Resumen del análisis de conglomerados

Evelyn Alejandra Morales Remache

2024-02-20

Índice

Introdución

El presente resumen es realizado en base a la lectura del libro de Análisis multivariante aplicado con R, de Aldás y Urien (2017). El análisis de conglomerados también conocido como análisis de cluster, busca dividir una población en subgrupos. Los cuales son distintos entre ellos y las observaciones son homogéneas dentro de un mismo grupo, es decir, cada observación es lo mas parecida a todas las que están dentro de ese grupo. El análisis de cluster se diferencia de otras técnicas puesto que no se conoce a priori los grupos, sino que hay que derivarlos de las observaciones.

1 Secuencia lógica que se sigue para realizar un análisis de conglomerados

- El investigador en primera instancia dispone de la información de n observaciones sobre las K variables.
- Después se establece un indicador que permite saber la (di)similaridad de cada par de observaciones.
- En base a esta medida de similaridad se forman los grupos, con las observaciones que mas se parecen entre si.
- Por ultimo se describen los grupos obtenidos para lo cual se usa la media de las k variables para cada grupo obtenido ($g \leq n$).

2 Medidas de similaridad

Caso de ejemplo

El caso hipotético presenta información de 8 empresas sobre las variables Inversión publicitaria y Ventas.

```
library(readxl)
datos <- read_excel("ejemplo.xlsx")
print(datos)
```

```
## # A tibble: 8 x 3
##   NombreEmpresa InversionPublicitaria `Ventas(Millardos)`
##   <chr>                         <dbl>               <dbl>
## 1 E1                               16                  10
## 2 E2                               12                  14
## 3 E3                               10                  22
## 4 E4                               12                  25
## 5 E5                               45                  10
## 6 E6                               50                  15
## 7 E7                               45                  25
## 8 E8                               50                  27
```

Los grupos que se pueden formar con este conjunto de datos van en relación a resultado obtenido por la inversión en publicidad que esta reflejado en las ventas. Para la idea intuitiva de los grupos que se pueden formar se presenta continuación un gráfico de dispersión.

```
library(ggplot2)


ggplot(datos, aes(x = InversionPublicitaria, y =`Ventas(Millardos)`))+
geom_point(size = 2, color ="#1874CD")+
geom_text(aes(InversionPublicitaria, `Ventas(Millardos)`, label =NombreEmpresa))+
ggtitle("Inversion en publisidad y ventas de 8 empresas")
```

Mediante este gráfico se puede observar que existen 4 grupos donde comparten características similares como lo es E1 y E2. Esta idea intuitiva del análisis de conglomerados se puede ver en este gráfico debido a que solo se tienen 2 variables pero cuando se tienen k variables es necesario conocer la similaridad de las variables para crear los grupos.

2.1 Medidas de similaridad para variables métricas

En e caso que las variables que describen las observaciones sean métricas tal como de intervalo o de razón existen 4 opciones de medidas de similaridad.

2.1.1 Distancia euclídea

Esta distancia euclídea entre un par de observaciones se denomina D_{ij} y esta conformada por la raíz de la sumatoria de la distancia al cuadrado de la observación i respecto de la j paras las K variables.

$$D_{ij} = \sqrt{\sum_{p=1}^{k}(x_{ip} - x_{jp})^2}$$

Con esto obtenemos la distancia de euclídea entre la observación i y j

2.1.2 Distancia euclídea al cuadrado

La distancia euclídea puede llegar a necesitar una capacidad del ordenador exigente debido a la raíz, por lo cual una de las formas de reducir los cálculos es considerar la distancia euclídea al cuadrado como se muestra en la siguiente expresión.

$$D_{ij} = \sum_{p=1}^{k} (x_{ip} - x_{jp})^2$$

2.1.3 Distancia de Minkowsky

Esta es la generalización de la distancia euclídea y la distancia euclídea al cuadrado.

$$D_{ij} = \left[\sum_{p=1}^{k} |x_{ip} - x_{jp}|^n \right]^{\frac{1}{n}}$$

Donde si $n = 2$ se obtiene la distancia euclídea.

2.1.3 Distancia de Manhattan

Esta se obtiene cuando n=1 en la función de Minkowski, a esta se la conoce como city block, que hace referencia a la distancia de un punto hacia otro sin usar el camino c de la hipotenusa.

$$D_{ij} = \sum_{p=1}^{k} |x_{ip} - x_{jp}|$$

2.2 Medidas de similaridad para datos binarios

En este tipo de datos para un atributo se asigna 1 que representa la presencia de dicho atributo caso contrario se asigna cero.Por lo cual para calcular las medidas de similaridad se elabora una matriz 2x2 para cada par de observaciones a comparar. Esta matriz recoge coincidencias y divergencias entre las distintas variables.

