V. 8, n.º 3 (setembre 1984) pp. 113-120

SOBRE LA NO UNICIDAD DE LA CLASIFICACIÓN JERÁRQUICA ASOCIADA A UNA DISIMILARIDAD POR LOS MÉTODOS DEL MÁXIMO Y UPGMA

A. ARCAS I M. SALICRÚ

UNIVERSIDAD DE BARCELONA

A partir de una distancia definida en el conjunto de disimilaridades sobre un conjunto finito podemos medir la distancia entre dos ultramétricas asociadas a una disimilaridad inicial a partir de la aplicación de un método de clasificación. En este artículo tratamos el problema de la no unicidad de la clasificación jerárquica asociada para los métodos del máximo y UPGMA estudiando los siguientes puntos:

a) Búsqueda de una definición de la disimilaridad asociada a cada nivel de la jerárquía.

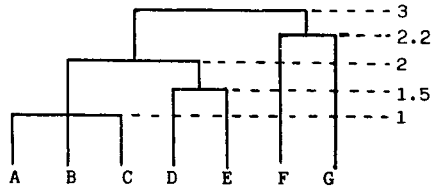
b) Cálculo de una cota para la distancia entre dos ultramétricas asociadas a la misma disimilaridad inicial.

Keywords: ULTRAMETRIC DISTANCE; CLUSTER ANALYSIS; COMPLETE LINKAGE METHOD; UPGMA METHOD.

1. INTRODUCCION

Uno de los principales objetivos del análisis multivariante consiste en la representación geométrica de objetos a partir de una disimilaridad inicial definida entre los mismos. - Así, si queremos representar los objetos de forma métrica en dimensión reducida (por ejemplo en el plano) podemos utilizar técnicas como el análisis de coordenadas principales o multidimensional scaling y si por el contrario queremos representarlos de modo que queden de manifiesto las relaciones de semejanza entre ellos formando una colección de clases naturales o deseamos observar su estructura evolutiva, entonces debemos utilizar técnicas como la taxonomía numérica o árboles aditivos.

La representación geométrica a través de la taxonomía numérica viene dada por un determinado tipo de grafos orientados simplemente conexos sin ciclos llamados dendogramas /5/



construidos directamente a través de la jerarquía indexada que obtenemos al realizar la clasificación.

Así pues, la taxonomía numérica intenta la construcción de una clasificación jerárquica de un conjunto finito X a partir de una disimilaridad inicial definida sobre el mismo. - Por comodidad nosotros estudiaremos las clasificaciones jerárquicas como funciones $\varphi: R^+ \longrightarrow P$ (siendo P el conjunto de particiones de X) cumpliendo ciertas propiedades.

El problema así planteado radica por lo general en que la disimilaridad inicial no satisface el axioma ultramétrico

$$d(x_i,x_j) \leq \max \{d(x_i,x_k),d(x_j,x_k)\} \quad \text{para } x_i,x_j,x_k \in X$$

y por tanto no es posible construir una clasificación jerárquica por el algoritmo fundamental de clasificación /4/, /7/, /2/. Así la dificultad de fondo consiste en la transformación de la disimilaridad inicial en otra que verifique el axioma ultramétrico, surgiendo así distintos métodos de clasificación (mí

- A. Arcas i M. Salicrú - Dptn. de Bioestadística de la Facultat de Biología de la Universitat de Barcelona
Av. Diagonal, 645 - Barcelona 08028

nimo(Johnson, 1967), máximo (Johnson, 1967), mediana (Gower, 1967), UPGMA (Sokal y Michener, 1958),...). Diversas propiedades de estos métodos son estudiadas extensamente en /1/, /6/, /7/ y /11/.

Otro problema que se nos plantea proviene de que sólo el método del mínimo proporciona siempre una única ultramétrica asociada a la disimilaridad inicial, resultando de especial interés la comparación de los dendogramas asociados a varias ultramétricas que provienen de la misma disimilaridad. En nuestro trabajo nos centramos en el estudio de este último problema para los métodos del máximo y UPGMA. Relacionamos la disimilaridad entre dos clases en un determinado nivel con la disimilaridad inicial y acotamos la distancia que separa a dos ultramétricas asociadas a una misma disimilaridad inicial -- tras haber definido una métrica en el conjunto de las disimilaridades.

