文章编号:1008-0570(2006)11-3-0244-03

高斯混合模型聚类中 EM 算法及初始化的研究

Algorithm EM and Its Initialization in Gaussian-Mixture-Model Based Clustering

(江南大学)岳　佳[1]　王士同[2]
YUE,JIA　WANG SHITONG

摘要:EM 算法是参数估计的重要方法,其算法核心是根据已有的数据来迭代计算似然函数,使之收敛于某个最优值。EM 算法收敛的优劣很大程度上取决于其初始参数。运用 EM 算法来实现高斯混合模型聚类,如何初始化 EM 参数便成为一个关键的问题。在比较其他的初始化方法的基础上,引入"binning"法来初始化 EM。实验结果表明,应用 binning 法来初始化 EM 的高斯混合模型聚类优于其它传统的初始化方法。

关键词:极大似然;高斯混合模型;EM 算法;初始化;聚类分析
中图分类号:TP181　　文献标识码:A

技术创新

Abstract:EM algorithm is an important method of parameter estimation. Its core idea is to literately compute the likelihood function until it converges to some optimal value for the given data, thus its performance heavily depends on the initial values of the parameters in EM. When EM is utilized to realize Gaussian-Mixture-Model based clustering, how to initialize it becomes a pivotal issue. In this paper, the binning method is adopted to initialize EM on the base of comparison other methods. Our experimental results demonstrate that Gaussian-mixture-model based clustering using EM with the binning method based initialization outperforms those with other classical initialization methods.

Keywords:maximum likelihood;gaussian mixture model;EM algorithm;initialization;clustering analysis

1 引言

聚类分析又称为数据分割,需要把一个数据对象分组,使得每个组内部对象之间的相关性比与其他组对象之间的相关性更加紧密。基于模型的聚类方法就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成最佳拟和。高斯混合模型是我们常用的一个数学模型,它是聚类,模式识别以及多元密度估计的一个有力的框架。

通常,我们是通过极大化似然函数对混合模型进行参数估计的。而对于混合模型进行极大似然估计的一个很好的工具就是 EM 算法。EM 算法又称期望最大(Expectation Maximization)算法,最初由 Dempaster,Laird 和 Rubin 提出的。它是一种在观测数据为不完全数据时求解极大似然估计的迭代算法。大大降低了极大似然估计的计算复杂度。

然而,EM 算法有个显著的缺陷,就是收敛速度比较慢,有时会收敛于局部最小值,而不能得到全局最优解,使得聚类效果受到影响。近年来,国外很多人在改进 EM 的收敛方面提出了不少新的思路和策略。同样,我们也可以在保持 EM 算法本身迭代的简单性的前提下,通过细化 EM 初始值的方法,来改善 EM 的收敛,从而获得更好的聚类效果。

国外对 EM 的初始化已有一些研究,本文将密度估计的 binning 法应用于高斯混合模型聚类的 EM 算法的初始化。比较了 binning 法与随机中心法,层次聚类法,Kmeans 法的实际操作性能,以及在最终聚类效果上的差别。

2 EM 算法

EM 算法是在观察数据为具有隐含变量(即为不完全数据)时,对观测数据进行极大似然估计,通过多步迭代,使得似然值收敛于某个最优值的迭代算法。

2.1 高斯混合模型

假设有一系列观测值由某混合分布 P 产生,该分布又是由 G 个成分构成,每一个成分都代表一个不同的类别(cluster)。假设观测样本 $X=(X_1,.....,X_n)$,每个向量 X_i 都是 P 维的。$f_k(X_i|\theta_k)$表示 X_i 是第 k 类的密度函数,θ_k 是相应的参数。π_k 表示某一观察值属于第 k 类的概率,即权重。最大化混合似然函数:

$$L_M(\theta_1,.....,\theta_G;\gamma_1,.....,\gamma_n|x)=\prod_{i=1}^{n}\sum_{k=1}^{G}\pi_k f_k(x_i|\theta_k)\qquad(\pi_k\geq 0;\sum_{k=1}^{G}\pi_k=1)\quad(1)$$

如果 $f_k(X_i|\theta_k)$是多元正态分布,即高斯分布,则此混合聚类的模型即为高斯混合模型(GMM),G 个成分即 G 个独立的高斯分布。参数 θ_k 由均值 μ_k 和协方差矩阵 Σ_k 组成。密度函数 $f_k(X_i|\theta_k)$如下:

岳佳:硕士研究生
基金项目:模式识别国家重点实验室开放课题

$$f_k(x_i \mid \mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\}}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \quad (2)$$

该分布 P 可由 G 个高斯密度函数的加权平均所表示的概率密度函数描述如下：

$$P(x \mid \theta) = \sum_{k=1}^{G} \pi_k f_k(x_i \mid \mu_k, \Sigma_k) \quad (3)$$

2.2 聚类的 EM 算法

假设存在一个完整数据集 Y=(X,Z)，X={x_1,.....,x_n}是不完整的数据集，Z_i 是引入的隐含变量。$Z_i \in \{1,2,.....,M\}$，M 是给定的有限整数。于是 Y={(x_1,z_1),.....,(x_n,z_n)}则完整数据的似然函数为：

$$L(\theta \mid X, Z) = p(X, Z \mid \theta) = \prod_{i=1}^{n} p(x_i, z_i \mid \theta) \qquad Z = \{z_1, \ldots\ldots, z_n\} \quad (4)$$

该似然函数的期望值：

$$E(L(\theta \mid X, Z)) = \int_z p(X, Z \mid \theta) f(Z) d_z \quad (5)$$

采用 EM 算法的基本思想是对于上述的不完整数据集 Y，假设这些数据独立同分布于我们已知的某一个模型，如 GMM，而我们知道该模型的参数，因此可以根据该模型推出属于每个成分的各数据点的概率，然后修改每个成分的值，重复该过程直到收敛到结束条件。

E-step：

$$Q(\theta, \theta^{(i-1)}) = E(\log L(\theta \mid X, Z)) = \int_z \log L(\theta \mid X, Z) f(Z \mid, \theta^{(i-1)}) d_z \quad (6)$$

显然，辅助函数 $Q(\theta,\theta^{(i-1)})$的值就是 log(L(θ|X,Z))的期望值，并且是 θ 的函数，$\theta^{(i-1)}$是上一步迭代运算求得的参数值。

M-step:

$$\theta^* = \theta^{(i)} = \arg\max Q(\theta, \theta^{(i-1)}) \quad (7)$$

求解 θ^*，使得 $Q(\theta^*,\theta^{(i-1)})$得到极大值。可以看出，随机向量 Z 的分布是由 X 和 $\theta^{(i-1)}$决定的，若 θ^*_i 表示第 i 次迭代的最大似然函数值，θ^*_{i-1} 表示第 i-1 次迭代的最大似然函数值，可知证明，EM 算法能够保证 $\theta^*_i \geq \theta^*_{i-1}$，并且算法是收敛的。

高斯混合模型(GMM)里，假设完整数据为 $y_i=(x_i,z_i)$，x_i 为可观测变量，z_i 为隐含变量，$z_i=(z_{i1}.....z_{iG})$

$$Z_{ik} = \begin{cases} 1 & if\ \ x_i\ \ belongs\ \ to\ \ group\ \ k \\ 0 & otherwise \end{cases} \quad (8)$$

设 z_i 是独立同分布于 G 类，其概率分别为 π_1,.....,π_G，并且由 x_i 给出的 z_i 的密度为：$\prod_{k=1}^{G} f_k(x_i \mid \theta_t)^{Z_{ik}}$

完整数据的 log 似然函数为：

$$L(\theta_k, \pi_k, z_{ik} \mid x) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}[\log \pi_k f_k(x_i \mid \theta_k)] \quad (9)$$

算法是在 E-step 和 M-step 之间迭代。在 E-step，由可观测变量 x 和当前的参数估计，计算出完整数据 log 似然的条件期望值 z_{ik}。M-step 中，根据 E-step 的值，计算使得 log 似然函数值最大的参数估计。

3 初始化方法

EM 算法的初始化就是使用某一算法，来获得 EM 迭代的初始值。在高斯混合模型聚类中，我们要得到的 EM 初始值就是参数 权重 π_0，均值 μ_0，协方差 Σ_0。