En este caso existen distintas medidas de similitud como por ejemplo la de Jaccard (1901)

$$\sqrt{1 - \left[\frac{a}{a+b+c} \right]}$$

- Donde a es el numero de **coincidencias** entre la observación 1 y 2 del valor binario (1).
- b es el numero de **coincidencias** entre la observación 1 y 2 del valor binario (0).
- c el numero de divergencias entre la observación 1 con el valor binario (1) y la observación 2 con el valor binario (0).
- d es el numero de divergencias entre la observación 1 con el valor binario (0) y la observación 2 con el valor binario (1).

2.3 Estandarizacion de los datos

Al medir la distancia entre un par de observaciones es posible que las medidas de disimilaridad sean sensibles a las unidades en que están medidas las variables. Por ejemplo si se tienen dos variables el número de trabajadores y el tamaño de activos si se efectúa el análisis con las unidades originales el análisis de conglomerados mostrarán que la influencia de la variable números trabajadores es casi nula en la obtención de los conglomerados. Esto no es cierto lo cual para evitar la alta influencia de la variable tamaño de activos respecto al número de trabajadores, generada por la unidad en que están recopilados los datos. Es importante estandarizar los datos.Existen diferentes formas de realizarlo:

- **Puntuaciones Z** aquí los datos son estandarizados restando la media de una variable y dividiendo para la desviación típica.
- **Rango 1** el valor de una variable se divide para la variación de la variable.
- **Rango 0 a 1** se estandariza considerando el valor mínimo de esa variable.

3 Formacion de grupos analisis jerarquico de conglomerados

Después de haber obtenido el la matriz de distancias entre las observaciones, se tiene que tomar dos decisiones:

- Selección del algoritmo de clasificación.
- El numero de grupos dado los datos.

Estas decisiones no son fáciles puesto que existen diversos algoritmos de agrupamiento, donde si al aplicar algunos de estos se obtienen resultados similares significa que hay una agrupación natural objetiva, caso contrario es importante analizar trabajos precedentes y el marco teórico sobre los datos.

Exiten don grandes enfoques de algoritmos de agrupación:

Métodos Jerárquicos Este supone $n-1$ decisiones de agrupación en este enfoque existen dos tipos:

- Jerárquicos *aglomerativos* en este metodo en su inicio cada observacion (individuo) es un grupo, y se va formando grupos de mayor tamaño añadiendo los individuos mas cercanos y al final todos confluyen en un solo grupo.
- Jerarquicos *desagregativos* es lo contrario del proceso aglomerativo, es decir iniciamos con que todos lo individuos son un solo grupo y se termina donde cada individuo es un grupo.

Métodos No Jerárquicos En este método se establecen en un inicio grupos a priori donde los individuos se clasifican en esos grupos. conservando que cada grupo es lo mas distinto de los demas grupos y internamente las observaciones de cada grupo son parecidas.

3.1 Metodo del centroide

El método del centroide de Sokal y Michener está implementado en la función hclust{stats} en *R* el insumo principal para realizar este método es la matriz de distancia que se puede calcular con el método distancia euclidea al cuadrado. Este método comienza uniendo las dos observaciones más cercanas en un grupo, luego este grupo es sustituido por una observación que lo representa y en la que las variables toman los valores medios de las observaciones que constituyen el grupo representado a esto se lo denomina (centroide). En el caso de ejemplo de este documento la empresa E3 y E4 son sustituidas por una empresa promedio que se la puede llamar E(3-4).

Seguidamente se recalcula la matriz de distancias reemplazando E3 y E4 por su centroide E(3-4) luego se unen entonces aquellas dos observaciones que están de nuevo más cerca. En nuestro ejemplo en este segundo paso las empresas E7 y E8 son las que están más cercas las cuales serán sustituidas por su centroide E(7-8) Y nuevamente se repite el mismo proceso se recalcó la matriz de distancias y se unen las observaciones que están más cercanas. El proceso termina cuando todas las empresas están en un solo grupo.

