则方法确定的更合理,避免了医学图像的过拟合问题。

关键词 高斯混合模型,模型选择,特征函数

中图法分类号 TP391.4 文献标识码 A

Estimating the Number of Components of Mixture Models for Medical Image

XIE Cong-hua^{1,2} SONG Yu-qing² CHEN Jian-mei² CHANG Jin-yi¹

(School of Computer Science and Engineering, Changshu Institute of Technology, Suzhou 215500, China)¹

(School of Computer Science and Telecommunication, Jiangsu University, Zhenjiang 212013, China)²

Abstract Estimating the number of mixture models is the key part of clustering analysis and density estimation for medical image. In order to overcome the over-fitting problem of the method of information criteria, we proposed a new estimation method which is based on a feature function of Gaussian mixture models (GMMs). First, the feature function of medical image was defined on the GMMs. Second, constructed a new criterion with the feature function to estimate the number of components of the mixture models. At last, we proposed an algorithm to compute the new criterion. Our new criterion uses a parameter to adjust the value of log-feature function and to keep the balance effect of the penalized function. Experiments on the simulate data and real CT image show our criterion can determine a more reasonable number of components K than others information model selection criteria and avoid the over-fitting problem of the medical image.

Keywords Gaussian mixture models, Model selection, Feature function

1 引言

有限混合模型,是一个非常灵活的统计建模工具,尤其是高斯混合模型广泛应用于图像聚类^[1]、图像分割^[2]、图像特征提取^[3]和图像归类和检索^[4]等。高斯混合模型的参数估计的主要方法是最大似然参数估计(如最经典的 EM^[5]算法),还有指导的在线学习法^[6]。最大似然参数估计要把混合成份数 K 当作先验知识,面对复杂的数据,即使专家也很难确定合适的 K 值。许多的医学图像聚类分析算法,如 K-mean 等都不能确定其成份数 K 。可见确定混合成份数是医学图像处理和分析的一个关键问题,到目前为止也没有一个满意的方法。

Mclachlan G 总结了假设检验准则、信息准则、分类准则和其他多种方法^[1]。一个比较直接的方法是通过连续的假设检验法,假设有 s 类和对立的 $s+1$ 类,使用著名的似然比率检验在两者之间作出选择,然后混合模型数一个一个地增加^[7]。由于基于假设检验法的计算量太大,目前主要使用有

惩罚的对数似然函数准则,这种似然函数随着混合模型的增加而增加的准则称为信息准则。基于分类准则是利用信息比矩阵的性质的模型选择准则。其他的估计方法有交叉验证准则^[8]、分裂合并法^[9]等。

对于真实数据,混合成份的真实数目是未知的,必须从数据中推出来,目前还没有找到一个对所有情况都是最优的方法。信息准则因计算简单而应用于不同领域,最常用的信息准则有贝叶斯推理准则(BIC^[10])、最小描述长度(MDL^[11])、AIC(Akaike's Information Criterion)等。但对于医学图像,因为其内容复杂、数据量大等特点,这些信息准则的对数似然函数远大于惩罚函数,导致了医学图像的过拟合问题。

为此,本文提出了一种新的医学图像高斯混合模型选择准则。通过定义图像高斯混合模型的特征函数,构造了模型选择准则。特征函数的参数可以调控对数特征函数的值,使得惩罚函数能起到平衡的作用,并且可以避免医学图像出现过拟合。模拟数据和真实数据实验表明基于特征函数的医学

到稿日期:2009-11-18 返修日期:2010-01-20
资助。

本文受国家自然科学基金项目(60841003),江苏省软件和集成电路专项基金(2009[100])

谢从华(1978—),男,博士生,讲师,主要研究方向为医学图像处理和模式识别, E-mail: x7c8h5@yahoo.com.cn; 宋余庆(1959—),男,教授,博士生导师,主要研究方向为医学图像处理; 陈健美(1962—),女,博士,副教授,主要研究方向为医学图像识别。

