案例 2.7

使用 R 对本科生助教进行类型分析

一、 拟求解的问题

某大学实行了本科生助教制度，即高年级的本科生作为助教参与教学工作。为了评价助教的工作，全体学生对 37 名助教通过下面 7 个方面的问题进行了学生对助教的满意度调查：

(1) X1-助教每周为学生提供帮助的工作时间(包括线上、线下)，满分 5 小时

(2) X2-助教实践课辅导的贡献情况，满分 1 分

(3) X3-助教课后答疑的贡献情况，满分 1 分

(4) X4-助教指导小组讨论的贡献，满分 1 分

(5) X5-助教批改作业的贡献，满分 1 分

(6) X6-助教其他方面的贡献，满分 1 分

(7) Y-对助教工作的总体满意度，满分 5 分

37 名助教的平均得分情况如表 1 所示。

表 1 学生对助教满意度评价表

TA_ID	X1	X2	X3	X4	X5	X6	Y
1	2.65	0.53	0.82	0.12	0.29	0.00	4.24
2	2.64	0.88	0.52	0.17	0.30	0.09	4.34
3	3.33	1.00	1.00	0.50	0.67	0.08	4.75
4	3.26	0.95	0.80	0.29	0.56	0.05	4.90
5	2.87	0.94	0.59	0.53	0.47	0.00	4.53
6	3.42	0.97	0.83	0.48	0.63	0.03	4.80
7	3.11	0.97	0.94	0.64	0.58	0.03	4.82
8	2.93	0.97	0.72	0.59	0.59	0.03	4.62

9	3.52	0.97	0.84	0.47	0.61	0.03	4.82
10	3.11	1.00	0.81	0.47	0.67	0.02	4.74
11	2.86	1.00	0.95	0.67	0.57	0.00	4.76
12	5.00	1.00	1.00	1.00	0.00	0.00	5.00
13	2.94	0.95	0.93	0.58	0.60	0.03	4.78
14	3.75	1.00	1.00	0.00	0.00	0.00	5.00
15	2.50	1.00	1.00	0.00	1.00	0.00	4.00
16	3.04	0.99	0.74	0.26	0.67	0.01	4.68
17	3.81	0.95	0.95	0.45	0.68	0.09	4.91
18	3.52	0.96	0.74	0.22	0.59	0.04	4.70
19	3.06	1.00	0.93	0.44	0.44	0.04	4.74
20	3.44	1.00	0.60	0.10	0.70	0.00	4.65
21	3.05	0.88	0.71	0.12	0.50	0.03	4.56
22	3.34	0.78	0.91	0.51	0.61	0.06	4.75
23	3.36	0.80	0.84	0.38	0.42	0.13	4.49
24	3.24	0.84	0.94	0.22	0.47	0.13	4.53
25	3.47	0.89	0.89	0.28	0.39	0.06	4.83
26	3.37	0.84	0.88	0.39	0.43	0.12	4.53
27	3.40	1.00	0.97	0.53	0.59	0.06	4.94
28	3.97	0.95	0.96	0.65	0.74	0.07	4.91
29	2.50	0.98	0.53	0.26	0.55	0.04	4.40
30	2.55	0.89	0.81	0.22	0.22	0.00	4.70
31	3.00	0.90	0.85	0.45	0.60	0.10	4.80
32	3.87	0.97	0.94	0.56	0.66	0.06	4.81
33	3.54	0.98	0.87	0.53	0.43	0.00	4.81
34	2.76	1.00	0.58	0.25	0.21	0.08	4.46
35	3.58	0.93	0.98	0.43	0.73	0.00	4.86
36	3.30	0.78	0.91	0.53	0.61	0.05	4.76

37	3.15	1.00	0.83	0.30	0.83	0.00	4.78

通过对表 1 中的数据进行聚类分析，看看助教是否可以明显地分为不同的类型，分析各类型助教的特征。

二、 预备知识

2.1 关于 R

关于 R 的简介、如何安装 R 和 RSudio、 R 的包、R 的数据类型、RStudio 的基本操作、等内容请参考案例 2.6 预备知识相关内容。

2.2 数据挖掘与线性回归分析简介

（1）数据挖掘

数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

（2）聚类分析

聚类分析(cluster analysis)是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类与分类的不同，聚类所要划分的类是未知的。聚类分析是一种探索性的分析，在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用方法的不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。而且聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。聚类分析还可以作为其他算法（如分类和定性归纳算法）的预处理步骤。

聚类分析的主要方法有 K-Means 划分法、层次聚类法和 DBSCAN 密度法。

（3）K-Means 划分法

K 表示聚类算法中类的个数，Means 表示均值算法，K-Means 即是用均值算法把数据分成 K 个类的算法。

K-Means 算法的目标是把 n 个样本点划分到 k 个类中，使得每个点都属于离它最近的**质心**（一个类内部所有样本点的均值）对应的类，以它作为聚类的标准。