2. CONCEPTOS PREVIOS.

2.1. CLASIFICACION JERARQUICA

Una clasificación jerárquica en el conjunto X será una función

$$\varphi : R^+ \longrightarrow P$$

siendo P el conjunto de particiones de X, verificando las siguientes condiciones:

$$\begin{aligned} &a)\quad \varphi(0) \supset \{\{x_1\},\dots,\{x_n\}\} \\ &b)\quad r<r' \Rightarrow \varphi(r)\subset\varphi(r') \\ &c)\quad \exists\, r\in R^+ \mid \varphi(r) = \{X\} \end{aligned} \qquad (1)$$

Así por ejemplo, la clasificación jerárquica que representa el dendograma de la figura 1 sería:

$$\varphi(\alpha)=\begin{cases} \{A\},\{B\},\{C\},\{D\},\{E\},\{F\},\{G\} & \text{si } \alpha\in[0,1) \\ \{A,B,C\},\{D\},\{E\},\{F\},\{G\} & \text{si } \alpha\in[1,1.5) \\ \{A,B,C\},\{D,E\},\{F\},\{G\} & \text{si } \alpha\in[1.5,2) \\ \{A,B,C,D,E\},\{F\},\{G\} & \text{si } \alpha\in[2,2.2) \\ \{A,B,C,D,E\},\{F,G\} & \text{si } \alpha\in[2.2,3) \\ \{A,B,C,D,E,F,G\} & \text{si } \alpha\geq 3 \end{cases}$$

E- el fondo esta aplicación asigna a cada nivel una partición de X, de forma que al incrementar la variable se obtienen particiones menos finas llegando un momento en que la partición tiene una sola clase.

2.2. JERARQUIA INDEXADA ASOCIADA A φ

Sea la jerarquia de partes de X definida por

$$J = \{T \mid \exists\, r,\ T\in\varphi(r)\}$$

Si consideramos $i : J \longrightarrow R^+$ de forma que:

$$i(T)=\inf\{r\in R^+ \mid T\in\varphi(r)\} \qquad ,$$

resulta que i es un índice de jerarquía. Así (J,i) será la jerarquía indexada asociada a φ y la distancia ultramétrica asociada a dicha jerarquía /4/ vendrá definida asignando a $d(x_i,x_j)$ el índice de la clase minimal de la jerarquía que contiene a x_i y x_j.

2.3. METODO DE CLASIFICACION

Tal como se ha indicado, un método de clasificación consiste en un sistema o algoritmo para la construcción de una clasificación jerárquica a partir de una disimilaridad inicial. Dicha clasificación jerárquica nos proporciona de forma única la jerarquía indexada y la distancia ultramétrica asociada.

Para nuestro trabajo sólo consideraremos aquellos métodos de clasificación que se obtengan mediante la siguiente recurrencia:

Si

$$l_1=\min\{d(x_i,x_j) \mid x_i\neq x_j,\ x_i,x_j\in X\}=d(x_{io},x_{jo})$$

es la mínima distancia entre los elementos de X, definimos

$$\varphi(l_1)=\{\{x_1\},\dots,\{x_{io}\}\cup\{x_{jo}\},\dots,\{x_n\}\}$$

es decir, tomamos el conjunto formado por los elementos anteriores habiendo juntado los más próximos, y consideramos una función $f_1 : R^3 \longrightarrow R$ de forma que

$$f_1(d(x_{io},x_{jo}),d(x_{io},x_k),d(x_{jo},x_k))\geq l_1 \quad ,\ \forall\, x_k\in X$$

Así, a $\varphi(l_1)$ le asociamos la disimilaridad d_1 definida como sigue:

$$d_1(\{x_i\},\{x_j\})=d(x_i,x_j) \text{ si } \{x_i,x_j\}\cap\{x_{io},x_{jo}\}=\emptyset$$

$$d_1(\{x_{io}\}\cup\{x_{jo}\},\{x_k\}) =$$

$$= f_1(d(x_{io},x_{jo}),d(x_{io},x_k),d(x_{jo},x_k))\geq l_1$$

de forma que la distancia entre los elementos que no se han juntado se mantiene invariante y la distancia que separa la unidad formada por los dos elementos que se han juntado con un tercer elemento viene determinado por la función f_1. Por este proceso hemos pasado del conjunto X con disimilaridad d al conjunto $\varphi(l_1)$ con disimilaridad d_1, y de forma análoga podemos pasar a $\varphi(l_2),\ldots,\varphi(l_s)$ considerando sucesivamente funciones $f_2,\ldots,f_{s-1}$.