图像模型选择准则是有效的。

2 高斯混合模型的相关工作

定义 1(有限混合模型) 设数据集 $X = \{X_1, X_2, \dots, X_N\}$, 其中 $X_i \in R^d$, 且 X 由 K 个独立同分布的模型构成, 其混合分布概率密度函数可以写成:

$$P(X|\Theta) = \sum_{r=1}^K \alpha_r f(X|\varphi_r) \quad (1)$$

式中, $\Theta = (\alpha_r, \varphi_r) (r=1, 2, \dots, K)$, α_r 为模型 r 的权重, φ_r 为模型 r 的均值和协方差, $f(X|\varphi_r)$ 为 X 属于模型 r 的密度函数, α_r 满足 $\sum_{r=1}^K \alpha_r = 1 (\alpha_r > 0)$ 。

定义 2(高斯混合模型) 如果 X 的有限混合模型的成份 r 服从高斯分布, 且模型的均值为 μ_r , 协方差为 Σ_r , 其密度函数为:

$$f(X|\mu_r, \Sigma_r) = \exp\left[-\frac{1}{2}(X-\mu_r)^T \Sigma_r^{-1}(X-\mu_r)\right] / ((2\pi)^{d/2} |\Sigma_r|^{1/2}) \quad (2)$$

则 $G(X) = \sum_{r=1}^K \alpha_r f(X|\mu_r, \Sigma_r)$ 为 X 的高斯混合模型。

高斯混合模型的参数估计最经典的算法是 EM^[5], 而模型选择即确定 K 值, 通常采用基于惩罚的对数似然函数的信息准则。对数似然函数为:

$$\log L(X|\Theta) = \sum_{j=1}^N \log\left(\sum_{i=1}^K \alpha_i f(X_j|\mu_i, \Sigma_i)\right) \quad (3)$$

惩罚函数一般使用 K 个模型中(自由)参数的个数。如果 $X \in R^d$, 则模型 r 的参数 $\alpha_r, \mu_r, \Sigma_r$ 分别有 $1, d, (d^2+d)/2$ 个自由参数, K 个模型的惩罚函数^[12]为:

$$P(K) = K(d^2 + 3d)/2 + K - 1 \quad (4)$$

Akaike 提出最常用的模型选择准则 AIC:

$$AIC(K) = -2\log L(X|\Theta) + 2P(K) \quad (5)$$

Bozdgoan^[13]对 AIC 修正为 AIC₃:

$$AIC_3(K) = -2\log L(X|\Theta) + 3P(K) \quad (6)$$

另一个常用的准则是 Schwarz 提出的基于贝叶斯的信息准则 BIC:

$$BIC(K) = -2\log L(X|\Theta) + \log N * P(K) \quad (7)$$

对于医学图像, 这些准则的对数似然函数远大于惩罚函数, 从而导致过拟合问题。

3 我们的方法

3.1 高斯混合模型的特征函数

我们提出的准则的理论基础是定义了一个高斯混合模型的特征函数。在物理振动波理论中, 模型的特征函数可以解释为一个振荡器, e^{jT_x} 为单位谐波, 特征函数是多个简谐波 $f_r(x)e^{jT_x} dx$ 运动的叠加。而这个解释与混合模型的基本思想是一致的。从图像信号而言, 特征函数是该模型所有数据的密度叠加。

定义 3(特征函数) 设图像 X 的第 r 个模型为 ξ_r , 则 ξ_r 的特征函数为:

$$\varphi_r(t) = E(e^{jT_{\xi_r}}) = \int_{-\infty}^{+\infty} f_r(x) e^{jT_x} dx \quad (8)$$

($t = (t_1, t_2, \dots, t_d)^T$)

式中, $j = \sqrt{-1}$, $E(\cdot)$ 是一个期望函数。

定理 1 如果模型 ξ_r 是均值为 μ_r , 协方差为 Σ_r 的高斯

分布, 即 $\xi_r \sim N_r(\mu_r, \Sigma_r)$, 则该模型的特征为 $\varphi_r(t) = \exp\{jt^T \mu_r - \frac{1}{2} t^T \Sigma_r t\}$ 。

