|  |  |
| --- | --- |
| 成 绩 |  |
| 评卷人 |  |

|  |  |
| --- | --- |
| 姓 名 | 胡月 |
| 学 号 | 2017113025 |

华 中 师 范 大 学

研 究 生 课 程 论 文

论文题目 EM算法在高斯混合模型聚类中的应用

完成时间 2018/01/15

课程名称 统计学习理论

专 业 计算机应用技术

年 级 2017级

School of Computer Since, Central China Normal University E-mail:huyue2017@mails.ccnu.edu.cn

©华中师范大学 计算机学院所有. Tel: +86-15527634383

# EM算法在高斯混合模型聚类中的应用

胡 月

（华中师范大学计算机学院,武汉 湖北）

通讯作者: 胡月, E-mail:huyue2017@ccnu.edu.cn

**摘 要:** 聚类分析在客户细分中十分重要.我们平时常用的有三类聚类模型,K-means聚类,层次（系统）聚类,最大期望EM算法.相比于其他两种算法来说,EM模型是基于概率的且能够拟合多种形状的类的聚类方法,因此常被用作工具算法.高斯混合模型（GMM）聚类中,利用概率密度假设,使EM算法进行迭代求解参数的极大似然估计,便可以将无监督学习问题转化为概率论的经典参数估计问题.

**关键词:** 聚类分析,高斯混合模型,EM算法,极大似然

**EM Algorithm in Gaussian Mixture Model Clustering**

Hu Yue

(School of Computer Since, Central China Normal University, Wu Han, China)

**Abstract:** Clustering analysis is extremely important in customer segmentation. There are three types of cluster models commonly used, K-means clustering, hierarchical (system) clustering, the maximum expected EM algorithm. Comparing to the other two algorithms, EM model is based on probability and can fit a variety of shapes of the clustering method, so it is often used as a tool algorithm. In Gaussian Mixture Model (GMM) clustering, the probability density hypothesis is used to make the EM algorithm carry out the maximum likelihood estimation of the iterative solution parameters, so that the unsupervised learning problem can be transformed into the classical parameter estimation problem of probability theory.

**Key words**: clustering analysis, Gaussian Mixture Model, EM algorithm, maximum likelihood

# 1 引言

聚类分析（cluster analysis）是一组将研究对象分为相对同质的群组的统计分析技术.聚类分析与分类分析的区别在于它可以是无监督的学习.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.利用已有的数学模型,通过逐渐逼近的方法,使给定数据集和数据模型之间达成最佳拟合的方法称为基于模型的聚类方法.高斯混合模型就是我们常用的一个数学模型,在聚类,模式识别以及多元密度估计中都将其作为一个有力的框架.

通常,我们对混合模型进行参数估计时都是利用极大化似然函数,而对于混合模型进行的极大似然估计的一个很好的工具就是EM算法.EM算法是一种迭代算法,又称为期望最大（Expectation Maximization）算法,由Dempster等人总结提出.它用于含有隐变量的概率模型参数的极大似然估计,并大大降低极大似然估计的计算复杂度.在聚类中使用期望最大化算法,本质是将数据点按照所选数量的簇进行分类,这个数量和预期生成的不同分布的数量是相同的,而且分布也必须是相同的类型.

然而,EM算法有个显著的缺陷,就是收敛速度比较慢,当优化的不是凸函数时,会收敛于局部最小值,而不能得到全局最优解,这会使得聚类效果受到影响.而且,EM算法对初始值十分敏感,所以确定初始值也十分重要.本文采用了K-means的初始化方法进行实验。

# 2 高斯混合模型及EM算法

EM算法是自收敛的分类算法,既不需要事先设定类别,也不需要数据间的两两比较合并等操作.在观察数据为具有隐含变量（即不完全数据）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

高斯混合模型（GMM）是指多个高斯分布函数的线性组合,理论上GMM可拟合出任意类型的分布,通常用于解决同一集合下的数据包含不同分布的情况.高斯混合模型基于这样一个概率密度假设:所有的数据点X={x1 , x2 ,…, xn}是从同一个概率密度函数中独立抽样得来的.这个概率密度函数为几个高斯函数的线性组合[1]: (1)

其中,加权系数需要满足:

 (2)

将可以这里的看成是数据点产生于第i个高斯成员的先验概率,也可看做由多个高斯密度的加权平均所表示的密度函数.(1)式中为高维高斯函数,表达式为:

(3)

构建公式(1)中的概率密度函数后,接下来就由观测的数据点求混合模型中的参数:.其中,.我们用EM算法求解θ的极大似然值.

