

**《多元统计分析》实验报告（二）**

姓名-学号

庞浩哲 201813870023

曹行健 201813870028

杨钦 201813870052

龙福渊201813230008

指导教师 来 鹏

**二Ｏ二Ｏ年 四月二日**

一 摘要

工业指标是经济指标中重要一项，在我国第二产业占比极大。所以工业指标往往能够反映经济社会发展状况。

聚类算法往往对分布不加太多假设，有一点“非参”的感觉，因此也有了很广阔的应用前景，算法对大样小样都有很适合的模型。

因为算法是无监督的，我们也可以通过聚类看到一点未曾思考过的规律，而对新的规律的解释往往是科学的开始

关键词：industry，heirarchical-clustering，k-medios，DBSCAN

二 引言

聚类分析作为一种无指导无监督的机器学习.

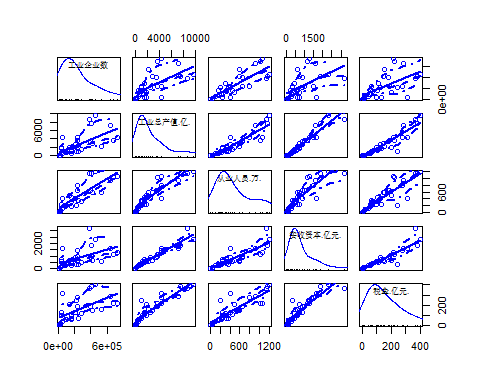
聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类分析的目标就是在相似的基础上收集数据来分类

本文通过工业企业数、工业总产值、从业人员年末人数、实收资本、本年应交税金总额五项指标，研究全国省市之间的分类特征，分析数据所体现的工业企业现状。

也利用不同的算法，研究其计算复杂度和聚类结果。

三 描述性统计

1针对全体数据



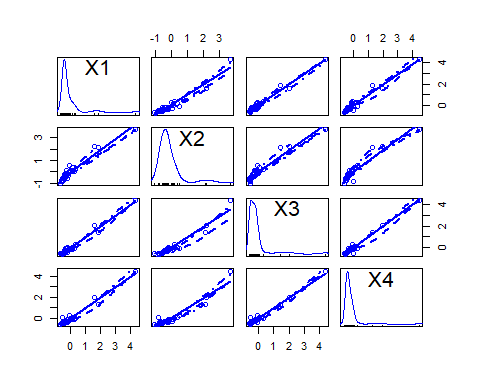
## 工业企业数 工业总产值.亿. 从业人员.万. 实收资本.亿元.   
## Min. : 7411 Min. : 7.78 Min. : 3.69 Min. : 12.99   
## 1st Qu.: 57086 1st Qu.: 803.26 1st Qu.: 209.22 1st Qu.: 334.83   
## Median :188234 Median :1745.83 Median : 386.49 Median : 592.39   
## Mean :241994 Mean :2683.99 Mean : 478.91 Mean : 827.20   
## 3rd Qu.:394841 3rd Qu.:3722.92 3rd Qu.: 657.58 3rd Qu.:1102.79   
## Max. :732957 Max. :9645.30 Max. :1202.21 Max. :3210.65   
## 税金.亿元.   
## Min. : 0.90   
## 1st Qu.: 58.84   
## Median :128.05   
## Mean :154.79   
## 3rd Qu.:219.25   
## Max. :402.17

结论：

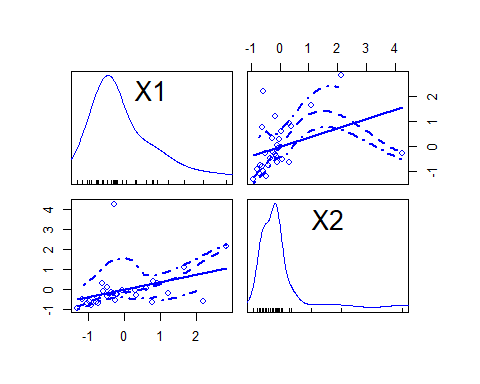
1由最大值最小值的差值和上分位但和下分位点的巨大差异，可以看出中国工业发展不平衡  
2平均值小于中位数，可见发达地区对总体影响大

2我们希望用多个视角看一组数据，由于数据是连续的，因此可以定义一些有实际经济意义的比值变量

2.1工厂规模（由各经济指标和工厂数量的比例定义）



#工厂效益由各经济指标和从业人数的比例定义）



结论：1左偏态严重，大多数地区不发达

2各指标相关性强（税金，企业数量）

四 层次聚类：Ward

引言：

层次聚类是一种贪婪的适合小样本的算法，能够比较好的划分层次结构。比如此例中，分成两类，就是以腾冲--漠河一线为界（工业发展与地形人口有着密切关联）。如果分地更细，会看到长三角·珠三角被划分在一起，作为腾冲--漠河以东的重要一类，这也是我国经济最发达的地区。

