Tema 32

Daniel Peña
César Pérez
E. Uriel y J. Aldás

1. Análisis de conglomerados.

2. Medidas de disimilaridad.

Construcc. de jerarquías

3. Métodos jerárquicos aglomerativos: el dendrograma.

4. Métodos jerárquicos divisivos.
(≡ disociativos) pg 432 César Pérez
(≡ de división) pg 233 Daniel Peña → Sebor 1984

Met. de partición →

5. Métodos no jerárquicos de clasificación.

1.- ANÁLISIS DE CONGLOMERADOS

El análisis de conglomerados (cluster) tiene por objeto agrupar elementos en grupos homogéneos en función de las semilitudes o similaridades entre ellos. Normalmente se agrupan las observaciones, pero el análisis de conglom. puede también aplicarse para agrupar variables.
Otros nombres asignados al mismo concepto son: métodos de clasificación automática o no supervisada, de reconocimiento de patrones sin supervisión, ...
(no supervisados para distinguirlos del anál. discriminante).

El análisis de conglomerados estudia tres tipos de problemas:

- Partición de los datos: se dividen los datos en un número prefijado de grupos, de manera que:
  - cada elemento pertenezca a un sólo grupo
  - todo elemento quede clasificado
  - cada grupo sea internamente homogéneo
- Construcción de jerarquías: se estructuran los ~~datos~~ elementos de un conjunto de forma jerárquica por su similitud. La jerarquía ~~obtenida~~ construida permite también obtener una partición de los datos en grupos
- Clasificación de variables: las variables pueden clasificarse en grupos o estructurarse en una jerarquía.

los mét. de partición de datos utilizan la matriz de datos
los algor. jerárquicos utilizan la matriz de distancias o similitudes entre datos

Para agrupar variables → var. continuas → matriz de correlación
→ var. discretas → distancia χ^2

2.- MEDIDAS DE DISIMILARIDAD

Los métodos jerárquicos parten de una matriz de distancias (=disimilaridades) o similaridades entre los elementos de la muestra y construyen una jerarquía basada en estas distancias.
La distancia o similaridad utilizada depende del tipo de variables que se analizan: v. continuas, v. discretas o ambas a la vez.

Distancias para variables métricas

- Distancia euclídea : $d_{ij} = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}$
- Distancia euclídea al cuadrado : $d_{ij} = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2$
- Distancia de Minkowski : $d_{ij} = \left(\sum_{h=1}^{p} |x_{ih} - x_{jh}|^n \right)^{1/n} = M_{ij}$
- Distancia de Manhattan o City-block : $d_{ij} = \sum_{h=1}^{p} |x_{ih} - x_{jh}| = B_{ij}$
- Distancia de Chebychev : $d_{ij} = \max_h |x_{ih} - x_{jh}| = C_{ij}$

Distancias para var. cualitativas (no dicotómicas) : χ^2

Medidas de similaridad para datos binarios

Se utilizan cuando las variables utilizadas en el análisis contemplan únicamente la ausencia (0) o presencia (1) del atributo considerado, para ello se construyen las tablas de asociación entre elementos:

Matriz de datos

	X_1	X_2	X_3	...	X_p
A	0	1	1	...	1
B	1	0	1	...	0
C	0	0	0	...	1
⋮	⋮	⋮	⋮		⋮

Tabla de asociación AB

A \ B	1	0
1	a	b
0	c	d

- Proporción de coincidencias : $s_{ij} = \dfrac{a+d}{a+b+c+d}$
  (≡ Parejas simples)

- Proporción de apariciones : $s_{ij} = \dfrac{a}{a+b+c}$
  (≡ Russel y Rao)

- Hamann (para probab. condicionales) : $s_{ij} = \dfrac{(a+d)-(b+c)}{a+b+c+d}$

- Q de Yule (de predicción) : $Q_{ij} = \dfrac{ad-bc}{ad+bc}$

Existen otras medidas de similaridad que pueden ser utilizadas según el criterio del investigador (ver D. Peña pgs. 422-423)
~~Medidas de la~~

