Advances in Applied Mathematics 应用数学进展, 2021, 10(11), 3865-3872
Published Online November 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1011410

Hans汉斯

EM聚类分析法在大数据时代的应用

周志慧，刘瑞银*，杜　欢

沈阳师范大学数学与系统科学学院，辽宁 沈阳

收稿日期：2021年10月17日；录用日期：2021年11月7日；发布日期：2021年11月19日

摘　要

在大数据的时代背景下，各行各业的数据信息在应用和共享上都有着极大的便利，所以对数据的处理和分析手段显得尤为重要。大数据具有数据量庞大、分析效率低下、非结构化等特点，由于这样的复杂性，对同一组数据，可以从不同的角度去分析。使用传统的单维聚类方法不再适合大数据，本文研究了一个探索性的聚类方法——EM聚类分析法，具体分类过程是基于R语言中的Mclust()函数。将该方法应用于两个不同的实例，对于第一个例子中的小数据集我们采取了加入噪声的方式扩充其为大数据集，并在大数据集中得到了更好的分类效果，经过研究发现EM聚类分析法在大数据中得到的聚类效果更好，也适用于多维数据分析，最后对1900个基因在六个不同时间点的观测数据应用，给出了具体分类结果。

关键词

EM算法，EM聚类分析法，大数据，R语言

Application of EM Clustering in the Era of Big Data

Zhihui Zhou, Ruiyin Liu*, Huan Du

School of Mathematics and Systems Science, Shenyang Normal University, Shenyang Liaoning

Received: Oct. 17th, 2021; accepted: Nov. 7th, 2021; published: Nov. 19th, 2021

Abstract

In the era of big data, the application and sharing of data information in all walks of life are greatly convenient, so the means of data processing and analysis are particularly important. Big data is characterized by a large amount of data, low analysis efficiency and unstructured data. Due to such complexity, the same set of data can be analyzed from different perspectives. Using the tradi-

*通讯作者。

文章引用: 周志慧, 刘瑞银, 杜欢. EM 聚类分析法在大数据时代的应用[J]. 应用数学进展, 2021, 10(11): 3865-3872.
DOI: 10.12677/aam.2021.1011410

tional unidimensional clustering method is not suitable for large data, so this paper studies an exploratory clustering method, EM clustering analysis. The specific classification process is based on the Mclust() function in R language. This method was applied to two different examples. For the first example of small data set, we have taken the way to add noise to extend it to a large data set, and got a better classification effect in the large data set. During the study, we found it is better to use the EM clustering analysis method in the large data, and it can also be applied to multidimensional data analysis. At last, the specific classification results of 1900 genes at six different time points are given.

Keywords

EM Algorithms, EM Cluster Analysis, Big Data, R Language

1. 引言

聚类分析是统计中的常见问题，各种聚类方式各有千秋，在现在常用的众多聚类方法中，k 均值算法[1]简单快捷，但是 k 需要提前给定，k 值的选定在实际数据中是非常难以估计的。很多时候，我们并不知道给定的数据集应该分成多少个类别才最合适。以距离为分类核心的系统聚类法[2]需要计算距离矩阵，在数据集很大的情况下，伴随的计算量也很大。而本文研究的 EM 聚类分析法是属于探索性的聚类，不用提前知道原始类别，也不用预知分类数，而且在本文的研究过程中发现，对于小的数据集，EM 聚类往往需要加噪声才能达到更好的分类效果，通过对具体实例应用，发现 EM 聚类算法适用于大数据，不但计算量没那么大，而且操作简便，对于大数据集能得到较好的分类效果。

2. EM 算法

2.1. EM 算法介绍

设样本数据 $(x_1, x_2, \cdots, x_n)$ 间互相独立，每个样本对应的类别 z_i 未知，我们的目的是确定样本所属类别使得 $p(x_i, z_i)$ 最大化，则其似然函数为：

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

取对数：

$$\tau(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta) = \sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i; \theta) \tag{1}$$