$$l_1 \longrightarrow \varphi(l_1)$$
$$l_2 \longrightarrow \varphi(l_2)$$
$$\vdots$$
$$l_s \longrightarrow \varphi(l_s)$$

que se generaliza a $\varphi: R^+ \longrightarrow P$ de la forma siguiente:

a) Si $\alpha\in[l_{m-1},l_m)$ se toma $\varphi(\alpha)=\varphi(l_{m-1})$

b) Si $l_1\neq 0$ y $\alpha\in[0,l_1)$

se toma $\varphi(\alpha)=\{\{x_1\},\ldots,\{x_n\}\}$

c) Si $\alpha\geq l_s$, $\varphi(\alpha)=\{X\}$

Los principales métodos clásicos de clasificación jerárquica pueden ser construidos a través de dicha recurrencia.

2.4. FUNCION DE CLASIFICACION ASOCIADA A φ

Consideremos D_X el conjunto de disimilaridades definidas sobre X y tomemos un método de clasificación definido en 2.3. Podemos entonces definir la función multívoca /11/.

$$f: D_X \longrightarrow D_X$$

de forma que f(d) sea el conjunto de distancias ultramétricas asociadas a las jerarquías indexadas correspondientes a la disimilaridad inicial por aplicación del método de clasificación fijado.

A la función f la llamaremos función de clasificación asociada al método de clasificación elegido.

2.5. DEFINICIÓN DEL MÉTODO DEL MÁXIMO

El método del máximo es el método de clasificación definido en 2.4 que proviene de tomar la sucesión de funciones (f_n) constante de la forma

$$f_n(\alpha,\beta,\gamma) = \max\{\beta,\gamma\}$$

En este caso se verifica que si

$$\varphi(l_m) = \{A_1,\ldots,A_{n_m}\}$$

y $l_{m+1}=d_m(A_{io},A_{jo})$ mínimo valor para d_m, resulta por aplicación del método de clasificación general

$$\varphi(l_{m+1}) = \{A_1,\ldots,A_{io}\cup A_{jo},\ldots,A_{n_m}\}$$

verificándose que

$$d_{m+1}(A_{io}\cup A_{jo},A_k)=\max\{d(A_{io},A_k),\ d(A_{jo},A_k)\}$$

$$d_{m+1}(A_i,A_j)=d_m(A_i,A_j) \text{ si } \{i,j\}\cap\{io,\ jo\} = \emptyset \tag{2}$$

La idea geométrica de este método es la deformación de un triángulo hasta obtener dos lados iguales que coincidan con el mayor de los lados que no son base (/4/). La ultramétrica que nos proporciona la aplicación del método del máximo es un elemento minimal dentro de las ultramétricas superiores a la disimilaridad inicial.

2.6. DEFINICIÓN DEL MÉTODO UPGMA

En este caso si

$$\varphi(l_m) = \{A_1,\ldots,A_m\}$$

y $l_{m+1}=d_m(A_{io},A_{jo})$ mínimo valor para d_m, resulta que

$$\varphi(l_{m+1}) = \{A_1,\ldots,A_{io}\cup A_{jo},\ldots,A_{n_m}\}$$

verificándose que f_m inducirá d_{m+1} de la forma

$$d_{m+1}(A_{io} \cup A_{jo}, A_k) =$$

$$= \frac{N_i}{N_i + N_j} \cdot d_m(A_{io}, A_k) + \frac{N_j}{N_i + N_j} \cdot d_m(A_{jo}, A_k) \qquad (3)$$

siendo $N_i = |A_{io}|$ (Cardinal de A_{io}) y $N_j = |A_{jo}|$.

La idea geométrica de este método es la deformación de un triángulo hasta obtener los lados iguales que coincidan con la media de los dos lados mayores ponderada respecto los cardinales de las clases (/4/).

Este método tiene importantes propiedades con respecto al coeficiente de correlación cofenética (coeficiente de correlación entre la disimilaridad inicial y la ultramétrica asociada) /6/, /1/. Es uno de los principales métodos utilizados en aplicaciones prácticas.

2.7. MÉTRICA EN D_X

Definimos la siguiente distancia en D_X:

$$\Delta_o(d_1, d_2) =$$

$$= \max\{|d_1(x,y) - d_2(x,y)| \,|\, (x,y) \in X \times X\} \qquad (4)$$

Al ser X conjunto finito resulta trivial que Δ_o es una distancia.

