ANALISIS DE CONGLOMERADOS

Jorge Galbiati R.

Consiste en buscar grupos (conglomerados) en un conjunto de observaciones de forma tal que aquellas que pertenecen a un mismo grupo se parecen, mientras que aquellas que pertenecen a grupos distintos son disímiles, según algún criterio de distancia o de similitud.

Los algoritmos de formación de conglomerados se agrupan en dos categorías:

Algoritmos de partición: Método de dividir el conjunto de observaciones en k conglomerados, en que k lo define inicialmente el usuario.

Algoritmos jerárquicos: Método que entrega una jerarquía de divisiones del conjunto de elementos en conglomerados.

Un método jerárquico aglomerativo parte con una situación en que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas están en un único conglomerado.

Un método jerárquico divisivo sigue el sentido inverso: Parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado distinto.

DISTANCIAS ENTRE ELEMENTOS: DISIMILARIDADES

Se dispone de una matriz de datos $X_{n\times p} = (x_{ic})$ en que i representa una observación y c una variable.

Una matriz de disimilaridades o distancias $D_{n\times n}$ es una matriz tal que su elemento i, j es una disimilaridad $d(ij)$ tal que para todoi, j, k:

1. $d(i, j) \geq 0$
2. $d(i, i) = 0$
3. $d(i, j) = d(j, i)$
4. $d(i, j) \leq d(i, k) + d(k, j)$

D es simétrica y su diagonal está formada por ceros.

La disimilaridad $d(i, j)$ representa una medida de la diferencia entre dos observaciones x_i y x_j y constituyen la base para la formación de conglomerados.

A continuación se muestra una colección de las principales medidas de disimilaridad, según el tipo de escala de medida de las variables.

Algunas medidas de disimilaridad Hay varias medidas de disimilaridad o distancia, apropiadas para diferentes tipos de escala en que se miden los datos: escala numérica lineal, numérica no lineal, ordinales, nominales y nominales binarios.

Escalas numéricas

1. Distancia Euclidea:

$$d(i,j) = \sqrt{\sum_{c=1}^{p}(x_{ic} - x_{jc})^2}$$

2. Distancia Manhatan o City Block

$$d(i,j) = \sum_{c=1}^{p} |x_{ic} - x_{jc}|$$

3. Distancia de Minkowski

Es una generalización de las anteriores:

$$d(i,j) = (\sum_{c=1}^{p} |x_{ic} - x_{jc}|^q)^{\frac{1}{q}}$$

en que q es cualquier número real mayor o igual que1.

4. Distancia de correlación

El coeficiente de correlación es una medida de proximidad o similitud entre dos series de datos. Por lo tanto, a partir de él se puede definir una medida de disimilaridad:

$$d(i,j) = (1 - corr(i,j))/2$$

Esta medida tiene un rango de valores entre 0 y 1.

5. Estandarización de variables:

La unidad de medida de las variables afecta el resultado. Si las variables tienen órdenes de magnitud muy distintas, es conveniente estandarizarlas previamente:

$$z_{ic} = \frac{x_{ic} - m_c}{s_c}$$

en que m_c y s_c son medidas muestrales de centro y dispersión respectivamente, ambas medidas en la misma escala de x_{ic}.

Por ejemplo, las más conocidas, media y desviación standard muestrales.

$$m_c = \frac{1}{n}\sum_{i=1}^{n} x_{ic} \qquad \text{y} \qquad s_c = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ic} - m_c)^2}$$

Medidas con escala númericas no lineales

6. Son medidas efectuadas en escalas no lineales, por ejemplo exponencial, cuadratica, etc .(Ej: un crecimiento bacteriano con función de crecimiento Ac^{Bt}, o un índice porcentual), se tratan como ordinales o bien se les aplica una transformación para linealizarlos, y se aplica cualquier medida para escalas lineales.

Escalas ordinales

Se conoce el orden pero no la magnitud de las observaciones.

7. Se obtiene una medida de disimilaridad mediante el siguiente procedimiento:

a. Reemplazar x_{ic} por su rango $r_{ic} \in \{1, ..., M_c\}$ dentro de la columna.

b. Transformar a la escala entre 0 y 1, haciendo: $z_{ic} = \frac{r_{ic}-1}{M_c-1}$

c. Calcular las disimilariaddes como en el caso de las escalas de numéricas.

Escalas nominales

Por ejemplo, resultados de una encuesta, en que cada encuestado responde a una serie de preguntas (variables) en escalas {a, b, c, ...}. La medida de disimilaridad entre dos encuestados es la proporción de respuestas en que difieren.

9. Caso general.

$$d(i,j) = \frac{\text{N° de variables con valores diferentes}}{\text{p}}$$

Variables en escala nominal binaria

Son variables con dos valores, se pueden codificar con 0 y 1. La siguiente es la tabla de contingencia para las observaciones "i" y "j".

$i \setminus j$	1	0
1	a	b
0	c	d

10. Si las variables son simétricas (ambos valores igualmente importantes), se define una medida de disimilaridad como

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

11. Si las variables son asimétricas (una de ellas, codificada 1, es más importante que la otra, codificada 0). Por ej., la presencia o ausencia de un cierto atributo.Se define una medida que sólo considera el universo de aquellos en que el atributo está presente, llamado coeficiente de Jaccard:

$$d(i,j) = \frac{b+c}{a+b+c}$$

Se excluye d, el número de comparaciones en que ambas variables valen 0.

