Introducción a los clusters jerárquicos

El problema de la clasificación

Clasificación o identificación es el proceso o acto de asignar un nuevo objeto u observación en su lugar correspondiente dentro de un conjunto de categorías. Los primeros intentos de clasificación científica proceden de las ciencias biológicas.

La segunda mitad de este siglo ha visto un gran aumento en el número de técnicas numéricas de clasificación disponibles. Este crecimiento ha ido paralelo con el desarrollo de los ordenadores, que son necesarios para poder realizar el gran número de operaciones aritméticas que se precisan. Asimismo, un desarrollo similar ha tenido lugar en las áreas de aplicación. Actualmente tales técnicas son usadas en campos como la arqueología, psiquiatría, astronomía e investigación de mercados. El uso actual se debe a la aportación de Sokal y Sneath en 1963: "Principios de taxonomía numérica".

Se han usado varios nombres para referirse a estas técnicas, entre ellos: Q-análisis (en psicología), Reconocimiento de patrones (informática), clusters y conglomerados (matemáticas y estadística).

En este curso veremos los más importantes usados para la IA. Veremos las librerías de R en las cuáles se implementan.

Introducción

Objetivo:

dividir un conjunto de individuos u objetos, descritos a través de un conjunto de p variables o con el conocimiento de las distancias o proximidades entre ellos, en subgrupos que difieran de alguna manera significativa (pequeñas variaciones dentro de los grupos en relación a las variaciones entre los grupos).

Similarmente, se puede plantear un análisis de conglomerados (clusters) de variables.

Se aplica a dos tipos de problemas:

- Partición de los datos: se dispone de un conjunto de datos, que se sospechan heterogéneos, y se desea dividirlo en grupos (a veces en número prefijado) de manera que cada elemento pertenezca a uno y solo uno de los grupos y que cada grupo sea internamente homogéneo.
- Clasificación de variables: agrupar variables con objeto de reducir la dimensión.

De esta forma, podemos distinguir entre clusters de datos u observaciones y clusters de variables.

Formalmente hablando, partimos de una matriz de datos m x n (m=número de individuos; n=número de variables), o bien de una matriz m x m cuyos elementos correspondan a las distancias entre los n individuos. Así, cada uno de los m individuos, está representado por un vector n-dimensional con los valores de ese individuo en las distintas variables.

$$\text{muestra } \Xi \text{ de } m \text{ individuos, } X_1, \ldots, X_m$$

$$X_j = (x_{j1}, x_{j2}, \ldots, x_{jn})', \; j = 1, \ldots, m$$

Debemos encontrar una partición de la muestra en "c" regiones o clusters:

$$\omega_1, \ldots, \omega_c$$

De forma que:

$$\bigcup_{i=1}^{c} \omega_i = \Xi$$

$$\omega_i \cap \omega_j = \emptyset \quad ; \quad i \neq j$$

Matricialmente, lo que tenemos es una matriz que proporciona los valores de las variables para cada individuo:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

De esta forma, los procedimientos que buscan clasificar individuos se realizan sobre esta matriz, mientras que los que clasifican variables, se aplican sobre la matriz traspuestas. Es decir, todos los procedimientos, técnicamente, podrían aplicarse a ambos tipos de clusterización. Estas técnicas son muy necesarias, pues una búsqueda exhaustiva de todas las

particiones suele ser prohibitiva. Viene dada por un número de Stirling de segunda clase. Si queremos clasificar m observaciones en k grupos, el número de Stirling es:

$$\mathbb{S}_m^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^m$$

Si encima, el número de grupos es desconocido, para m observaciones, tendríamos que el número de posibles clusters sería:

$$\sum_{j=1}^{m}\mathbb{S}_m^{(j)}$$

De esta forma, podría clasificarse como una técnica exploratoria 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.

Además, los resultados obtenidos con el análisis de conglomerados, se puede aprovechar para determinados análisis posteriores como regresiones. También es importante decir que distintos procedimientos clusters pueden generar soluciones diferentes sobre el mismo conjunto de datos.

Tipos de Clusters

Existen dos grandes familias de algoritmos de cluster: métodos jerárquicos y métodos no jerárquicos.

