

基于近似密度初始化的医学图像混合模型聚类

张树伟, 宋余庆, 陈健美, 谢从华

(江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

摘 要: 基于 EM(Expectation Maximization) 的混合模型聚类的效果与参数的初始值存在密切的关系. 提出了一种基于近似密度的 EM 参数初始化方法, 该方法用近似密度估计聚类样本点, 再根据每个聚类统计 EM 的混合比、均值、协方差参数的初始值. 并应用于人体腹部医学图像数据的高斯混合模型聚类分析, 实验结果表明该方法比 Kmeans 随机初始化方法有更好的聚类效果.

关键词: EM; 密度估计; 参数初始化; 模型聚类

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2010)09-0168-04

The Gaussian Mixture Model Clustering of Medical Image Based on Initialization of Approximate Density Function

ZHANG Shu-wei, SONG Yu-qing, CHEN Jian-mei, XIE Cong-hua

(School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhengjiang 212013, China)

Abstract: The performance of EM algorithm heavily depends on the initial values of the parameters in EM. In this paper, The approximate density function is adopted to initialize EM. The method estimate samples by the approximate density function and statistics mixture rate, mean value, and covariance. The application of these parameters in analysis of Gaussian Mixture Density Mode based on real human abdomen medical images and the results of experiments show that it can achieve better effect than Kmeans and random initialization.

Key words: EM algorithm; density estimation; initialization; clustering analysis

1 引言

聚类分析已成为模式识别和图像处理领域的一种关键技术. 基于模型的聚类方法是利用已知的数学模型, 通过逐渐逼近的方法, 使得给定数据集和数据模型之间达到最佳拟合^[1]. 在医学图像聚类分析中, 高斯混合模型是一个常用的数学模型, 通常是通过极大化似然函数对混合模型进行参数估计. 而 EM 算法是进行极大似然估计的一个有力工具, 主要用于不完全数据参数估计, 通过假设隐含变量的存在, 极大简化了似然函数方程, 解决了方程求解问题^[2].

EM 算法收敛的效果在很大程度上取决于其初

始参数, 运用 EM 算法实现高斯混合模型聚类, 如何初始化 EM 参数是一个关键问题. 国内外对 EM 的初始化主要的方法有随机初始化^[3]、层次聚类初始化、Kmeans 初始化^[4]. 随机初始化的优点是过程简单、操作方便, 但是太粗糙. 层次聚类初始化的显著缺点是时间效率低. Kmeans 初始化操作方便, 执行效率也较高, 但是结果不稳定. 为了克服上述几种方法的缺点, 针对医学图像语义难以表达的特点, 文中将近似密度函数应用于高斯混合模型聚类的 EM 算法的初始化.

2 高斯混合模型及核密度估计的相关工作

设 d 维数据空间 R^d 中的数据集 $X = \{x_1, x_2,$

$\dots, x_n\}$. 如果 X 的分布近似椭球体, 则可用单一高斯密度函数 $p(x; \mu, \Sigma)$ 描述这些数据的概率密度函数为

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (1)$$

其中 μ 为高斯分布的中心, 即均值, Σ 为高斯分布的协方差阵. 如果 X 的分布不符合椭球体, 则单一高斯分布密度函数无法正确描述这些数据的概率密度函数. 为此, 提出了多个单一高斯密度函数的加权平均表示这些数据点的分布, 由多个高斯密度函数的线性组合构成的函数叫做高斯混合密度模型 (Gaussian Mixture Density Model).

假定由 G 个高斯密度函数的混合来表示 X 的分布, 则高斯混合密度模型为

$$f(x, \theta) = \sum_{k=1}^G \alpha_k f_k(x, \theta_k) = \sum_{k=1}^G \alpha_k p(x; \mu_k, \Sigma_k) \quad (2)$$

其中 α_k 表示数据点属于第 k 个高斯分布的先验概率, 即混合密度模型的混合比例, 且满足 $\sum_{k=1}^G \alpha_k = 1$, $0 \leq \alpha_k \leq 1$; $p(x; \mu_k, \Sigma_k)$ 表示第 k 个高斯密度函数.

高斯混合模型的参数估计属于不完全数据的参数估计, 而 EM 算法是解决不完全数据参数估计的有效方法, 算法的每一次迭代都是由一个期望步 (E-step) 和极大化步 (M-step) 组成, 通过多步迭代, 使得似然值收敛于某个最优值.