定义类别变量 z_i 满足某一分布 Q_i，并且该分布(离散分布)满足以下条件：

$$\sum_{z_i} Q_i(z_i) = 1, Q_i(z_i) \geq 0 \tag{2}$$

因此，利用詹森不等式对公式(1)变形得到：

$$\sum_{i=1}^{n}\log p(x_i;\theta)=\sum_{i=1}^{n}\log\sum_{z_i}p(x_i,z_i;\theta)$$
$$=\sum_{i=1}^{n}\log\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$$
$$\geq\sum_{i=1}^{n}\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$$

因为$\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$是$\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$的期望，所以由詹森不等式可推导出如下：

$$f(E[X])=\log\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}\geq E[f(X)]=\sum_{z_i}Q_i(z_i)\log\frac{p(x_i,z_i;\theta)}{Q_i(z_i)} \tag{3}$$

至此，通过公式(3)，得到了似然函数$\tau(\theta)$的下界，如果θ是已知的，那么似然函数的值就取决于两个概率$Q_i(z_i)$，$p(x_i,z_i)$，因此若想使公式(3)由不等式变为等式，可以调整$Q_i(z_i)$，$p(x_i,z_i)$的值，以这种方式来逼近似然函数$\tau(\theta)$的值。由詹森不等式可以知道，当且仅当x为常量时，不等式取等号，于是有

$$\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}=C \tag{4}$$

其中C是常量，对于一系列不同的z_i进行求和得到：

$$\frac{\sum_{z_i}p(x_i,z_i;\theta)}{\sum_{z_i}Q_i(z_i)}=c$$

又由公式(2)，得：

$$\sum_{z_i}p(x_i,z_i;\theta)=C$$

代入公式(4)，并且引入条件概率公式可得：

$$Q_i(z_i)=\frac{p(x_i,z_i;\theta)}{C}=\frac{p(x_i,z_i;\theta)}{\sum_{z_i}p(x_i,z_i;\theta)}=\frac{p(x_i,z_i;\theta)}{p(x_i,\theta)}=p(z_i|x_i;\theta)$$

上式显示了有关类别的分布，接下来就是将似然函数$\tau(\theta)$最大化的过程了：

给定初始值θ，循环重复下列步骤，直到收敛：

(E 步)记对于每个x_i，计算$Q_i(z_i)=p(z_i\mid x_i;\theta)$

(M 步)计算

$$\theta=\arg\max_{\theta}\sum_{i=1}^{n}\sum_{z_i}\log\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$$

2.2. EM 聚类思想

EM 算法的总体思想是通过引入一个潜在变量 Z，通过添加的 Z 得到的后验分布密度函数，对 Z 积掉 Z，得到一个新的，由新的可以得到新的，如此迭代下去，求出一个离真实值最近的。

EM 聚类的算法设计是基于统计分布的，它认为全体观测值可被分成 k 个自然小类，其中每一类的观测值数据都对应一个不同的分布，可以理解为 EM 算法将具有相同分布的观测值聚为一类，这样就将总体分成了包含 k 个不同分布的混合样本。在 EM 聚类的过程中，分类数 k 是未知的，即便确定了分类数

为 k，这 k 个总体分布的参数也是未知的，用观测值可以进行估计，EM 算法在潜变量(如样本归属某个小类的概率)和分布参数(如某小类总体分布参数)未知的情况下，通过迭代方式最大似然函数来实现，在确定分类数为 k 的条件下，我们可以逐一代入 $k=1,2,3,\cdots$，随后利用一定的标准(如选取 BIC 值较小的)，挑选最佳的分类数 k，EM 聚类的迭代思路是：设有关于分类参数 z 和成分参数的两个参数集合 Z 和，初始步，从集合中指定一个值作为 t 时刻成分参数的估计值，记作：第一步，在的基础上找到 t 时刻使联合概率最大的分类参数；第二步，在基础上计算使对数似然函数最大的成分参数，记作 $\theta^{(t+1)}$。

记 $\theta^{(i)}$ 为第 $i+1$ 次迭代开始时后验众数的估计值，则第 $i+1$ 次迭代

E-step：

引入一个潜在变量 Z，$P(\theta|Y,Z)$ 表示添加数据 Z 后得到的关于 θ 的后验分布密度函数，将 $P(\theta|Y,Z)$ 或 $\log P(\theta|Y,Z)$ 关于 Z 的条件分布求期望 $Q\left(\theta|\theta^{(i)},Y\right)$，从而把 Z 积掉。