Métodos jerárquicos

Estos métodos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan el análisis con tantos grupos como individuos haya en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que, al final del proceso, todos los casos están englobados en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso al anterior. Empiezan con un conglomerado que engloba a todos los individuos. A partir de este grupo inicial se van formando, a través de sucesivas divisiones, grupos cada vez más pequeños. Al final del proceso se tienen tantos grupos como

individuos en la muestra estudiada. Independientemente del proceso de agrupamiento, hay diversos criterios para ir formando los conglomerados; todos estos criterios se basan en una matriz de distancias o similitudes.

Dentro de estos métodos, destacan:

- Amalgamiento simple (simple linkage)
- Amalgamiento completo (complete linkage)
- Amalgamiento promedio (average linkage)
- Amalgamiento por centroide ponderado
- Amalgamiento por centroide mediana
- Método de Ward

Métodos No-Jerárquicos

Tienen por objetivo realizar una sola partición, generalmente en un número de clusters previamente establecidos, aunque también existen métodos que no necesitan esta información previa. Existen cuatro grandes grupos, si bien, solo vamos a entrar en detalle de los dos primeros grupos y vamos a mencionar los métodos más comunes. Existen otros que encajan también en esta clasificación que son variaciones de los anteriores.

- Métodos de Reasignación
    a. K-Means
    b. Forgy

- Métodos de Densidad

    a. Basados en densidad topológica: Mean Shift y DBSCAN
    b. Basados en densidad de probabilidad: EM-GMM

- Métodos directos. Buscan agrupar simultáneamente individuos y variables: Block clustering

- Métodos reductivos. Específico para variables: Factorial Q

Etapas en el proceso de análisis de clusters

A. Elección de las variables. Son las características que define a cada individuo. Hay que evitar tomar demasiadas y tener en cuenta las unidades en las que están.

B. Elección de la medida de asociación que se va a usar para determinar la semejanza de observaciones o variables. Normalmente, para individuos se usan las distancias, mientras que para variables, las correlaciones. El siguiente apartado se centra en analizarlas más exhaustivamente.

C. Elección de la técnica de cluster.

D. Valoración de los resultados.

Medidas de Asociación

Una vez considerado que el objetivo del Análisis de clúster consiste en encontrar agrupaciones naturales del conjunto de individuos de la muestra, es necesario definir qué se entiende por agrupaciones naturales y, por lo tanto, con arreglo a qué criterio se puede decir que dos grupos son más o menos similares.

Definiciones preliminares

Distancia:

Definición 2.1 *Sea U un conjunto finito o infinito de elementos. Una función $d : U \times U \longrightarrow \mathbb{R}$ se llama una distancia métrica si $\forall x, y \in U$ se tiene:*

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)\ ,\ \forall z \in U$

Similitud:

Definición 2.2 *Sea U un conjunto finito o infinito de elementos. Una función $s : U \times U \longrightarrow R$ se llama similaridad si cumple las siguientes propiedades: $\forall x, y \in U$*

1. $s(x, y) \leq s_0$
2. $s(x, x) = s_0$
3. $s(x, y) = s(y, x)$

donde s_0 es un número real finito arbitrario.

Basadas en distancias

(Las dimensiones correspondientes a las categorías, es decir, las variables que describen a los individuos son "no dicotómicas"):

Distancia euclídea

$$d_2(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{(x_i - x_j)' (x_i - x_j)} = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}$$

Distancia de Minkowski

$$d_p (x_i, x_j) = ||x_i - x_j||_p = \left(\sum_{l=1}^{n} |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}}$$

Dos variantes de esta distancia son la City Block y la de Chebyshev:

1. Distancia d_1 o distancia ciudad (City Block) $(p = 1)$

$$d_1 (x_i, x_j) = \sum_{l=1}^{n} |x_{il} - x_{jl}|$$

2. Distancia de Chebychev o distancia del máximo $(p = \infty)$

$$d_\infty (x_i, x_j) = \underset{l=1,\ldots,n}{\text{Max}} |x_{il} - x_{jl}|$$

Distancia de Mahalanobis

$$D_S(x_i, x_j) = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

Donde la matriz S es la matriz de varianzas covarianzas de la matriz de datos.

