RESUMEN ANÁLISIS CLUSTER

RESUMEN ANÁLISIS CLUSTER

1. Introducción

A veces, los datos se muestran como masas informes difíciles de organizar.

La técnica de análisis cluster o análisis de conglomerados consiste en clasificar a los individuos en estudio formando grupos o conglomerados (cluster) de elementos, tales que los individuos dentro de cada conglomerado presenten cierto grado de homogeneidad en base a los valores adoptados sobre un conjunto de variables.

En el análisis cluster, a diferencia del análisis discriminante (donde los grupos están establecidos a priori y la función discriminante permite reasignar los elementos a los grupos), los conglomerados son **desconocidos** y el proceso consiste en su **formación** de modo óptimo, aglutinando unidades homogéneas.

Está claro que los grupos formados vendrán determinados por las múltiples variables usadas en el estudio, pero el interés está en caracterizar y resumir entre esa espesura de características observables, algo inherente a cada grupo. Tras el resultado del agrupamiento surge la necesidad de encontrar respuestas a esas agrupaciones.

Queremos encontrar la posible agrupación "natural" existente entre los datos analizados; es decir, estructuras latentes no detectadas explícita y directamente a través de las variables observadas. Luego, el investigador tratará de especificar la configuración de los grupos encontrados en el conjunto de datos, tratando de **explicar** dicha ordenación con argumentos, generalmente, ajenos a la técnica en sí misma (conocimiento teórico de materia estudiada, conexión con otros estudios, etc.). Por tanto, una vez establecida empíricamente la clasificación, para que ésta sea útil, puede ser analizada detenidamente con objeto de descubrir las claves o propiedades que han producido tal agrupamiento. La aparición de esta estructura puede llevar al investigador a aprehender aspectos o propiedades de los individuos que de otro modo habrían pasado inadvertidos. Lo que puede conducir, a su vez, a plantear nuevas hipótesis de trabajo y nuevas investigaciones desde diferentes perspectivas.

Por ejemplo, puede clasificarse un grupo de consumidores de un producto según ciertas características personales y socioeconómicas (salario, sexo, nivel cultural, etc) lo que proporciona una segmentación de mercado. En base a este conocimiento, el vendedor se dirigirá con diferentes estrategias de marketing a ellos, aprovechando mejor los recursos, haciendo uso de ese conocimiento.

2. Los datos

Se tiene un conjunto de N individuos sobre los que se observa una serie de variables X_1 , X_2 , ..., X_p . Los datos se estructuran en una matriz **X** de orden Nxp, constituida por N casos (filas) y p variables (columnas).

Veremos que algunos métodos trabajan directamente con esta información (X), y otros derivan de ella una nueva matriz de similaridades/distancias de orden NxN, que

establece la similitud o distancia entre pares de objetos mediante el cálculo de ciertos estadísticos (por ejemplo, distancia euclídea), que determinan su proximidad o lejanía.

A veces interesa cambiar el papel que juegan los individuos y las variables. Puede tener interés agrupar las variables (en vez de los casos) en un intento de buscar, por ejemplo, redundancias en la información recogida. En la matriz de datos se cambian los papeles de filas y columnas. Este tipo de análisis se encuadra en las denominadas R-técnicas (estudio de la similitud entre variables: cluster de las variables) frente a las Q-técnicas (estudio de la similitud entre los casos: cluster de los individuos).

La matriz de similitudes de orden pxp está formada por coeficientes de asociación o correlación entre variables.

Por tanto, se puede agrupar casos o agrupar variables, según el objetivo perseguido.

2.1. Transformación de los datos

No sólo es importante elegir las variables sino también el tipo de escala usado.

Las variables binarias no suelen transformarse. Las variables categóricas se convierten en binarias (presencia/ausencia) para las distintas modalidades.

Las variables cuantitativas requerirán, en muchos casos, una transformación previa al tratamiento. No pueden darse reglas fijas sobre cómo deben transformarse los datos, ni siquiera de cuándo es conveniente o no hacerlo.

