Universidad de Granada
Escuela Internacional de Posgrado
Máster en Estadística Aplicada
Materia: Técnicas Estadísticas Multivariantes.
Alumno: Francisco Javier Márquez Rosales

Análisis Cluster:

Ejercicios:

Diciembre, 2022

1. Realizar un breve resumen de la teoría del Análisis Cluster, ocupando un máximo de 3 páginas.

El Análisis Cluster es una técnica estadística multivariante que se utiliza para agrupar objetos o individuos a modo de partición (cada objeto pertenece a un grupo y solamente a uno), los elementos más parecidos formen parte del mismo grupo y los elementos menos parecidos queden en grupos separados.
El punto de partida del análisis cluster es una matriz que contiene medidas de varias variables en un conjunto de individuos y el objetivo es descubrir la estructura de las categorías en las que encajan las observaciones.

Preparación de los datos de inicio

El punto de partida es un examen de los datos a analizar y la naturaleza de las variables de entrada ya que hay casos en los que son recomendables transformaciones en las mismas que mejoren la información a utilizar para el agrupamiento de los individuos. Estas transformaciones son, por ejemplo, la normalización de las variables cuantitativas o la conversión a dicotómicas de las variables categóricas nominales.

Tipo de Análisis de interés individuos o variables

Debe definirse si el agrupamiento buscado corresponde a individuos o variables, luego de esto se define la medida que se utilizará para decidir si los individuos se ubican en el mismos grupos, esta medida es una distancia o similaridad. Habitualmente se miden distancias en el caso de clasificación de individuos y similaridades en el caso de la clasificación de variables, sin embargo, cualquiera de las medidas puede aplicarse en los casos sólo considerando la matriz traspuesta.

Distancias entre Individuos

Podemos definir diferentes distancias entre las medidas de los individuos, las cuales representan la cercanía o lejanía de éstos. Las medidas más habituales son: Distancia euclídea, Distancia de Minkowski, Distancia de bloque, ciudad o Manhattan, Distancia de Chebychev, etc.

Distancias entre Variables

Es habitual utilizar medidas de similaridad basadas en la asociación o relación entre las Variables. En el caso de dicotómicas, se pueden obtener medidas de similaridad una vez se determina los conteos de las frecuencias de clase. Entre las más comunes están: Medida de Ochiai, Medida Φ, Medida de Russel y Rao, Parejas simples, etc.

Para el caso de variables ordinales, primero debe definirse y cuantificarse las parejas empatadas, parejas concordantes y parejas discordantes en tablas de doble entrada. Las parejas concordantes son aquellas parejas de individuos que presentan valores menores en una de las variables y mayores en otra (en las mismas). Las parejas discordantes son aquellas en las que uno de los individuos presenta valores menores en una variable mientras que el otro individuo presenta valores mayores en esa variable. En los pares empatados, al menos una de las variables no presenta valores mayores ni menores. Con base en estas definiciones se pueden obtener medidas tales como: Coeficiente τ de Kendall, Coeficiente de concordancia $\tau - c$ de Kendall y Coeficiente de correlación de Spearman.

Métodos Jerárquicos

Los métodos jerárquicos toman la información de la matriz de datos y la convierten en una matriz de distancias o similaridades, cuadrada y simétrica, donde se establece la medida existente entre cada dos objetos.

Los métodos jerárquicos se dividen en dos tipos: Aglomerativos y Disociativos. En los Aglomerativos cada observación se asigna a su propio clúster. Luego, se calcula la similitud (o distancia) entre cada uno de los clusters y los dos clusters más similares se fusionan en uno. Finalmente, los pasos 2 y 3 se repiten hasta que solo quede un grupo conteniendo todos los objetos inciales. En los disociativos se realiza el procedimiento antes explicado de forma inversa: partiendo de un solo cluster con todos los objetos, se llega a tantos clusters como objetos. Algunos de los métodos más usados:

Método del vecino más próximo (Linkage simple)

considera que la distancia entre dos grupos será la mínima que exista entre elementos de ambos grupos (en caso de similaridad, la máxima), es decir, se toma como distancia de todo el grupo la de aquellos objetos de los grupos que estén más cerca (los vecinos más cercanos).

Método del vecino más lejano (Linkage completo)

considera que la distancia entre dos grupos será la máxima (si hablamos de similaridad, la mínima) entre los elementos de ambos grupos.

