聚类模型案例：癌细胞基因表达数据

吴翔

2019-03-20

概述

我们通过R语言 ISLR 包中64个细胞的6830个基因表达数据作为案例，来阐述如何使用如下聚类模型：

- 系统聚类
- K均值聚类

数据集 NCI60 包含基因表达数据矩阵 NCI60$data 以及癌细胞系的类型 NCI60$labs 。

```
# clean the work directory
rm(list = ls())

# set seeds
set.seed(123)

# read dataset
suppressMessages(library(ISLR))
suppressMessages(library(tidyverse))
data("NCI60")
# display the variables
str(NCI60)
```

```
## List of 2
##  $ data: num [1:64, 1:6830] 0.3 0.68 0.94 0.28 0.485 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:64] "V1" "V2" "V3" "V4" ...
##   .. ..$ : chr [1:6830] "1" "2" "3" "4" ...
##  $ labs: chr [1:64] "CNS" "CNS" "CNS" "RENAL" ...
```

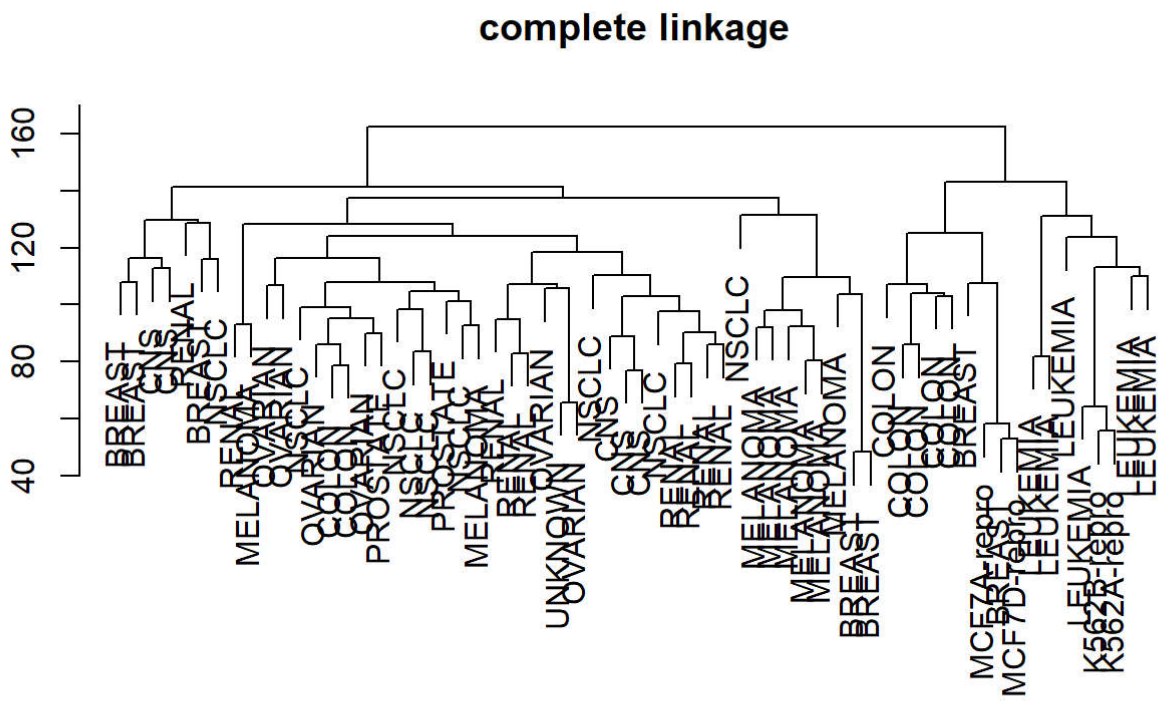
```
# summary of dataset
ncidat <- NCI60$data
ncilab <- NCI60$labs
# types of cancer cells
table(ncilab)
```

```
## ncilab
##      BREAST         CNS       COLON K562A-repro K562B-repro    LEUKEMIA 
##           7           5           7           1           1           6 
## MCF7A-repro MCF7D-repro    MELANOMA       NSCLC     OVARIAN    PROSTATE 
##           1           1           8           9           6           2 
##       RENAL     UNKNOWN 
##           9           1
```

系统聚类

```
# scaling
ncidat <- scale(ncidat)

# clustering
nci.dist <- dist(ncidat, method = "euclidean", p = 2)
hc.out <- hclust(nci.dist)
plot(hc.out, labels = ncilab, main = "complete linkage", xlab = "", ylab = "")
```