<u>Medidas de similaridad para var métricas y no métricas</u>

Cuando en una muestra existen variables continuas y atributos, si se utiliza la distancia euclídea (u otra para var. métricas) será 0 o 1 en las variables dicotómicas y sin embargo en las var. continuas puede llegar a ser muy alta.
Cuando, por la naturaleza del problema, esto no sea aceptable, la solución es trabajar con similaridades, para lo cual se define la similaridad global entre dos elementos (i,j) como el coeficiente propuesto por Gower:

$$s_{ij} = \frac{\sum_{h=1}^{p} w_{hij}\, s_{hij}}{\sum_{h=1}^{p} w_{hij}}$$

" s_{hij} ≡ similaridad según la var h entre los elementos muestrales (i,j)

donde $w_{hij} = \begin{cases} 1 & \text{si la comparación entre los elem. } i,j \text{ tiene sentido para la var } h \\ 0 & \text{'' '' '' '' '' '' '' no '' '' '' '' '' ''} \end{cases}$

los coficiente de similaridad para una variable continua se construyen mediante:

$$S_{hij} = 1 - \frac{|X_{hi} - X_{hj}|}{\text{rango}(X_h)}$$

que cumple función no negativa y simétrica:

$S_{hii} = 1$

$0 \leq S_{hij} \leq 1$

$S_{hij} = S_{hji}$

Una vez obtenida la similaridad entre los elementos muestrales, la podemos transformar en distancias:

$$d_{ij} = \sqrt{2(1 - S_{ij})}$$ (verifica la propiedad triangular)

Obs: Si se define $d_{ij} = 1 - S_{ij}$ puede no verificar la propiedad triangular.

~~3.2.4. Métodos jerárquicos~~

Estandarización de los datos

Si se analizan las medidas de distancia/similaridad, se comprueba que todas ellas están basadas en la sustracción, para cada par de observaciones, de los valores de las variables utilizadas.

Por ello, estas medidas son muy sensibles a las unidades en que se miden las variables. Para evitar esta influencia no deseable de una var. debida exclusivamente a la unidad de medida, es necesario corregir el efecto de los datos estandarizando las variables (puntuaciones Z, rango 1, rango 0,1 ...)

3./4.- METODOS JERÁRQUICOS

Dada una matriz de distancias o de similitudes (similaridades o disimilaridades) se desea clasificar los elementos en una jerarquía. los algoritmos existentes funcionan de manera que los elementos son sucesivamente asignados a los grupos y la asignación es irrevocable. los algoritmos son de dos tipos:

1.- De aglomeración (M. aglomerativos): Parten de los elementos individuales y los van agregando en grupos
2.- De división (M. divisivos o disociativos): Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales

3.- METODOS JERÁRQUICOS AGLOMERATIVOS

los algoritmos jerárquicos que se utilizan tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Dicha estructura es:

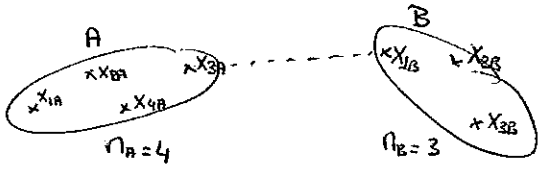
1.- Comenzar con tantos grupos como elementos, n. Las distancias entre grupos son las distancias entre los elementos originales.
2.- Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos un grupo.
3.- Sustituir los dos elementos utilizados en 2.- para definir el grupo por un nuevo elemento que represente el grupo construido. la distancia entre este nuevo elemento

y los autores se calcula con uno de los criterios que se comentan a continuación.

4.- Volver a 2.- y repetir 2. y 3.- hasta que se tengan todos los elementos asignados a un único grupo.

Métodos para definir distancias entre grupos

Método del vecino más cercano (o encadenamiento simple) la distancia entre dos grupos es la distancia entre sus dos puntos más próximos



$$d_{AB} = \min \{ d_{x_{iA}, x_{jB}} \}$$
$$i = 1, 2, \ldots n_A$$
$$j = 1, 2, \ldots n_B$$

El método consiste en agrupar los individuos que tienen menor distancia (o mayor similitud).
Se mide los elementos H y K si:

$$d_{H,K} = \min \{ d_{g,g'} \} \quad g = 1, 2 \ldots G \quad g' = 1, 2 \ldots G \quad g \neq g'$$

donde $d_{g,g'} = \min \{ d_{x_{ig}, x_{jg'}} \}$ $i = 1, 2 \ldots n_g$ $j = 1, 2 \ldots n_{g'}$