Otros métodos

no consideran la distancia entre clusters como la mínima o la máxima de entre las medidas de sus respectivos objetos, sino que pueden promediar dichas medidas, de forma ponderada según el número de objetos que contenga cada grupo

o no ponderada, o establecer la medida entre clusters como la medida entres los centroides de cada cluster. Entre los más utilizados se encuentra el método de Ward y el de los Promedios.

Idoneidad del método

Pueden aplicarse criterios basados principalmente en alguno de los siguientes coeficientes:

El coeficiente de aglomeración: es una medida del ajuste de la aplicación de un método de clasificación jerárquico aglomerativo. se calcula realizando la media aritmética de todos los valores $1 - \mu(i)$. Este coeficiente toma valores entre 0 y 1. Cuanto más cercano al 1, mejor será el ajuste del método.

Coeficiente de correlación cofenético. Este coeficiente es un coeficiente de correlación lineal de Pearson entre dos matrices; la de medidas entre objetos inicial y la matriz cofenética. Esta ultima es la matriz que se va utilizando en las etapas del análisis sin reducir la dimensionalidad de la matriz original. Cuando se tengan valores cercanos a 1 o -1, será indicativo de que la distribución en grupos de los datos es adecuada.

Métodos No Jerárquicos

están diseñados para la clasificación de individuos y no de variables, en k grupos o clusters. La metodología básicamente consiste en realizar una partición y a partir de ella hacer una reasignación de individuos a clusters para obtener una partición mejor.

Método de las k-medias

El método de las k-medias consiste en partir de k semillas y asignar, según la proximidad de cada individuo a las semillas, cada individuo al cluster con semilla más cercana. Tras cada asignación, la semilla se recalcula como el centroide del cluster formado.

Otros métodos no jerárquicos

Además del K-medias, se encuentran métodos como Quick-Cluster analysis, método de Forgy o método de las nubes dinámicas. Siempre, un individuo asignado en un paso determinado a un cluster puede ser asignado
a otro grupo en un paso posterior si con ello se consigue una mejor clasificación.

2. Utilizando el fichero de datos del tema (de los dos archivos, el que contiene más información), donde se recogen los valores para las comunidades autónomas, acerca de las aficiones culturales de los habitantes, realizar una selección de entre 5 y 7 variables para usarlas en un análisis cluster que permita agrupar las comunidades autónomas, aplicando al menos 5 métodos diferentes entre los cuales debe haber al menos uno jerárquico y uno no jerárquico. Seleccionar el mejor método en base al coeficiente de correlación cofenético. Interpretar en un pequeño informe los resultados obtenidos. Esta parte debe ocupar un máximo de 10 páginas.

En primer lugar, hacemos la lectura de la data con la siguiente sintaxis:

```
data<-read.csv("datos_culturales2.csv",header = T)
names (data)[1] = "comunidad"
```

Examinamos preliminarmente los datos:

```
str(data)
```

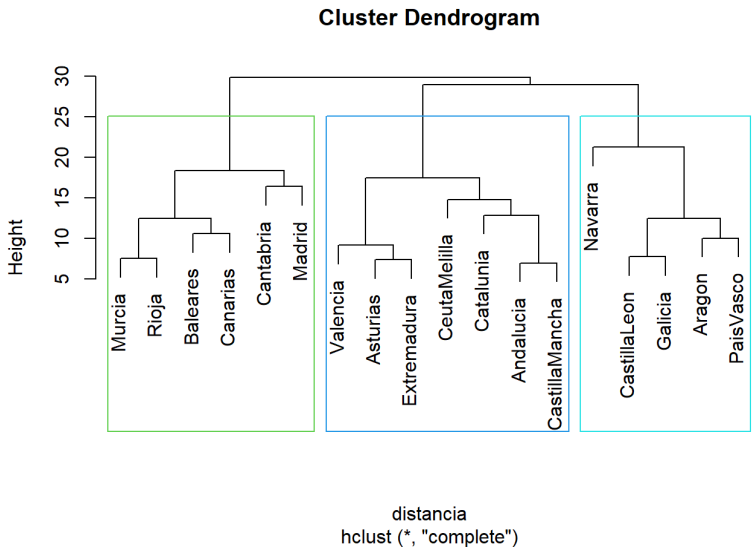
```
## 'data.frame':    18 obs. of  7 variables:
##  $ comunidad: chr  "Andalucia" "Aragon" "Asturias" "Baleares" ...
##  $ libros   : num  88.4 96.4 90.7 95.1 91.6 96.9 93.9 92.3 85.7 94.2 ...
##  $ movil    : num  85 85.5 84 88.3 86.5 85.8 77.5 84.3 77.9 83.9 ...
##  $ concierto: num  8.3 9.7 9 9.3 6.6 8.3 9.2 11.4 7 10.1 ...
##  $ radio    : num  76.7 82.8 70 76.4 84.3 77.3 76.5 76.6 72.9 71.2 ...
##  $ cine     : num  40.4 34.9 28.9 43.2 43.9 49.6 34.2 40.3 42.4 35.7 ...
##  $ deporte  : num  36.3 33.3 34.7 29.5 26.4 19.4 30.8 32.3 36.7 31.5 ...
```