3. OBTENCION DE LA DISIMILARIDAD EN CADA NIVEL A PARTIR DE LA INICIAL.

Obtendremos una expresión de la distancia asociada a cada nivel a partir de la disimilaridad inicial cuando realizamos una clasificación jerárquica por los métodos del máximo y UPGMA. De esta forma relacionamos directamente la distancia ultramétrica que caracteriza la jerarquía con la disimilaridad inicial, lo cual es muy útil para estudiar las propiedades de las clasidicaciones por niveles.

3.1. MÉTODO DEL MAXIMO.

PROPOSICION 3.1.- Si $\varphi(l_s) = \{A_1, \ldots, A_{n_s}\}$ y utilizamos el método del máximo, se verifica

$$d_s(A_i, A_j) = \max\{d(x_i, x_j) \,|\, x_i \in A_i, x_j \in A_j\}$$

DEMOSTRACION:

La demostración la realizamos por inducción sobre los niveles $l_1, \ldots, l_s$ de la jerarquía. Así pues, sea

$$l_1 = \min\{d(x_i, x_j) \,|\, x_i, x_j \in X,\ x_i \neq x_j\} = d(x_{io}, x_{jo})$$

resultándonos

$$\varphi(l_1) = \{\{x_o\}, \ldots, \{x_{io}\} \cup \{x_{jo}\}, \ldots, \{x_n\}\}$$

y por aplicación del método del máximo

$$d_1(\{x_{io}\} \cup \{x_{jo}\}, \{x_k\}) =$$

$$= \max\{d(x_{io}, x_k),\ d(x_{jo}, x_k)\}$$

$$d_1(\{x_i\}, \{x_j\}) = d(x_i, x_j)$$

$$\text{para } \{i,j\} \cap \{io, jo\} = \emptyset$$

con lo que se demuestra la proposición para el primer nivel.

Supongamos cierta la hipótesis hasta el nivel l_{k-1} y vamos a demostrar la proposición para l_k; sea

$$l_k = \min\{d_{k-1}(A_i, A_j) \,|\, A_i \neq A_j,\ A_i, A_j \in \varphi(l_{k-1})\}$$

se tendrá

$$\varphi(l_k) = \{A_1, \ldots, A_{io} \cup A_{jo}, \ldots\}$$

por hipótesis de inducción tenemos que

$$d_k(A_i, A_j) = d_{k-1}(A_i, A_j) =$$

$$= \max\{d(x_i, x_j) \,|\, x_i \in A_i,\ x_j \in A_j\}$$

$$\text{para } \{i,j\} \cap \{io, jo\} = \emptyset$$

y por aplicación del método del máximo (2)

$$d_k(A_{io} \cup A_{jo}, A_t) =$$

$$= \max\{d_{k-1}(A_{io}, A_t),\ d_{k-1}(A_{jo}, A_t)\} =$$

$$= \max\{\max\{d(x_i, x_t) \mid x_i \in A_{io}, x_t \in A_t\},$$
$$\max\{d(x_j, x_t) \mid x_j \in A_{jo}, x_t \in A_t\}\} =$$

$$= \max\{d(x_i, x_t) \mid x_i \in A_{io} \cup A_{jo},\ x_t \in A_t\}$$

tal y como queríamos demostrar.

Así pues, la distancia entre dos clases de la jerarquía a un mismo nivel es la máxima de las distancias entre elementos de clases distintas.

3.2. METODO UPGMA.

PROPOSICION 3.2. Si $\varphi(l_s) = \{A_1, \dots, A_{n_s}\}$ y utilizamos el método UPGMA, se verifica

$$d_s(A_i, A_j) = \sum_{\substack{x_i \in A_i \\ x_j \in A_j}} \frac{d(x_i, x_j)}{|A_i| \cdot |A_j|}$$

DEMOSTRACION:

Realizamos también la demostración por inducción sobre $l_1, \dots, l_s$, niveles de la jerarquía. Sea,

$$l_1 = \min\{d(x_i, x_j) \mid x_i, x_j \in X,\ x_i \neq x_j\} = d(x_{io}, x_{jo})$$

obteniendo pues

$$\varphi(l_1) = \{\{x_1\}, \dots, \{x_{io}\} \cup \{x_{jo}\}, \dots, \{x_n\}\}$$

y de esta manera

$$d_1(\{x_i\}, \{x_j\}) = d(x_i, x_j) =$$

$$= \frac{d(x_i, x_j)}{|\{x_i\}| \cdot |\{x_j\}|} \quad \text{para } \{i,j\} \cap \{io, jo\} = \emptyset$$

y al unir $\{x_{io}\}$ y $\{x_{jo}\}$ tenemos por (3)

$$d_1(\{x_{io}\} \cup \{x_{jo}\}, \{x_t\}) = \tfrac{1}{2} \cdot d(x_{io}, x_t) + \tfrac{1}{2} \cdot d(x_{jo}, x_t)$$

y queda demostrada la proposición para el caso k=1.

Supongamos cierta la proposición hasta k-1 y vamos a demostrarla para k.

Sea

$$l_k = \min\{d_{k-1}(A_i, A_j) \mid A_i, A_j \in \varphi(l_{k-1}),\ A_i \neq A_j\} =$$

$$= d_{k-1}(A_{io}, A_{jo})$$

entonces

$$d_k(A_i, A_j) = d_{k-1}(A_i, A_j) = \sum_{\substack{x_i \in A_i \\ x_j \in A_j}} \frac{d(x_i, x_j)}{|A_i| \cdot |A_j|}$$

para $\{i,j\} \cap \{io, jo\} = \emptyset$

y por (3)

$$d_k(A_{io} \cup A_{jo}, A_t) = \frac{|A_{io}|}{|A_{io}| + |A_{jo}|} \cdot d_{k-1}(A_{io}, A_t) +$$

$$+ \frac{|A_{jo}|}{|A_{io}| + |A_{jo}|} \cdot d_{k-1}(A_{jo}, A_t) =$$

$$= \frac{|A_{io}|}{|A_{io}| + |A_{jo}|} \cdot \sum_{\substack{x_i \in A_{io} \\ x_t \in A_t}} \frac{d(x_i, x_t)}{|A_{io}| \cdot |A_t|} +$$

$$+ \frac{|A_{jo}|}{|A_{io}| + |A_{jo}|} \cdot \sum_{\substack{x_j \in A_{jo} \\ x_t \in A_t}} \frac{d(x_j, x_t)}{|A_{jo}| \cdot |A_t|} =$$

$$= \sum_{\substack{x_s \in A_{io} \cup A_{jo} \\ x_t \in A_t}} \frac{d(x_s, x_t)}{(|A_{io}| + |A_{jo}|) \cdot |A_t|}$$

quedando pues demostrada la proposición por inducción.

Hemos obtenido para cada uno de los métodos anteriores una expresión general de la disimilaridad asociada a un determinado nivel a partir de la disimilaridad inicial, expresión común para todas las jerarquías asociadas a la misma.

4. ACOTACION DE LA DISTANCIA ENTRE DOS ULTRAMETRICAS ASOCIADAS A d

En este capítulo tratamos de buscar una acotación de la distancia Δ_o entre dos ultramétricas asociadas a la misma disimilaridad d. Así, si en el nivel l_s

$$\varphi(l_s) = \{A_1, \dots, A_m\}$$

y el mínimo de las distancias se obtiene a la vez para dos parejas, es decir,

$$\alpha = d_s(A_i, A_j) = d_s(A_k, A_l) =$$
$$= \min\{d_s(A_r, A_t) \mid A_r, A_t \in \varphi(1_s)\}$$

por definición del método de clasificación obtenemos φ_1 y φ_2

de forma que

$$\varphi_1(\alpha) = \{A_1, \ldots, A_i \cup A_j, \ldots, A_m\} \quad y$$
$$\varphi_2(\alpha) = \{A_1, \ldots, A_k \cup A_l, \ldots, A_m\}$$

con d_1' y d_2' distancias ultramétricas asociadas a las jerarquías correspondientes a φ_1 y φ_2.