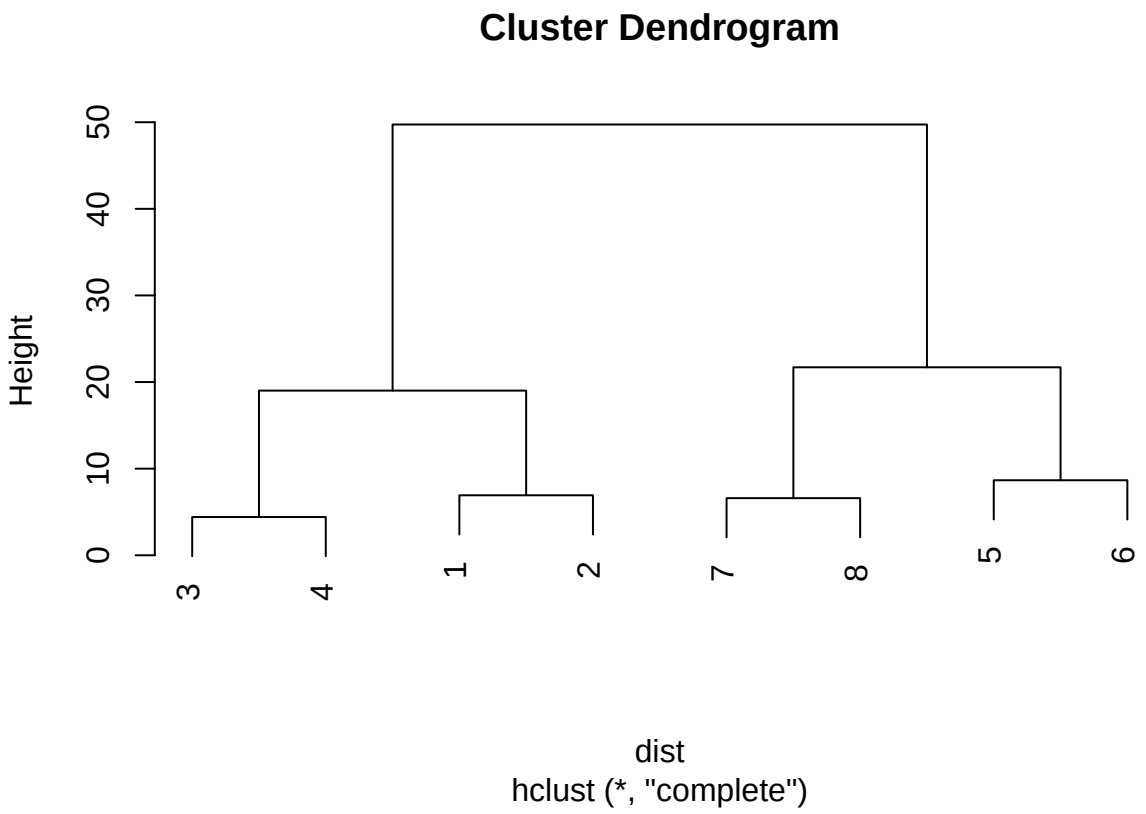
Las distancias entre los grupos que se van formando se recopilan en el historial de conglomeración el cual tiene una traducción gráfica denominada **dendrograma** es de gran utilidad para determinar el número de grupos que se deben retener. Como bien sabemos el análisis de conglomerados jerárquicos comienza considerando a cada individuo como un grupo independiente y luego va añadiendo más individuos al grupo pero en cada etapa une individuos más distantes. El gráfico de dendograma tienen el eje (y) la distancia (height) entre grupos y en el eje (x) las observaciones, mediante este gráfico para saber el numero de grupos a retener, se corta donde existe un gran Salto es decir la distancia entre un grupo y otro es es alta por lo que no se las puede fusionar esos grupos.

```
# Matriz de distancias
dist <- dist(datos)
```

```
## Warning in dist(datos): NAs introduced by coercion
```

```
# Clúster jerárquico
hc <- hclust(dist)

# Dendrograma
plot(hc)
```



3.2 Método del vecino mas cercano

Método también se lo conoce bajo la etiqueta de vinculación simple de los autores Florek et al, 1951; y Michener, 1958) hace referencia a la distancia entre dos grupos que se da entre dos miembros más cercanos de esos grupos.

3.3 Método del vecino mas lejano

Este método también es conocido en algunos textos como la vinculación completo es similar al anterior pero en este se calcula la distancia entre dos grupos entre sus miembros más alejados.

3.4 Método de la vinculación promedio

En este método la distancia de dos grupos se obtiene el calcular el promedio de las distancias de todos los pares de observaciones de dos grupos.

3.5 Método de Ward

Este método no calcula la distancia entre dos grupos para para decidir si fusionarlos o no. El objetivo de este método es maximizar la homogeneidad dentro de cada grupo (conglomerado). Para este método se calcula los centroides de los grupos resultantes de las *posibles fusiones* (alternativas de grupos a fusionar), luego se obtiene la **distancia euclidia** al cuadrado al **centroide** de todas las observaciones del grupo resultante de una posible fusión (suma de cuadrados total). La solución en que se tenga una menor suma de cuadrados es la que garantiza una mayor homogeneidad por lo cual será la elegida.

Conclusión

En síntesis de esta parte sobre el análisis de conglomerados se entiende que este análisis busca obtener grupos derivados de un conjunto de datos, donde las observaciones dentro de cada uno de estos grupos son parecidas entre sí, para lo cual es importante calcular la distancia que hay entre un par de observaciones mediante diferentes medidas de similaridad, para con esta información generar los conglomerados sea de forma jerárquica o no jerárquica.

Bibliografía

Aldás, J., & Uriel, E. (2017). Análisis multivariante aplicado con R (Segunda edición). Paraninfo.(pp. 77-97).

Jesús mi capitán