证明: 因为 $\xi_r \sim N_r(\mu_r, \Sigma_r)$, 作一个变化 $\xi_r = \Sigma_r^{1/2} y + \mu_r$, 那么 y 的密度函数为:

$$g(y) = f(\Sigma_r^{1/2} y + \mu_r) \left| \frac{\partial \xi_r}{\partial y} \right| = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2} y^T y\right\}$$

也就是说 y 为标准高斯分布 $y \sim N_d(0, I)$ 。根据特征函数的定义, 有

$$\begin{aligned} \varphi_r(t) &= E(e^{jT_{\xi_r}}) = E(e^{jT(\Sigma_r^{1/2} y + \mu_r)}) \\ &= e^{jT \mu_r} \prod_{i=1}^d e^{-\frac{1}{2} (\Sigma_r^{1/2})_{ii}^2} = \exp\{jt^T \mu_r - \frac{1}{2} t^T \Sigma_r t\} \end{aligned}$$

证毕。

为了书写方便, 我们记

$$a_r(t) = t^T \mu_r, b_r(t) = -\frac{1}{2} t^T \Sigma_r t \quad (9)$$

则特征函数为:

$$\varphi_r(t) = \exp\{a_r(t)j + b_r(t)\} (1 \leq r \leq K) \quad (10)$$

定理 2 高斯混合模型的特征函数为 K 个相互独立模型的加权特征函数之积。

证明: 因图像 X 的混合模型由 K 个模型线性组合, 即 $\alpha_1 \xi_1 + \alpha_2 \xi_2 + \dots + \alpha_K \xi_K$, 根据模型特征函数的定义, 图像 X 的特征函数为 $\varphi(t) = E(e^{jT_X}) = E(e^{jT(\alpha_1 \xi_1 + \alpha_2 \xi_2 + \dots + \alpha_K \xi_K)}) = \varphi(\alpha_1 t, \alpha_2 t, \dots, \alpha_K t)$ 。又这 K 个模型相互独立, 所以 $\varphi(\alpha_1 t, \alpha_2 t, \dots, \alpha_K t) = \varphi_1(\alpha_1 t) \varphi_2(\alpha_2 t) \dots \varphi_K(\alpha_K t)$ 。因此, $\varphi(t) = \prod_{r=1}^K \varphi_r(\alpha_r t)$ 。证毕。

由于特征函数具有良好的可控性质和简洁的形式, 我们提出了基于特征函数的模型选择准则。

3.2 基于特征函数的模型选择准则

对于复杂的医学图像, 其对数似然函数值要远远大于惩罚函数, 一般的信息准则都会导致过拟合问题。而参数 t 可以控制对数特征函数, 选取合适范围的 t , 惩罚函数就能起到平衡的作用。因对数特征函数是一个复数, 惩罚函数为实数, 为了统一, 我们定义了对数特征函数的频谱函数来代替它:

$$|\log \varphi(t)| = \sqrt{(a_r a_r(t))^2 + (a_r b_r(t))^2} \quad (11)$$

通过这个频谱函数, 我们构造了一个新的模型准则函数:

$$FFAIC(t, K) = |\log \varphi(t)| + P(K) \quad (12)$$

t 的合适范围定义为频谱函数与惩罚函数的数量级在 10 倍以内, 即

$$0.1 \leq |\log \varphi(t)| / P(K) \leq 10 \quad (13)$$

预先设置 K 的范围为 $2 \leq K \leq K_{\max}$ (一般 $K_{\max}$ 容易确定), 准则函数取得最小值的 K 就是混合模型的成份数。

综上所述, 根据特征函数, 设计了医学图像的高斯混合模型准则的计算算法。

算法 1 医学图像的高斯混合模型准则计算算法

输入: 图像 X , K 的最大可能值 $K_{\max}$

输出: 混合模型成份数 K

(1) $K=2$;

(2) While($K \leq K_{\max}$)

(3) { 用 K-mean 初始划分图像 X 为 K 个聚类, 并初始化模型参数 $\Theta = (\alpha_r, \varphi_r) (r=1, 2, \dots, K)$;