Por la propia definición de las clasificaciones, existirán s_k y s_r (mínimos) tales que

$$\varphi_1(s_k) = \varphi_2(s_r) = \{E_1, \ldots, E_{n_s}\}$$

de forma que $\varphi_1(r) = \varphi_2(r)$ para $r > \max\{s_k, s_r\}$

PROPOSICION 4.1

Para el método del máximo y UPGMA

$$\Delta_o(d_1', d_2') \leq \max\{s_k, s_r\} - \alpha \tag{5}$$

DEMOSTRACION:

$$\Delta_o(d_1', d_2') =$$
$$= \max\{|d_1'(x_i, x_j) - d_2'(x_i, x_j)|, x_i, x_j \in X\} =$$
$$= \max\{\alpha_1, \alpha_2, \alpha_3\}$$

siendo

$$\alpha_1 = \max\{|d_1'(x_i, x_j) - d_2'(x_i, x_j)|, x_i, x_j \in X \text{ con } d_1'(x_i, x_j) < \alpha\}$$

$$\alpha_2 = \max\{|d_1'(x_i, x_j) - d_2'(x_i, x_j)|, x_i, x_j \in X \text{ con } d_1'(x_i, x_j) \in [\alpha, s_k)\}$$

$$\alpha_3 = \max\{|d_1'(x_i, x_j) - d_2'(x_i, x_j)|, x_i, x_j \in X \text{ con } d_1'(x_i, x_j) \geqslant s_k\}$$

Como $\alpha_1 = \alpha_3 = 0$ y $\alpha_2 \leq \max\{s_k - \alpha, s_r - \alpha\}$

se tiene

$$\Delta_o(d_1' d_2') \leq \max\{s_k - \alpha, s_r - \alpha\}$$

c.q.d.

PROPOSICION 4.2

Para el método UPGMA

$$\Delta_o(d_1', d_2') \leq \sum_{\substack{x_i, x_j \in X \\ i<j}} \frac{d(x_i, x_j)}{n-1} - \alpha \tag{6}$$

DEMOSTRACIÓN:

Veamos que $\varphi\left(\sum_{\substack{x_i, x_j \in X \\ i<j}} \frac{d(x_i, x_j)}{n-1}\right) = \{X\}$

Sea $\varphi(1_s) = \{F_1, F_2\}$ y

$$\beta = d_s(F_1, F_2) = \sum_{\substack{x_i \in F_1 \\ x_j \in F_2}} \frac{d(x_i, x_j)}{|F_1| \cdot (n - |F_1|)}$$

Por $F_1 \cap F_2 = \emptyset$ y al alcanzar $h(r) = (n-r).r$ el mínimo en $r=1$ y $r=n-1$ en el intervalo $[1, n-1]$ se tendrá

$$\beta \leq \sum_{\substack{x_i, x_j \in X \\ i<j}} \frac{d(x_i, x_j)}{|F_1| \cdot (n - |F_1|)} \leq \sum_{\substack{x_i, x_j \in X \\ i<j}} \frac{d(x_i, x_j)}{n-1}$$

resultando

$$\varphi\left(\sum_{\substack{x_i, x_j \in X \\ i<j}} \frac{d(x_i, x_j)}{n-1}\right) = \{X\}$$

y aplicando la proposición 4.1 se tiene

$$\Delta_o(d_1', d_2') \leq \sum_{\substack{x_i, x_j \in X \\ i<j}} \frac{d(x_i, x_j)}{n-1} - \alpha$$

c.q.d.

5. EJEMPLOS.

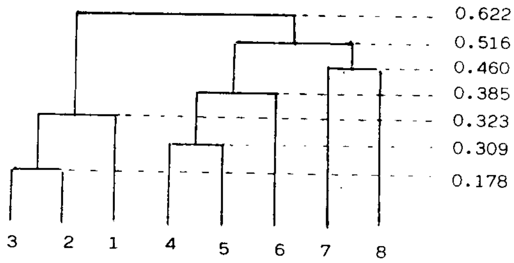
Sea d la distancia entre poblaciones de Drosophila subobscura definida por Prevosti /9/ que tiene por matriz asociada

$$
d = \begin{pmatrix}
0.000 & & & & & & & \\
0.276 & 0.000 & & & & & & \\
0.370 & 0.187 & 0.000 & & & & & \\
0.465 & 0.529 & 0.469 & 0.000 & & & & \\
0.657 & 0.693 & 0.606 & 0.309 & 0.000 & & & \\
0.643 & 0.680 & 0.581 & 0.357 & 0.413 & 0.000 & & \\
0.489 & 0.636 & 0.531 & 0.489 & 0.514 & 0.574 & 0.000 & \\
0.699 & 0.846 & 0.802 & 0.485 & 0.429 & 0.604 & 0.460 & 0.000
\end{pmatrix}
$$

siendo las poblaciones

1.Dalkeith 2. Groningen 3. Viena
4.Barcelona 5. Formia 6. Silifke
7.Chalus 8. Las Mercedes