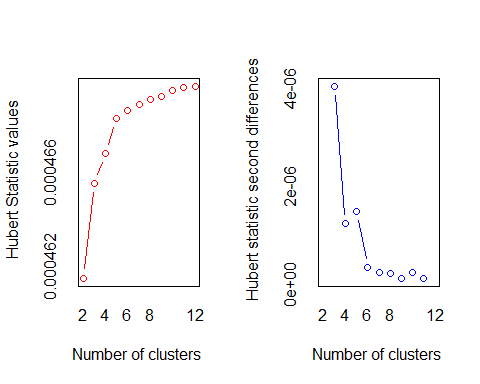
Ward方法也是聚类效果较好且鲁棒性较好的算法。

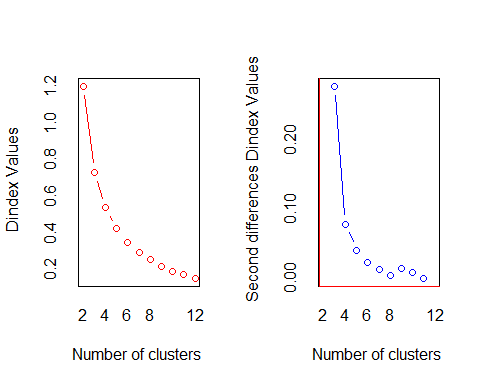
（上图为聚四类结果）

注：



4.1评价聚类数目  
！！！一种“新”的计算方法：加入总体（一个很大的项，如果舍弃，计算遭遇奇异矩阵），我们希望通过加入总体来得出判别结果（每次总体都被归于一类，故其本质上不参与系统聚类过程）





综合NbCluster评价结果分两类或者三类结果较好（9个指标支持2类，3个指标支持3类.且Hubert统计量也在2，3之间增长明显）结合期望聚类个数，我们倾向于聚三类。

|  |  |
| --- | --- |
| 类别 | 地区 |
| 1 | 北京，天津，山西，内蒙古，吉林，黑龙江，江西，广西，海南，贵州，西藏，陕西，甘肃，青海，宁夏，新疆 |
| 2 | 江苏，浙江，山东，广东 |
| 3 | 河北，辽宁，上海，安徽，福建，河南，湖北，湖南，四川，云南 |

4.2检查几种不同视角下，聚类结果相似程度

1:各地总的体量：2：各地工业企业规模：3：各地工业企业经济效益

各次分类两两之间（同一地区被分到不同类的个数/地区总数）的平均#注：我们一直把好的类放在后面（体量大，效率高）#我们发现s值接近0.4，这是一个很大的数

## Group.1 工业企业数 工业总产值.亿. 从业人员.万. 实收资本.亿元. 税金.亿元.  
## 1 1 7259822.0 80519.610 14367.2300 24815.9900 4643.62000  
## 2 2 122426.5 1165.508 252.1284 426.1916 89.41158  
## 3 3 461709.0 4062.921 758.3600 1164.1862 224.40000  
## 4 4 413348.7 8623.863 1169.9667 2468.2867 383.20667

aggregate(dq,by=list(tt1),mean)

## Group.1 工业企业数 工业总产值.亿. 从业人员.万. 实收资本.亿元. 税金.亿元.  
## 1 1 33751.0 1745.825 235.4450 660.8700 117.190  
## 2 2 264876.4 2692.112 499.4578 809.2981 151.637  
## 3 3 40656.0 4340.950 410.9700 1643.2000 315.060

aggregate(xy,by=list(tt2),mean)

## Group.1 V1 V2  
## 1 1 1.4786432 0.3677072  
## 2 2 -0.4579596 -0.3100568  
## 3 3 -0.2753904 4.2473002

分析差异大的原因：

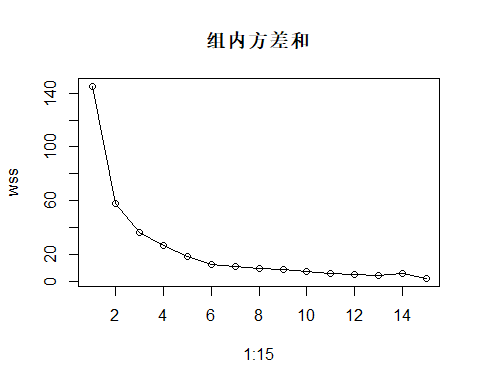
结论：经分析，我们发现，北京天津上海等地区，虽然总体体量不大，但是效益却很好。山东，东北老工业基地，大企业多，体量也不小，但是效率不高。因此，这些样品在三次聚类中归类不同