Variables mixtas

12. Para observaciones constituidas por combinaciones de variables con escalas diferentes hay medidas de distancia que combinan medidas de los tipos anteriores, según el tipo de variable, ponderadas de manera conveniente.

DISTANCIAS ENTRE CONGLOMERADOS

Las distancias entre los conglomerados son funciones de las distancias entre observaciones, y hay varias formas de definirlas: Sean A y B dos conglomerados.

Vecino más cercano

$$d(A,B) = \min_{\substack{i\in A \\ j\in B}} d(i,j)$$

Vecino más lejano

$$d(A,B) = \max_{\substack{i\in A \\ j\in B}} d(i,j)$$

Promedio de grupo

$$d(A,B) = \frac{1}{n_A n_B} \sum_{\substack{i\in A \\ j\in B}} d(i,j)$$

Centroide

$$d(A, B) = d(\overline{x}_A, \overline{x}_B)$$

en que $\overline{x}_A$ y $\overline{x}_B$ son los respectivos centroides de los conglomerados A y B.

El siguiente gráfico ilustra las distancias entre conglomerados: Vecino más cercano, vecino más lejano, promedio del grupo y centroide, respectivamente.

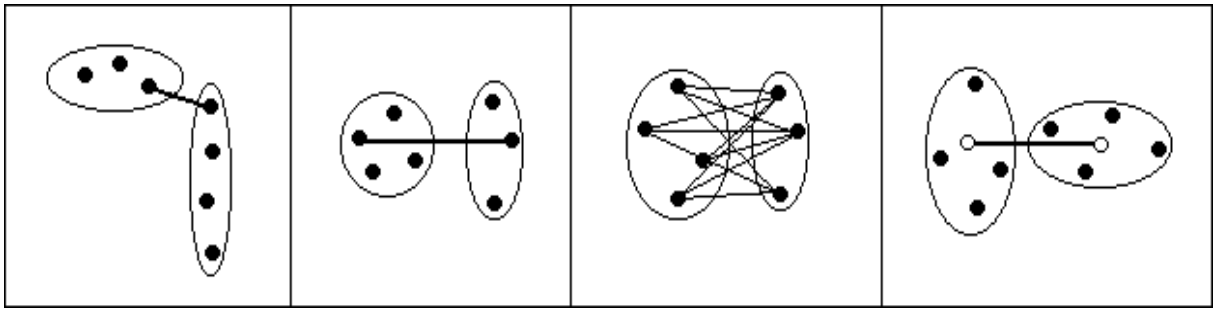

Figura 1: Distancias entre conglomerados: Vecino más cercano, más lejano, promedio, centro gravedad.

Medoide

Es la distancia entre los medoides de los grupos.

El vecino más cercano tiende a formar conglomerados más alargados.

El vecino más lejano forma conglomerados más esfericos.

El promedio de grupo y el centroide son más robustos que los demás.

El medoide es la más robusta de las distancias entre conglomerados.

METODOS DE PARTICION.

Se mostrarán dos métodos de análisis de conglomerados no jerérquicos, o de partición, el de las k-medias y el de ás k-medoides, de los cuales el métododе las k-medias es el más conocido, y otros dos derivados de los anteriores.

1. Método K-medias. Es un método iterativo que consiste en los siguiente procedimiento. El usuario debe proporcionar el númro k de conglomerados que desea tener. También se debe definir una medida de distancia:

a. Particionar el conjunto de observaciones en k grupos iniciales *arbitrarios*.

b. Recorrer todas las observaciones, asignándolas al conglomerado cuyo centroide esté a menor distancia. Cada vez que se reasigna una observación a un conglomerado distinto del que la contenía s deben recalcular los centroides del conglomerado que pierde la observación y del que la recibe.

Si el conglomerado A (que consiste en n_A observaciones) pierde la observación x_i y si el conglomerado B (con n_B observaciones) recibe a x_i , los centroides respectivos $\overline{x}_A$ y $\overline{x}_B$ se modifican de la siguiente forma:

$$\underline{\overline{x}}'_A = \frac{1}{n_A - 1}(n_A \underline{\overline{x}}_A - \underline{x}_i)$$

$$\underline{\overline{x}}'_B = \frac{1}{n_B + 1}(n_B \underline{\overline{x}}_B + \underline{x}_i)$$

c. Repetir el paso b hasta que no haya más reasignaciones.

EJEMPLO:

Se tienen 4 observaciones cuya matriz de datos está dada a continuación:

$$\begin{bmatrix} 0 & 3 & 9 & 12 \\ 4 & 1 & 6 & 10 \\ 10 & 7 & 3 & 4 \\ 10 & 10 & 3 & 1 \end{bmatrix}$$

Se usará el método de las k-medidas para formar dos conglomerados. También se utilizarán las distancias euclidea.