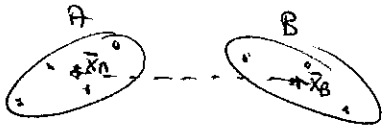
Método del vecino más lejano (o encadenamiento completo)
La distancia entre dos grupos es la distancia entre sus dos puntos más lejanos.

$$d_{AB} = \max \{ d_{x_{iA}, x_{jB}} \}$$

El método consiste en agrupar los individuos que tienen menor distancia.
H y K se unen si $d_{H,K} = \min \{ d_{g,g'} \}$ $g = 1, 2 \ldots G$ $g' = 1, 2 \ldots G$ $g \neq g'$

Método del centroide

la distancia entre dos grupos es la distancia entre sus centros de gravedad.



$$d_{AB} = d_{\bar{x}_A \bar{x}_B}$$

H,K se unen $\Leftrightarrow$ $d_{HK} = \min \{d_{gg'}\}$, $g = 1,2,\ldots,G$ $\quad g \neq g'$
$g' = 1,2,\ldots,G$

Método de la vinculación promedio (o media de grupos)

la distancia entre dos grupos es la media ~~ponderada~~ de las distancias entre todas las pares de observaciones que pueden formarse tomando un individuo de un grupo y otro individuo del otro grupo.

H,K se unen $\Leftrightarrow$ $d_{HK} = \min \{d_{gg'}\}$

Método Ward

Este método no calcula distancias entre grupos, ya que su objetivo es maximizar la homogeneidad dentro de cada grupo. Para ello define la medida global de la heterogeneidad como:

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)' (x_{ig} - \bar{x}_g) \quad , \ \bar{x}_g = \text{media del grupo } g$$

Se unen los grupos que produzcan un incremento mínimo de W. Puede demostrarse que, en cada etapa, los grupos que deben unirse para minimizar W son aquellos tales que:

$$\min \frac{n_A n_B}{n_A + n_B} (\bar{x}_A - \bar{x}_B)' (\bar{x}_A - \bar{x}_B)$$

Comparación

Es difícil dar reglas generales que justifiquen un método sobre otro. Lo recomendable es analizar qué criterio es más razonable para los datos que se quieren agrupar y, en caso de duda, probar con varios y comparar los resultados.

El dendograma

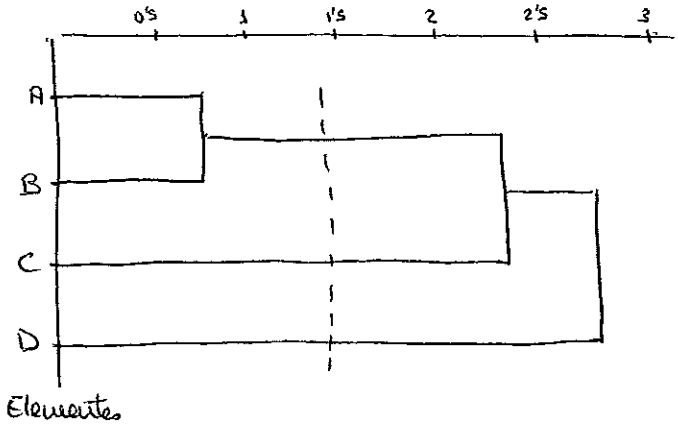
El dendograma, o árbol jerárquico, es una representación gráfica del resultado de la agrupación en forma de árbol.

Los criterios presentados para definir distancias tienen la propiedad ultramétrica ($d(A,C) \leq \max\{d(A,B), d(B,C)\}$)

El dendograma, por tanto, es la representación de una ultramétrica, y se construye:

1. En la parte izquierda del gráfico se colocan los n elementos iniciales.
2. Las uniones entre elementos se indican por tres líneas rectas. Dos perpendiculares al eje de los elementos dirigidas a este, y una paralela a este eje que une las dos anteriores y que se sitúa al nivel (distancia) en que se unen. Este nivel (distancia) se representa mediante un eje perpendicular al eje de elementos en la parte superior del gráfico.
3. El proceso se repite hasta que todos los elementos están conectados por líneas rectas.

Obs: la colocación de los elementos en el eje se hace de forma que las líneas rectas no se corten entre sí.



Dendograma.

Si se corta el dendograma a un nivel de distancia dado, se obtiene una clasificación del número de grupos existentes a ese nivel y los elementos que los forman (---) ⟶ (Regla de parada)

El dendograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se interpreta mecanicamente.