五 划分聚类

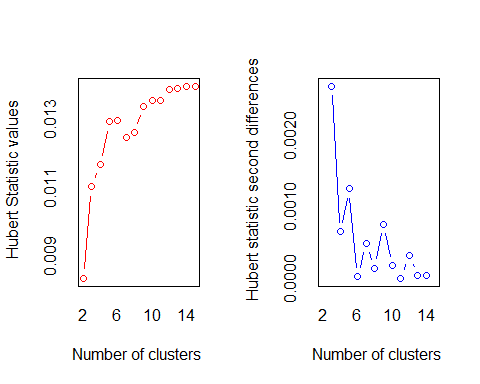
5.1 Kmeans

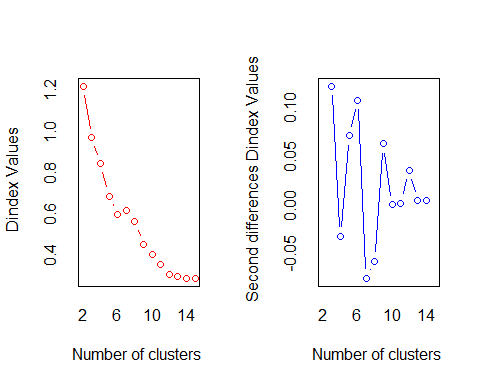
1因为Kmeans对初始值点敏感，因此选定25组初始值点，选取最好一个结果

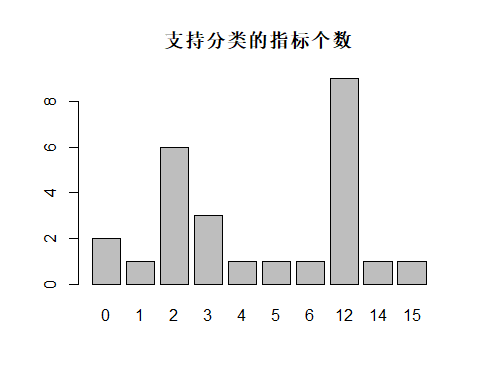
2选定聚类数量



由图看出两到三组，组间方差下降快速





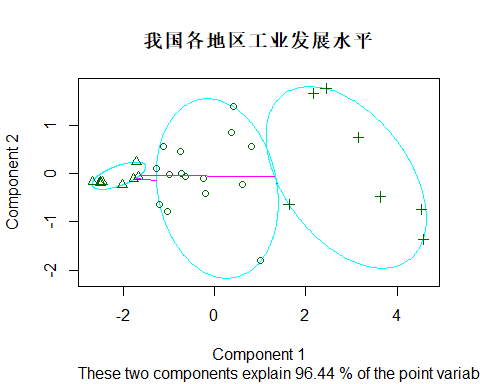


支持两组，三组的比较多，选定三组簇中心:  
## 工业企业数 工业总产值.亿. 从业人员.万. 实收资本.亿元. 税金.亿元.  
## 1 0.7924487 0.3595978 0.6625851 0.3571937 0.4928844  
## 2 -0.5945951 -0.6339786 -0.6739588 -0.5916345 -0.6118077  
## 3 1.0907806 2.1337083 1.7076445 1.9479680 1.7673657  
##   
分类结果:  
## 北京 天津 河北 山西 内蒙古 辽宁 吉林 黑龙江 上海 江苏 浙江   
## 2 2 1 2 2 1 2 2 1 3 3   
## 安徽 福建 江西 山东 河南 湖北 湖南 广东 广西 海南 四川   
## 1 2 2 3 1 1 1 3 2 2 1   
## 贵州 云南 西藏 陕西 甘肃 青海 宁夏 新疆   
## 2 2 2 2 2 2 2 2   
##   
## Within cluster sum of squares by cluster:  
## [1] 15.637514 14.485931 5.222456  
## (between\_SS / total\_SS = 75.6 %)

5.2 PAM

PAM是动态聚类，与kmeans不同，它以代表性样本点为类中心，结果唯一，适合小样，以靠近簇中心的点为代表中心的，对待异常值有好的鲁棒性，但是聚类效果不一定好于kmeans。