K-Means 算法的计算步骤：

①取得 k 个初始质心：从数据中随机抽取 k 个点作为初始聚类的中心，来代表各个类

②把每个点划分进相应的类：根据欧式距离最小原则，把每个点划分进与质心距离最近的类中

③重新计算质心：根据均值等方法，重新计算每个类的质心

④迭代计算质心：重复第二步和第三步，迭代计算

⑤聚类完成：聚类质心不再发生移动

图 1 为 K-Means 聚类过程示意图。

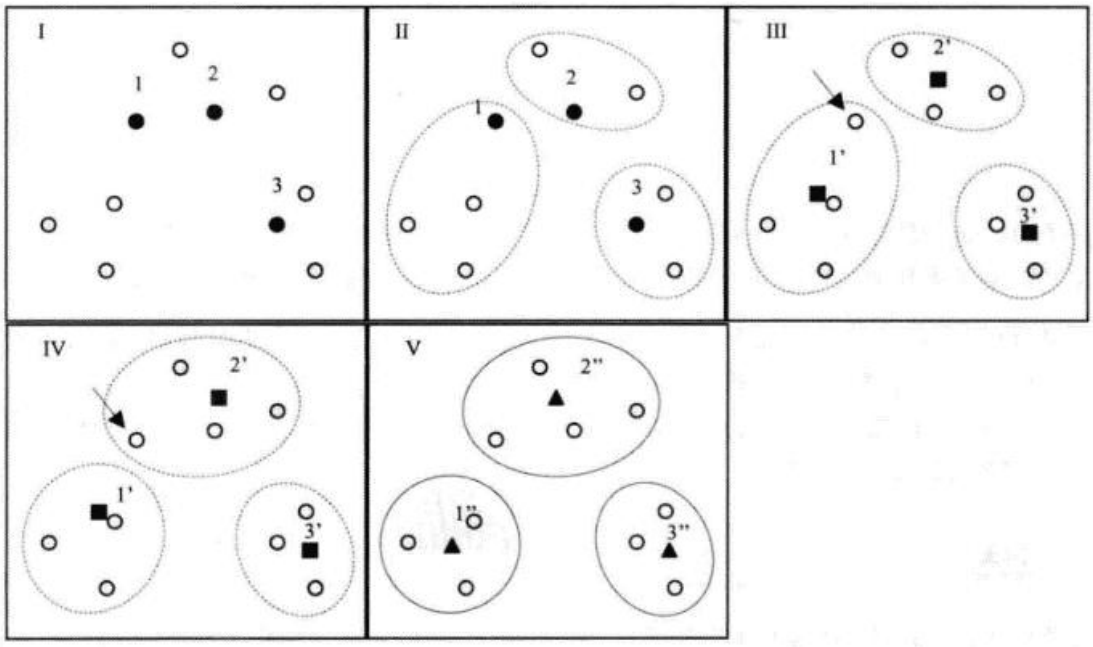

图 1 K-Means 聚类过程示意图

三、 使用 R 进行聚类分析

根据表 11 的数据，使用 K-Means 聚类分析方法探究助教的类型及各类型的特征。

下面使用 R 中的 fpc 包的 pamk()函数探索聚类的数量 K，用 kmeans()函数进行聚类，并查看聚类结果和 K 个簇的中心；然后将聚类中心结果写入本地 csv 文件中，用 excel 雷达图来描述聚类因子特征，找出每类助教的的特征；最后，把聚类结果添加到原始数据中，对各助教属于的类别进行标识。

3.1 建立数据集文件并导入到 R 中

R 可以接受多种格式的文件，其中包括 Excel 文件。在 Excel 中将录入表 1 中的数据，并存储为 TA.xlsx。**注意**：可以直接使用数字资源中的 TA.xlsx 文件，表示各项信息的标题用英文表示。

【操作 1】在 RStudio 中设置自己的工作目录，如“D:\MyR”。直接使用数字资源 TA.xlsx，首先将该文件拷贝到工作目录下，然后导入到 R 中的 TAdata 对象中。选择包含要参与聚类的属性集，并进行数据标准化处理。

①在 RStduio 的编辑窗口中执行下面的代码设置工作目录，单击 run 依次执行每一条代码（#后面的内容为对代码的解释）：

```
getwd()              #查看当前的工作目录
dir.create("D:/MyR")#创建你自己的工作目录，如果已经有此文件夹，忽略此步骤
setwd("D:/MyR")      #将默认工作目录设置为你的工作目录
getwd()              #查看现在的工作目录
```