En forma de vectores, las cuatro observaciones (filas) son:

$$\underline{x}_1 = \begin{bmatrix} 0 \\ 3 \\ 9 \\ 12 \end{bmatrix} \qquad \underline{x}_2 = \begin{bmatrix} 4 \\ 1 \\ 6 \\ 10 \end{bmatrix} \qquad \underline{x}_3 = \begin{bmatrix} 10 \\ 7 \\ 3 \\ 4 \end{bmatrix} \qquad \underline{x}_4 = \begin{bmatrix} 10 \\ 10 \\ 3 \\ 1 \end{bmatrix}$$

Definimos arbitrariamente dos conglomerados iniciales . Sean

$$A = \{\underline{x}_1\} \quad \text{y} \quad B = \{\underline{x}_2, \underline{x}_3, \underline{x}_4\}$$

Sus centroides respectivos son:

$$\overline{\underline{x}}_A = \begin{bmatrix} 0 \\ 3 \\ 9 \\ 12 \end{bmatrix} \qquad y \qquad \overline{\underline{x}}_B = \begin{bmatrix} 8 \\ 6 \\ 4 \\ 5 \end{bmatrix}$$

Algoritmo Iterativo:

Se deben calcular las distancias de cada observación a los centroides de cada conglomerado. Si una observación está a menor distancia del conglomerado vecino, se cambia de conglomerado, se recalculan los centroides y se pasa a la siguiente iteración.

Iteración 1

Cuadro de distancias euclideas (al cuadrado) de las observaciones a los centroides, partiendo por $\overline{\underline{x}}_1$

centroide / observación	$\overline{\underline{x}}_A$	$\overline{\underline{x}}_B$
$\underline{x}_1$	0	147
$\underline{x}_2$	33	70

Cambia $\underline{x}_2$ del conglomerado B a A y termina la iteración 1. No es necesario seguir probando con $\underline{x}_3$ ni $\underline{x}_4$.

Iteración 2

Nuevos centroides, recalculados. Ahora $A = \{\underline{x}_1, \underline{x}_2\}$ y $B = \{\underline{x}_3, \underline{x}_4\}$

$$\overline{\underline{x}}_A = \begin{bmatrix} 2 \\ 2 \\ 7{,}5 \\ 11 \end{bmatrix} \qquad \overline{\underline{x}}_B = \begin{bmatrix} 10 \\ 8{,}5 \\ 3 \\ 2{,}5 \end{bmatrix}$$

Cuadro de distancias al cuadrado, partiendo de $\underline{x}_3$:

centroide / observación	$\overline{\underline{x}}_A$	$\overline{\underline{x}}_B$
$\underline{x}_3$	158.25	4.5
$\underline{x}_4$	248.25	4.5
$\underline{x}_1$	8.25	256.5
$\underline{x}_2$	8.25	157.5

Las cuatro observaciones quedaron bien clasificadas, luego ya no hay más cambios, por lo tanto los dos conglomerdos resultantes son:

$$A = \{\underline{x}_1, \underline{x}_2\} \qquad \text{y} \qquad B = \{\underline{x}_3, \underline{x}_4\}$$

2. Método K-Medoides.

Es como el k-medias, pero usa los medoides en lugar de los centroides.

El medoide es el punto tal que sus coordenadas son las medianas de las variables respectivas.

3. Conglomerados para conjuntos grandes.

La matriz de distancias es de orden n^2, por lo que en un conjunto muy grande de observaciones, estos métodos resultan impracticables.

En tal caso se puede hace una simplificación, que lleva a resultados no óptimos, como los entregados por los métodos anteriores, pero que buscan acercarse al óptimo.

El más común consiste en extraer una muestra aleatoria de casos, con tamaño más adecuado al procedimiernto que se desea utilizar.

A esta muestra se le aplica un método de conglomerados, como el k-medias o el k-medoide. Una vez finalizado, cada observación que no está en la muestra, es asignada al conglomerado cuya media (o medoide) es más cercano. Una medida de bondad de conglomeración es obtenida mediante el promedio de las distancias entre cada observación y el medoide de su conglomerado.

Es conveniente repetir el procedimiento anterior, partiendo de diversas mientras. Luego de ésto se selecciona la que ya tenga la mejor medida de conglomeración. Se recomienta usar 5 muestras distintas.

4. Análisis Fuzzy (difuso)

Es una variante de los métodos k-medias y k-medoides. En lugar de asignar un objeto a un grupo en forma deterministica, entrega probabilidades de pertenencia de cada observación a los distintos conglomerados, en base a sus distancias a los centros de estas. Por ejemplo, pueden ser proporcionales a las distancias. Se reasigna una observación por sorteo, de acuerdo a las probabilidades definidas.

METODOS JERARQUICOS.

Son métodos que parten de tantos conglomerados como casos hay, y en cada etapa siguiente van juntando conglomerados, hasta llegar a uno solo (método aglomerativo). O bien, partiendo de uno, van subdividiendo conglomerados hastya llegar a un caso por conglomerado (método divisivo).

5. Aglomerativo

Inicialmente cada observación es un conglomerado.

Luego en cada paso se unen los conglomerados que están a menor distancia y se calcula la distancia del nuevo conglomerado con todos los demás, formándose una nueva matriz de distancias.

El algoritmo termina cuando queda un conglomerado con todas las observaciones.

EJEMPLO

Se tiene una muestra de siete entrevistados que responden a una encuesta de diez preguntas, cada una con respuestas entre las alternativas a, b, c, d y e

.La matriz de datos de las respuesta es la siguiente:

encuestado \ pregunta	1	2	3	4	5	6	7	8	9	10
1	a	b	b	c	a	b	b	a	a	d
2	a	c	b	c	d	e	e	a	b	c
3	c	b	b	c	d	a	b	c	a	d
4	a	b	e	c	a	d	b	a	a	c
5	c	c	b	b	d	a	b	c	d	d
6	a	c	e	c	d	c	e	a	e	d
7	b	b	c	a	a	a	b	c	a	b

Se usará como distancia entre casos el número (o la fracción, dividiendo el número por 10) de respuestas diferentes, y la distancia entre conglomerados, la del vecino más próximo.