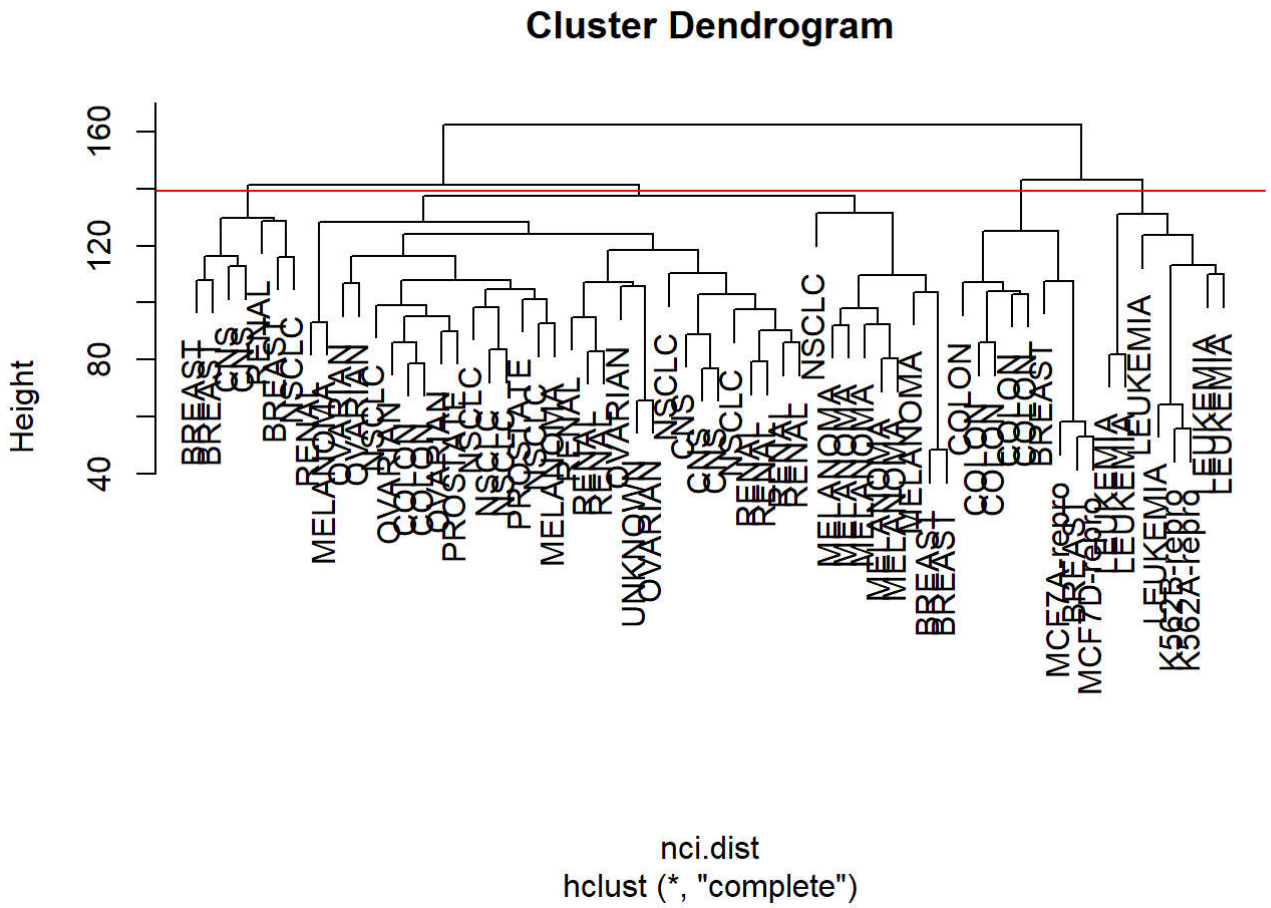


在以上谱系图上某个高度切割，可以得到指定类数的聚类，例如类别数量$s = 4$。

```
# four clusters
table(cutree(hc.out, 4))
```