M-step：

将 $Q\left(\theta|\theta^{(i)},Y\right)$ 极大化，即找一个点 $\theta^{(i+1)}$，使

$$Q\left(\theta^{(i+1)}|\theta^{(i)},Y\right)=\max_{\theta}Q\left(\theta|\theta^{(i)},Y\right)$$

如此形成了一次迭代 $\theta^{i}\to\theta^{(i+1)}$。重复计算 E 步和 M 步，直到

$$\left\|\theta^{(i+1)}-\theta^{(i)}\right\|$$

充分小时停止[3]。

3. 实例研究

3.1. 数据来源

以 2017 年全国 30 个地区房地产数据为例，数据来源于国家统计局网站。数据集有 30 行 6 列，六列数据分别是商品房平均销售价格 X_1，住宅商品房平均销售价格 X_2，别墅，高档公寓平均销售价格 X_3，办公楼商品房平均销售价格 X_4，商业营业用房平均销售价格 X_5，其他商品房平均销售价格 X_6。

3.2. 结果分析

全过程进行了两次聚类，第一次是用程序包 mclust 中的函数 Mclust()对已知的 30 个地区房地产数据直接进行聚类，第二次是考虑到样本数据量较小，适当加入了一些均匀分布随机数据点，得到进一步聚类结果。

第一次聚类结果图如图 1。

在第一次聚类的结果中我们可以看到，在 BIC 值为−3108.931 的时候，挑选出了最佳分类值 k=8，将 30 个地区分为了 8 类，样本大小分别为 1，1，11，9，1，2，4，1，各个类别以不同的表示方法在图中分别标示，如红色圆圈，蓝圆点，绿色叉，蓝色三角等等，占比依次为 0.0333，0.0333，0.3667，0.3001，0.0333，0.0665，0.1333，0.0333，得到每个地区属于每一类的概率，即类别变量 z_i，八个类的均值分别为：

(32,140, 34,117, 49,926, 34,539, 36,370, 9385), (15,331, 15,139, 17,951, 18,327, 17,291, 14,117), (5815.127, 5517.203, 8904.254, 8767.730, 8235.651, 4689.247), (5998.003, 5646.239, 10,460.344, 9140.791, 9659.226, 4462.933), (23,804, 24,866, 54,399, 31,753, 26,249, 6024), (11,029.313, 11,255.139, 14,896.581, 12,295.227, 12,772.177, 5282.641), (8726.50, 8333.25, 12,642.50, 16,800.25, 11,610.25, 6596.25), (11,837,

11,381, 18,008, 17,334, 18,918, 24,187)，同时输出八类数据对应的协方差矩阵以及似然函数值等。然后给出了样本 x_i 各属于哪一类别。