PAM簇中心：  
## ID 工业企业数 工业总产值.亿. 从业人员.万. 实收资本.亿元. 税金.亿元.  
## 山西 4 -0.3766567 -0.4110990 -0.07330535 -0.3237123 -0.4242518  
## 宁夏 29 -1.0316533 -0.9626633 -1.18219851 -1.0077011 -1.2107480  
## 浙江 11 1.9257737 1.7140572 1.21794877 1.0650336 1.1482224  
PAM类别：  
## 北京 天津 河北 山西 内蒙古 辽宁 吉林 黑龙江 上海 江苏 浙江   
## 1 1 1 1 2 3 1 1 1 3 3   
## 安徽 福建 江西 山东 河南 湖北 湖南 广东 广西 海南 四川   
## 1 1 1 3 3 1 1 3 1 2 3   
## 贵州 云南 西藏 陕西 甘肃 青海 宁夏 新疆   
## 2 1 2 1 2 2 2 2   
 主成分图



Silhouette统计量



分析：1两个成分解释了96.44%的可能性，可见数据有明显的共线性，

2由图看出分得不太开，Silhouette值也不太大，也可见聚类结果不很好。

3可见数据的分布基本是球状的。

5.3比较PAM和Kmeans算法的差异

## RI   
## 0.4473468

分析：

pam效果和Kmeans不太一致，选择kmeans(因为没有标准答案故这里使用兰德指数而非调整兰德指数)

六 基于密度的聚类分析

DBSCAN

引言：

前边的k-Means和k-Mediods算法比较适用于簇为球型的，对于非球型的，一般需要基于密度的聚类，比如DBSCAN。

且DBSCAN不需要划分个数。

缺点是维数太大时会造成维数灾难，不过此例中明显不必担心。

6.1

主成分聚类图



6.2 比较密度聚类和kmeans

## RI   
## 0.8479408

结论：我们发现两者基本一致，这里基于密度的聚类结果也是不错的。

可见DBSCAN算法对球形数据效果也不错。

(因为没有标准答案故这里使用兰德指数而非调整兰德指数)

七 总结与收获：

1我们清楚了各种聚类算法的优劣（复杂度，鲁棒性，适用的分布类型）。

2明白了如何用各种统计量去评估结果好坏。

3知道如何设定的聚类算法和算法参数，以获得更好的结果。

4在方法论上知道了如何设计实验，从哪里获取优质的数据。

5明白了如何利用数据特点，看到更多信息（连续型数据的使用）

6在聚类前学习一些有关领域的先验知识作为指导

7熟悉了如何用R语言去实现算法，并交互式的展示分享我们的结果

\*最大的疑惑

聚类分析作为一种无指导无监督的机器学习，我们难解释最终结果，且各个指标在计算样本点距离前被标准化了，但是它们的影响大小（提供的信息大小）不应该被粗略的认为是一样的，我们希望清楚（针对不同的目的，比如重视就业或者税收）该如何给各个指标赋权

Appendix

1计算数据

数据来源：国家统计局<https://www.cnblogs.com/think90/p/7133753.html>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 工业企业数 | 工业总产值(亿） | 从业人员（万） | 实收资本（亿元） | 税金（亿元） |
| 全国总计 | 7259822 | 80519.61 | 14367.23 | 24815.99 | 4643.62 |
| 北京 | 33217 | 1624.91 | 234.42 | 733.56 | 132.53 |
| 天津 | 34285 | 1866.74 | 236.47 | 588.18 | 101.85 |
| 河北 | 227531 | 3646.32 | 651.12 | 1002.18 | 187.1 |
| 山西 | 162553 | 1609.97 | 451.83 | 590.28 | 105.69 |
| 内蒙古 | 97968 | 677.4 | 214.83 | 365.8 | 54.24 |
| 辽宁 | 236437 | 4435.26 | 801.25 | 1485.18 | 275.95 |
| 吉林 | 162226 | 1242.88 | 330.95 | 568.06 | 102.39 |
| 黑龙江 | 135083 | 1914.55 | 494.06 | 705.71 | 204.49 |
| 上海 | 40656 | 4340.95 | 410.97 | 1643.2 | 315.06 |
| 江苏 | 453857 | 9645.3 | 1202.21 | 2327.83 | 402.17 |
| 浙江 | 648161 | 7162.05 | 928.79 | 1606.68 | 287.67 |
| 安徽 | 418746 | 2833.53 | 598.46 | 698.07 | 144.73 |
| 福建 | 213914 | 2611.62 | 411.83 | 878.38 | 123.57 |
| 江西 | 243888 | 1138.63 | 362.01 | 329.81 | 63.7 |
| 山东 | 406779 | 7554.17 | 1159.56 | 1866.38 | 381.46 |
| 河南 | 732957 | 4296.57 | 1055.55 | 1272.09 | 213.18 |
| 湖北 | 399985 | 3510.79 | 659.73 | 877.45 | 182.88 |
| 湖南 | 531174 | 2175.77 | 588.27 | 594.5 | 154.46 |
| 广东 | 379410 | 8672.12 | 1148.13 | 3210.65 | 365.99 |
| 广西 | 284083 | 1490.21 | 308.67 | 472.06 | 114.78 |
| 海南 | 17315 | 167.59 | 38.71 | 95.56 | 9.02 |
| 四川 | 685556 | 3748.45 | 1023.86 | 1136.32 | 221.27 |
| 贵州 | 161221 | 492.06 | 169.92 | 244.52 | 57.22 |
| 云南 | 220136 | 1048.25 | 207.35 | 349.91 | 243.85 |
| 西藏 | 7411 | 7.78 | 3.69 | 12.99 | 0.9 |
| 陕西 | 152973 | 1038.98 | 316.25 | 426.47 | 70.76 |
| 甘肃 | 84708 | 724.69 | 177.69 | 296.06 | 56.58 |
| 青海 | 15307 | 125.2 | 32.28 | 75.34 | 11.66 |
| 宁夏 | 24407 | 168.98 | 42.23 | 89.68 | 14.67 |
| 新疆 | 47878 | 547.9 | 106.13 | 273.09 | 43.82 |