设 Y 是观测数据, Z 是隐含变量, Φ 是待估计参数, 用 $p(\Phi | Y)$ 表示 Φ 的基于观测数据 Y 的后验分布密度函数. $p(\Phi | Y, Z)$ 表示添加数据 Z 后得到的关于 Φ 的后验分布密度函数, $p(Z | \Phi, Y)$ 表示在给定 Φ 和观测数据 Y 下潜在数据 Z 的条件分布密度函数.

假设 $\Phi^{(t)}$ 为第 t 次迭代得到的最大似然估计 Φ 的估计值, 则第 $t+1$ 次最大似然估计 Φ 的估计值由下面的两个迭代步得到:

E-step: 求 $p(\Phi | Y, Z)$ 或 $\ln p(\Phi | Y, Z)$ 关于 Z 的条件概率求期望, 即

$$\begin{aligned} Q(\Phi | \Phi^{(t)}, Y) &= E(\ln p(\Phi | Y, Z) | \Phi^{(t)}, Y) \\ &= \int \ln p(\Phi | Y, Z) p(Z | \Phi^{(t)}, Y) dZ \end{aligned} \quad (3)$$

M-step: 极大化 $Q(\Phi | \Phi^{(t)}, Y)$, 使 $Q(\Phi^{(t+1)} | \Phi^{(t)}, Y) = \max_{\Phi} Q(\Phi | \Phi^{(t)}, Y)$

重复执行上述的两个迭代步, 直到 Φ 值收敛为止, 即 $\|\Phi^{(t+1)} - \Phi^{(t)}\| \leq \epsilon$ (ϵ 是一个充分小的数).

上述 EM 迭代算法的初始值直接影响聚类效果, 在用 Kmeans 方法对 EM 进行初始化的高斯混合模型聚类结果会随着 Kmeans 中分类数 k 的取值不同而呈现出较大的差异, 并且语义表达效果不佳. EM 算法的初始化问题就是使用某一方法, 来获得 EM 迭代的初始值. 在高斯混合模型聚类中, 要得到的 EM 初始值是混合比 α_0 , 均值 μ_0 , 协方差 Σ_0 . 一般的初始化方法有随机中心、层次聚类、Kmeans 等. 在 EM 作为一种参数估计的方法提出以后, 多数情况下都是用随机中心和 Kmeans 初始化方法. 层次聚类初始化: 首先将 n 个样本分成 n 类, 使得每一类正好含有一个样本. 然后将样本凝集成 $n-1, n-2, \dots, k$ 类. 在模型聚类的方法里, 多采用最大似然标准作为凝集的准则. Kmeans 初始化则是一种基于距离的聚类, 之后对聚类结果进行统计得出初始参数.

3 基于核密度估计的 EM 初始化应用于混合模型聚类

核密度估计问题可以描述如下: 设 ξ 为分布特征未知的连续型随机变量, $\{X_i\}_{i=1,2,\dots,n}$ 为取自随机变量 ξ 的独立样本集合, 通过 $\{X_i\}_{i=1,2,\dots,n}$ 求分布密度函数 $f(X)$ 的近似估计 $\hat{f}(X)$.

定义 1 (核估计) 数据集 $D = \{X_1, \dots, X_n\} \subset F^d$ 是独立同分布随机变量集合, 变量所服从的分布密度函数为 $f(X)$, 定义函数:

$$\hat{f}_\sigma(X) = \frac{1}{n\sigma} \sum_{i=1}^n K\left|\frac{X_i - X}{\sigma}\right|, X \in R \quad (4)$$

称 $\hat{f}_\sigma(X)$ 为密度函数 $f(X)$ 的核估计. $K(\square)$ 称为核函数, σ 为预先给定的正数, 通常称为窗宽或光滑参数.

定义 2 影响函数和近似密度函数^[5-9]. 设数据集 $D = \{X_1, \dots, X_n\} \subset F^d$ 中一点 X_0 生成的影响函数 $f_B^{X_0}(X): F^d \rightarrow R_0^+$ 是由基函数 f_B 定义的如下函数:

$$f_B^{X_0}(X) = f_B(X_0, X), X \in F^d \quad (5)$$

且 D 所生成的空间上 F^d 的密度函数定义为

$$f_B^D(X) = \sum_{i=1}^N f_B^{X_i}(X), X \in F^d \quad (6)$$

分布密度函数 $f(X)$ 的核估计函数 $\hat{f}_o(X)$ 不仅与给定的样本点集合有关,还与核函数的选择和窗宽参数的选择有关.在理论上,任何函数均可以作为核函数,但为了密度函数估计的方便与合理性,通常要求核函数满足以下条件:

- (1) 对称性: $K(-u) = K(u)$
- (2) 有限性: $\sup |K(u)| < \infty$,

$$\int_{-\infty}^{\infty} K(u) du = 1$$

选用的核函数为高斯核函数时,密度函数可以表示为:

$$f_{\text{Gauss}}^D = \frac{1}{N\sigma^d(2\pi)^{\frac{d}{2}}} \sum_{i=1}^N \exp \left| -\frac{1}{2\sigma^2} (X - X_i)^T (X - X_i) \right| \quad (7)$$

为了提高运算速度,可以只计算其周围点对其影响函数之和,距离较远的点对其产生的影响较小,可忽略不计,于是近似密度函数可以表示为:

$$\hat{f}_{\text{Gauss}}^D(X) = \sum_{X_i \in \text{near}(X)} \exp \left| -\frac{d(X - X_i)^2}{2\sigma^2} \right| \quad (8)$$

医学图像中含有大量的噪声,所以数据点分布的规律性比较差,如果仅靠统计一个点周围点的个数作为密度,其可靠性较差.在上述定义的基础上,采用高斯核定义的密度来减少噪声点的影响,另外,医学图像是一个二维平面,每一个数据点的属性都有坐标值 (x, y) 和灰度值 (gray) . near 函数为 $\text{near}(X) = \{X^* / d(X, X^*) < k\sigma, k \in \mathbb{Z}\}$, 实验表明,取 $k = 4$ 足够表示相邻数据的影响.综上所述,医学图像数据的近似密度函数可构造为:

$$\hat{f}_{\text{Gauss}}^D(X) = \sum_{X_i \in \{X^* / d(X, X^*) < 4\sigma\}} \exp((-d(X(x, y, \text{gray}), X_1(x, y, \text{gray})))^2 / 2\sigma^2) \quad (9)$$

定义 3 (密度吸引子) F^d 中的点 X^* 称为局部吸引子,当且仅当近似密度函数 $\hat{f}_{\text{Gauss}}^D(X)$ 在 X^* 处取得局部极大值.数据点 $X \in D$ 称为被密度吸引子 X^* 吸引的,如果存在 $X = X_0, X_1, \dots, X_k$, $d(X_k, X^*) \leq \epsilon$ 使得:

$$X_i = X_{i-1} + \delta \frac{\Delta f(X_{i-1})}{\|\Delta f(X_{i-1})\|}, i = 1, 2, \dots, k \quad (10)$$

其中, $\Delta \hat{f}_{\text{Gauss}}^D(p)$ 称为近似密度函数在点 X 处的梯度.

定义 $\hat{f}_{\text{Gauss}}^D(p)$ 的数据点的近似密度函数为图像聚类特征值.

Kmeans 是一种单纯的基于距离的聚类方法,这

就使得它在解决特殊问题时遇到困难.医学图像的像素点之间的关系是比较复杂的,而不只是单纯的距离关系,所以, Kmeans 在解决医学图像的聚类中有一定的局限性,由此文中提出运用近似密度估计法来进行医学图像高斯混合模型聚类的 EM 算法初始化,用爬山算法提取出聚类模式^[10]. 近似密度函数能够较好的表达医学图像像素点之间的关系,这在临床诊断中有非常重要的意义.

初始化 EM 的任务就是找到最优的聚类中心,这个问题可看作近似密度估计问题,根据原始数据集,最好的聚类中心可能就是概率密度函数最稠密的部分.于是利用近似密度方法来找到数据最稠密的部分,其初始化过程的算法描述如下:

输入:医学图像 G , 阈值 δ, ξ

输出:聚类结果

- ① read a medical image G ;
- ② preprocess G ;
- ③ $p = \text{select samples from } G \text{ by random sampling}$;
- ④ Compute density $\hat{f}_{\text{Gauss}}^D(p)$;
- ⑤ Compute gradient $\Delta \hat{f}_{\text{Gauss}}^D(p)$;
- ⑥ Search for attractor p^* of p by hill-climbing scheme directed by gradient at p ;
- ⑦ if $\hat{f}_{\text{Gauss}}^D(X^*) > \xi$ do begin
- ⑧ mark p, p^* and all points on the way of searching to cluster attracted by p^* ;
- ⑨ write x, y and to $\hat{f}_{\text{Gauss}}^D(X)$ cluster feature database;
- ⑩ end
- ⑪ else
- ⑫ mark all point from ptop^* as no-cluster;
- ⑬ for each marked point in G do begin
- ⑭ replace related pixel attribute in G by mark pixel;
- ⑮ rebuilt image G' ;
- ⑯ end
- ⑰ end
- ⑱ $\pi = \text{classdata} / \text{total samples}$;
- ⑲ compute classmean σ
- ⑳ compute classdata covariance Σ ;
- ㉑ if $(\sigma > 0 \& \& \Sigma > 0 \& \& \pi > 0)$ do EM algorithm;
- ㉒ save finalrate finalmean and finalcovariance;
- ㉓ clustering medical image by Gaussian mixture

model;

②4 show clustering result.