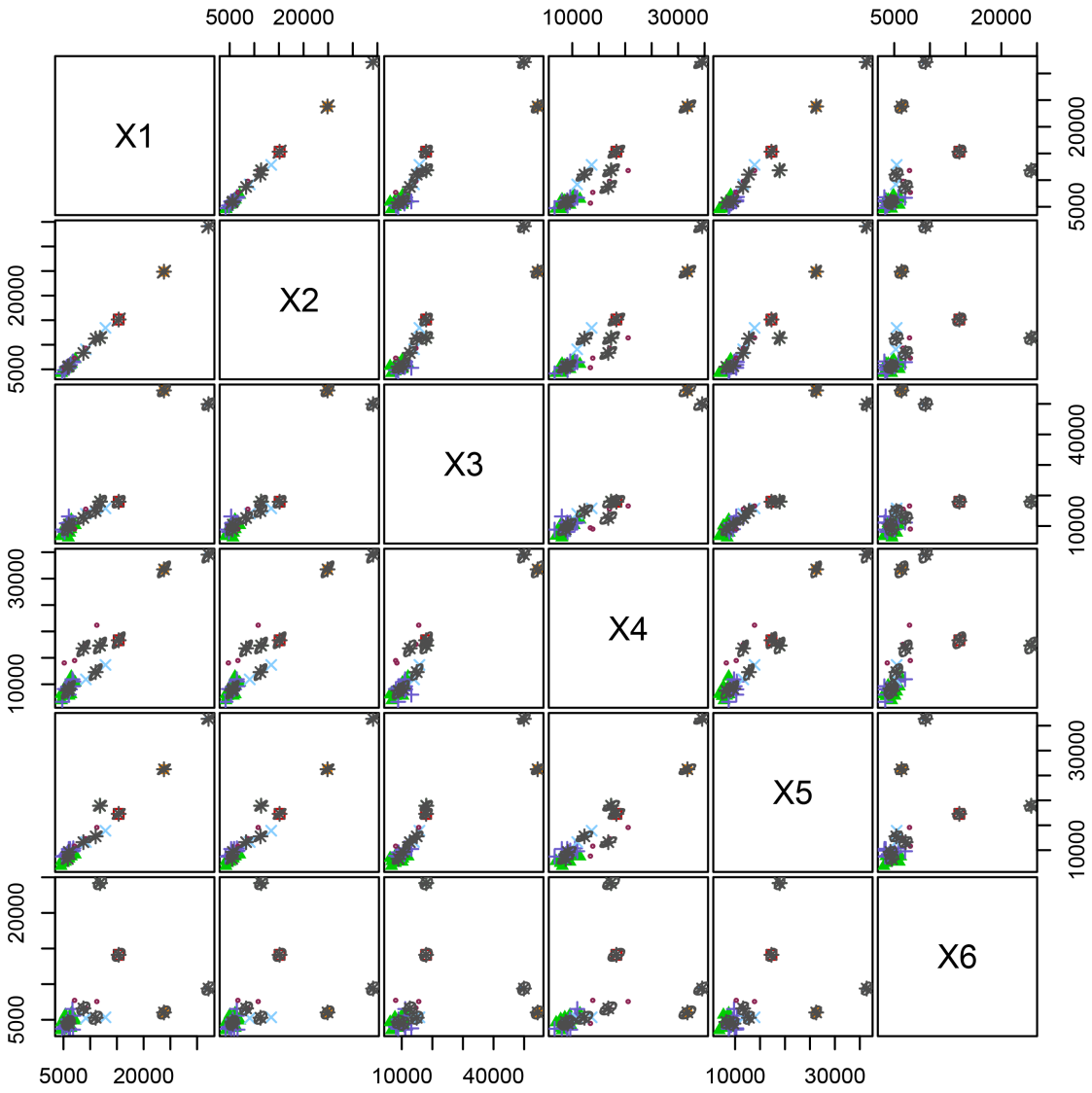

Figure 1. The first classification result diagram
图 1. 第一次分类结果图

第二次聚类结果图如图 2。

在第二次聚类的结果中，我们可以看到，在 BIC 值为−18,163.46 的时候，挑选出了最佳分类值 $k = 2$，将添加噪声后的 150 个样本分成了两类，样本大小分别为 28,122。占比依次为 0.1865 和 0.8135。两次分类结果得到了两个分类模型，从模型参数 BIC 来看，加入噪声后的大数据集模型 BIC 值远远超过第一个小数据集的 BIC 值，说明第二次分类结果远好于第一次分类结果。这也说明 EM 聚类分析法在大数据中应用得到的分类结果更好。(补充说明：BIC 准则，又称贝叶斯信息准则，是检验模型好坏的判断参数，一般，值越小的模型越好)

(以上过程实现的具体 R 程序[4]如附录程序一)

