第 10 章　聚类分析

1. 比较常用的几种聚类方法的原理及使用方法的异同点。

【参考答案】常用的聚类方法有 k-均值算法、k-中心点算法及层次聚类方法。k-均值算法、k-中心点算法是基于划分的方法，即把一堆需要聚类的样本点聚类成“类内的点都足够近，类间的点都足够远”的一个个类。首先确定样本点要聚成几类，然后挑选几个点作为初始中心点，再依据预先定好的启发式算法(heuristic algorithms)给数据点做迭代重置(iterative relocation)，最后达到“类内的点都足够近，类间的点都足够远”的目标效果。这里聚类的不同启发式算法形成了 k-means 算法及其变体包括 k-medoids、k-modes、k-medians、kernel k-means 等算法。层次聚类方法通过某种特征相似性测度计算样本点之间的相似性，并按相似度由高到低排序，逐步重新连接样本点，得出聚类结果。

2. 对聚类结果评估主要有哪些方法？

【参考答案】评估聚类结果的质量是另一个重要的阶段。聚类是一个无管理的程序，也没有客观的标准来评价聚类结果，它通过一个类有效索引来评价。一般来说，几何性质、类间的分离及类内部的耦合，一般都用来评价聚类结果的质量。类有效索引在决定类的数目时经常扮演了一个重要角色。类有效索引的最佳值被期望从真实的类数目中获取，一个通常的决定类数目的方法是选择一个特定的类有效索引的最佳值。这个索引能否真实地得出类的数目是判断该索引是否有效的标准。很多已经存在的标准对于相互分离的类数据集合都能得出很好的结果，但是对于复杂的数据集，却通常行不通。

当我们采用一种聚类方法时，如何评估该聚类的结果的好坏？聚类评估主要包括以下任务：

①估计聚类趋势：仅当数据中存在非随机数据时，评估数据集是否存在随机数据。②确定数据集中的簇数：在使用聚类算法之前，需要估计簇数。③测定聚类质量：在数据集上使用聚类方法之后，需要评估簇的质量。

3. 通过 rattle 包获取葡萄酒中 13 种化学成分的数据集 wine，再使用 k-均值聚类来处理数据集，并把处理的结果用图形展示出来。

【参考答案】思路：首先加载 rattle 包，查看 wine 数据集，并进行标准化。然后判断适合的聚类数目。经过判断，适合的聚类数目为 3。最后调用 kmeans 函数完成聚类。聚类分析结果如图 10-1 所示，聚类个数结果如图 10-2 所示。

```
> data(wine,package = "rattle")
> head(wine)
  Type Alcohol Malic  Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids
1    1   14.23  1.71 2.43       15.6       127    2.80       3.06          0.28
2    1   13.20  1.78 2.14       11.2       100    2.65       2.76          0.26
3    1   13.16  2.36 2.67       18.6       101    2.80       3.24          0.30
4    1   14.37  1.95 2.50       16.8       113    3.85       3.49          0.24
5    1   13.24  2.59 2.87       21.0       118    2.80       2.69          0.39
6    1   14.20  1.76 2.45       15.2       112    3.27       3.39          0.34
  Proanthocyanins Color  Hue Dilution Proline
1            2.29  5.64 1.04     3.92    1065
2            1.28  4.38 1.05     3.40    1050
3            2.81  5.68 1.03     3.17    1185
4            2.18  7.80 0.86     3.45    1480
5            1.82  4.32 1.04     2.93     735
6            1.97  6.75 1.05     2.85    1450
```

```
> df<-scale(wine[-1])

 > library(NbClust)
 > set.seed(1234)
 > devAskNewPage(ask=TRUE)
 > nc<-NbClust(df,min.nc = 2,max.nc = 15,method = "kmeans")

> barplot(table(nc$Best.nc[1,]),xlab = "Number of Clusters",ylab = "Number of Criteria",
+         main="Number of Clusters chosen by 26 Criteria")

 > set.seed(1234)
 > fit.km<-kmeans(df,3,nstart = 25)
 > fit.km$size
 [1] 62 65 51
 > fit.km$centers
      Alcohol      Malic        Ash Alcalinity   Magnesium     Phenols  Flavanoids
 1  0.8328826 -0.3029551  0.3636801 -0.6084749  0.57596208  0.88274724  0.97506900
 2 -0.9234669 -0.3929331 -0.4931257  0.1701220 -0.49032869 -0.07576891  0.02075402
 3  0.1644436  0.8690954  0.1863726  0.5228924 -0.07526047 -0.97657548 -1.21182921
   Nonflavanoids Proanthocyanins      Color        Hue   Dilution    Proline
 1   -0.56050853       0.57865427  0.1705823  0.4726504  0.7770551  1.1220202
 2   -0.03343924       0.05810161 -0.8993770  0.4605046  0.2700025 -0.7517257
 3    0.72402116      -0.77751312  0.9388902 -1.1615122 -1.2887761 -0.4059428
 > aggregate(wine[-1],by=list(cluster=fit.km$cluster),mean)
   cluster  Alcohol    Malic      Ash Alcalinity Magnesium  Phenols Flavanoids Nonflavanoids
 1       1 13.67677 1.997903 2.466290   17.46290 107.96774 2.847581  3.0032258     0.2920968
 2       2 12.25092 1.897385 2.231231   20.06308  92.73846 2.247692  2.0500000     0.3576923
 3       3 13.13412 3.307255 2.417647   21.24118  98.66667 1.683922  0.8188235     0.4519608

   Proanthocyanins    Color       Hue Dilution   Proline
 1        1.922097 5.453548 1.0654839 3.163387 1100.2258
 2        1.624154 2.973077 1.0627077 2.803385  510.1692
 3        1.145882 7.234706 0.6919608 1.696667  619.0588
```