Vemos que todos los datos corresponden a los porcentajes de las aficiones culturales de los habitantes de las comunidades, por lo que las unidades pueden ser comparables. No es necesaria ninguna transformación de los datos previa a la aplicación del análisis.

Iniciamos el estudio considerando los métodos jerárquicos, obtenemos la matriz de distancias, utilizando la distancia euclídea y aplicamos el algoritmo de cluster jerárquico con el método '*linkaje completo*'.

```
row.names(data)<-as.character(data$comunidad)
distancia<-dist(data)
```

```
jerarbas<-hclust(distancia, method = "complete")
plot(jerarbas)
rect.hclust(jerarbas, k = 3, border = 3:5)
```



Al observar el dendrograma resultante, parece haber tres grupos naturales con relación al predominio de las aficiones culturales entre las comunidades:

el primero: Baleares, Canarias, Cantabria, Madrid, Murcia, Rioja.
el segundo: Andalucia, Asturias, Castilla LaMancha, Cataluña, Valencia, Extremadura y Ceuta/Melilla.
el tercero: Aragón, Castilla y Leon, Galicia, Pais Vasco.

Procedemos ahora a calcular el coeficiente de correlación cofenético basado en el método aplicado.

```
cor(x = distancia, cophenetic(jerarbas))
```

```
## [1] 0.4839919
```

Dado que el valor no es tan cercano a -1 ó 1, no es concluyente para indicar que la distribución en grupos de los datos es adecuada.

A continuación, obtenemos el valor del coeficiente cofenético para los modelos jerárquicos: Promedio, Simple (vecino más cercano) y Ward.

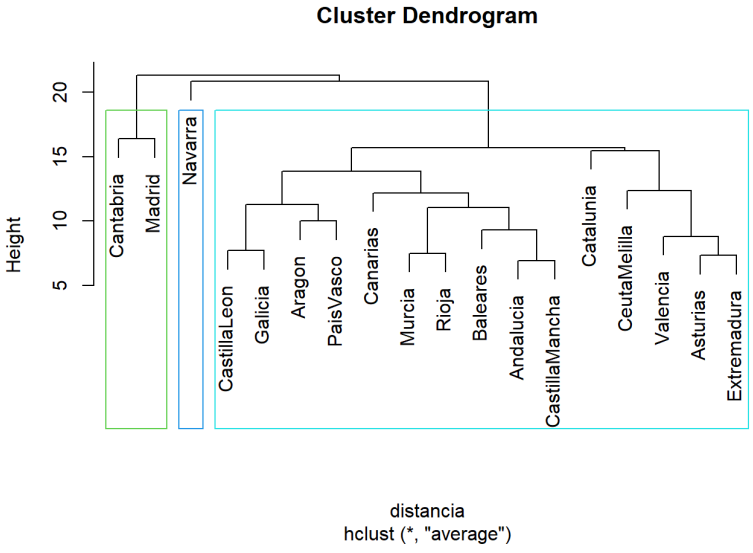
```
jeraravg<-hclust(distancia, method = "average")
jerarsin<-hclust(distancia, method = "single")
jerarwar<-hclust(distancia, method = "ward")


ac1 <-  cor(distancia, cophenetic(jeraravg))
ac1
ac2 <-  cor(distancia, cophenetic(jerarsin))
ac2
ac3 <-  cor(distancia, cophenetic(jerarwar))
ac3
```