3.1 一般的初始化

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，多数情况下我们都是使用随机初始化方法。即 在我们的数据集里随机抽取 n 个点作为聚类中心，n 为聚类(类别)数。然后，对数据集里的每个数据点(除了这 n 个点以外)，计算其与这 n 个点的每个点的欧氏距离，根据距离最短原则，把每个点放入 n 类中的某一类。于是，我们对于每一类的数据，可以计算出其权重，均值，协方差。层次聚类初始化，即 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 类，n-2 类，直到所有的样本都凝集成我们所需要的 k 类为止。在模型聚类的方法里，我们多采用最大似然标准作为凝集的准则。Kmeans聚类即 K 均值聚类，属于聚类分析方法中一种基本的且应用最广泛的划分算法。目标就是要找到 k 个均值向量，k 就是我们的聚类数目。基于给定的聚类目标函数(或者说是聚类效果判别准则)，算法采用迭代更新的方法，每一次迭代过程都是向目标函数值减小的方向进行，最终的聚类结果使目标函数值取得极小值，达到较优的聚类效果。

3.2 Binning 初始化

Binning 法中文意思是装箱法，可以想象成把数据空间在各维上划分成一个个的箱子，再把数据点投射到对应的箱子里去。在统计学里，它是用于密度估计的一种方法。

在这里，初始化 EM 的任务就是找到最优聚类中心。我们可以把这个问题视为密度估计的问题。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，可以通过 binning 法来寻找概率密度函数最稠密的部分。我们根据一定的 bin 宽，将每一维上的整个数据空间分成若干个 bin，然后把每个数据的每一维放到对应的 bin 里面，再计算每一个 bin 里所含的数据点的个数。含的点数多的则为概率密度相对大的区域。也就是聚类中心最可能存在的区域。

在得到了每一个 bin 里的数据点的个数后，可以作进一步的优化：

1 给每一个数据点一个向量标记，表示其各维所在的 bin 位置。

2. 计算出 bin 里数据点的均值。把所有小于均值的 bin 去除考虑范围。

3 按照 bin 的数目，大致估计聚类中心的位置。

4 按照相似度标准和已估计的聚类中心，针对每一个数据，比较其相似度。

5 根据相似度重新排列数据点，相似度高的为同一类。

6 给相同的类的点以相同的类标。

这里的相似度标准是按照每一个向量标记，用相同

技术创新

位置上且值也相同的分量的个数除以总的维数。

Binning法的一个关键点就是找到每一维上最优或者近似最优的 bin 宽。文献给出了高斯分布的一元数据的最优 bin 宽。应用里的推导,我们假设:每一维上的数据都近似一个高斯分布的概率密度函数。对于和高斯分布相差很大的每一维的数据,里给出的 bin 宽可能只能只是一个次优值。

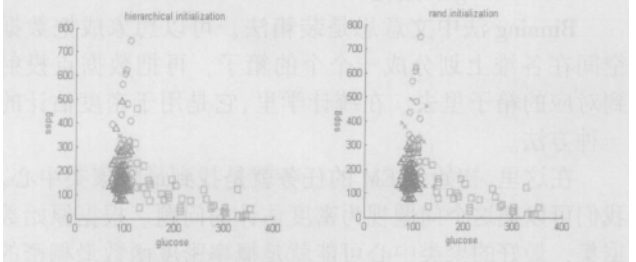
4 实验

基于模型的聚类方法有了很广泛的实际应用。包括字符识别,组织细胞分割,纺织品瑕疵的鉴别,医疗数据分析以及大量数据的分类等等。在这里,我们通过医疗诊断数据集(diabetes)和植物开花数据集(Iris data)来实现基于模型的 EM 算法。比较随机中心,层次聚类,Kmeans 和 binning 四种初始化 EM 算法的操作性能和对聚类效果的影响。

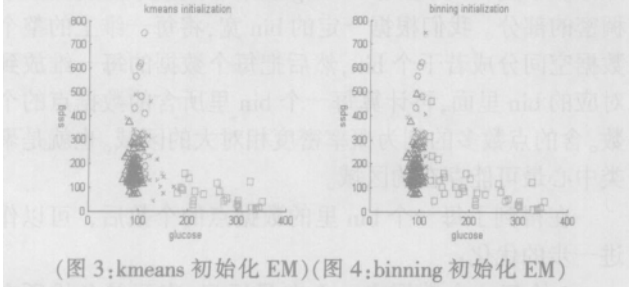
4.1 实验一

使用 diabetes data。数据集包含 145 个 3 维的数据点。数据集在临床上被分为 3 类。把这三维的数据点用三个二维的平面图形来表示,(1,2),(1,3)和(2,3)。选择(1,3)作为代表进行分析