La transformación más popular es la estandarización. Con ello se evita la influencia de la unidad de medida. No hay una postura clara sobre las ventajas y desventajas de la estandarización. Es un hecho que variables con alto poder discriminante pueden ver mermada su capacidad diferenciadora tras una estandarización. No obstante, cuando las variables usadas vengan en la misma escala, tales como, por ejemplo, puntuaciones de ítem en un cuestionario o, por ejemplo, si todas las variables se refieren a porcentajes, etc., no es aconsejable la estandarización. Ésta se reserva, fundamentalmente, cuando se observa que, por ejemplo, unas determinadas variables pueden tener un peso mayor que otras, simplemente porque la unidad de medida en que aparecen dan lugar a puntuaciones con valores relativamente altos en comparación con los de las otras, de tal modo, que pueden, incluso, llegar a anular la influencia de otras hasta el extremo de que dé igual incluirlas o no.

Por tanto, cuando proceda, se pueden **transformar** los valores de los datos, para los casos o las variables, antes de calcular las proximidades o matriz de distancias. Las opciones disponibles son: **puntuaciones estandarizadas Z** (se obtienen restando la media y dividiendo por la desviación típica de los valores originales de cada variable o de cada caso, según corresponda al análisis); **rango -1 a 1** (se obtiene dividiendo los valores originales por el rango de cada variable o caso, según corresponda al análisis); **rango 0 a 1** (se obtiene restando el mínimo y dividiendo por el rango de cada variable o caso, según corresponda); **magnitud máxima de 1** (se obtiene dividiendo los valores originales por el máximo de cada variable o caso, según corresponda al análisis); **media de 1** (se obtiene dividiendo los valores originales por la media de cada variable o caso,

según corresponda al análisis) y **desviación típica 1** (se obtiene dividiendo los valores originales por la desviación típica de cada variable o caso, según corresponda al análisis).

3. Medidas de distancias

A partir de la matriz de datos X de orden Nxp se construye la matriz **S** de distancias de orden NxN, donde cada coeficiente de S, s_{ij} representa el valor de un coeficiente de disimilitud para los casos i y j, que mide el grado de disimilitud/distancia de los individuos. Esta matriz será simétrica, dado que $s_{ij} = s_{ji}$

Estas matrices pueden variar considerablemente para los mismos datos según la medida de disimilitud/distancia usada y según haya o no transformado o estandarizado las variables originales.

R proporciona varias medidas de distancia. Una de las más usadas es la euclidea (la raíz cuadrada de la suma de cuadrados de las diferencias) . Vea ayuda para la función **dist** en R ("euclidean", "maximum", "manhattan", "canberra", "binary" o "minkowski"). La función **daisy** en el package cluster permite usar métricas euclidea, manhattan o distancia de Gower, esta función es útil cuando las variables usadas no son numéricas (nominales, binarias, ordinales o incluso combinaciones de ellas). Para más información use ayuda de R.

Variables cuantitativas:

La medida más importante es la distancia euclídea y derivaciones de ella mediante ponderaciones.

Sean X_1, X_2, ..., X_p las variables observadas. Notamos con X_{ij} valor observado en el caso i-ésimo en la variable j-ésima. Dados dos casos i e i', se definen las siguientes medidas de proximidad:

- **Distancia euclídea**:

Es la raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de los elementos. Ésta es la medida por defecto que suele usarse para datos de intervalo.

$$d_{ii'} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$

Depende de la escala de medida. Sus valores no están acotados.

Variables binarias (presencia/ausencia de atributo):

Existe una gran diversidad de medidas. Basadas en si están o no presentes las modalidades de las variables binarias.

Supongamos que la tabla siguiente resume la información para un par de casos i e i'. Cada caso vendrá dado por una p-upla de unos y ceros, donde: 1 indica la presencia y 0 la ausencia de cierto atributo.

		Caso i	
		1	0
Caso i'	1	a	b
	0	c	d

a = total de atributos presentes en el caso i y en el caso i'
d = total de atributos ausentes en ambos casos
b = total de atributos ausentes en i y presentes en i'
c = total de atributos ausentes en i' y presentes en i

Distancia euclídea:

$$d_{ii'} = \sqrt{b+c}$$

Es sencillamente la distancia euclídea para el caso particular en que las variables toman sólo los valores 1 y 0.

Dependiendo de las características de los datos, unos coeficientes serán más o menos apropiados que otros. Se puede llegar a designar a dos casos como iguales o distintos, dependiendo de la medida usada.

Entre los de uso más frecuente destaca el coeficiente de concordancia simple. Unas veces será importante considerar el recuento de ausencias y otras no. En algunos análisis, tan importante es considerar la concordancia de presencias como de ausencias.[1]

Es preciso advertir que si se realizan varios análisis con medidas diferentes y los resultados del agrupamiento son similares, esto no garantiza siempre que se haya encontrado la verdadera estructura de los datos, dado que muchos de los coeficientes especificados están relacionados entre sí y, en consecuencia, la concordancia de resultados puede responder, en muchos casos, a las relaciones existentes entre las medidas de similaridad usadas, más que al carácter de la estructura del agrupamiento.

