CAPÍTULO 9

Análisis Jerárquico de Agrupaciones

III. 9.1 Introducción.

Esta técnica es conocida por una diversidad de términos, esto es, además de análisis de conglomerados se puede encontrar como análisis de cluster, análisis de grupos y todos sus correspondientes términos en inglés.

Pertenece, al igual que el escalamiento multidimensional, a los métodos de interdependencia dentro del conjunto de técnicas multivariantes.

Tal y como se mencionó anteriormente en el apartado de introducción al análisis multivariante, el análisis de conglomerados se fundamenta en que dado un conjunto de individuos u objetos (productos, marcas,…) descritos por un cierto número de características (variables), se tratará de buscar una partición de este conjunto de objetos en un determinado número de grupos o tipos de forma tal que respecto a la distribución de los valores de las variables:

- Cada conglomerado sea lo más homogéneo posible en su interior, es decir, los objetos que configuran un grupo sean lo más similares posible respecto de sus características.

- Los conglomerados sean muy distintos entre sí.

III. 9.1.1 Idea conceptual básica del análisis de conglomerados.

- La heterogeneidad de una población constituye la materia prima del análisis cuantitativo...

- ... sin embargo, en ocasiones, el individuo u objeto particular, aislado, resulta un "recipiente" de heterogeneidad demasiado pequeño,...

- la unidad de observación es demasiado reducida con relación al objetivo del análisis...

- ... en estos casos, se trata entonces de agrupar a los sujetos originales en grupos, centrando el análisis en esos grupos, y no en cada uno de los individuos...

- ... si existe una "taxonomía" ya diseñada que resulte útil, ajustada al objetivo de análisis, se recurre a ella,

- ... pero si no es así, deberemos crearla, generando una nueva "agrupación" que responda bien a las dimensiones de nuestro análisis.

III. 9.2 Definición.

El análisis cluster es una técnica cuyo objetivo principal es obtener grupos de objetos de forma que, por un lado, los objetos pertenecientes a un mismo grupo sean muy semejantes entre sí, es decir, que el grupo esté cohesionado internamente y, por el otro, los objetos pertenecientes a grupos diferentes tengan un comportamiento distinto con respecto a las variables analizadas, es decir, que cada grupo esté aislado externamente de los demás grupos.

Es una técnica eminentemente exploratoria puesto que la mayor parte de las veces, no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación. Se la podría calificar como una técnica de aprendizaje no supervisado, es decir, una técnica muy adecuada para extraer información de un conjunto de datos sin imponer restricciones previas en forma de modelos estadísticos, al menos de forma explícita y, por ello, puede llegar a ser muy útil como una herramienta de elaboración de hipótesis acerca del problema considerado sin imponer patrones o teorías previamente establecidas.

Conviene, sin embargo, estar siempre alerta ante el peligro de obtener, como resultado del análisis, no una clasificación de los datos sino una disección de los mismos, en distintos grupos que sólo existen en la memoria del ordenador. El conocimiento que el analista tenga acerca del problema decidirá cuáles de los grupos obtenidos son significativos y cuáles no.

A continuación vamos a detallar las distintas etapas necesarias para llevar a cabo un análisis cluster, si bien, posteriormente concretaremos todas las etapas para el caso en estudio.

II. 9.3 Etapas de un análisis cluster.

Las etapas que componen cualquier análisis cluster son las siguientes:

1. Planteamiento del problema y selección de la muestra de datos.
2. Selección y transformación de las variables a utilizar.
3. Selección de concepto de distancia o similitud y medición de las mismas.
4. Selección y aplicación del criterio de agrupación o clasificación.
5. Interpretación de los resultados obtenidos.
6. Validación de la solución.

Cada una de estas etapas serán explicadas en detalle a continuación.

0. Planteamiento del problema y selección de la muestra de datos.

Esta etapa comprende los dos primeros pasos mencionados en el apartado del análisis multivariante, esto es, *objetivos del análisis* y *diseño del análisis*.