使用不同的初始化 EM 方法的聚类结果如下图表示,'叉'表示错误的分类点。



(图 1:层次聚类初始化 EM)(图 2:随机中心初始化 EM)



(图 3:kmeans 初始化 EM)(图 4:binning 初始化 EM)

表 1　初始化方法对 diabetes 聚类效果,错误率的比较

初始化方法	Normal	Chemical	Overt	错误率(百分比)
随机	73	25	29	12.4138
层次聚类	70	30	33	8.2759
Kmeans	75	21	21	19.3103
Binning	73	26	33	8.9655

4.2 实验二

使用 Iris data。数据集包含 150 个 4 维的数据点,3 类,每一类有 50 个数据点,通过四种不同的初始化 EM 的算法进行对聚类效果进行比较。实验结果如下:

表 2　初始化方法对 Iris data 聚类效果,错误率的比较

初始化方法	Class-1	Class-2	Class-3	错误率(百分比)
随机	47	43	46	9.3333
层次聚类	49	44	47	6.6667
Kmeans	50	44	50	4.0000
Binning	49	46	48	4.6667

4.3 实验分析

使用了四种初始化 EM 的方法对两组数据集做了分析。得到了每一类数据可正确聚类的数目和总体的聚类的错误率。

随机中心初始化操作方便,但是由于随机性,实际效果有较大的偏差,这里取的是多次情况下最好的结果。层次聚类初始化可以聚类到我们指定的类别数,但是迭代的次数较多。在数据规模较大的情况下,效率很低,即耗时和需要较大的存储空间。Kmeans 初始化很常见,但是结果不稳定,每次实验结果不完全一致。Binning 法相对操作方便,也达到了较好的聚类效果。

5 结束语

本文作者的创新点在于:针对 EM 算法在收敛上的缺陷,在保持算法迭代的简单性的前提下,通过改变赋 EM 的初始值的方法,来优化 EM 算法,达到更好的聚类效果。使用的 binning 法作为初始化 EM 的一种方法,比较了传统的初始化方法,实验表明,该方法有着较高的可操作性和较好的实际聚类效果。

Binning 法为 EM 算法的初始化提供了很好的可行性措施,即 EM 算法的一种优化。尤其可以推广到对高维数据的聚类。该方法的一个关键的问题就是 bin 宽的选择。在本文的实验里,我们用的是里的推导。它对于维上非高斯分布的数据只是一个次优解。如何选择一个更好的 bin 宽将是我们以后的研究内容。

参考文献:

[1]汤晓琴 数据挖掘中聚类分析的技术方法[J].微计算机信息.2003,1:23-27

[2]Dempster, A.P, Laird, N.M, Rubin, D.B. Maximum likelihood for incomplete data via the EM algorithm.[J] .J.R. Stat. Soc,1977, B,39:1-38.

[3]Liu C, Sun D X. Acceleration of EM Algorithm for Mixtures Models using ECME[J]. ASA Proceedings of the Stat. Comp. Session, 1997, 109-114.

[4]Christophe Biernacki.Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures [J]. Statistics and Computing, 2004, 14, 3:267-279.

[5]Patricia McKenzie, Michael Alder. Initializing the EM Algorithm for use in Gaussian Mixture Modelling [J]. Amsterdam Esevier Science BV, 1994:91-105. (下转第 302 页)

点,像素的坐标作为力臂,从而以各阶矩的形式来表示区域特征。设图像各像素的质量为 1,即 1—像素的质量就等于它的像素值;S 为图形面积;i,j为图形内像素坐标。矩的公式可表示为:

$$M(p,q)=\sum_{(i,j)\in s} i^p j^q f(i,j)$$

式中,M 为 p,q 值下的图形的矩;p=0,1,2,…,q=0,1,2…,f(i,j)相当于一个像素质量,当 p,q 取值不同,可得阶数不同的矩。若 p=0,q=0 时:

$$M(0,0)=\sum_{(i,j)\in s} f(i,j)$$

即为图像中 1-像素之和,也就是图像的面积。

利用函数:[Y,bw]=bwlabel(G)将图 4 中的白色斑点进行标注,可以区分出多块白色区域。根据实验统计数据,在本段时期单条鱼病变或死亡时所暴露的白色区域面积为 M(0,0)=1200±200,将用标注好的多块白色区域中最小区域分别与值下限和上限比较,若值落在两者之间,则出现鱼出现不适或死亡现象,系统报警,否则正常。