Iteración 1

La matriz de distancias entre los encuestados es la siguiente, siendo cada caso un conglomerado:

$D_1 =$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	0	6	4	3	7	6	6
(2)	6	0	7	6	7	4	10
(3)	4	7	0	6	3	7	5
(4)	3	6	6	0	9	6	6
(5)	7	7	3	9	0	7	7
(6)	6	4	7	6	7	0	10
(7)	6	10	5	6	7	10	0

Inicialmente se unen 1 con 4 y 3 con 5 a la distancia 3.

Iteración 2.

La nueva matriz de distancias entre conglomerados queda:

$D_2 =$

	$(1,4)$	(2)	$(3,5)$	(6)	(7)
$(1,4)$	0	6	4	6	6
(2)	6	0	7	4	10
$(3,5)$	4	7	0	7	5
(6)	6	4	7	0	10
(7)	6	10	5	10	0

Se unen $(1,4)$ con $(3,5)$ y (2) con (6) a la distancia 4.

Iteración 3.

La matriz de las distancias entre conglomerados queda:

$D_3 =$

	$(1,3,4,5)$	$(2,6)$	(7)
$(1,3,4,5)$	0	6	5
$(2,6)$	6	0	10
(7)	5	10	0

Se unen $(1,3,4,5)$ con (7) a la distancia 5. Obsérvese que las distancias de unión van aumentando con cada paso. Es decir, cada vez se unen observaciones más disímiles.

Ultima matriz de distancias entre conglomerados:

$D_4 =$

	$(1,3,4,5,7)$	$(2,6)$
$(1,3,4,5,7)$	0	6
$(2,6)$	6	0

Se unen todos en un sólo conglomerdo, a la distancia 6.

El gráfico siguiente es un *dendograma*. Ilustra la forma cómo se fueron uniendo los conglomerados hasta formar uno solo. La escala horizontal corresponde a la distncia en que produjeron las uniones, en cada caso.
De este gráfico se desprende que si deseamos tener dos conglomerados, serían (1,3,4,5,7) y (2,6). Si deseamos tener tres, serían (7), (1,3,4,5) y (2,6). Si queremos 5, éstos serían (1,4), (3,5), (2), (6) y (7).

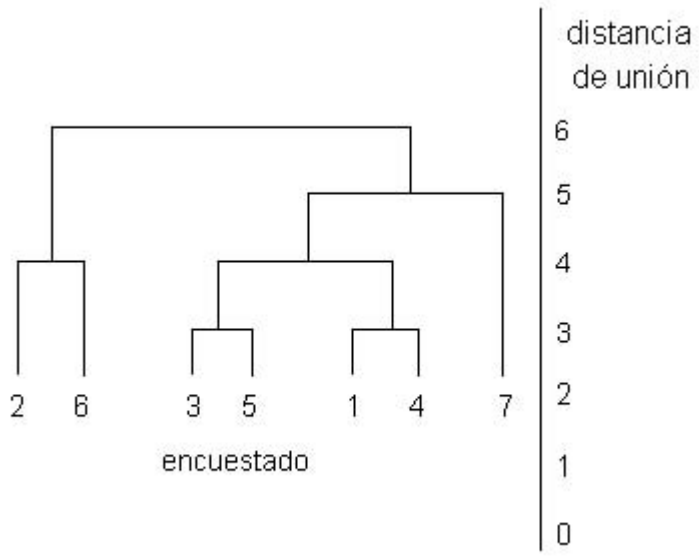

Figura 2: Dendograma.

6. Método Divisivo.

Comienza con un grupo que contiene todas las observaciones, y en sucesivos pasos lo va dividiendo hasta quedar cada observaciòn en un conglomerado diferente.

Sin embargo mientras en el paso inicial el método aglomerativo tiene $\binom{n}{2} = \frac{n(n-1)}{2}$ posibilidades de unir los primeros dos conglomerados, el método divisivo parte con $2^{n-1} - 1$ posibilidades del división del conglomerado inicial.

Este número es muchísimo mayor. En efecto, si hay n=10 observaciones, $\frac{n(n-1)}{2} = 45$ mientras $2^{n-1} - 1 = 511$.

Eso hace poco atractivo este método. Para avitar considerar todas las posibles divisiones, se diseñó el siguiente algoritmo:

a) Encontrar el objeto más discimil, el que tiene mayor distancia promedio con todos los demás. Este da origen a un grupo "disidente".

b) Por cada observación fuera del grupo disidente D, calcular:

$$V_i = promedio_{j \notin D} d(i,j) - promedio_{j \in D} d(i,j)$$

Para encontrar la observación h para la cual esta diferencia es mayor.

c) Si $V_h > 0$, h está en promedio más cerca del grupo disidente que a su complemento, por lo que se debe agregar al primero.

d) Repetir b y c, hasta que todos los V_h sean negativos.

De este modo, el conjunto queda partido en dos conglomerados.

e) Seleccionar el conglomerado de mayor diámetro (el diámetro es la distancia mayor entre dos objetos de él). Dividirlo como en los pasos a,b,c,d.

f) Repetir e hasta que todos los conglomerados contienen solo un objeto.