Coeficiente de Bray-Curtis

Dados dos individuos:

$$x_i = (x_{i1}, \ldots, x_{in})'$$

$$x_j = (x_{j1}, \ldots, x_{jn})'$$

El coeficiente de Bray-Curtis viene dado por:

$$D_{i,j} = \frac{\sum_{l=1}^{n} |x_{il} - x_{jl}|}{\sum_{l=1}^{n} (x_{il} + x_{jl})}$$

Adaptadas de otros conceptos estadísticos

Encontramos dos tipos principales: Derivadas de la correlación ycDerivadas de la contingencia cuadrática χ^2:

Derivadas de la correlación

Como ya se introdujo, la matriz de datos donde las filas son individuos y las columnas los valores que estos adquieren en las diferentes variables puede usarse de manera indistinta trasponiéndola para variables y observaciones. De esta manera, es posible obtener el coeficiente de correlación entre dos individuos de manera habitual. Sean dos individuos "i" y "j" como:

$$r_{ij} = \frac{\sum_{l=1}^{n} (x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{s_i s_j}$$

Derivadas de la contingencia cuadrática χ^2

El estadístico χ^2 puede considerarse también una medida de similitud entre dos individuos (tal y como lo es entre dos variables), de forma que, a más tendente a cero, más proporcionalidad y por tanto, se podría interpretar que dos individuos tienen perfiles similares.

Medidas para datos en binario

Partiendo de una tabla de contingencia sobre individuos, no variables:

Ind. I\Ind. J	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$n=a+b+c+d$

Donde “a”, por ejemplo, representa el número de variables donde el individuo “i” y “j” presentan simultáneamente el mismo valor (1), “b” el número de variables donde “i” presenta el valor 1 y “j” el valor 0, etc… Es posible definir un amplio conjunto de medidas de asociación. Sin embargo, para mejorar la comprensión, se presentarán a continuación aplicadas sobre variables, si bien su significado puede extenderse a este caso.

Medidas de Asociación entre variables

Para poder agrupar variables es necesario poder establecer medidas que muestren la similitud entre elles. Veamos algunas de las más importantes, pero hay que tener en cuenta que las que presentamos miden directamente la similaridad, de forma que, a mayor valor, más similaridad (en distancia es al revés, mayor distancia implica menos similaridad.

Medidas para variables dicotómicas

$X_i \backslash X_j$	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$m=a+b+c+d$

En la anterior tabla se tiene:

1. a representa el número de individuos que toman el valor 1 en cada variable de forma simultánea.
2. b indica el número de individuos de la muestra que toman el valor 1 en la variable X_i y 0 en la X_j.
3. c es el número de individuos de la muestra que toman el valor 0 en la variable X_i y 1 en la X_j.
4. d representa el número de individuos que toman el valor 0 en cada variable, al mismo tiempo.
5. $a+c$ muestra el número de veces que la variable X_j toma el valor 1, independientemente del valor tomado por X_i.
6. $b+d$ es el número de veces que la variable X_j toma el valor 0, independientemente del valor tomado por X_i.
7. $a+b$ es el número de veces que la variable X_i toma el valor 1, independientemente del valor tomado por X_j.
8. $c+d$ es el número de veces que la variable X_i toma el valor 0, independientemente del valor tomado por X_j.

Las medidas más comunes, también aplicables al caso de la similitud entre individuos con datos binarios, son:

Medida de Russell y Rao

$$\frac{a}{a+b+c+d} = \frac{a}{m}$$

Medida parejas simples

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{m}$$

Medida de Jaccard

$$\frac{a}{a+b+c}$$

Medida de Rogers-Tanimoto

$$\frac{a+d}{a+d+2(b+c)}$$

Medida de Kulcynsky

$$\frac{a}{b+c}$$

Medidas para variables cuantitativas

La más extendida, que mide la similitud entre dos variables es, obviamente, el coeficiente de correlación:

$$r = \frac{\mathrm{Cov}(x_i, x_j)}{(\mathrm{Var}(x_i)\,\mathrm{Var}(x_j))^{\frac{1}{2}}} = \frac{\sum_{l=1}^{m}(x_{li} - \overline{x}_i)(x_{lj} - \overline{x}_j)}{\left(\sum_{l=1}^{m}(x_{li} - \overline{x}_i)^2 \sum_{l=1}^{m}(x_{lj} - \overline{x}_j)^2\right)^{\frac{1}{2}}}$$