Se deben determinar el tamaño muestral, las ecuaciones a estimar y las distancias a calcular. Una vez determinado todo esto se proceden a observar los datos.

Como norma general el problema tendrá el siguiente esquema:

- Serán $X_1,...,X_p$ p variables numéricas observadas en n objetos.
- Sea x_{ij} = valor de la variable X_j en el i-ésimo objeto i=1,...,n; j=1,...,p.

1. Selección y transformación de variables.

En primer lugar vamos a enumerar ciertos criterios que hacen referencia a la idoneidad en la elección de las variables:

1. *No elegir variables indiscriminadamente*: cada estructura se manifiesta en una serie de variables y cada grupo de variables revela, sólo, una determinada estructura.

2. El resultado final es muy sensible a la inclusión de variables irrelevantes, por tanto es muy importante un análisis previo de cada variable.

3. La inclusión indiscriminada de variables aumenta la probabilidad de obtener resultados atípicos (outliers).

Transformación de variables.

La transformación de variables hace referencia a la realización de diversas modificaciones de éstas previo a la utilización de las técnicas estadísticas que vamos a emplear sobre ellas.

Antes de realizar estas transformaciones es necesario estudiar a las variables en profundidad para determinar si las transformaciones son o no necesarias, puesto que llevarlas a cabo puede tener algunos efectos. A continuación, se muestran algunas consideraciones que hay que tener en cuenta para la transformación de las variables de cualquier problema:

- **Depende / Afecta a** muchas decisiones posteriores (medida de distancia / similitud empleada, por ejemplo).

- **Estandarización por variable:** aunque resulta útil para mediciones posteriores de distancia puede afectar al resultado del análisis y no se recomienda si las diferencias de medidas reflejan alguna cualidad natural de interés conceptual.

- **Estandarización por encuestado:** singular, pero en baterías de indicadores elimina patrones de respuesta en los sujetos, ofreciendo la importancia relativa de cada indicador.

- **Factorización:** puede resultar interesante factorizar previamente las variables y realizar el Cluster con factores en lugar de con variables.

- **El tipo de escala de medida** afectará a fases posteriores del procedimiento.

Existen numerosas transformaciones de variables, pero entre las más importantes destacan las siguientes:

- Transformación Z. Esta transformación es de las más utilizadas. Es útil cuando las variables de nuestro problema tienen unidades de medida muy diversas. El procedimiento para realizar esta transformación es de sobra conocido, puesto que es el mismo que se utiliza para las funciones estadísticas Normales o de Gauss. Consiste en restar a la variable su media y dividirlo posteriormente por su desviación típica, esto es:

$$X_N = \frac{X - m_x}{\sigma_x}$$

II. Rango [-1,1]. Tras la aplicación de esta transformación todos los valores de la variable están comprendidos entre los valores -1 y 1, por tanto, se consigue una nueva variable con media 0.

II. Rango [0,1]. Al contrario que en la transformación anterior, aquí la media no es 0, puesto que no hay valores negativos. Con esta

transformación se pretende que el máximo valor de la variable sea la unidad.

II. Media distinta de 0. En algunas ocasiones puede ser deseable tener medias distintas de 0, para ello basta que los valores mínimo y máximo de la variable no sean simétricos respecto al 0.

II. Magnitud máxima de 1. Esta transformación es similar a Rango [-1,1], pero en esta ocasión se obliga a que el máximo tome el valor [-1,1].

III. 9.4 Medidas de proximidad y de distancia.

Una vez establecidas las variables y los objetos a clasificar, y realizadas las transformaciones si éste fuera el caso, el siguiente paso consiste en establecer una medida de proximidad o de distancia entre ellos que cuantifique el grado de similitud entre cada par de objetos.