7. Análisis monotético.

Se utiliza cuando todas las variables son binarias ,(0 o 1). Es un método divisivo.

a) Se elige la variable con mayor asociación con las demás, de la siguiente forma: considere las variables f y g, y sea la siguiente tabla de contingencia para estas variables, dentro del conglomerado que se va a dividir:

f \ g	1	0
1	a	b
0	c	d

La asociación entre f y g se define como

$$A_{fg} = |ad - bc|$$

La asociación total entre f y las demás variables se define como:

$$A_f = \sum_{g \neq f} A_{fg}$$

La variable t que satisface

$$A_t = \text{máx}_{1 \leq f \leq p} A_f$$

es seleccionada.

b) Usando está variable, se divide el conglomerado en dos, uno en que ésta toma el valor 0, y otro en que toma el valor 1.

c) Se repite a y b, en los dos conglomerados resultantes.

d) Se detiene el proceso cuando todos los conglomerados tienen un sólo objeto o bien tienen objetos idénticos.

8. Conglomerados jerárquicos basados en modelos

Asume que todos los datos son generados por una mezla de distribuciones probabilisticas subyacentes. Si hay G poblaciones diferentes y la densidad de una observación x de la k-esima población es $f_k(x;\theta)$ para algún vector de parámetros θ desconocido.

Dados los datos: $X = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ . \\ . \\ \underline{x}_n' \end{bmatrix}$ sea $j = \begin{bmatrix} j_1 \\ j_2 \\ . \\ . \\ j_n \end{bmatrix}$

el vector de rótulos tales que si x_i proviene de la k-esima población, entonces $j_i = k$.

El método de máxima verosimilitud busca θ y j tales que se maximice la verosimilitud

$$L(X;\theta;j) = \Pi_{i=1}^{n} f_{j_i}(x_i;\theta)$$

Existen diferentes casos para $f_k(x_i;\theta)$. Se suele asumir que es normal multivariante $N(\underline{\mu}_k, \Sigma_k)$.

Si además se asume que $\Sigma_k = \sigma_k^2 I$, los conglomerados resultan de forma hiperesférica.

Si Σ_k tiene cualquier forma, sus valores propios especifican la orientación que tiene el n-esimo conglomerado y el mayor valor propio es una medida de su tamaño o varianza, $\underline{\mu}_k$ da su posición.

9. Algoritmo Genético

Este algoritmo de conglomeración no puede clasificarse como jerárquico. Tiene su origen de la informática, y son aplicables al análisis de conglomerados. El siguiente método se basa en estos algoritmos.

Suponga que se desea particionar un conjunto $\{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\}$ de observaciones en 3 grupos.

Se debe disponer de una medida de Bondad de Conglomeración que permite discriminar cual esquema de distribución de observaciones en un grupo de conglomerados es mejor .

Por ejemplo: puede ser el coeficiente silueta definido anteriormente o un cuociente entre cuadros medios entre conglomerados y cuadrados medidas dentro de los conglomerados

Un vector de rótulos es un vector de coordenadas enteras, que indican el conglomerado al que pertenece cada respectivo elemnto muestral. Por ejemplo,

$$c' = [1\ 1\ 2\ 2\ 2\ 3\ 1\ 1\ 3\ 3\ 2]$$

que indica que $\underline{x}_1, \underline{x}_2, \underline{x}_7, y\ \underline{x}_8$, están en un conglomerado,$\underline{x}_3, \underline{x}_4, \underline{x}_5$ y $\underline{x}_{11}$ están en otro, y$\underline{x}_6, \underline{x}_9$,y $\underline{x}_{10}$ están en un tercero.

El método parte de un conjunto de "cromosomas", que son vectores de rótulos,$\{\underline{c}_1, \underline{c}_2, ...\underline{c}_k\}$

Estos cromosomas son arbitrarios, así como el número de ellos. Por ejemplo: pueden ser 18 cromosomas, seis cuyos elementos son todos 1, seis cuyos elementos son 2, y seis cuyos elementos son 3, de la forma

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}, ..., \begin{bmatrix} 2 \\ 2 \\ . \\ . \\ . \\ 2 \end{bmatrix}, ..., \begin{bmatrix} 3 \\ 3 \\ . \\ . \\ . \\ 3 \end{bmatrix} \right\}$$

Cada uno de los números del cromosoma es un "gen". Este conjunto forma la "primera generación".

Para formar la segunda generación se forman pares, relacionados al azar. Estos son los "Padres" con un determinado número de "hijos", cuyos cromosomas se forman eligiendo cada gen, uno entre los dos de ambos padres, que ocupan la misma posición, seleccionado al azar, Por ejemplo, la siguiente ilustración muestra un caso posible, en que dos parejas de padres tienen tres hijos de cada uno.

Padres (Generación k)

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

Hijos (generación k+1)

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 3 \\ 1 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 3 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 3 \\ 3 \\ 1 \\ 3 \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 1 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \\ 2 \\ 1 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

Supongamos en nuestro ejemplo que cada pareja tiene 3 hijos. Entonces la segunda generación habrían 27 individuos.

Sea M_k ($\underline{c}$) el valor de la medida individuo de conglomeración aplicada al definido por el cromosoma $\underline{c}$, en la k-ésima generación.

Se ordenan todos los individuos de la última generación de acuerdo a sus medidas de conglomeración.