这样就可以将无监督学习的问题转化为概率论中经典的参数估计问题.对参数的估计,一般采用极大似然估计方法.

## 2.2 聚类的EM算法

EM算法的基本出发点是假设数据集是完整的[2].这里将完整的数据集记为{X,Y},即假设存在所谓的标记数据为,其中是一个2值化向量且只有一个元素为1,其余都为0.即如果第i个数据点是由GMM模型中第p个高斯成员产生的那么.其中q=1,…,K,.然后基于假设得到的完整数据集建立辅助Q函数（E-step）,实际上是求基于完整数据集的似然函数关于隐藏变量Y的条件期望.完整数据的似然函数为:

 (4)

该似然函数的期望值:

 (5)

采用EM算法的基本思想是对上述的不完整数据集{X,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:

 (6)

显然,辅助函数的值就是的期望值,并且是θ的函数,是上一步迭代运算求得的参数值.

1. step:

迭代M步是求函数对θ的极大值,即求新一轮迭代的模型参数:

 (7)

可以看出随机向量Y的分布是由X和决定的,若表示第i次迭代的最大似然函数值,表示第i-1次迭代的最大似然函数值,可以证明,EM算法能够保证,并且算法是收敛的[2].

高斯混合模型里,假设完整数据为为可观测变量,为隐含变量【4】,

 (8)

设是独立同分布于G类,其概率分别为,并且由给出的密度为:

完整数据的log似然函数为:

 (9)

算法在E-step和M-step之间迭代计算出,直到对数似然函数值不再有明显的变化为止.

将估计高斯混合模型参数的EM算法总结如下:

输入:观测数据,高斯混合模型；

输出:高斯混合模型参数

1. 初始EM,利用K-means找到最优聚类中心,取优化后的参数的初始值开始迭代[3]；
2. 确定想要的高斯模型,并根据第(1)步生成数据
3. E步:求期望.依据当前模型参数,计算分模型K对观测数据的响应度；
4. M步:求极大.求Q函数对θ的极大值,即计算新一轮迭代的模型参数:
5. 重复第(2)步和第(3)步,直到收敛,即对数似然函数值θ不再有明显的变化为止.

# 3 初始化方法

通常以K-means算法和GMM的EM解法相比较,发现二者具有很强的相似性.K-means算法对数据点的聚类进行了"硬分配",即每个数据点只属于唯一的聚类[5].而GMM的EM解法则基于后验概率分布,对数据点进行"软分配",即每个单独的高斯模型对数据聚类都有贡献价值,不过贡献价值并不对等.从解法可以看到,K-means算法其实是GMM的EM解法在高斯分量协方差时的一个特例.实际应用中,K-means通常需要重复一定次数后取最好的结果,但由于GMM每一次迭代的计算量比K-means大许多,使用GMM时,更多的是用K-means得到一个粗略的结果,然后将其作为初值,即将K-means所得的聚类中心传给GMM,再用GMM进行细致迭代[6].流程具体如下:

1. 初始化:为选取初值；
2. 固定,关于最小化目标函数；
3. 固定各数据点类别,关于最小化目标函数,K-Means迭代步骤与EM算法的关系:（2）（3）中分别更新分别对应EM算法中的E步和M步.

