Algoritmo PAM

Omar Sanchez Hernandez

4/6/2022

Introducción

El algoritmo ***PAM*** es una mejora del algoritmo K-Means para resolver el problema de la interferencia de ruido. El enfoque adoptado por este problema es la mejora lograda al igualar el peso de cada dato para lograr asi un mejor agrupamiento de los datos.

En el presente ejercicio se llevara a cabo el desarrollo de esta tecnica usando la matriz de datos state.x77 que se encuentra en la libreria de datasets que contiene informacion sobre indicadores de 50 estados de Estados Unidos.

Cargar la matriz de datos

```
library(knitr)
library(cluster)
X<-as.data.frame(state.x77)
kable(X)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	0	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	126	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	122	36097
Iowa	2861	4628	0.5	72.56	2.3	59.0	140	55941
Kansas	2280	4669	0.6	72.58	4.5	59.9	114	81787
Kentucky	3387	3712	1.6	70.10	10.6	38.5	95	39650
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12	44930
Maine	1058	3694	0.7	70.39	2.7	54.7	161	30920

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Maryland	4122	5299	0.9	70.22	8.5	52.3	101	9891
Massachusetts	5814	4755	1.1	71.83	3.3	58.5	103	7826
Michigan	9111	4751	0.9	70.63	11.1	52.8	125	56817
Minnesota	3921	4675	0.6	72.96	2.3	57.6	160	79289
Mississippi	2341	3098	2.4	68.09	12.5	41.0	50	47296
Missouri	4767	4254	0.8	70.69	9.3	48.8	108	68995
Montana	746	4347	0.6	70.56	5.0	59.2	155	145587
Nebraska	1544	4508	0.6	72.60	2.9	59.3	139	76483
Nevada	590	5149	0.5	69.03	11.5	65.2	188	109889
New Hampshire	812	4281	0.7	71.23	3.3	57.6	174	9027
New Jersey	7333	5237	1.1	70.93	5.2	52.5	115	7521
New Mexico	1144	3601	2.2	70.32	9.7	55.2	120	121412
New York	18076	4903	1.4	70.55	10.9	52.7	82	47831
North Carolina	5441	3875	1.8	69.21	11.1	38.5	80	48798
North Dakota	637	5087	0.8	72.78	1.4	50.3	186	69273
Ohio	10735	4561	0.8	70.82	7.4	53.2	124	40975
Oklahoma	2715	3983	1.1	71.42	6.4	51.6	82	68782
Oregon	2284	4660	0.6	72.13	4.2	60.0	44	96184
Pennsylvania	11860	4449	1.0	70.43	6.1	50.2	126	44966
Rhode Island	931	4558	1.3	71.90	2.4	46.4	127	1049
South Carolina	2816	3635	2.3	67.96	11.6	37.8	65	30225
South Dakota	681	4167	0.5	72.08	1.7	53.3	172	75955
Tennessee	4173	3821	1.7	70.11	11.0	41.8	70	41328
Texas	12237	4188	2.2	70.90	12.2	47.4	35	262134
Utah	1203	4022	0.6	72.90	4.5	67.3	137	82096
Vermont	472	3907	0.6	71.64	5.5	57.1	168	9267
Virginia	4981	4701	1.4	70.08	9.5	47.8	85	39780
Washington	3559	4864	0.6	71.72	4.3	63.5	32	66570
West Virginia	1799	3617	1.4	69.48	6.7	41.6	100	24070
Wisconsin	4589	4468	0.7	72.48	3.0	54.5	149	54464
Wyoming	376	4566	0.6	70.29	6.9	62.9	173	97203

Sobre la matriz de datos

```
dim(X)
```

```
## [1] 50  8
```

```
str(X)
```

```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

La matriz de datos esta compuiesta por 50 observaciones y 8 variables numericas.

Transformacion de las variables x1,x3 y x8 con la funcion de logaritmo

Se aplica el logaritmo a esas variables con la finalidad de reducir los valores de la escala y que esto no resulte tan influyente al momento de aplicar el metodo PAM.

```
X[,1]<-log(X[,1])
colnames(X)[1]<-"Log-Population"

X[,3]<-log(X[,3])
colnames(X)[3]<-"Log-Illiteracy"

X[,8]<-log(X[,8])
colnames(X)[8]<-"Log-Area"
```

Metodo PAM

Estandarizacion univariante.