R permite efectuar un análisis cluster jerárquico usando como entrada una matriz de datos o la de distancias. Esta opción es interesante porque a veces no se dispone de los datos originales o, bien, porque se desea manipular previamente la matriz de distancias.

4. Métodos: aspectos generales

La creación de los cluster puede efectuarse de muy diversos modos en función de los

[1] Por ejemplo, individuos que responden "si/no estoy de acuerdo con una determinada afirmación", pudiendo tener el mismo valor una respuesta afirmativa que negativa. Cuando las variables binarias se generan a partir de una variable cualitativa de más de dos modalidades, es conveniente el uso de índices de similaridad en los que no interviene el recuento de las concordancias de las ausencias (d) .

algoritmos de cálculo (operaciones recursivas y repetitivas) utilizados. Pero, métodos distintos darán, en general, soluciones distintas para los mismos datos.

Entre los más usados destacan los **métodos jerárquicos** aglomerativos (**hclust** del package stats y **agnes** del package cluster) y divisivos (como **diana** del package cluster) y los de **partición iterativa** como **kmeans** del package **stats** . Cada uno de los cuales da lugar, a su vez, a una gama de posibilidades diversas, según las reglas usadas para el agrupamiento.

Usaremos R para agrupar según dos métodos: análisis jerárquico aglomerativo y análisis k-medias con la distancia euclidea. El método k-medias puede usarse con varios algoritmos. El que usa R por defecto es "Hartigan-Wong". Vea ayuda de R para más información.

En análisis jerárquico se pueden estandarizar los datos desde el proceso si se usa la función **agnes** del package **cluster**. En k-medias es preciso efectuarlo previamente si se decide su uso.

5. El método jerárquico aglomerativo

Se dice que los cluster forman una jerarquía, están anidados, cada cluster formará parte de otro mayor, hasta el último paso que engloba a todos.

A diferencia de los métodos de reasignación, sólo se efectúa una inspección para agrupar, de modo que los resultados no podrán modificarse en pasos sucesivos. En el cluster jerárquico, una vez que dos individuos se funden en un cluster permanecen ahí hasta el final del proceso.

Los métodos jerárquicos usan una matriz de proximidades (distancias o similaridades) como base para el agrupamiento. Comienzan con tantos cluster como objetos y terminan con un solo cluster.

Nota: Las medidas de distancia se pueden generar mediante la función dist o la función daisy del package cluster. Estas matrices pueden usarse, tal como hemos dicho en párrafos anteriores, como input para un análisis cluster jerárquico.

5.1 Algoritmos más usados

El método se diversifica según las reglas usadas para el agrupamiento. permite 7 formas de agrupación: "ward", "single", "complete", "average", "mcquitty", "median" o "centroid".

Entre los más usados destacan el método AVERAGE y el de WARD, seguido de COMPLETE y SINGLE.

Si diversos algoritmos producen resultados similares, posiblemente estemos ante una situación ideal de cluster compactos y perfectamente delimitados.

5.2 El dendrograma

El dendrograma es un gráfico usado en el procedimiento jerárquico que permite *visualizar el proceso de agrupamiento* de los cluster en los distintos pasos, formando un diagrama en árbol.

El dendrograma es una valiosa herramienta visual que puede ayudar a decidir el número de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en que se van anidando los cluster y la medida de similitud a la cual lo hacen. Cortando transversalmente, a una distancia determinada, las ramas del gráfico se obtiene una partición.

Pero, además, puede ser que el investigador no esté tan interesado en encontrar el número de grupos en que dividir los casos sino, más bien, en "seguir la pista" de formación de los distintos cluster, que van englobándose o anidándose, hasta resumirse en sólo uno. Así, por ejemplo, en biología es corriente tener mayor interés en desvelar las distintas categorías en que van clasificándose los individuos estudiados, desde los grupos más particulares a los más generales.

5.3 Establecimiento del número de grupos

No existe una norma fija para establecer cuántos grupos pueden considerarse. El dendrograma puede servir de ayuda visual para determinar dicho número. Dependiendo del coeficiente de proximidad usado (además de ser posible obtener una clasificación diferente) el aspecto visual variará. Todo esto hace que las decisiones sean confusas e, incluso, a veces, arbitrarias si no hay un soporte teórico que apoye la solución.