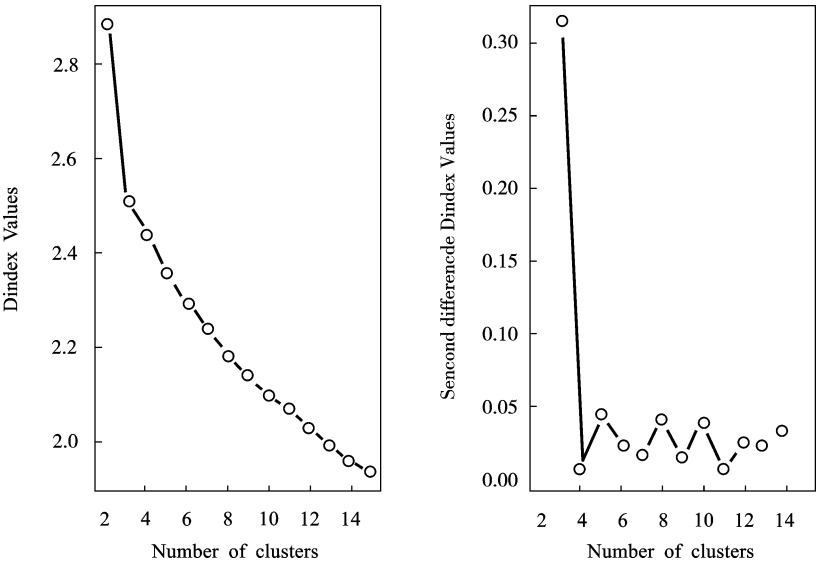

图 10-1　聚类分析图

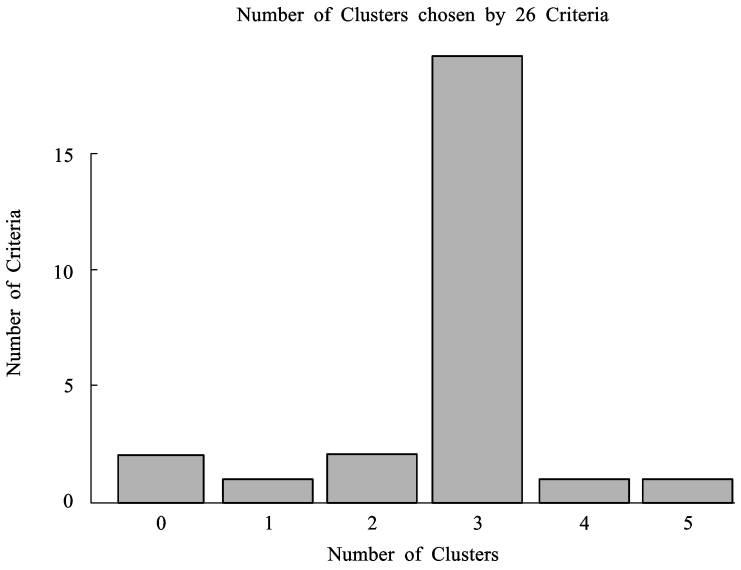

图 10-2　聚类个数图

4. 为了研究世界各国森林、草原资源的分布规律，共抽取了 21 个国家的数据，每个国家 4 项指标，分别为森林面积(万公顷)、森林覆盖率(%)、林木蓄积量(亿立方米)、草原面积(万公顷)，原始数据见表 10-1。使用该原始数据以国别进行聚类分析。