Se selecciona un grupo de las mejores, que forma la "elite", y sus genes se copian en la siguiente generación. (se "clonan"). Por ejemplo, la elite pueden estar formadas por los tres primeros.

Después se seleccionan los mejores en igual número que al inicio, (se incluyen los de la elite), y se repite todo el proceso. Es decir, se forman parejas al azar, tienen hijos cuyos genes resultan de la combinación, al azar, de los correspondientes genes de sus padres. Se ordenan de acuerdo a la medida de bondad de conglomeración, se obtiene una elite que se clona en la siguiente generación, por ejemplo con los 18 mejores se seleccionan 9 parejas, etc.

En el ejemplo, cada generación tiene 3 que pertenecen a la elite de la generación anterior, más 3 hijos por cada una de las 9 parejas de la generación anterior, son 30 individuos en cada generación.

Se repite el proceso por un número alto de generaciones mejorándose progresivamente la medida de bondad de conglomeración, optimizándose el proceso.

Falta un elemento para completar el proceso. Hasta el momento el procedimiento apunta a buscar un óptimo. Sin embargo, puede ser que estemos tratando de mejorar en el entorno de un óptimo local. Se debe tratar de explorar, paralelamente, otras zonas del espacio de posibles esquemas de conglomeración, en busca de óptimos locales que superen al óptimo local actual. Esto se hace de la siguiente forma:

En cada generación se introduce una pequeña fracción de "mutantes", éstos son individuos que cambian espontáneamente un gen. Esta fracción es pequeña por ejemplo, un 10 % de individuos. En nuestro ejemplo serían 3. En cada generación se seleccionan al azar estos individuos mutantes, y se les selecciona al azar un gen, al que se les asigna un valor, también al azar.

Estos mutantes permiten que la exploración se extienda a otras zonas, donde podrían haber óptimos locales que superen el ya encontrado.

Resumen del algoritmo genético aplicado al análisis de conglomerados:

Siguiendo con los valores dados en el ejemplo, que pueden variar, en la practica. Además, la forma presentada aquí es una de varias posibles variantes del algoritmo.

Generación k-esima: Recibe de la generación anterior: La elite de la generación anterior formada por los 3 mejores, más 27 hijos (incluidos 3 mutantes) = 30 individuos.

Los 3 mejores (elite) pasan idénticos a la siguente generación.

Además, entre los 18 mejores (incluyendo los de la elite) se forman 9 parejas al azar.

Cada pareja enjendra 3 hijos. De estos 27 hijos, 3 son mutantes.

Entrega a la generación siguiente:

Los 3 de la elite más los 27 hijos (incluidos 3 mutantes)

MEDIDAS DE BONDAD DE CONGLOMERACION.

Se debe disponer de una medida de Bondad de Conglomeración que permite discriminar cual esquema de distribución de observaciones en un grupo de conglomerados es mejor .

La Figura 3 siguiente muestra un conjunto de 12 observaciones bidimensionales, con tres esquemas de conglomerados, donde, a simple vista, el de tres conglomerados es el que mejor separa las observaciones en grupos.

Si se dividen en dos conglomerados, el primero contiene observaciones muy distantes entre sí. Si se separa en cuatro, aparecen dos conglomerados muy próximos.

1. Coeficiente Silueta

Se define, para una observación i, el valor:

$a(i)$ = promedio de las disimilitudes de i con todos los demás objetos del conglomerado A al cual pertenece i.

Sea C otro conglomerado, $C \neq A$. Sea $d(i, C)$ = promedio de distancias de$i$a todos los elementos de$C$.

Sea $b(i) = \text{mín}_{C \neq A}\, d(i, C)$

El conglomerado B que alcanza el mínimo, es decir, tal que $d(i, B) = b(i)$ se denomina vecindad del objeto i. B es el segundo mejor conglomerado para i.

La silueta del objeto i se define como:

$$s(i) = \frac{b(i)-a(i)}{\text{máx}\{a(i),b(i)\}}$$

observar que $-1 \leq s(i) \leq 1$

Interpretación:

$s(i) \approx 1$, el objeto i está bien clasificado

$s(i) \approx 0$, el objeto i está entre dos conglomerados

$s(i) \approx -1$, el objeto i está mal clasificado.

El coeficiente silueta es el promedio a través de todas las observaciones. Mientras más grande, mejor es la distribución de conglomerados.

La Figura 3 muestra que el mayor valor, de 0,69, corresponde al esquema de conglomerados que a simple vista parece mejor, con tres conglomerados.

2. Cuadrado medio dentro de los conglomerados

Es el promedio de la suma de cuadrados de las distancias de cada observación hasta el centroide del conglomerado a que pertenecen.

$$CMD = \frac{1}{d} \sum_j \left(\sum_i \left\| \underline{x}_{ij} - \overline{\underline{x}}_j \right\|^2 \right)$$

en que $\underline{x}_{ij}$ es la observación $i - ésima$ del conglomerado j , $\overline{\underline{x}}_j$ es el vector promedio del conglomerados j, y $d = \sum n_j - G$ es el divisor ("grados de libertad"), donde n_j es el numero de observaciones en el conglomerado j. Mientras más pequeño, mejor.

Cada elemento $\sum_i \left\| \underline{x}_{ij} - \overline{\underline{x}}_j \right\|^2$ es una medida de bondad dentro del respectivo conglomerado. Si uno de ellos es muy grande, indica que el correspondiente conglomerado tiene elementos muy discímiles.