Paso de distancia a similitud

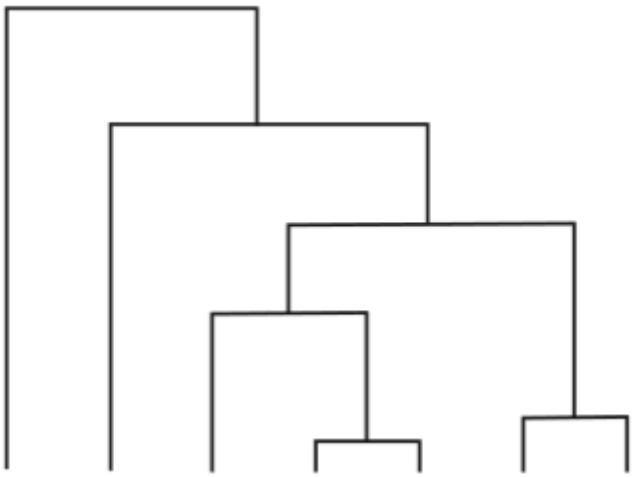
Es posible pasar de distancia a similitud mediante la expresión

$$dist(\underline{x}_i, \underline{x}_j) = \sqrt{2\left[1 - simil(\underline{x}_i, \underline{x}_j)\right]}$$

Métodos de clusters jerárquicos

En una clasificación no jerárquica se forman grupos homogéneos sin establecer relaciones entre ellos. Por el contrario, en una clasificación jerárquica, los grupos se van fusionando (o subdividiendo) sucesivamente, siguiendo una prelación o jerarquía, decreciendo la homogeneidad conforme se van haciendo más amplios. Como se desprende, pueden ser aglomerativos (ascendentes o forward), si el método empieza considerando cada observación aisladamente y empieza a conglomerar hasta llegar a un solo cluster que engloba a todas las observaciones, o bien, disociativos (descendentes o backwards), si parten de la consideración de un solo cluster que engloba a todas las observaciones y empieza a dividirlo hasta llegar al punto donde cada observación aislada pertenece al cluster formado por ella misma.

La dinámica de estos métodos se representa muy claramente con el dendograma:



Si prestamos atención, la parte superior representa a un solo cluster que engloba todas las observaciones, mientras que la inferior representa a cada observación como un cluster. Si usamos el camino de abajo a arriba, estaremos ante un método aglomerativo. Si usamos el camino de arriba abajo, estaremos ante uno disociativo.

Métodos jerárquicos aglomerativos

Se diferencian en la estrategia que decide cuándo dos individuos o clusters son similares para llevar a cabo una fusión entre ellos, es decir, en cómo define la distancia entre dos individuos o clusters.

Simple linkage (amalgamiento simple o distancia mínima)

Según este método, la distancia o similitud entre dos clusters viene dada, respectivamente, por la mínima distancia (o máxima similitud) entre sus componentes.

Así, tras la etapa K-ésima, habrá formados n-K clusters. La distancia entre los clusters C_i (que tiene n_i elementos) y C_j (con n_j elementos), vendrá dada por:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Min}} \{d(x_l, x_m)\} \quad l = 1, \ldots, n_i \; ; \; m = 1, \ldots, n_j$$

En caso de usar medidas de similitud:

$$s(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Max}} \{s(x_l, x_m)\} \quad l = 1, \ldots, n_i \; ; \; m = 1, \ldots, n_j$$

De esta manera, se unirán en la etapa K+1 los clusters con menor distancia o mayor similitud según la definición anterior.

Complete linkage (amalgamiento completo o distancia máxima)

Según este método, la distancia o similitud entre dos clusters viene dada, respectivamente, por la distancia entre sus componentes más dispares, es decir, entre los que tienen mayor distancia o mínima similitud.