(4) 采用 EM 算法计算后验概率, 并更新权重、均值和协方差矩阵;

(5) 根据式(10)计算每一个模型的特征函数;

- (6) 根据定理 2 计算混合高斯模型的特征函数值;
- (7) 如果对数似然函数收敛, 则 $K=K+1$, 转入(2); 否则转入(4);
- (8) 根据式(4)计算 $P(K)(2 \leq K \leq K_{\max})$;
- (9) 根据式(13)计算参数 t ;
- (10) 根据式(12)计算准则函数, 并确定 K 。

设 s_1 为 K-mean 迭代的次数, 第(4)~(7)步要迭代 s_2 次, 则算法 1 的时间复杂度为 $O(s_1KN+s_2KK_{\max}N^2d^2)$ 。同 AIC, AIC_3 , BIC 相比较, 本文通过特征函数的参数让惩罚函数能真正地起到平衡的作用, 有效地避免了图像的过拟合问题。与其他的高斯混合模型密度估计方法^[1,4]相比, 本文构建了一个特征函数空间选择模型, 我们的方法更适合于医学图像这类特殊的数据。

4 实验结果

为了验证我们的混合模型选择准则的有效性, 同 AIC, AIC_3 , BIC 准则在模拟和真实的医学图像数据上进行了比较实验。

4.1 模拟 2D 图像

模拟的 2D 图像如图 1 所示, 共有 300,000 个数据, 9 个高斯模型, 模型的均值、协方差和权重如表 1 所列。对不同的 K 值, AIC, AIC_3 , BIC 准则的函数值和变化趋势如图 2 所示。因为模拟图像的数据量较大, 对数似然函数是惩罚函数的 1000 倍, 对数似然函数起着决定性作用, 惩罚函数起不到平衡的作用, 所以准则函数单调递减。从图 2 可知, 当这些准则函数最小时, K 都肯定大于 9, 出现了过拟合问题。而我们的 FFAIC 准则函数最小时, $K=9$, 正确地确定了模型数目。

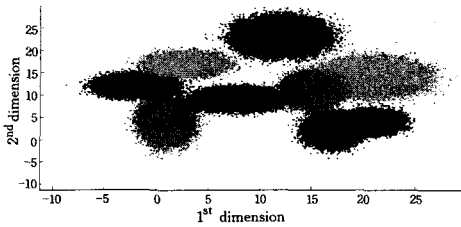


图 1 模拟图像

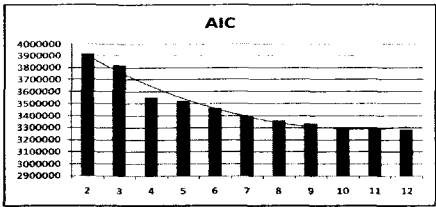
表 1 数据模型参数

成份	均值	协方差	权重
1	(17,2)	{1,0;0.2}	0.1
2	(8,9)	{3,0;0.1}	0.1
3	(3,17)	{2,0;0.1}	0.1
4	(12,23)	{3,0;0.3}	0.1
5	(21,4)	{1,0;0.1}	0.1
6	(1,4)	{1,0;0.4}	0.1
7	(-2,12)	{2,0;0.1}	0.1
8	(21,14)	{-4,0;0.3}	0.1
9	(15,11)	{1,0;0.2}	0.2

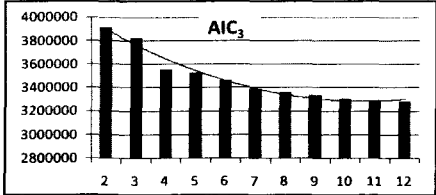
4.2 真实的 CT 腹部图像

图 3 是来自江苏大学附属医院的一幅 CT 腹部图像, 大小为 512×512 , 除去背景后有 141,281 个数据。各种模型选择的准则函数及趋势图如图 4 所示, AIC, AIC_3 , BIC 准则对图 3 确定的模型数 K 大于 30, 明显出现了过拟合现象, 而我们的 FFAIC 准则确定为 21。图 5 为图 3 的密度函数, 图像的组织器官对应着密度函数的山峰, 密度函数可以有效地表达医学图像内容, 从图 5 中可以定性分析 $K=21$ 是比较合理的。我们的方法能有效避免过拟合问题, 主要是由于通过参