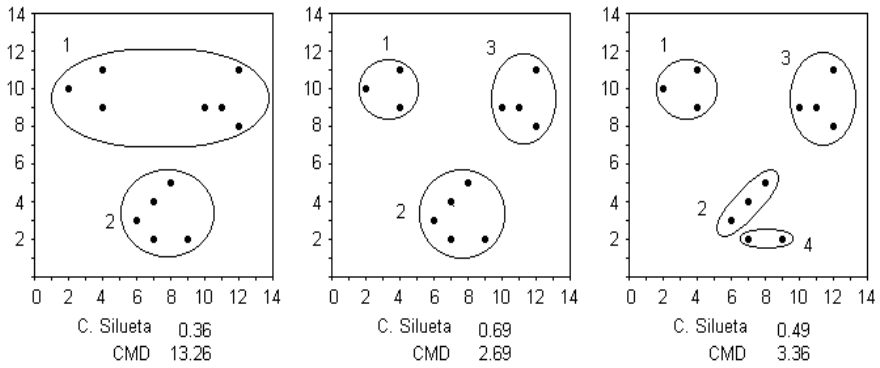

Figura 3: Tres esquemas de conglomerados para un mismo conjunto de puntos. El del centro es el mejor.

3. Coeficiente F

Es el un cuociente entre los cuadrados medidas dentro de los conglomerados (CMD) y los cuadros medios entre (CME) conglomerados, donde

$$CME = \frac{1}{G-1} \sum_j \left\| \overline{\underline{x}}_j - \overline{\underline{x}} \right\|^2$$

en que $\overline{\underline{x}}_j$ es el vector promedio del conglomerados $j - ésimo$, $\overline{\underline{x}}$ es el vector promedio global, G el número de conglomerados, y $\|\|$ indica *norma vectorial.*

4. Coeficiente aglomerativo

Es una medida global de conglomeración, asociada a los métodos jerárquicos. Se aplica a todo el procedimiento, no a un determinado número de conglomerados.

Por cada elemento i, sea $d(i)$ su distancia al primer conglomerado con que se une, dividida por la distancia de los últimos conglomerados en unirse. El coeficiente aglomerativo es

$$CA = 1 - \frac{\sum i=1nd(i)}{n}$$

5. Otros indicadores.

Se pueden definir varios otros indicadores, como por ejemplo, el cuociente o la diferencia entre la distancia máxima o distancia promedio dentro de los conglomerados (que se espera sea pequeña), y la distancia mínima o la distancia promedio entre conglomerados (que se espera sea grande). se pueden obtener otros indicadores como variantes de estos.

Sea $\underline{x}_j$ una observación. Se define $d(j)$ como el cuociente entre la distancia en que $\underline{x}_j$ se une por primera vez a otro conglomerado y la distancia en que se produce la última fusión de todos en un sólo gran conglomerado. El coeficiente aglomerativo es el promedio de $(1 - d(j))$ a través de todas las observaciones.

Obsérvese que si este coeficiente es grande, significa que los $d(j)$ tienden a ser pequeños, es decir, que la mayoría las fusiones se produjeron a distancias relativamente pequeñas.

GRAFICOS ASOCIADOS AL ANALISIS DE CONGLOMERADOS

1. Silueta

La silueta de un conglomerado es una representación gráfica de los coeficientes silueta $s(i)$ para todas las observaciones $i = 1, 2, ..., n$, rangueados en orden descendente dentro de su conglomerado. La proporción de superficie contenida en las barras, respecto del área del ancho 1, corresponde al coeficiente silueta. Mientras más largas las barras, mejor. Ver Figura 4.

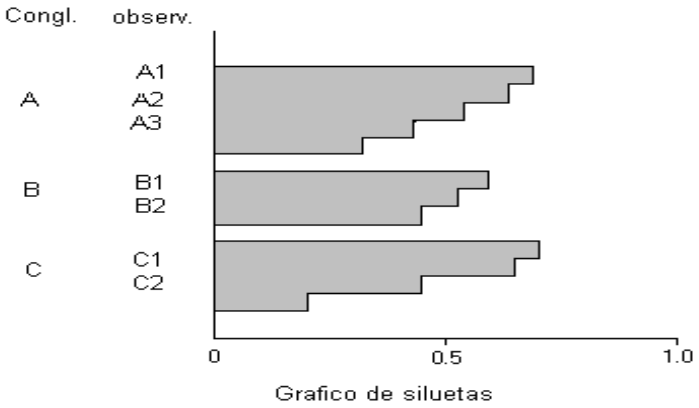

Figura 4: Gráfico de silueta.

2. Dendograma.

Es un gráfico que muestra cómo se fueron uniendo los conglomerados hasta formar uno solo. La escala vertical corresponde a la distancia en que produjeron las uniones, en cada caso. Ver Figura 5.