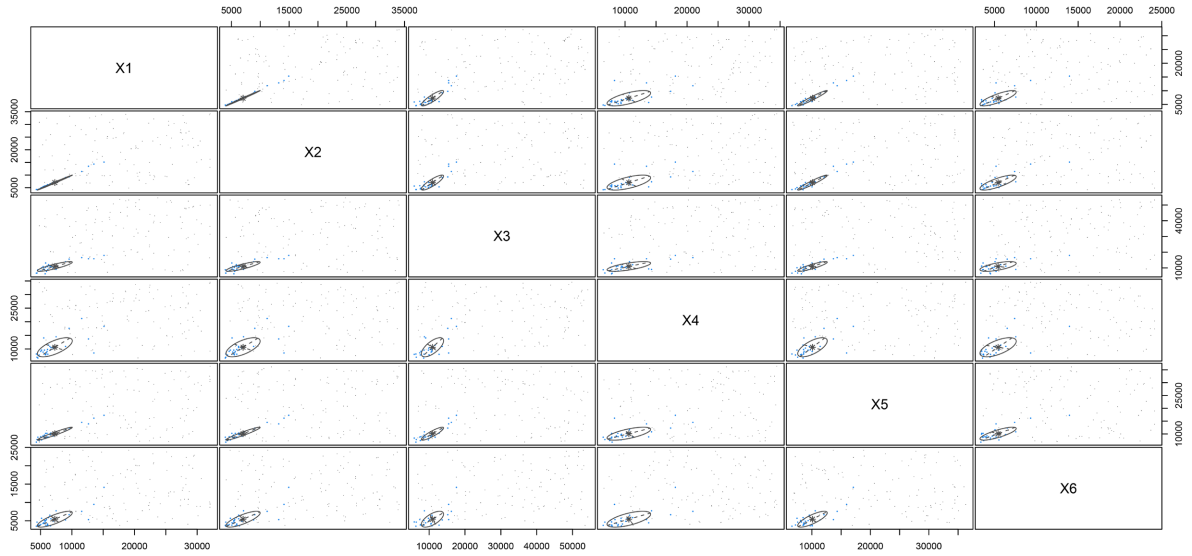

Figure 2. The second classification result diagram
图 2. 第二次分类结果图

4. 实例应用

一种新药的问世需要经过多个试验阶段，而新药的药效水平也有很多检测的方式，利用药物对基因的作用，观察基因在不同时间的不同表达是研究药效的一种重要方式，可以根据基因的表达水平对药物进行分类[5]，本例收集到了 1900 个基因在六个不同观测时间的表达数据，然后对得到的数据实行 EM 聚类(具体 R 程序算法如附录程序二)：

得到的图形如图 3 所示。

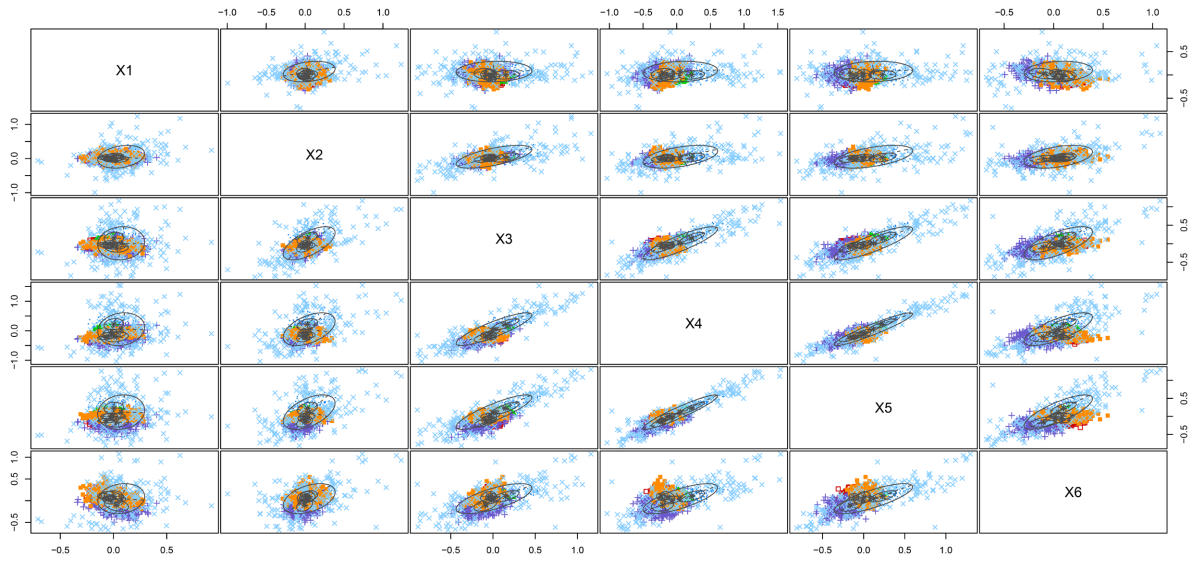

Figure 3. Map of gene classification results
图 3. 基因分类结果图

在聚类的结果中我们可以看到，在 BIC 值为 16,280.52 的时候，挑选出了最佳分类值 k = 6，将全部 1900 个基因分为了 6 类，并计算出了每个基因属于每一类的概率，即类别变量 z_i，然后给出了样本 x_i 各