表 10-1　抽样数据表

国家	森林面积/万公顷	森林覆盖率/%	林木蓄积量/亿立方米	草原面积/万公顷
中国	11978	12.5	93.5	31908
美国	28446	30.4	202.0	23754
日本	2501	67.2	24.8	58
德国	1028	28.4	14.0	599
英国	210	8.6	1.5	1147
法国	1458	26.7	16.0	1288
意大利	635	21.1	3.6	514
加拿大	32613	32.7	192.8	2385
澳大利亚	13.9	10.5	45190	10700
前苏联	92000	41.1	841.5	37370
捷克	458	35.8	8.9	168
波兰	868	27.8	11.4	405

续表10-1

国家	森林面积/万公顷	森林覆盖率/%	林木蓄积量/亿立方米	草原面积/万公顷
匈牙利	161	17.4	2.5	129
南斯拉夫	929	36.3	11.4	640
罗马尼亚	634	26.7	11.3	447
保加利亚	385	34.7	2.5	200
印度	6748	20.5	29.0	1200
印尼	2180	84.0	33.7	1200
尼日利亚	1490	16.1	0.8	2090
墨西哥	4850	24.6	32.6	7450
巴西	57500	67.6	238.0	15900

【参考答案】

(1)读取数据 forest.txt，查看数据基本信息。

```
> forest <- read.table('forest.txt')
> head(forest)
    V1    V2   V3    V4    V5
1 中国 11978 12.5  93.5 31908
2 美国 28446 30.4 202.0 23754
3 日本  2501 67.2  24.8    58
4 德国  1028 28.4  14.0   599
5 英国   210  8.6   1.5  1147
6 法国  1458 26.7  16.0  1288
> str(forest)
'data.frame':   21 obs. of  5 variables:
 $ V1: Factor w/ 21 levels "澳大利亚","巴西",..: 21 10 15 5 20 6 17 7 1 14 ...
 $ V2: num  11978 28446 2501 1028 210 ...
 $ V3: num  12.5 30.4 67.2 28.4 8.6 26.7 21.1 32.7 10.5 41.1 ...
 $ V4: num  93.5 202 24.8 14 1.5 ...
 $ V5: int  31908 23754 58 599 1147 1288 514 2385 10700 37370 ...
```

(2)数据预处理。

```
> my.forest <- forest[,-1]
> row.names(my.forest) <- forest[,1]
> colnames(my.forest) <- c('森林面积','森林覆盖率（%）','林木蓄积量','草原面积')
> head(my.forest)
     森林面积 森林覆盖率（%） 林木蓄积量 草原面积
中国    11978            12.5       93.5    31908
美国    28446            30.4      202.0    23754
日本     2501            67.2       24.8       58
德国     1028            28.4       14.0      599
英国      210             8.6        1.5     1147
法国     1458            26.7       16.0     1288
```

(3)层次聚类代码如下，结果如图 10-3 所示。

```
> my.forest.scaled <- scale(my.forest)
> my.forest.d <- dist(my.forest.scaled)
> my.forest.ave <- hclust(my.forest.d, method = 'average')
> plot(my.forest.ave,hang = -1, cex=0.8)
> rect.hclust(my.forest.ave, k=3)
```

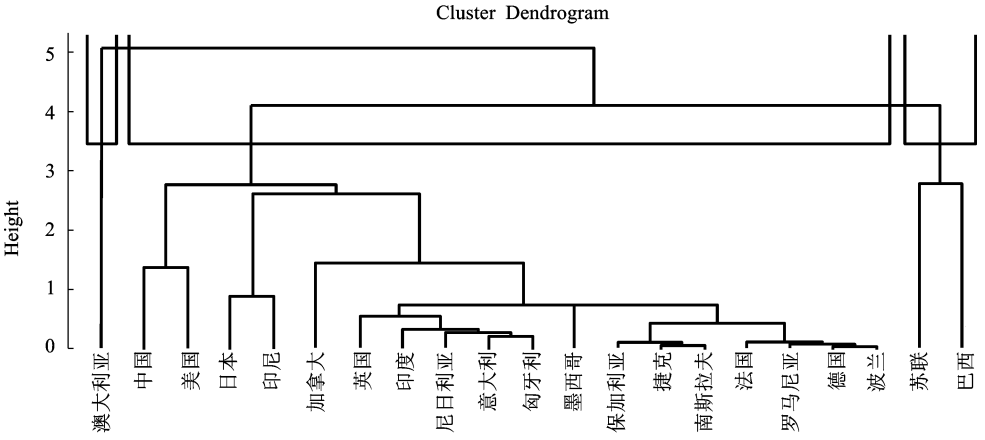

图 10-3　层次聚类分析图

(4)也可以尝试采用 k-means 和 k-中心聚类。