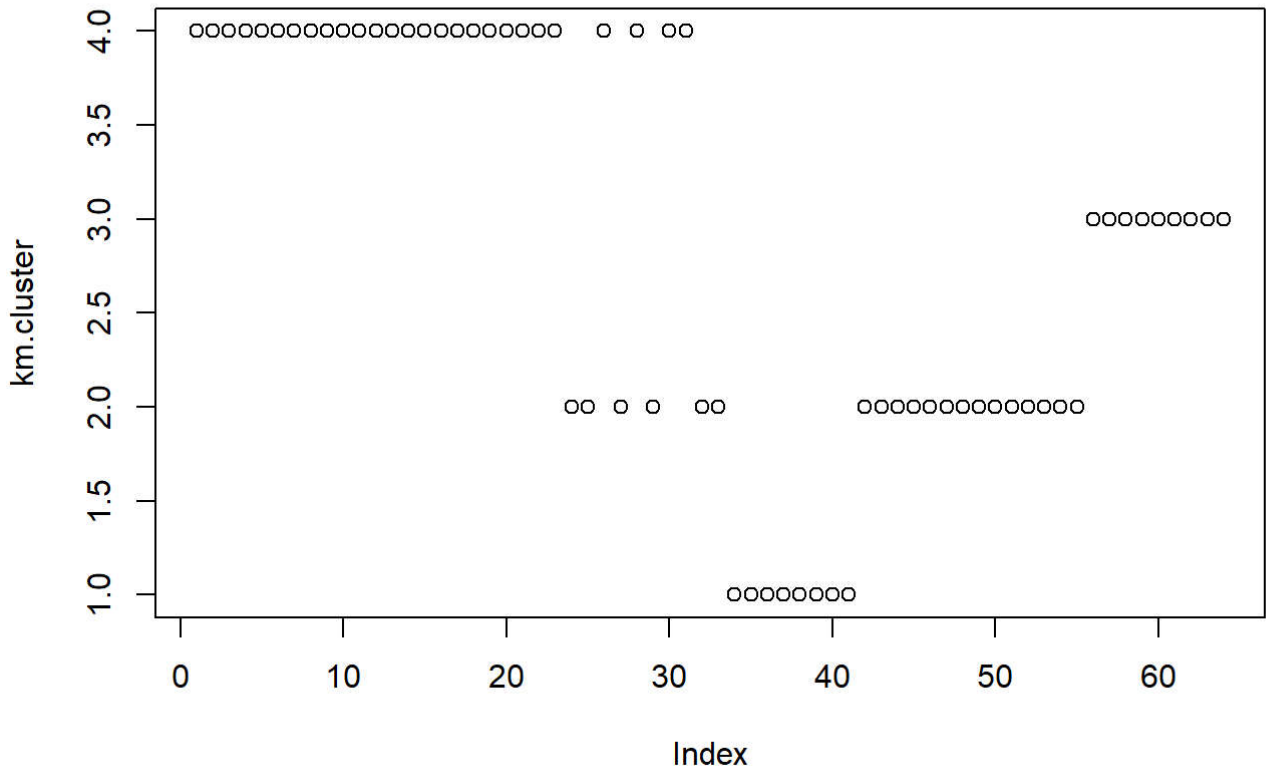
```
##
##  1  2  3  4
## 40  7  8  9
```

```
# plot
plot(hc.out, labels = ncilab)
abline(h = 139, col = "red")
```



K均值聚类

```
# kmeans
km.out <- kmeans(ncidat, 4, nstart = 20)
km.cluster <- km.out$cluster
names(km.cluster) <- ncilab
plot(km.cluster)
```




```
km.cluster
```

```
##         CNS         CNS         CNS       RENAL      BREAST         CNS
##           4           4           4           4           4           4
##         CNS      BREAST       NSCLC       NSCLC       RENAL       RENAL
##           4           4           4           4           4           4
##       RENAL       RENAL       RENAL       RENAL       RENAL      BREAST
##           4           4           4           4           4           4
##       NSCLC       RENAL     UNKNOWN     OVARIAN    MELANOMA    PROSTATE
##           4           4           4           4           4           2
##     OVARIAN     OVARIAN     OVARIAN     OVARIAN     OVARIAN    PROSTATE
##           2           4           2           4           2           4
##       NSCLC       NSCLC       NSCLC    LEUKEMIA K562B-repro K562A-repro
##           4           2           2           1           1           1
##    LEUKEMIA    LEUKEMIA    LEUKEMIA    LEUKEMIA    LEUKEMIA       COLON
##           1           1           1           1           1           2
##       COLON       COLON       COLON       COLON       COLON       COLON
##           2           2           2           2           2           2
## MCF7A-repro      BREAST MCF7D-repro      BREAST       NSCLC       NSCLC
##           2           2           2           2           2           2
##       NSCLC    MELANOMA      BREAST      BREAST    MELANOMA    MELANOMA
##           2           3           3           3           3           3
##    MELANOMA    MELANOMA    MELANOMA    MELANOMA
##           3           3           3           3
```