Las **medidas de proximidad, similitud o semejanza** miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (resp. menor) es su valor, mayor (resp. menor) es el grado de similitud existente entre ellos y con más (resp. menos) probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo.

Las **medidas de disimilitud, desemejanza o distancia** miden la distancia entre dos objetos de forma que, cuanto mayor (resp. menor) sea su valor, más (resp. menos) diferentes son los objetos y menor (resp. mayor) la probabilidad de que los métodos de clasificación los pongan en el mismo grupo.

En la literatura existen multitud de medidas de semejanza y de distancia, éstas dependen del tipo de variables y de los datos considerados.

Las variables pueden tener muy diversas procedencias y valores, pero siempre podrán considerarse los siguientes tipos de datos:

1. **De intervalo:** se trata de una matriz objetos **x** variables donde todas las variables son cuantitativas, medidas en escala intervalo o razón.

2. **Frecuencias:** las variables analizadas son categóricas de forma que, por filas, tenemos objetos o categorías de objetos y, por columnas, las variables con sus diferentes categorías. En el interior de la tabla aparecen frecuencias.

3. **Datos binarios:** se trata de una matriz objetos **x** variables pero en la que las variables analizadas son binarias de forma que 0 indica la ausencia de una característica y 1 su presencia.

III. 9.4.1 Medidas de proximidad.

a) Medidas para variables cuantitativas.

1) Coeficiente de congruencia.

$$c_{rs} = \frac{\sum_{j=1}^{p} X_{rj} X_{sj}}{\sqrt{\sum_{j=1}^{p} X_{rj}^2} \sqrt{\sum_{j=1}^{p} X_{sj}^2}}$$

que es el coseno del ángulo que forman los vectores $(x_{r1},...,x_{rp})'$ y $(x_{s1},...,x_{sp})'$.

2) Coeficiente de correlación.

$$r_{rs} = \frac{\sum_{j=1}^{p} \left(X_{rj} - \overline{X}_r\right)\left(X_{sj} - \overline{X}_s\right)}{\sqrt{\sum_{j=1}^{p} \left(X_{rj} - \overline{X}_r\right)^2} \sqrt{\sum_{j=1}^{p} \left(X_{sj} - \overline{X}_s\right)^2}}$$

donde $\overline{X}_r = \frac{\sum_{j=1}^{p} X_{rj}}{p}$ y $\overline{X}_s = \frac{\sum_{j=1}^{p} X_{sj}}{p}$

Si los objetos r y s son variables, r_{rs} mide el grado de asociación lineal existente entre ambas.

Estas dos medidas se utilizan, preferentemente, para clasificar variables, siendo en este caso, invariantes por cambios de escala y, en el caso del coeficiente de correlación, invariante por cambio de origen. Por esta razón es más conveniente utilizar el coeficiente de congruencia con variables tipo razón en las cuales el origen está claramente definido.

Conviene observar además, que tanto c_{rs} como r_{rs} toman valores comprendidos entre -1 y 1 pudiendo tomar, por lo tanto, valores negativos. Dado que, en algunos casos, (por ejemplo, si los objetos a clasificar son variables), los valores negativos cercanos a -1 pueden implicar fuerte

semejanza entre los objetos clasificados, conviene en estas situaciones utilizar como medida de semejanza sus valores absolutos.

b) Medidas para datos binarios.

En este caso se construyen, para cada par de objetos r y s, tablas de contingencia de la forma:

Objeto s \ Objeto r	0	1
0	a	b
1	c	d

Tabla 7. Medidas para datos binarios en HCA.

donde:

"a" = número de variables en las que los objetos r y s toman el valor 0.

"b" = número de variables en las que el objeto r toma el valor 1 y el objeto s el valor 0.

"c" = número de variables en las que el objeto s toma el valor 1 y el objeto r el valor 0.

"d" = número de variables en las que los objetos r y s toman el valor 1.

"p" = a+b+c+d.