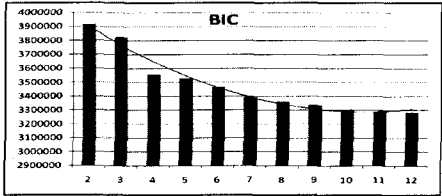
数 t 进行了合理调控, 让惩罚函数起到了平衡作用。



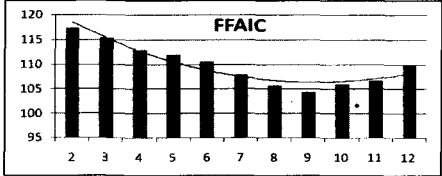
(a) AIC 准则



(b) AIC_3 准则



(c) BIC 准则



(d) FFAIC 准则

图 2 模拟数据的不同模型选择准则及其趋势图

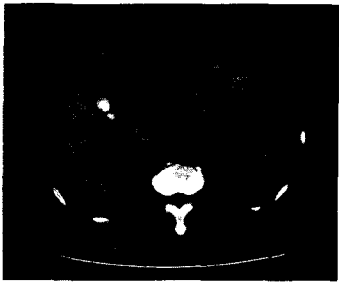
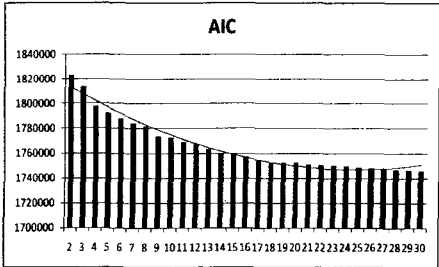
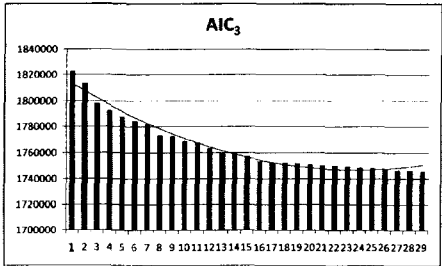


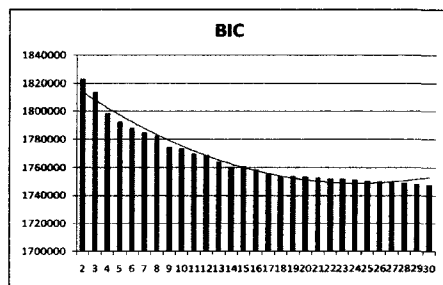
图 3 CT 腹部医学图像



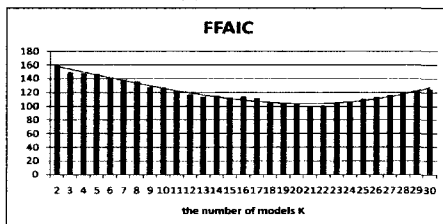
(a) AIC 准则



(b) AIC_3 准则



(c) BIC准则



(d) FFAIC准则

图4 图3的不同模型选择准则及其趋势图

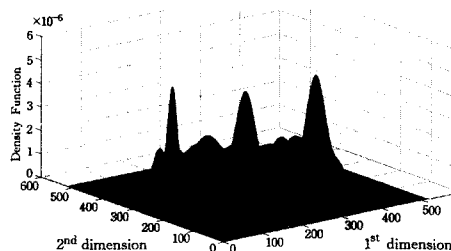


图5 图3的高斯混合密度函数(K=21)

结束语 本文提出了一种基于特征函数的医学图像高斯混合模型选择准则,克服了目前广泛采用的基于似然函数的高斯混合模型选择存在的过拟合问题。基于特征函数混合模型选择准则将会对统计学、机器学习和模式识别等领域有着一定的理论和应用价值。我们下一步工作将研究基于该准则的医学图像高斯混合模型聚类 and 分割等方面的应用。