Así, tras la etapa K-ésima, habrá formados n-K clusters. La distancia entre los clusters C_i (que tiene n_i elementos) y C_j (con n_j elementos), vendrá dada por:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Max}} \{d(x_l, x_m)\} \quad l = 1, \ldots, n_i \; ; \; m = 1, \ldots, n_j$$

En caso de usar medidas de similitud:

$$s(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Min}} \{s(x_l, x_m)\} \quad l = 1, \ldots, n_i \; ; \; m = 1, \ldots, n_j$$

De esta manera, se unirán en la etapa K+1 los clusters con menor distancia o mayor similitud según la definición anterior.

Average linkage (amalgamiento promedio)

Según este método, la distancia o similitud entre dos clusters viene dada, respectivamente, por el promedio de la distancia (o de la similitud) entre sus componentes.

Así, tras la etapa K-ésima, habrá formados n-K clusters. La distancia entre los clusters C_i (que, supongamos, es el que ha experimentado la última unión entre dos elementos C_{i1} y C_{i2} en la etapa K-1), y C_j (con n_j elementos), vendrá dada por:

$$d(C_i, C_j) = \frac{d(C_{i_1}, C_j) + d(C_{i_2}, C_j)}{2}$$

En caso de usar medidas de similitud, la formulación sería idéntica, y de esta manera, se unirán en la etapa K+1 los clusters con menor distancia o mayor similitud según la definición anterior.

Nota: es posible la versión ponderada de este método, de forma que, si n_{i1} y n_{i2} son el número de elemento de C_{i1} y C_{i2} respectivamente, la distancia vendría dada por:

$$\frac{n_{i_1} d(C_{i_1}, C_j) + n_{i_2} d(C_{i_2}, C_j)}{n_{i_1} + n_{i_2}}$$

Métodos basados en centroides

En estos métodos, la semejanza entre dos clusters viene determinada por la semejanza entre sus centroides, es decir, los puntos que tienen por coordenadas las medias de las variables medidas sobre los individuos que forman parte del cluster. Dentro de estos métodos se distinguen dos: centroide ponderado y centroide no ponderado.

Método del centroide ponderado.

Para medir la distancia entre los clusters C_i (que, supongamos, es el que ha experimentado la última unión entre dos elementos C_{i1} y C_{i2} en la etapa K-1, donde n_{i1} y n_{i2} son el número de elemento de C_{i1} y C_{i2} respectivamente y m^{i1} y m^{i2} sus respectivos centroides), y C_j (con n_j elementos y centroide m^j), vendrá dada por la distancia entre los centroides de C_i (denotado por m^i) y de C_j (denotado por m^j), donde:

$$m^i = \frac{n_{i_1} m^{i_1} + n_{i_2} m^{i_2}}{n_{i_1} + n_{i_2}}$$

De esta manera, se unirán en la etapa K+1 los clusters con menor distancia o mayor similitud medida a través de la distancia o similitud de sus centroides.

Método del centroide no ponderado o de la mediana.

Si los tamaños de los sub-clusters que componen un cluster (n_{i1} y n_{i2}) son muy diferentes, el centroide del cluster que los aglomera (C_i) estará demasiado influenciado por el tamaño del cluster con más elementos. Por ello, la estrategia del cluster mediano, no pondera los centroides de cada sub-cluster por su tamaño, sino que considera $n_{i1} = n_{i2}$. Salvo esta diferencia, la estrategia es análoga a la anterior. Este método por lo general es más recomendable que el anterior. Si desarrollamos la formulación de las distancias usando los centroides, tendremos que:

$$d(C_i, C_j) = \frac{1}{2}\left[d(C_{i_1}, C_j) + d(C_{i_2}, C_j)\right] - \frac{1}{4} d(C_{i_1}, C_{i_2})$$

De esta manera, se unirán en la etapa K+1 los clusters con menor distancia o mayor similitud medida a través de la distancia o similitud de sus centroides.

Método de Ward

Ward propuso que la pérdida de información que se produce al integrar los distintos individuos en clusters puede medirse a través de la suma total de los cuadrados de las desviaciones entre cada punto (individuo) y la media del cluster en el que se integra, considerando como tal media, su centroide. Para que el proceso de clusterización resulte óptimo, en el sentido de que los grupos formados no distorsionen los datos originales, proponía la siguiente estrategia:

En cada paso del análisis, considerar la posibilidad de la unión de cada par de grupos y optar por la fusión de aquellos dos grupos que, al unirse, menos incrementen la suma de los cuadrados de las desviaciones con el centroide resultado de la unión.