# 4 实证研究

本文以2017年全球178个国家的总值GDP和人均GDP为例,数据来源于世界经济信息网.抽取前10的具体数据如表1所示:

表 1 2017年全球GDP及人均GDP

|  |  |  |
| --- | --- | --- |
| 国家 | GDP（亿美元） | 人均GDP（美元） |
| 美国 | 195558.74 | 60014.895 |
| 中国 | 131735.85 | 9481.881 |
| 日本 | 43421.6 | 34486.474 |
| 德国 | 35954.06 | 43793.547 |
| 英国 | 32322.81 | 49104.498 |
| 印度 | 26074.09 | 1964.973 |
| 法国 | 25865.68 | 39914.885 |
| 意大利 | 19329.38 | 31505.041 |
| 巴西 | 17592.67 | 8471.814 |
| 加拿大 | 16823.68 | 46172.119 |

根据贝叶斯信息准则选择聚类的类数,由GDP和人均GDP将178个国家分为四个部分.实验前数据效果如下:

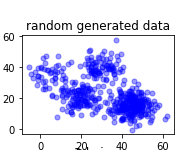


图 1原始数据图

经过算法聚类之后实验结果图2:

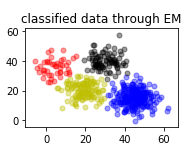


图 2聚类二维效果图

用3d效果图可表示为高斯模型的形状,如图3:

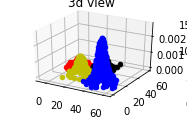


图 3 三维效果图

用四个高斯分布来分别表示178个国家的国力指标.从GDP总量来看,尽管有的国家的指标高,可是人均GDP还是很低,有的国家人均GDP很高可是国家GDP还是并不高,但总体上来看人均和总值持恒的国家还是占大多数.从实验结果来看,中国在总值排名中占第二名,可是在人均方面排到了70名,这样的发展是不太平衡的.在分类中也只能排到第2类国家中.

使用Iris data数据集,包含了150个四维的数据点,3类,每一类有50个数据点,通过三种不同的初始化EM的算法进行对聚类效果比较.实验结果表2:

表 2 不同初始方法对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 初始化方法 | Class-1 | Class-2 | Class-3 | 错误率（%） |
| 随机 | 47 | 43 | 46 | 9.3333 |
| 层次聚类 | 49 | 44 | 47 | 6.6667 |
| K-means | 50 | 44 | 50 | 4.0000 |

# 5 结论

本文采用高斯混合模型对178个国家的GDP和人均GDP进行了聚类分析,运用EM算法估计出模型参数有均值为[[ 4.34447837 34.54033523][ 29.46915118 39.49975567] [ 20.32193712 20.13464259] [ 45.27105555 14.50100861]]估计的混合项系数为[0.25417456091458962, 0.25527259452113749, 0.24046485556408578, 0.25008798900018708],数据成功地被分成四类,且分类结果与实际情况相吻合.并且,使用3种初始化EM的方法对数据集做了分析,并得到总体的错误率.随机中心初始化操作方便,但是由于随机性,实际效果有比较大的偏差.层次聚类初始化可以聚类到我们指定的类别数,但是迭代次数较多,在数据规模较大的情况下,效率很低,不仅耗时,而且需要较大的存储空间.K-means初始化较常见,但是结果不是很稳定,每次实验结果也不完全一致,但是是三个方法中综合效率是较高的.

本文证实了算法及模型的有效性,可以应用到其他领域的聚类分析.

References:

1. Pernkopf F. Genetic － based EM algorithm for component selection and parameter estimation og Gaussian mixture models［J］. Pattern Analysis and Machine Intelligence,2005,27 ( 8) : 1346 －1348.s

# 中文参考文献:

1. 邱藤. 基于高斯混合模型的EM算法及其应用研究[D]. 电子科技大学硕士论文,2015.
2. 李航 . 统计学习方法［M］. 清华大学出版社,2015,158 － 165.
3. 岳佳,王士同 . 高斯混合模型聚类中 EM 算法及初始化的研究[J]. 微计算机信息,2006,22( 11) :244 －246.
4. 翟少丹 . 基于混合模型的算法研究［D］. 西北大学硕士论文,2009:13 －15.
5. 岳佳,王士同 . 双重高斯混合模型的 EM 算法的聚类问题研究［J］. 计算机仿真,2007,24( 11) :110 － 113