4 结论

本文创新点是利用计算机数字图像处理技术,应用 MATLAB 软件强大的图像处理功能,实时监视池塘现场的情况,经过计算机的处理,分析现场情况,使鱼类生长情况始终处于控制之中。还可以建立监控画面,进行实时视频监控,可以早期发现鱼类生长过程中的鱼变等变异情况。利用上述图像处理的方法对鱼塘进行监控,具有快速、简易、准确等优点,程序编制简单,程序代码少,而且可以方便地嵌入到 VC++系统,对水产养殖的环境进行监控,满足监控现场更高的要求,能有效地监视鱼类的生长情况,具有良好的应用前景。

本文作者创新点:介绍了一种可以对水产养殖中鱼的病变情况进行监控的系统。通过 MATLAB 软件中强大的图像处理功能,对读入的图像进行中值滤波、灰度处理和二值处理,然后根据收缩与扩张的原理,编写程序,对二值图像进行数次收缩与扩张,以去除杂质点,最后采用区域矩特征的概念,统计出白色像素的数目,与统计数据相比较,即可判断出是否出现鱼类病变等异常情况。利用上述图像处理的方法对鱼塘进行监控,具有快速、简易、准确等优点,程序编制简单,程序代码少,而且可以方便地嵌入到 VC++系统,对水产养殖的环境进行监控,满足监控现场更高的要求,能有效地监视鱼类的生长情况,还可以建立监控画面,进行实时视频监控,可以早期发现鱼类生长过程中的鱼变等变异情况,具有良好的应用前景。

参考文献:

[1]孙兆林 MATLAB6.X 图像处理[M]北京 清华大学出版社 2002 142-173 213-226

[2]张兆礼,赵春晖,梅晓丹 现代图像处理技术及 MATLAB 实现[M]北京 人民与邮电出版社 2001.83-122 240-260

[3]李了了,邓善熙 MATLAB 在图像处理技术方面的应用[J]微计算机信息,2003,2:65-67

[4]Maragos,P.;Differential morphology and image processing. Image Processing,IEEE Transactions on Volume 5,Issue 6,June 1996 Page(s):922 - 937

[5]飞思科技产品研发中心 MATLAB6.5 辅助图像处理[M]北京 电子工业出版社 2003 189-202

[6]何强 何英 MATLAB 扩展编程[M]北京 清华大学出版社 2002 159-230

作者简介:刘星桥(1960-),男,副教授,博士研究生,主要研究方向为农业电气化与自动化,图像处理。邮箱:xqliu@ujs.edu.cn;李娟(1981-),女,硕士研究生,主要研究方向为自动控制系统,图像处理。E-mail:coco9303@163.com。

Biography:Liu Xing-qiao (1960-),male,adjunct professor,doctor,research direction:agricultural electrization and automation,image processing.Email:xqliu@ujs.edu.cn.

(212013 镇江市江苏大学电气信息工程学院)刘星桥 李娟 张如通

(School of Electrical and Information Engineering,Jiangsu university,Zhenjiang,Jiangsu,212013)Liu Xingqiao Li Juan Zhang Rutong

通讯地址:(212013 镇江市江苏大学电气信息工程学院)刘星桥 李娟

(收稿日期 2006.4.28)(修稿日期 2006.5.26)

(上接第 246 页)

[6]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.

[7]Banfield J. D,Raftery A. E. Model-based Gaussian and non-Gaussian clustering [J]. Biometrics,1993,49:803-821.

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

[9]D W Scott. On optimal and data-based histograms [J]. Biometrika,1979,66:605-610.

[10]Fraley C Algorithms for model-based Gaussian hierarchical clustering [J]. SIAM J Sci Computer,1999,20:270-281.

作者简介:岳佳(1981-)女,硕士研究生,主要研究方向为模式识别 E-mail:jiajia510weiyi@yahoo.com.cn;王士同(1964-)男,教授,博士生导师,主要研究方向为人工智能、模式识别、生物信息学。

(214122 江苏无锡 江南大学信息工程学院)岳佳 王士同

(School of Information Technology,Southern Yangtze University,Wuxi 214122,China)Yue Jia Wang Shitong

通讯地址:(214122 江苏省无锡市 蠡湖大道 1800 号江南大学蠡湖校区桂园公寓 8 号楼 509 室)岳佳

(收稿日期 2006.4.28)(修稿日期 2006.5.26)