Es muy útil apoyarse en otros elementos del análisis cluster como son el historial de aglomeración y el diagrama de témpanos.

4.- MÉTODOS JERÁRQUICOS DIVISIVOS

Los metodos disociativos se dividen en dos:

1. Método monotético: cuando el criterio de división toma en consideración cada variable observada una a una
2. Método politético: cuando se toman en cuenta todas las variables.

Los metodos expuesta para la clasificación jerárquica aglomerativa tambien se utilizan ~~para~~ como métodos de clasificación disociativos (M. de Ward y del promedio son los más utilizados).

Se exponen dos métodos disociativos:

- Métodos monotéticos:

1.- Método asociativo de Williams y Lambert: Se construyen tablas de contingencia 2x2 para cada par de variables y se calcula χ^2 para cada tabla. El criterio de partición de los grupos se basa en la variable que maximiza la χ^2.

2. M. Detector automático de iteración (AID): los elementos básicos del AID son análogos a los de la regresión, es una variable dependiente y varias independientes. El objetivo del AID consiste en determinar qué variables independientes proporcionan la mayor diferencia a las distintas medias de la variable dependiente para los diferentes grupos.

- Métodos politéticos:

1.- El grupo inicial se divide en dos, separando "uno a uno" los elementos mediante el sig. criterio:

a) Se separa el individuo cuya distancia media al resto de los individuos sea mayor, para formar el grupo A.

b) Se estudia que elemento del grupo original es el que separa más para ir a formar parte del grupo A.

c) Este proceso se repite hasta que los individuos que queden estén más próximos del grupo ~~que del~~ original que del A.

d) Se repite el proceso para cada uno de los subgrupos que secuencialmente se vayan obteniendo.

5.- MÉTODOS NO JERÁRQUICOS DE CLASIFICACIÓN.

los métodos no jerárquicos, también se conocen como métodos partitivos o de optimización, tienen por objetivo realizar una sola partición de los individuos en K grupos. Esto implica que el investigador debe especificar a priori los grupos que deben ser formados. Esta es la principal diferencia respecto de los métodos jerárquicos. La asignación de los individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia radica en que estos métodos trabajan con la matriz de datos original y no requieren su conversión en matriz de proximidades.
los métodos no jerárquicos se agrupan, según Pedret, en cuatro familias:

1.- Met. de reasignación: permiten que un individuo asignado a un grupo en un determinado paso del proceso, sea reasignado a otro grupo en un paso posterior si esto optimiza el criterio de selección.
   - Met. de K-medias
   - Met. de McQueen
   - Quick Cluster Analysis } Met. Centroide o Centros de gravedad
   - Met. de Forgy
   - Met. de las nubes dinámicas (Diday).

2.- Met. de búsqueda de densidad: se distinguen dos tipos de métodos según la aproximación que presentan:
   - Aproximación tipológica: los grupos se forman buscando las zonas en las que se da una mayor concentración de individuos:
     - Análisis modal de Wishart
     - Met. de Taxmap de Carmichael y Sneath
     - Met. de Fortin.

- Aproximación probabilística: se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a una misma distribución.

  - Met. de las combinaciones de Wolf

3.- Met. directos: Permiten clasificar simultaneamente a los individuos y a las variables. Las entidades agrupadas, ya no son los individuos o las variables, sino que son las observaciones, es decir, las celdas que configuran la matriz de data.

4.- Met de reducción de dimensiones: Consiste en buscar factores en el espacio de individuos, correspondiendo cada factor a un grupo.

  - Análisis factorial de tipo Q.

Se describe el algoritmo más utilizado por investigadores:

Método de K-medias

Supongamos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado, K.
El algoritmo de K-medias requiere cuatro etapas:

1.- Seleccionar K puntos como centros de los grupos iniciales. (semillas)
Esto puede hacerse:
  a) Tomando como centros K individuos de forma aleatoria
  b) Tomando como centros los K puntos más alejados entre sí.
  c) Seleccionando centros a priori o construyendo grupos con información a priori.

2.- Calcular las distancias euclídeas de cada elemento a los centros de los K grupos, y asignar cada elemento al grupo de cuyo centro esté más cerca. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalcula el nuevo centro del grupo.

3.- Definir un criterio de optimalidad y comprobar si reasignando alguno de los elementos mejora el criterio.