El método de Ward es uno de los más utilizados en la práctica. Una investigación llevada a cabo por Kuiper y Fisher probó que este método era capaz de acertar mejor con la clasificación óptima que otros métodos (mínimo, máximo, promedio y centroide).

Fórmula de Lance y Williams

as distintas distancias entre grupos definidas en los métodos anteriores se pueden expresar a través de una única fórmula recurrente de cuatro parámetros; de forma que, para los distintos valores de éstos se generan las distintas distancias. En efecto, si consideramos el grupo formado por la fusión de los grupos I, J, (I,J) y el grupo exterior K, la distancia entre (I,J) y K puede expresarse como:

$$d((I,J),K) = a_I\, d(I,K) + a_J\, d(J,K) + b\, d(I,J) + g\,|d(I,K) - d(J,K)|$$

Diferentes valores de a_I, a_J, b y g determinarán las distancias según los diferentes métodos. A modo ilustrativo, en el caso del método amalgamiento simple:

$$a_I = a_J = 1/2;\ b = 0;\ g = -\,½$$

Métodos jerárquicos disociativos

Constituyen el proceso inverso al aglomerativo, partiendo de un único cluster que engloba todas las observaciones hasta llegar a un punto donde cada observación es un cluster en sí misma. Son bastante menos populares que los anteriores y su filosofía es exactamente la misma. Son los mismos métodos, con los mismos nombres, pero en este caso, de un grupo se elimina o saca del cluster aquel individuo cuya distancia sea mayor, o similaridad menor, al cluster que formarían el resto de los individuos del cluster. Tras la primera etapa, donde se tendrá un cluster unitario y otro con el resto de los individuos, se añadirá al cluster unitario aquel elemento cuya distancia (similaridad) total al resto de los elementos que componen su actual cluster menos la distancia (similaridad) al cluster anteriormente formado sea máxima (mínima). Cuando esta diferencia sea negativa dicho elemento no se añade y se repite el proceso sobre los dos subgrupos.

Consideraciones postanálisis

Los métodos jerárquicos tienen en común que no determinan el número de clusters óptimos, sino que es necesario que el investigador determine en qué nivel es conveniente situar la clusterización final. Por ello, no deja de someterse a un punto de vista objetivo. Sin embargo, adoptar otra óptica donde la “calidad” de un número de clusters sea mejor que la de otro número tampoco dejará de ser nunca subjetiva, aunque se determinen mediante métodos matemáticos, pues la definición de “calidad” siempre será subjetiva y puede referirse a multitud de magnitudes matemáticas para expresarla según el significado que se le quiera dar. Por ello, lo más recomendable es que el investigador conozca la naturaleza de los métodos a usar y del problema al que se enfrenta, teniendo siempre en mente la hipótesis a la que se enfrenta. A modo de ejemplo, Beale, propuso un contraste basado en la función F de Snedecor, donde compara la hipótesis de la existencia de c_2 clusters frente a la existencia de c_1 clusters, con $c_2 > c_1$. A partir de las siguientes medidas de desviación cuadrática:

$$DC_1 = \frac{1}{n - c_1} \sum_{i=1}^{c_1} \sum_{j=1}^{n_i} ||x_{ij} - \overline{x}_i||^2$$

$$DC_2 = \frac{1}{n - c_2} \sum_{i=1}^{c_2} \sum_{j=1}^{n_i} ||x_{ij} - \overline{x}_i||^2$$

A partir de ellas genera el estadístico

$$F(p(c_2 - c_1), p(n - c_2)) = \frac{\frac{DC_1 - DC_2}{DC_2}}{\left[\left(\frac{n - c_1}{n - c_2}\right)\left(\frac{c_2}{c_1}\right)^{\frac{2}{p}} - 1\right]}$$

Un resultado significativo implicaría que una división en c_2 clusters sería una mejoría frente a la existencia de c_1 clusters.