Utilizando dichas tablas algunas de las medidas de semejanza más utilizadas son:

- **Coeficiente de Jacard**: $\frac{d}{b+c+d}$
- **Coeficiente de acuerdo simple**: $\frac{a+d}{p}$

Ambas toman valores entre 0 y 1 y miden, en tanto por uno, el porcentaje de acuerdo en los valores tomados en las p variables, existente entre los dos objetos.

Difieren en el papel dado a los acuerdos en 0. El coeficiente de Jacard no los tiene en cuenta y el de acuerdo simple sí. Ello es debido a que en algunas situaciones las variables binarias consideradas son asimétricas en el sentido de que es más informativo el valor 1 que el valor 0. Así, por ejemplo, si el color de los ojos de una persona se codifica como 1 si tiene los ojos azules y 0 en caso contrario. En éste tipo de situaciones es más conveniente utilizar coeficientes tipo Jacard.

c) Medidas para datos nominales y ordinales.

Una generalización de las medidas anteriores viene dada por la expresión:

$$s_{rs} = \sum_{k=1}^{p} s_{rsk}$$

donde:

"s_{rsk}" es la contribución de la variable k-ésima a la semejanza total. Dicha contribución suele ser de la forma 1-d_{rsk} donde **"d_{rsk}"** es una distancia que suele tener la forma $\delta_{k\ell m}$ siendo ℓ el valor del estado de la variable X_k en el r-ésimo objeto y m el del s-ésimo objeto.

En variables nominales suele utilizarse $\delta_{k\ell m} = 1$ si ℓ=m y 0 en caso contrario. En variables ordinales suele utilizarse medidas de la forma $|\ell\text{-}m|^r$ con r>0.

III. 9.4.2 Medidas de distancia.

a) Medidas para variables cuantitativas.

Las más utilizadas son:

1) Distancia euclídea y distancia euclídea al cuadrado.

$$d = \sqrt{\sum_{j=1}^{p} \left(X_{rj} - X_{sj}\right)^2} \qquad d^2 = \sum_{j=1}^{p} \left(X_{rj} - X_{sj}\right)^2$$

2) Distancia métrica de Chebychev: $d = \max_i \left|x_{ri} - x_{si}\right|$

3) Distancia de Manhattan: $d = \sum_{i=1}^{p} \left|X_{ri} - X_{si}\right|$

4) Distancia de Minkowski:

$$d = \sqrt[q]{\sum_{i=1}^{p} \left(X_{ri} - X_{si}\right)^q}$$

con $q \in \mathbf{N}$.

Las tres primeras medidas son variantes de la distancia de Minkowski con q=2, ∞ y 1, respectivamente. Cuanto mayor es "q" más énfasis se le da a las diferencias en cada variable.

Todas estas distancias no son invariantes a cambios de escala por lo que se aconseja estandarizar los datos si las unidades de medida de las variables no son comparables. Además no tienen en cuenta las relaciones existentes entre las variables. Si se quieren tener en cuenta se aconseja utilizar la **distancia de Mahalanobis** que viene dada por la forma cuadrática:

$$d = \left(X_r - X_s\right)' S^{-1} \left(X_r - X_s\right)$$

donde:

$x_r = (x_{r1},\ldots,x_{rp})'$

$x_s = (x_{s1},\ldots,x_{sp})'$

S = Matriz de covarianzas del vector de variables $(x_{r1},\ldots,x_{rn})$

b) Medidas para tablas de frecuencias:

Suelen estar basadas en la χ^2 de Pearson. Algunas de las más utilizadas son:

$$\chi^2 = \sqrt{\sum_{i=1}^{p} \frac{\left(X_{ri} - E\left(X_{ri}\right)\right)^2}{E\left(X_{ri}\right)} + \sum_{i=1}^{p} \frac{\left(X_{si} - E\left(X_{si}\right)\right)^2}{E\left(X_{si}\right)}}$$

$$\varphi^2 = \sqrt{\frac{\sum_{i=1}^{p} \frac{\left(X_{ri} - E\left(X_{ri}\right)\right)^2}{E\left(X_{ri}\right)} + \sum_{i=1}^{p} \frac{\left(X_{si} - E\left(X_{si}\right)\right)^2}{E\left(X_{si}\right)}}{N}}$$

donde: $E\left(X_{ri}\right) = \frac{X_r X_i}{N}$ con $X_r = \sum_{i=1}^{p} X_{ri}$ y $X_i = X_{ri} + X_{si}$

es el valor esperado de la frecuencia x_{ri} si hay independencia entre los individuos r y s y las categorías 1,...,p de las variables y $N = x_{r.}+x_{s.}$ es el total de observaciones. La diferencia entre ambas medidas radica en la división por N en el caso de χ^2 para paliar la dependencia que tiene la χ^2 de Pearson respecto a N.

c) Medidas para datos binarios.

Las más utilizadas son:

Distancia euclídea al cuadrado: b+c

Lance y Williams: $\dfrac{b+c}{2d+b+c}$

Esta última ignora los acuerdos en 0.

d) Medidas para datos de tipo mixto.

Si en la base de datos existen diferentes tipos de variables: binarias, categóricas, ordinales, cuantitativas no existe una solución universal al problema de cómo combinarlas para construir una medida de distancia. Anderberg (1973) o Gordon (1990) sugieren las siguientes soluciones:

- Expresar todas las variables en una escala común, habitualmente binaria, transformando el problema en uno de los ya contemplados anteriormente. Esto tiene sus costes, sin embargo, en términos de pérdida de información si se utilizan escalas menos informativas como las nominales ú ordinales o la necesidad de incorporar información extra si se utilizan escalas más informativas como son las intervalo o razón.

- Combinar medidas con pesos de ponderación mediante expresiones de la forma:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} d_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

donde:

"d_{ijk}" es la distancia entre los objetos i y j en la k-ésima variable y w_{ijk} = 0 ó 1 dependiendo de si la comparación entre i y j es válida en la k-ésima variable

- Realizar análisis por separado utilizando variables del mismo tipo y utilizar el resto de las variables como instrumentos para interpretar los resultados obtenidos.

III. 9.5 Métodos de clasificación.

Entre los muchos tipos de métodos que existen en la literatura cabe destacar los siguientes:

- **Jerárquicos**: en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo.

- **Repartición:** tienen un número de grupos **"g"** fijado de antemano como objetivo y agrupa los objetos para obtener los **"g"** grupos. Comienzan con una solución inicial y los objetos se reagrupan de acuerdo con algún criterio de optimalidad.

- **Métodos tipo Q**: son similares al análisis factorial y utilizan como información la matriz **XX'** utilizando las variables como objetos y los objetos como variables.

- **Procedimientos de localización de modas:** agrupan los objetos en torno a modas con el fin de obtener zonas de gran densidad de objetos separadas unas de otras por zonas de poca densidad.

- **Métodos que permiten solapamiento:** permiten que los grupos tengan elementos en común.

En este proyecto se han utilizado métodos jerárquicos aglomerativos, así como el algoritmo de k-medias, que es un caso particular de métodos de repartición.

III. 9.5.1 Métodos jerárquicos.

Se caracterizan porque en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo.

Pueden ser, a su vez de dos tipos: aglomerativos y divisivos.

- *Los métodos aglomerativos* comienzan con n clusters de un objeto cada uno. En cada paso del algoritmo se recalculan las distancias entre los grupos existentes y se unen los 2 grupos más similares o menos disimilares. El algoritmo acaba con 1 cluster conteniendo todos los elementos.

- *Los métodos divisivos* comienzan con 1 cluster que engloba a todos los elementos. En cada paso del algoritmo se divide el grupo

más heterogéneo. El algoritmo acaba con n clusters de un elemento cada uno.