5.4 Tabla resumen de la aglomeración

R proporciona además del dendrograma una tabla (**merge**) que permite ver el orden del proceso de agrupamiento de los cluster que se van uniendo. Mediante ella puede seguirse la pista a los conglomerados formados.

5.5 Diagrama banner (de package cluster)

Otro modo gráfico de inspeccionar la estructura de conglomerados que va formándose en cada paso es mediante el diagrama horizontal que es similar a un gráfico de barras. Para más información use la ayuda de R.

El gráfico muestra la agrupación recogida indicando la distancia en la que se van uniendo los cluster.

Está disponible como gráfico de **agnes** en package **cluster:**

```
bannerplot(agnes(datos))
```

En el eje horizontal aparecen las alturas (height) y en el vertical los casos o elementos a agrupar. Espacios en blanco indican separación entre elementos. Espacios coloreados indican agrupamiento.

6. Partición iterativa: el método K-means

Proceso iterativo en el que se van usando los resultados de la partición anterior para mejorar la siguiente.

A diferencia de los métodos jerárquicos, aquí es necesario especificar a priori los grupos a formar[2] y se trabaja directamente con la matriz de datos original en vez de con la matriz de distancias. Esto último hace que el método k-means sea más idóneo para analizar gran número de casos, dado que no requiere tanta capacidad de memoria, pues no precisa del almacenamiento de la matriz de proximidades NxN para el establecimiento de los grupos.

Está incluido entre los métodos denominados de reasignación, dado que un caso puede ser asignado a un cierto cluster en un determinado paso y, luego, puede ser reasignado, en otro paso, a otro cluster diferente. Por el contrario, en el método jerárquico una vez que dos individuos se funden en un cluster permanecen ahí hasta el final del proceso.

El método consiste en dividir (en una clasificación inicial) los datos en un número k de clusters, especificado previamente.
Un modo corriente de asignar los individuos a los k grupos antes de iniciar el proceso, es clasificándolos de acuerdo con alguna variable resumen, por ejemplo, suma de todas las variables implicadas o un procedimiento más refinado como puede ser la primera componente principal resultante de un análisis de componentes principales. La variable en cuestión se divide en k intervalos, codificados de 1 a k, y cada individuo será asignado al correspondiente código o grupo, según el intervalo al que pertenezca. A veces se aprovecha una clasificación previa obtenida mediante otro procedimiento como, por ejemplo, análisis jerárquico. También puede usar unas estimaciones previas de los centroides y con ellas una clasificación inicial.
R dispone de varios algoritmos para el procedimiento de agrupación.

El mayor inconveniente de esta técnica es que, si la agrupación original es muy distinta de la realmente existente, se podría obtener una solución local también alejada de la óptima.

7. Funciones de R usadas

Usaremos las del package **stats**: hclust (análisis aglomerativo jerárquico) y kmeans (partición óptima)

hclust

Puede realizar el agrupamiento a partir de los datos o de la matriz de distancias.
Entre los métodos usados están:
Single: la distancia entre dos cluster es la que hay entre los dos miembros que más cercanos están
Average: la distancia entre dos cluster es el resultado de la media de distancias entre los pares formados con los miembros de cada grupo

[2] Cuando no se está muy seguro del número de cluster a usar, suelen ensayarse diversos valores.

Complete: la distancia entre dos cluster es el máximo de la distancia existente entre los pares de miembros de los grupos.

```
hclust(d, method = "complete")
```

d es la matriz de distancias resultante de aplicar a los datos.

Resultado de hclust:

```
plot(objeto)  # Realiza el dendrograma del objeto resultante de aplicar hclust.
```

Una tabla describiendo el proceso de agrupamiento (merge)

Otras funciones usadas con hclust

Función cutree()

```
cutree(objeto, k = NULL, h = NULL)
```

Puede usar el número de grupos (k) a efectuar o bien un valor de la medida de distanciamiento (h) para agrupar.

kmeans

Se usa preferentemente con variables continuas. Agrupa los casos en un número de grupos, k, previamente fijado tomando como punto de partida los elementos previamente asignados a dichos grupos. A partir de esta configuración inicial se pueden usar distintos criterios para optimizar el agrupamiento.

```
kmeans(datos, centers, iter.max = 10, nstart = 1,
     algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))
```

el argumento centers indica el número de cluster o bien los centros de los cluster.