Con los siguientes resultados:

```
Promedio: ## [1] 0.71799
Simple: ## [1] 0.6403901
Ward: ## [1] 0.4252559
```

El método de los promedios nos da el coeficiente de aglomeración más alto. Por ser un valor cercano a 1 puede considerarse que con este método la distribución en grupos de los datos es adecuada. Veamos su dendograma



De igual forma se presentan tres grupos, pero muy diferentes a los obtenidos en el análisis anterior. En este caso:

el primero: Cantabria y Madrid.
el segundo: Navarra.
el tercero: El resto de comunidades.

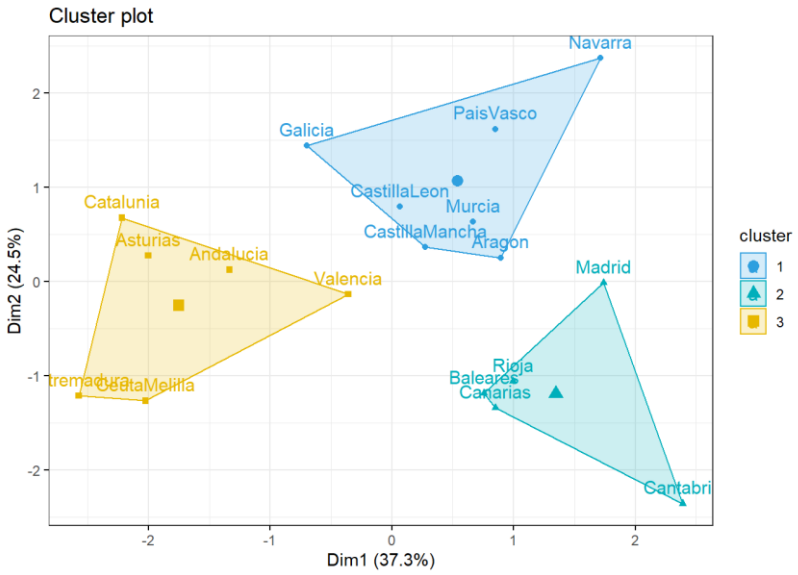
Ahora aplicamos el método No jerárquico de Kmedias básico, establecemos 3 clusters:

```
data2<-data[,-1]
res.km <- kmeans(data2,centers=3,nstart=25)
res.km
```

```
K-means clustering with 3 clusters of sizes 7, 6, 5
##
## Cluster means:
##     libros    movil concierto    radio     cine  deporte
## 1 93.41429 79.07143 10.671429 79.05714 37.28571 31.98571
## 2 90.05000 83.86667  7.733333 70.88333 36.98333 34.56667
## 3 94.50000 85.60000  8.920000 78.62000 46.70000 26.60000
##
## Clustering vector:
##      Andalucia         Aragon       Asturias       Baleares       Canarias
##              2              1              2              3              3
##      Cantabria   CastillaLeon CastillaMancha      Catalunia       Valencia
##              3              1              1              2              2
##    Extremadura        Galicia         Madrid         Murcia        Navarra
##              2              1              3              1              1
##      PaisVasco          Rioja   CeutaMelilla
##              1              3              2
##
## Within cluster sum of squares by cluster:
## [1] 458.8314 344.6717 312.6560
##  (between_SS / total_SS =  47.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Esta solución es muy parecida a la obtenida con el primer método evaluado, el del linkaje completo, porque identifica la conformación de tres grupos naturales de comunidades en cuanto a las aficiones culturales casi con los mismas comunidades por grupos. Una buena representación gráfica de estos grupos la podemos observar en el siguiente gráfico:

```
fviz_cluster(res.km, data = data2,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
             ellipse.type = "convex",
             ggtheme = theme_bw()
             )
```



De resultado obtenido, vemos como para este modelo la variabilidad explicada por el agrupamiento alcanza un 47% con relación al total, lo que no nos permite concluir que el agrupamiento aquí obtenido sea una representación concluyente de las aficiones culturales.

Finalmente, con base en los resultados anteriores, parece apropiado pensar que el modelo que mejor se ajusta para explicar el agrupamiento natural de las aficiones culturales de las comunidades es el modelo basado en el método de los promedios.