Para determinar qué grupos se unen o dividen se utiliza una función objetivo o criterio que, en el caso de los métodos aglomerativos recibe el nombre de *enlace.*

III. 9.5.1.1 Tipos de enlace.

Se utilizan con los métodos aglomerativos y proporcionan diversos criterios para determinar, en cada paso del algoritmo, qué grupos se deben unir. Cabe destacar los siguientes:

1. Enlace simple o vecino más próximo.

Mide la proximidad entre dos grupos calculando la distancia entre sus objetos más próximos o la similitud entre sus objetos más semejantes.

2. Enlace completo o vecino más alejado.

Mide la proximidad entre dos grupos calculando la distancia entre sus objetos más lejanos o la similitud entre sus objetos menos semejantes.

3. Enlace medio entre grupos.

Mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos o la media de las similitudes entre objetos de ambos grupos. Así, por ejemplo, si se utilizan distancias, la distancia entre los grupos r y s vendría dada por:

$$d_{rs} = \frac{1}{n_r n_s} \sum_{j \in r} \sum_{k \in s} d(j,k)$$

donde:

"d(j,k)" = distancia entre los objetos j y k.
"n_r, n_s" = son los tamaños de los grupos r y s, respectivamente.

4. Enlace medio dentro de los grupos.

Mide la proximidad entre dos grupos, utilizando la distancia media existente entre los miembros del grupo unión de los dos.

III. 9.5.1.2 Métodos del centroide y de la mediana.

Ambos métodos miden la proximidad entre dos grupos calculando la distancia entre sus centroides:

$$d_{rs}^2 = \sum_{j=1}^{p} \left(\overline{x}_{rj} - \overline{x}_{sj} \right)^2$$

donde:

"X_{rj} y X_{sj}" = son las medias de la variable X_j en los grupos r y s, respectivamente.

Los dos métodos difieren en la forma de calcular los centroides: el método del centroide utiliza las medias de todas las variables de forma que las coordenadas del centroide del grupo r = s ∪ t vendrán dadas por:

$$\overline{x}_{rj} = \frac{1}{n_r} \sum_{m=1}^{n_r} x_{rjm} \quad = \frac{n_s}{n_s + n_t} \overline{x}_{sj} + \frac{n_t}{n_s + n_t} \overline{x}_{tj}$$

j=1,…,p

En el método de la mediana el nuevo centroide es la media de los centroides de los grupos que se unen:

$$X_{rj} = \frac{1}{2} X_{sj} + \frac{1}{2} X_{tj}$$

III. 9.5.1.3 Método de Ward.

El método busca minimizar $\sum_r SSW_r$

donde **SSW_r** es, para cada grupo r, las sumas de cuadrados intragrupo que viene dada por:

$$SSW_r = \sum_{m=1}^{n_r} \sum_{j=1}^{p} \left(X_{rjm} - \overline{X}_{rj} \right)^2$$

donde:

"x_{rjm}" = el valor de la variable X_j en el m-ésimo elemento del grupo r.

$$SSW_t - SSW_r - SSW_s = \frac{n_r n_s}{n_r + n_s} d_{rs}^2$$

En cada paso del algoritmo une los grupos r y s que minimizan:

donde:

t = r $\cup$s y d_{rs}^2 la distancia entre los centroides de r y s.

III. 9.5.1.4 Comparación de los diversos métodos aglomerativos.

- El enlace simple conduce a clusters encadenados.
- El enlace completo conduce a clusters compactos.
- El enlace completo es menos sensible a outliers que el enlace simple.
- El método de Ward y el método del enlace medio son los menos sensibles a outliers.
- El método de Ward tiene tendencia a formar clusters más compactos y de igual tamaño y forma en comparación con el enlace medio.
- Todos los métodos salvo el método del centroide satisfacen la desigualdad ultramétrica:

$$d_{ut} \leq \min \{d_{ur}, d_{us}\} \qquad t = r \cup s$$

III. 9.5.1.5 Elección del número de grupos.

Existen diversos métodos de determinación del número de grupos. Algunos están basados en intentar reconstruir la matriz de distancias original, otros en los coeficientes de concordancia de Kendall y otros realizan análisis de la varianza entre los grupos obtenidos. No existe un criterio universalmente aceptado.