②将 TA.xlsx 拷贝到默认工作目录“D:\MyR”下。

③在 RStudio 的脚本编辑窗口中输入下面代码后，单击 run 依次执行每一条代码：

```
install.packages("readxl")          #安装导入 excel 文件的包“readxl”
library(readxl)                     #加载“readxl”包
TAdata<-read_excel("TA.xlsx")       #将 TA.xlsx 的数据导入到 TAdata 对
```

象中

注意：代码中的 TAdata 是变量，“<-”是 R 中的符号，可以理解为“=”号。后面跟着的是一个函数 read_excel，括号内是要导入的 excel 文件。由于我们已经设置了默认工作目录，因此，此时会直接到“D:\MyR”下导入 TA.xlsx 文件。也可以通过绝对路径导入该文件。

④单击 RStudio 的右上角“Environment”窗口中出现的 TAdata，可以查看导入的数据。

⑤使用除 TA_ID 以外的属性进行聚类，并对数据进行标准化处理（数据标准化处理方法很多，此处使用 scale 函数）。在 RStudio 的脚本编辑窗口中输入下面代码，单击 run 依次执行每一条代码：

```
TAd<-TAdata[,2:8]          #使用除 TA_ID 以外的属性进行聚类
TAd=scale(TAd)       #进行 kmeans 聚类前，使用 scale 函数进行数据标准化处理，
```

⑥单击 RStudio 的右上角“Environment”窗口中出现的 TAd，比较 TAd 与 TAdata 数据集有什么不同。

3.2 使用 R 中的函数进行聚类分析

在 R 语言，可以使用内置的 kmeans 函数进行 K-Means 聚类分析。 kmeans 函数的使用语法为：

```
kmeans(x, centers, iter.max = 10, nstart = 1,algorithm = c("Hartigan-Wong", "Lloyd", "Forgy","MacQueen"), trace=FALSE)
```

其中，

x 为进行聚类分析的数据集；

centers 为预设类别数 k；

iter.max 为迭代的最大值，且默认值为 10；

nstart 为选择随机起始中心点的次数，默认取 1；

algorithm 则提供了 4 种算法选择，默认 Hartigan-Wong 算法；

trace 为日志打印选项，默认为 False，即不打印。

【操作 2】对 Tad 数据集进行聚类及并查看聚类结果。

①用 fpc 包中的 pamk()函数探索 K-Means 聚类簇的数量 K。在 RStudio 的脚本编辑窗口中输入如下代码：

```
install.packages("fpc")     #安装 fpc 包
library(fpc)                #加载 fpc 包
result=pamk(TAd)            #调用 pamk 函数，结果放到 result 中
result$nc                   #显示 result 中的 nc 值，即聚类簇的数量
```

单击 run 依次执行每一条代码，最后显示的结果是 2。说明建议将助教分为两类。

②使用 kmeans 函数进行聚类。在 RStudio 的脚本编辑窗口中输入如下代码：

```
kmd=kmeans(TAd,centers = 2,iter.max = 100)  #进行聚类，两类，迭代次数 100
center=kmd$centers                          #聚类中心结果存储到 center 变量中
type=kmd$cluster                            #输出聚类结果
center                                      #显示聚类中心
table(type)                                 #输出聚类结果
```