属于哪一类别。同时得到了各个类别对应的均值，协方差矩阵，似然函数值等等。

5. 总结

本文研究了一个探索性的聚类方法 EM 聚类分析法，首先系统介绍了 EM 算法，然后叙述了用 EM 算法进行聚类的思想，随之应用于具体的实例当中探索 EM 聚类分析法能达到的分类效果，以及和其他聚类方法对比来看有哪些优缺点，具体分类过程是基于 R 语言中的 Mclust()函数。

在实例研究过程中，我们对收集到的 2017 年中国的 30 个省、市、自治区房地产业的相关统计数据进行聚类分析，展示了分类结果及得到的各项参数值，对于实例研究中的小数据集我们采取了加入噪声的方式扩充其为大数据集，并在大数据集中得到了更好的分类效果，发现 EM 聚类分析法在大数据中得到的聚类效果更好，也适用于多维数据分析。

最后在实例应用部分，我们对 1900 个基因在六个不同时间点的观测数据进行聚类，根据基因在不同时间点的表达水平不同对研究出的新药进行分类，得到不同类别下具有不同药效的药物，我们可以根据实际问题，去分析药效较好的类别中的药物成分，或者探索每一类别下对应的药物成分的适应功效，进而促进新药的研究。

参考文献

[1] 孙艺, 赵瑛珲, 王天棋, 马彦凯, 赵佳琪. 一种 K-均值优化算法的研究与改进[J]. 自动化技术与应用, 2021, 40(9): 1-5+11.

[2] 田兵. 系统聚类法及其应用研究[J]. 阴山学刊(自然科学版), 2014, 28(2): 11-16.

[3] 茆诗松, 王静龙, 濮晓龙. 高等数理统计[M]. 第 2 版. 北京: 高等教育出版社, 2006: 427-433.

[4] 薛薇. R 语言数据挖掘 R [M]. 北京: 中国人民大学出版社, 2019: 155-160.

[5] 刘瑞银. 基于趋势性的剂量反应研究[D]: [博士学位论文]. 长春: 东北师范大学, 2011.

附录

程序一：

```
library(mclust)#导入需要用的程序包
read.table(file="liti.txt",TRUE)#读入处理后的数据
w=read.table(file="liti.txt",TRUE)
EM1=Mclust(w)#聚类
summary(EM1,parameter=TRUE)#查看分类结果
plot(EM1,what="classification")#作出分类图
EM1$classification#具体查看每个地区所属类别
nNoise=120#添加 120 个噪声数据
Noise=apply(w,2,function(x)runif(nNoise,min=min(x)-0.1,max=max(x)+0.1))
data=rbind(w,Noise)
plot(w)
points(Noise,pch=16,cex=0.5)
NoiseInit=sample(c(TRUE,FALSE),size=nrow(w)+nNoise,replace=TRUE,prob=c(3,1)/4)
EM2=Mclust(data,initialization=list(noise=NoiseInit))#对添加噪声后的数据重新聚类
summary(EM2,parameter=TRUE)
plot(EM2,what="classification")
```

程序二：

```
library(mclust)#导入需要用的程序包
read.table(file="jiyin.txt",TRUE)#读入处理后的数据
w=read.table(file="jiyin.txt",TRUE)
EM1=Mclust(w)#聚类
summary(EM1,parameter=TRUE)#查看分类结果
plot(EM1,what="classification")#作出分类图
EM1$classification#具体查看每个基因所属类别
a=EM1$classification
a[a==1]#挑选出类别一的基因序号
type1=a[a==1]
type1
names(type1)[type1==1]
T1=names(type1)[type1==1]
T1
V1=w[T1,c(1:6)]#导出第一个类别中的基因对应数据
V1
```