Dado que la mayor parte de los paquetes estadísticos proporcionan las distancias de aglomeración, es decir, las distancias a las que se forman cada grupo, una manera de determinar el número de grupos consiste en localizar en qué iteraciones del método utilizado dichas distancias dan grandes saltos.

Utilizando dichas distancias se pueden utilizar criterios como el *criterio de Mojena* que determina el primer s$\in$**N** tal que s+1 > $\overline{\alpha}$ + ks$_\alpha$ si se utilizan distancias y < si son similitudes donde $\{\alpha_j ; j=1,...,n-1\}$ son las distancias de aglomeración, $\overline{\alpha}$, s$_\alpha$ su media y su desviación típica respectivamente y k una cte entre 2.5 y 3.5.

III. 9.5.1.6 Método de las k-medias.

Este tipo de método es conveniente utilizarlo cuando los datos a clasificar son muchos y/o para refinar una clasificación obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

Existen varias formas de implementarlo pero todas ellas siguen, básicamente, los siguientes pasos:

1. Se seleccionan k centroides o semillas donde k es el número de grupos deseado.
2. Se asigna cada observación al grupo cuya semilla es la más cercana.
3. Se calculan los puntos semillas o centroides de cada grupo.
4. Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos.

El método suele ser muy sensible a la solución inicial dada por lo que es conveniente utilizar una que sea buena. Una forma de construirla es mediante una clasificación obtenida por un algoritmo jerárquico

III. 9.6 Interpretación de los resultados.

Interpretar la clasificación obtenida por un Análisis Cluster requiere, en primer lugar, un conocimiento suficiente del problema analizado. Hay que estar abierto a la posibilidad de que no todos los grupos obtenidos tienen por qué ser significativos.

Algunas ideas que pueden ser útiles en la interpretación de los resultados son las siguientes:

- Realizar ANOVAS y MANOVAS para ver qué grupos son significativamente distintos y en qué variables lo son.
- Realizar Análisis Discriminantes.
- Realizar un Análisis Factorial o de Componentes Principales para representar, gráficamente los grupos obtenidos y observar las diferencias existentes entre ellos.
- Calcular perfiles medios por grupos y compararlos.

III. 9.7 Validación de la solución.

Una vez obtenidos los grupos e interpretados los resultados conviene, siempre que sea posible, proceder a la validación de los mismos con el fin de averiguar, por un lado, hasta qué punto los resultados obtenidos son extrapolables a la población de la que vienen los objetos seleccionados y, por el otro, por qué han aparecido dichos grupos. Esta validación se puede realizar de forma externa o interna.

III. 9.7.1 Validez interna.

Se puede establecer utilizando procedimientos de validación cruzada. Para ello se dividen los datos en dos grupos y se aplica el algoritmo de clasificación a cada grupo comparando los resultados obtenidos. Por ejemplo, si el método utilizado es el de las k-medias se asignaría cada objeto de uno de los grupos al cluster más cercano obtenido al clasificar los datos del otro grupo y se mediría el grado de acuerdo entre las clasificaciones obtenidas utilizando los dos métodos.

III. 9.7.2 Validez externa.

Se puede realizar comparando los resultados obtenidos con un criterio externo (por ejemplo, clasificaciones obtenidas por evaluadores independientes o analizando en los grupos obtenidos, el comportamiento de variables no utilizadas en el proceso de clasificación) o realizando un Análisis Cluster con una muestra diferente de la realizada.