2 代码实现

idu<-read.csv(choose.files(),header = T,row.names = "r")  
library(car)

## Loading required package: carData

scatterplotMatrix(idu[-1,])

summary(idu[-1,])

#工厂规模  
gm<-matrix(ncol = 4,nrow = 30)  
#工厂效益  
xy<-matrix(ncol = 2,nrow = 30)  
for(i in 2:5 )  
{gm[,i-1]<-dq[,i]/dq[,1]}  
xy[,1]<-dq[,4]/dq[,3]  
xy[,2]<-dq[,5]/dq[,3]  
gm<-scale(gm);xy<-scale(xy)  
scatterplotMatrix(gm)

scatterplotMatrix(xy)

row.names(xy)<-row.names(idu[-1,])  
row.names(gm)<-row.names(idu[-1,])

d<-dist(scale(idu))  
d1<-dist(gm)

aggregate(idu,by=list(tt),mean)

nc<-NbClust(d,distance = "euclidean",min.nc = 2,max.nc = 12,method = "ward.D")

tt1<-cutree(c1,k=3)  
tt2<-cutree(c2,k=3)

il<-table(nc$Best.nc[1,])  
barplot(il,main="支持分类的指标个数")

nc<-NbClust(dq,min.nc=2,max.nc=15,method = "kmeans")

## Warning in pf(beale, pp, df2): 产生了NaNs

rect.hclust(hc, k=3)

d2<-dist(xy)  
c<-hclust(d,method = "ward.D")  
c1<-hclust(d1,method ="ward.D" )  
c2<-hclust(d2,method = "ward.D")  
par(cex=.1,mar=c(2,2,2,2))  
hc<-c  
plot(c,hang=-1)

library(NbClust)

s=0  
for(i in 1:30 )  
{if(tt[i]!=tt1[i])s<-s+1;}  
#if(tt[i]!=tt2[i])s<-s+1;if(tt1[i]!=tt2[i])s<-s+1}  
s<-s/3/30  
cc2<-kmeans(dq,3,nstart = 25)  
cc2

dq<-scale(dq)  
wss<-(nrow(dq)-1)\*sum(apply(dq,2,var))  
for(i in 2:15)  
{wss[i]<-kmeans(dq,i)$tot.withinss  
}  
plot(1:15,type='o',wss,main="组内方差和"

library(cluster)  
pam(dq,3,stand =T)

cc<-pam(dq,3,stand=T)  
clusplot(cc,main = "我国各地区工业发展水平")

#install.packages("flexclust")  
library(flexclust)

cc<-table(cc2$cluster,cc$clustering)  
randIndex(cc)

library(fpc)

ds<-dbscan(dq,eps=.42,MinPts=3)

plotcluster(dq,ds$cluster)

3 参考文献

[1] <https://www.cnblogs.com/think90/p/7133753.html>

[2]王学明.应用多元分析[M].上海：上海财经大学出版社，1999.8

[3](美)卡巴科弗(Kabacoff,R.I)著.R语言实战[M].高涛，肖楠，陈钢译.北京：人民邮局出版社，2013.1