单击 run 依次执行每一条代码，在 RStudio 的窗口中首先看到 2 个类的中心信息，如图 2 所示。

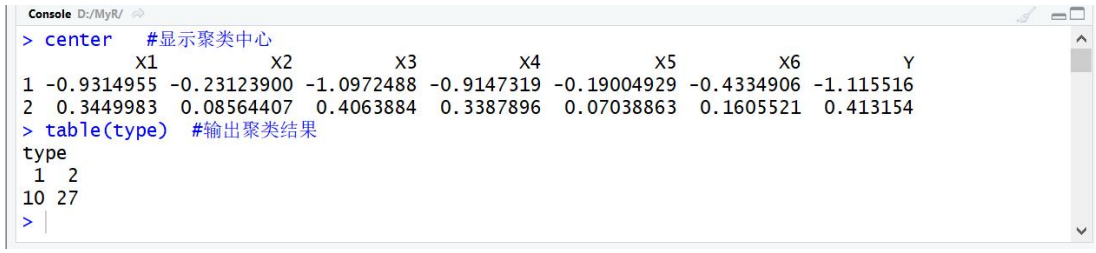

图 2 K-Means 聚类的两个中心和聚类结果

从图 2 中可以看出，两类的中心，第一类有 10 人，第 2 类有 27 人。

3.3 聚类结果分析

【操作 3】将聚类中心结果写入 excel 文件中，然后用 excel 雷达图来描述聚类因子特征，找出每类的优势特征和劣势特征。

①将聚类中心结果写入工作目录下的 TAcenter.csv 文件中。

```
write.csv(center, 'TAcenter.csv')
```

②用 Excel 打开 TAcenter.csv，选择所有数据，然后插入雷达图图表，如图 3 所示。

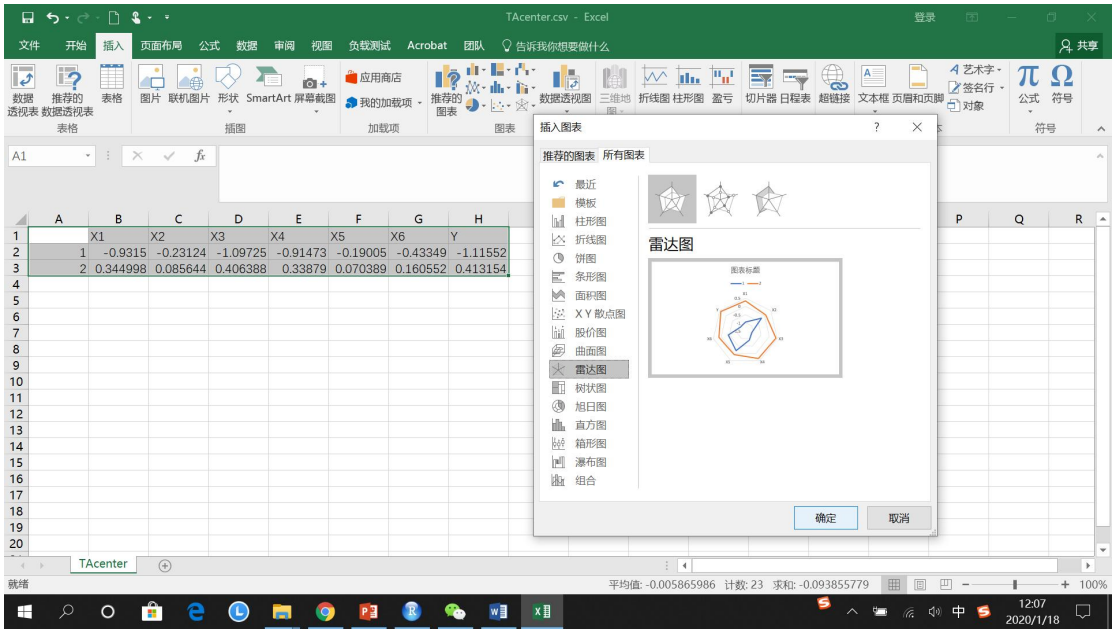

图 3 在 Excel 中做聚类中心的雷达图

单击图 3 中的“确定”后，生成两类助教特征的雷达图。用户可以编辑“图表标题”信息、还可以移动图中元素的位置，设置文字大小等，结果如图 4 所示。

从图 4 中可以看出，第 1 类助教在 X2（上机辅导）、X5（批改作业贡献因素）比第 2 类助教略低，其他各项贡献均明显低于第 2 类。第 2 类助教在 X1 至 X6 各方面均比较均衡，且都高于第 1 类助教。学生对第 2 类助教的综合满意程度 Y 也远远高于第 1 类助教。

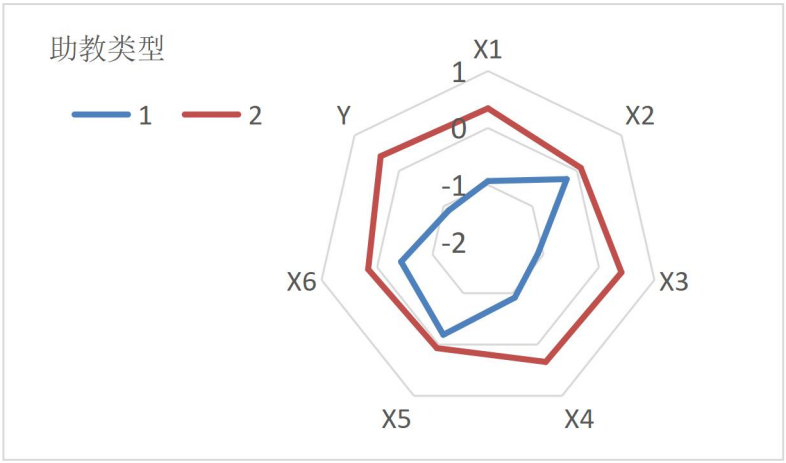

图 4 两类本科助教特征图

聚类结果说明，第 1 类助教（10 人，约占 27%）主要工作是上机辅导和批改作业，与学生面对面接触如课后答疑、讨论等很少，X1（每周工作时间）也远远小于第 2 类助教。第 2 类助教（27 人，约占 73%）能够在各方面对学生提供帮助，学生给出的 Y 值(综合满意度)也较高。可以看出，第 1 类助教属于不合格的助教，教学管理部门和主讲老师应注意对此类助教的筛选，同时加强监管。