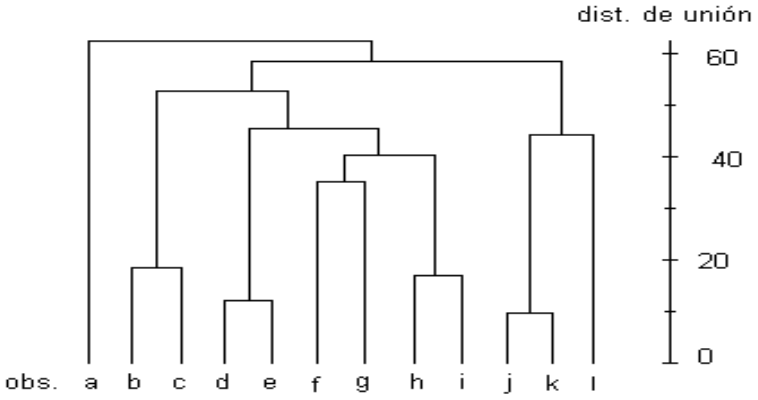

Figura 5: Dendograma.

3. Bandera (Banner)

También es sólo para métodos jerárquicos. Los objetos se listan de arriba hacia abajo a la izquierda.

Al lado de cada uno hay una linea horizontal. Las lineas se unen mediante trazos verticales, colocados a la distancia de unión.

La información que entrega este gráfico es la misma que el dendograma. Nótese que el coeficiente aglomerativo corresponde al la proporción de superficie del lado derecho del gráfico de bandera. Ver Figura 7. La situación ilustrada en el gráfico corresponde al mismo caso del dendograma de la Figura 6.

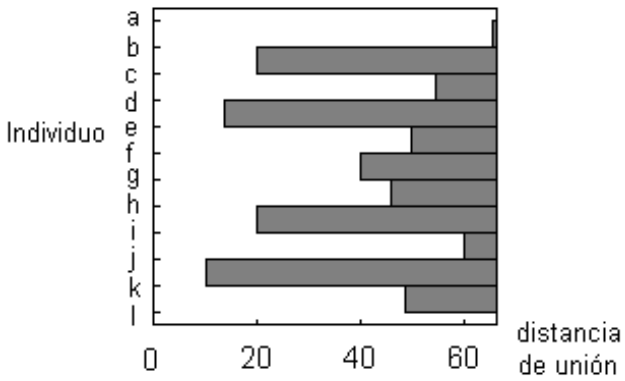

Figura 6: Gráfico de bandera.

4. Biplot

Es un plano formado con dos coordenadas, en que cada un representa una componente principal de los datos. El caso más usual es el que se compone de las componentes 1 y 2, y representa el plano en que las proyecciones de las observaciones aparecen más dispersas. Es posible visualizar los conglomerados en este plano. Ver Figura 7.

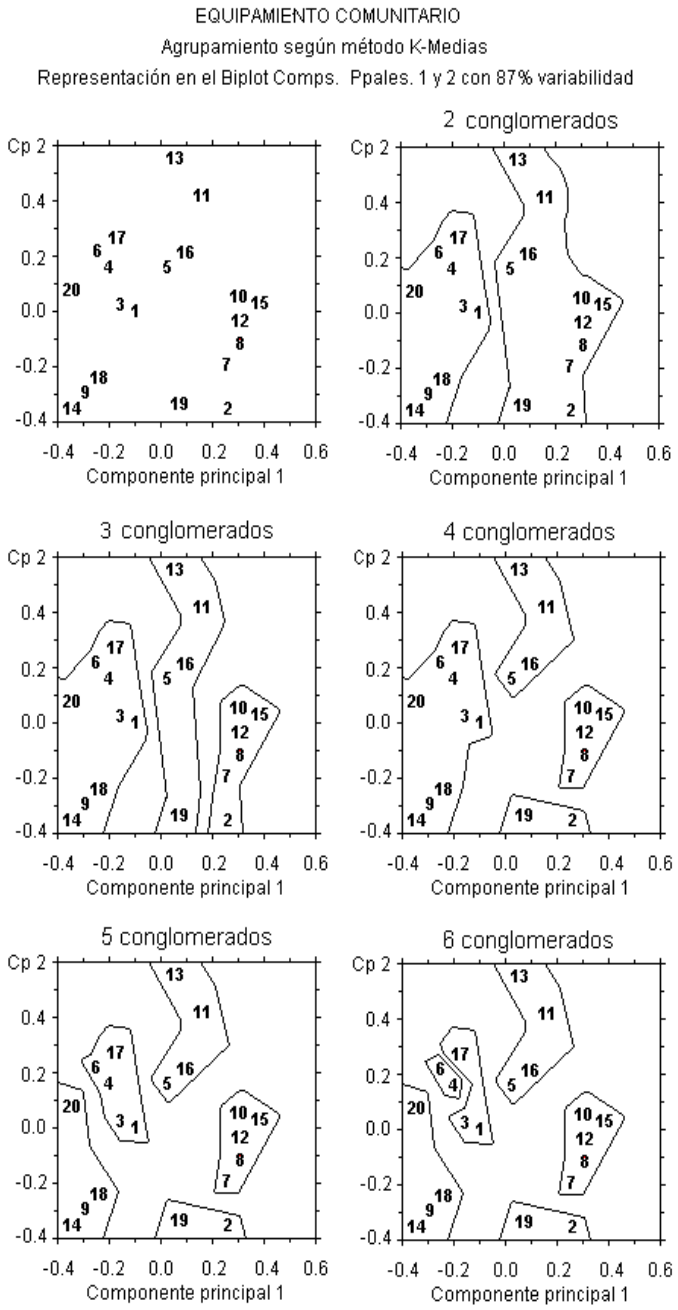

Figura 7: Gráficos Biplot.