Se realiza una estandarización sobre los datos con la finalidad de homogeneizar la escala de los datos sin perder la información de los sujetos.

```
n<-dim(X)[1]
p<-dim(X)[2]
X.s<-scale(X)
```

Aplicacion del algoritmo y sus clusters

```
pam.3<-pam(X.s,3)
cl.pam<-pam.3$clustering
cl.pam
```

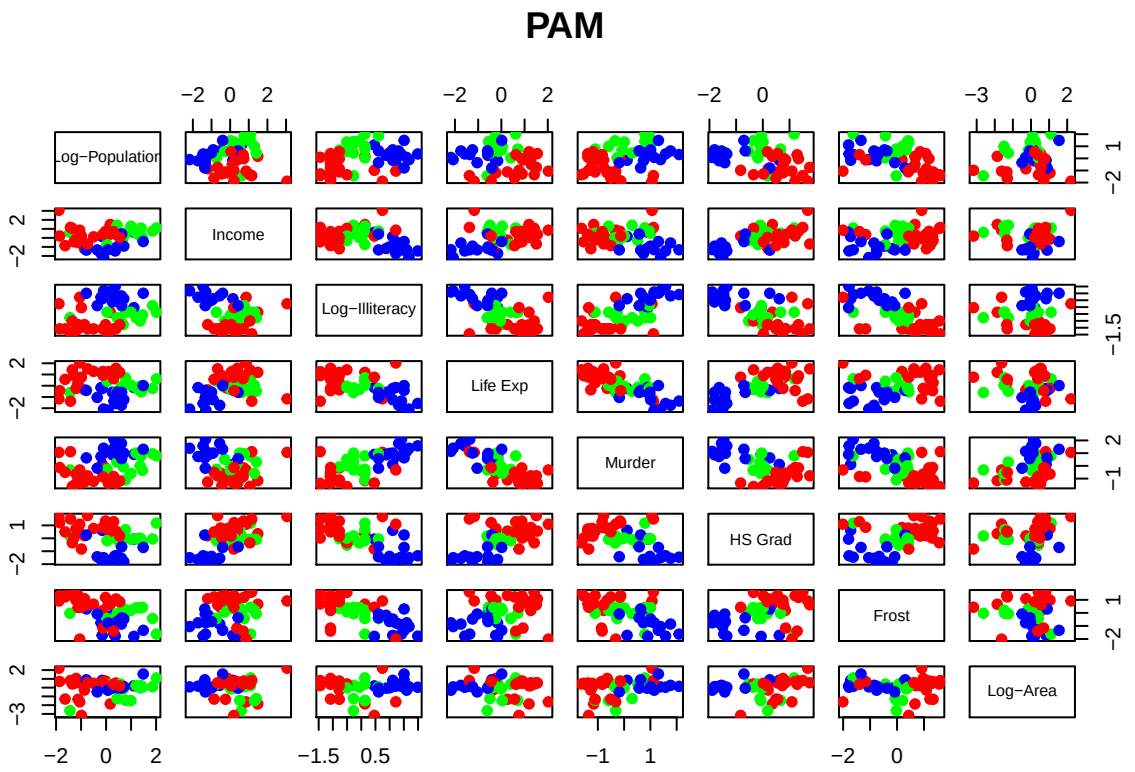
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              2              3              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              3              3              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              1              1              2              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              3              2              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              3
```

```
##      New Mexico        New York North Carolina   North Dakota           Ohio
##               1               3              1              2              3
##        Oklahoma          Oregon   Pennsylvania   Rhode Island South Carolina
##               3               2              3              2              1
##    South Dakota       Tennessee          Texas           Utah        Vermont
##               2               1              1              2              2
##        Virginia      Washington  West Virginia      Wisconsin        Wyoming
##               1               2              1              2              2
```

Se obtienen las etiquetas que el algoritmo le asigna a cada sujeto.

Scatter plot de la matriz con los grupos

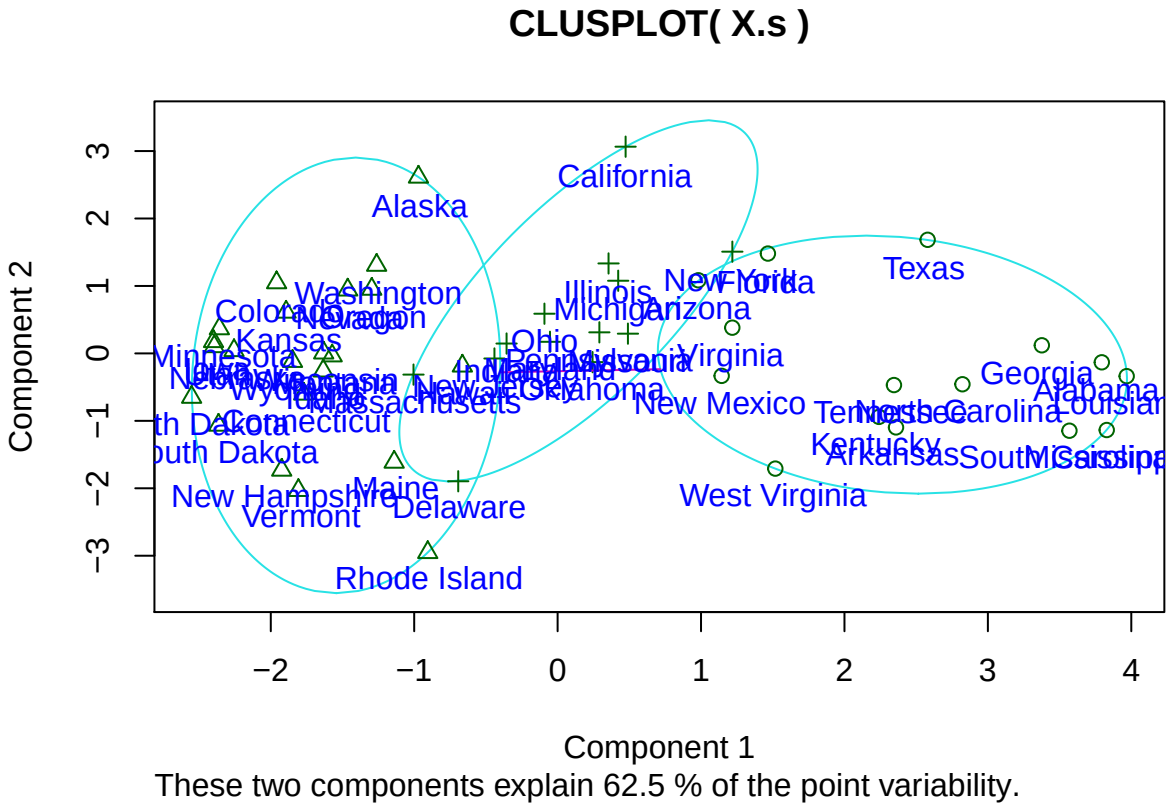
```
col.cluster<-c("blue","red","green")[cl.pam]
pairs(X.s, col=col.cluster, main="PAM", pch=19)
```



En el scatter plot se puede observar el comportamiento de los sujetos de manera bivariada pero estando identificados por la etiqueta que les asigno el algoritmo ***PAM***. De manera bivariada se observan traslapes entre los sujetos de cada grupo.

Visualizacion con Componentes Principales