参考文献

- [1] McLachlan G, Peel D. Finite Mixture Models. Wiley series in probability and statistics[M]. New Yourk: Wiley, 2000
- [2] Yang X Y, Shankar M K. Image Segmentation Using Finite

Mixtures and Spatial Information[J]. Image and Vision Computing, 2004(22): 735-745

- [3] Alata O, Quintard L. Is there a best color space for color image characterization or representation based on Multivariate Gaussian Mixture Model? [J]. Computer Vision and Image Understanding, 2009, 113: 867-877
- [4] Hayit G, Adi T P. Medical Image Categorization and Retrieval for PACS Using the GMM-KL Framework[J]. IEEE Transactions on Information Technology in BioMedicine, 2007, 11(2): 190-202
- [5] Dempster A P, Laird N M, Rubin D B. Maximum-likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, SERIES B, 1977, 39(1): 1-38
- [6] Kivinen J, Manfred K W. Additive versus exponentiated gradient updates for linear prediction[J]. Information and Computation, 1997, 132(1): 1-64
- [7] Durio A, Isaia E D. A quick procedure for model selection in the case of mixture of normal densities[J]. Computational Statistics & Data Analysis, 2007, 51: 5635-5643
- [8] Miloslavsky M, Van Der Laan J M. Fitting of mixtures with unspecified number of components using cross validation distance estimate[J]. Computational Statistics & Data Analysis, 2003, 41: 413-428
- [9] Wang H X, Luo B, Zhang Q B. Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm[J]. Pattern Recognition Letters, 2004, 25: 1799-1809
- [10] Schwarz G. Estimating the Dimension of a Model[J]. Annual Statistics, 1978(6): 461-464
- [11] Hansen M H, Yu B. Model selection and the principle of minimum description length [J]. Journal of the American Statistical Association, 2001, 96: 746-774
- [12] Paalanen P, Kamarainen J K, Ilonen J. Feature representation and discrimination based on Gaussian mixture model probability densities—Practices and algorithms [J]. Pattern Recognition, 2006, 39: 1346-1358
- [13] Bozdogan H. Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Information Measure of Complexity [M]. Dordrecht: Kluwer Academic Publishers, 1994

(上接第 224 页)

- [8] Hung W L. A note on entropy of intuitionsic fuzzy sets[J]. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 2003, 11(33): 627-633
- [9] 范平, 梁家荣, 李天志. Vague 集的新模糊熵[J]. 计算机工程与应用, 2007, 43(13): 179-181
- [10] 李凡, 徐章艳. Vague 集之间的相似度量[J]. 软件学报, 2001, 12(6): 922-927
- [11] 闫德勤. Vague 集的相似度量[J]. 计算机科学, 2006, 33(5): 195-196
- [12] 石玉强, 王鸿绪. 一种基于 Vague 集间距离的双向近似推理方法[J]. 小型微型计算机系统, 2007, 28(4): 661-665
- [13] Wang W Q, Xin X L. Distance measure between intuitionsic

fuzzy sets[J]. Pattern Recognition Letters, 2005, 26(13): 2063-2069

- [14] Li D F, Cheng C T. New similarity measures of intuitionsic fuzzy sets and application to pattern recognition[J]. Pattern Recognition Letters, 2002, 23(1): 221-225
- [15] 黄国顺, 刘云生. Vague 集相似度量及其在模式识别中的应用[J]. 复旦学报, 2004, 43(5): 869-873
- [16] 符海东, 卢正鼎. 基于 Vague 集距离的多评价指标模糊决策方法[J]. 华中科技大学学报, 2003, 31(8): 77-79
- [17] 要瑞璞, 沈惠璋. 基于 Vague 集投影及距离的模糊多指标决策[J]. 数学的实践与认识, 2009, 39(2): 19-22
- [18] 李凡, 徐章艳, 饶勇. Vague 集[J]. 计算机科学, 2000, 27(9): 12-14