在本算法中,密度聚类的效果由 δ, ξ 加以控制,用密度聚类对样本进行聚类并计算出 EM 算法的初始值后再用高斯混合模型对整幅图像进行聚类分析,聚类结果较稳定,并具有良好的语义效果.

4 实验结果和性能评估

为了比较 Kmeans 法和近似密度法对 EM 算法初始化的效果,分别将两种方法初始化后的参数应用于人体腹部医学图像的高斯混合模型聚类.文中重点对人体肿瘤发病率较高的肝脏和肾脏两个腹部器官进行聚类分析首先对图像进行预处理并抽取 2500 个图像样本.表 1 为使用不同方法进行 EM 初始化后用于聚类的两个参数比较.

表 1 使用不同方法的比较

初始化方法	脏器	正确率	误判率
Kmeans	肝脏	82.46	2.93
	肾脏	75.76	0.74
近似密度	肝脏	90.66	1.98
	肾脏	92.56	0.45

用两种初始化 EM 的方法对一幅人湾腹部医学图像进行混合模型聚类,并得到了一组正确率和误判率的参数比较.从实验结果可以看出,用近似密度法初始化 EM 得到的正确率要大于 Kmeans 法,错误率小于 Kmeans 方法,达到了更好的聚类效果.

5 结束语

文中针对人体腹部医学 CT 图像,提出了用近似密度估计法来初始化 EM,并将初始化结果应用于高斯混合模型聚类分析,由于近似密度估计能较好地表达医学图像的内容,因而得到了较好的聚类结果.采用近似密度方法初始化 EM,提高了高斯混合模型聚类分析的准确性,有深刻的临床意义.

参考文献:

[1] 罗述谦,周果宏.医学图像处理与分析[M].北京:科学出版社,2003.

[2] Fraley C, Raftery A E. How many clusters? Which clustering method? - Answers via model - based cluster analysis[J]. The Computer Journal,1998(41):578 - 588.

[3] Jeff C F WU. On the convergence properties of the EM algorithm[J]. The Annals of Statistics, 1983,11(1):95 - 103.

[4] Christophe Biernacki. Initializing EM using the properites of its trajectories in gaussian mixtures[J]. Statistics and Computing, 2005,14(3):267 - 269.

[5] Biernacki C, Celeux G, Govaert G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models[J]. Computational Statistics and Data analysis, 2002,24(5):603 - 619.

[6] Hinneburg A, Keim D. An efficient approach to clustering in large multimedia databases with noise[C]// Proc of the 1998 4th Int,l Conf on Knowledge Discovery and Data mining. New York: AAAI Press,1998:58 - 65.

[7]李艳玲,王加俊.基于高斯混合模型的纹理图像的分割[J].微电子学与计算机,2004,21(4):16 - 19.

[8] 苏锦旗,薛惠锋,詹海亮.基于划分的 K - 均值初始聚类中心优化算法[J].微电子学与计算机,2009,26(1):8 - 11.

[9] 宋余庆.医学图像数据挖掘若干技术研究[D].镇江:江苏大学,2007.

[10] 谢从华,朱峰,王立军,等.基于密度聚类的医学图像分割[J]. 计算机应用研究,2007(2):167 - 169.

作者简介:

张树伟 男,(1982 -),硕士研究生.研究方向为医学图像处理.

宋余庆 男,(1959 -),博士,教授.研究方向为数据挖掘、知识发现、图像数据库系统、模式识别等.

陈健美 女,(1962 -),博士,教授.研究方向为数据挖掘、医学图像数据库.

谢从华 男,(1978 -),博士研究生,讲师.研究方向为医学图像处理、数据挖掘、模式识别.

(上接第 167 页)

参考文献:

[1] 郑人杰.软件工程:高级[M].北京:清华大学出版社,2002.

[2] 称华军.J2EE 构件企业级应用解决方案 [M].北京:清华大学出版社,2008.

[3] 贾小云,陈桦,奚宗悦.企业应用程序中会话 Bean 的设计[J].微电子学与计算机,2004,21(8):52 - 55.

[4] 胡滨.自动化测试工具的研究[J].现代电子技术,2007(8):97 - 99.

[5] Roman ED. Mastering EJB[M]. New York, Wiley Longmen, 2002.

作者简介:

贾小云 女,(1971 -),教师,副教授.研究方向为软件工程.