```
clusplot(X.s,cl.pam)
text(princomp(X.s)$scores[,1:2],
     labels=rownames(X.s),pos=1, col="blue")
```



El agrupamiento que se consigue es ligeramente distinto al del algoritmo K-medias pero aun asi sigue logrando separar de una manera considerablemente bien a los sujetos.

Silhouette

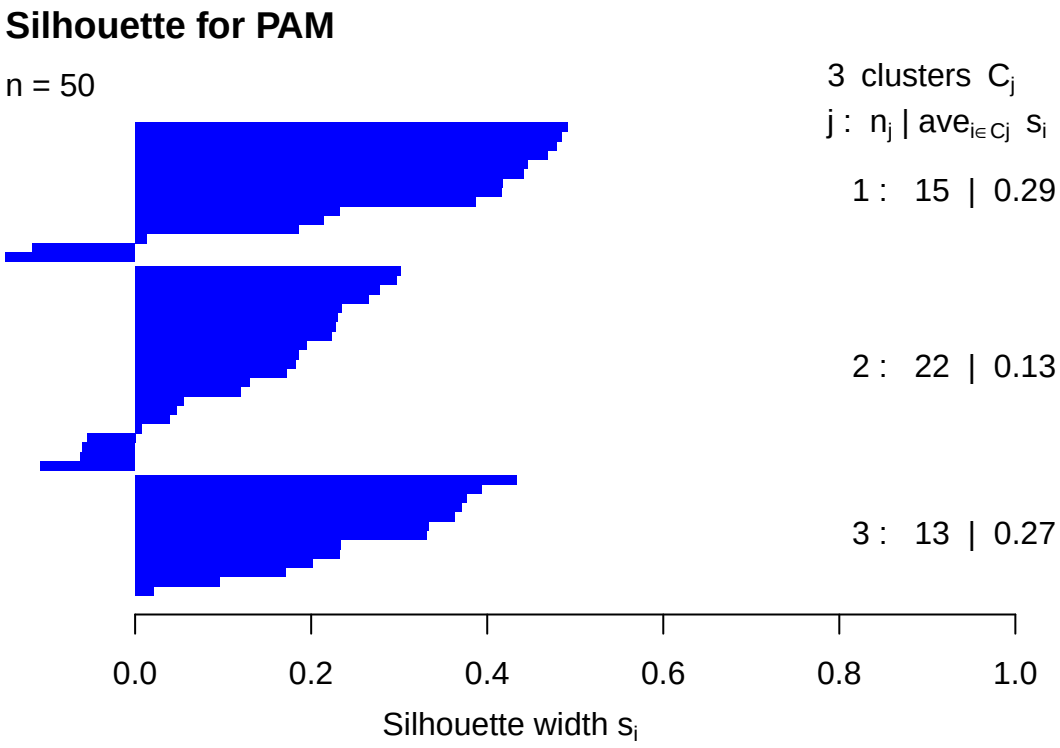
Representacion grafica de la eficacia de clasificacion de una observación dentro de un grupo.

Generacion de los calculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.pam<-silhouette(cl.pam, dist.Euc)
```

Generacion del gráfico

```
plot(Sil.pam, main="Silhouette for PAM",
     col="blue")
```



Average silhouette width : 0.22