4.- Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

El criterio de homogeneidad o de optimalidad, que se utiliza en el algoritmo de K-medias, es minimizar la suma de cuadrados dentro de los grupos (SCDG) para todas las variables:

$$SCDG = \sum_{g=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{jg})^2$$

Este criterio es equivalente a minimizar la suma ponderada de las varianzas de las variables en los grupos:

$$\min SCDG = \min \sum_{g=1}^{G} \sum_{j=1}^{p} n_g s_{jg}^2$$

El resultado del algoritmo puede depender de la elección inicial de las semillas y del orden de los elementos. Conviene siempre repetir el algoritmo con distintos valores iniciales y permutando los valores de la muestra (el efecto del orden suele ser pequeño). Este criterio supone var. cuantitativas, aunque puede aplicarse si el número de var. binarias es pequeño.

Número de grupos

En la aplicación habitual del algoritmo de K-medias hay que fijar el número de grupos, K.
Un procedimiento que se utiliza bastante (aunque sin mucha justificación) es realizar un test F de reducción de varibilidad, comparando la SCDG de K grupos con la de K+1 grupos. El test es:

$$F = \frac{SCDG(K) - SCDG(K+1)}{\frac{SCDG(K+1)}{n-K-1}}$$

Se compara F con $F_{p,\ p(n-K-1)}$. Pero esta regla no está muy justificada porque los datos no tienen por qué verificar las hipótesis necesarias para aplicar la distribución F.

Una regla empírica sugerida por Hartigan (1975), es introducir un grupo más si $F > 10$.

6.- Elección entre los distintos tipos de análisis de conglomerados

- A. Cluster jerárquico / no jerárquico
  Si se busca el número de grupos en que se agregan las observaciones ⇒ A. Cluster no jerárquico
  Si también se buscan las centroides de esos grupos.
  Si no, realizamos primero un A. Cluster jerárquico y con los resultados obtenidos se realiza un A. Cluster no jerárquico que permite maximizar la homogeneidad dentro de los grupos y la heterogeneidad entre conglomerados.

- Métodos de agrupación en el D. Cluster jerárquico:

Es difícil escoger uno u otro método, lo mejor es probar varios procedimientos en un mismo estudio y comparar los resultados.

De todas formas es interesante tener en cuenta los siguientes resultados:

a) El mét. del vecino más cercano es más sensible a la presencia de data aváricos y tiene tendencia a crear menos grupos que el met. del vecino más lejano.

b) El mét. del vecino más lejano identifica grupos muy homogéneos en los que los obsrv. son muy parecidas unas a otras.

c) El mét. Ward tiende a encontrar conglomerados muy compactos y de tamaño similar.

d) El mét. del centroide es el más robusto ante la presencia de data atípicos

e) El mét. ward junto con el del promedio los demostró ser mayor eficacia en estudios de simulación.

f) El mét. Ward requiere una dist. normal multivar. en las var. del estudio

ESTAD_T32

1_Análisis conglomerados
- Objetivo ≡ grupos < casos vars.
- Sin hipótesis, téc. descriptivas.
- Tipos problemas:
  - Partición datos
  - Constr. jerarquías
  - Clasif. variables.

2_Medidas de disimilaridad.
- Cuantit. (eucl. $eucl.^2$, Mink...)
- Cualit. (coincid, apar, ...)
- mixtas → similaridad global
- Estandarización

3_ Mt. jerárquicos aglom.
- Aglom / Divis
- Método
- Distancia
  - vecino + cercano → $d = mín$
  - vecino + lejano → $d = max$
  - centroide → $d = d(centros)$
  - V. promedio → $d = \overline{d}$
  - ward → $W = \sum\sum (x_{ij} - \bar{x}_j)' (x_{ij} - \bar{x}_q)$
- Dendograma
  - Descrip.
  - Tiempos etiológicos

4_ Mt. jerárquicos divisivos
- Mt. monotético
- Mt. politético

5_Mts. no jerárquicos.
- A priori K grupos, nº mín de datos
- Mts reasignación
- Mts. búsqueda densidad
- Mts. directos (obsen)
- Mts reducción dimensiones

* Mt K-medias (reasignación):
  1º. Elección K centros
  2º. Asignación y reasignación
  3º. Criterio de optimalidad (min SCDG)
  4º. Parada

6_ ELECCIÓN:
- Mt jerárq / no jerárq.
- Mt. agrupación