Por el método UPGMA, el dendograma resultante es



y al considerar la clasificación a nivel 0.460 obtenemos los grupos

G_1: -Dalkeith, -Groningen, -Viena
G_2: -Barcelona, -Formia, -Silifke
G_3: -Chalus, -Las Mercedes

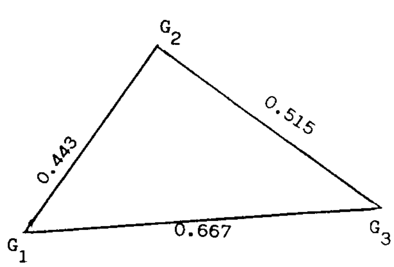
Así, en este nivel, d_5 vendrá definida por 3.2 como:

$$d_5(G_1,G_2) = \frac{0.465+0.529+0.469+0.657+0.693+0.606+0.643+0.680+0.581}{12} = 0.443$$

$$d_5(G_1,G_3) = \frac{0.489+0.636+0.531+0.699+0.846+0.802}{6} = 0.667$$

$$d_5(G_2,G_3) = \frac{0.489+0.514+0.574+0.485+0.429+0.604}{6} = 0.515$$

obteniendo gráficamente la siguiente representación en el plano



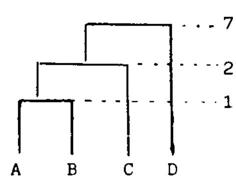
Observemos que la distancia d_5 la hemos calculado directamente de la distancia inicial, sin preocuparnos de las distancias d_2, d_3, d_4. Así, cuando el número de elementos sea elevado y cortemos en un nivel alto, la fórmula 3.2 será de gran utilidad para el estudio de las interdistancias entre los grupos resultantes y su posterior representación.

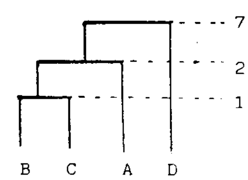
EJEMPLO 2

Como ya hemos referido en todo el artículo, el problema fundamental radica en la no unicidad de la clasificación asociada a una disimilaridad. Así, la disimilaridad d definida en $X = \{x_1, x_2, x_3, x_4\}$ por la matriz

$$
d = \begin{pmatrix}
0 & & & \\
1 & 0 & & \\
2 & 1 & 0 & \\
3 & 5 & 7 & 0
\end{pmatrix}
$$

nos proporciona por el método del máximo dos clasificaciones





con disimilaridades ultramétricas asociadas

$$
d_1' = \begin{pmatrix}
0 & & & \\
1 & 0 & & \\
2 & 2 & 0 & \\
7 & 7 & 7 & 0
\end{pmatrix}
\qquad
d_2' = \begin{pmatrix}
0 & & & \\
2 & 0 & & \\
2 & 1 & 0 & \\
7 & 7 & 7 & 0
\end{pmatrix}
$$

El problema así planteado reside en estudiar las "diferencias" entre las clasificaciones resultantes. Por esto, en 4 hemos acotado la distancia que separa d_1' y d_2', obteniendo en la proposición 4.1

$$\Delta_o(d_1', d_2') \leq \max\{s_k, s_r\} - \alpha = \max\{2,2\} - 1 = 1$$

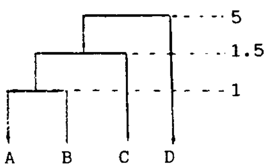
En segundo lugar, cabe preguntarnos por la validez de la cota hallada, siendo el ejemplo anterior quien nos saca de dudas a la -

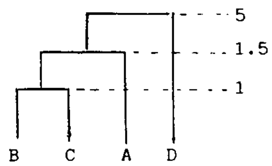
vista de

$\Delta_o(d_1', d_2') = 1$

y observando que la cota hallada no puede ser afinada.

Análogamente por el método UPGMA obtenemos las clasificaciones





con disimilaridad ultramétrica asociadas

$$d_1' = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 1.5 & 1.5 & 0 & \\ 5 & 5 & 5 & 0 \end{pmatrix} \qquad d_2' = \begin{pmatrix} 0 & & & \\ 1.5 & 0 & & \\ 1.5 & 1 & 0 & \\ 5 & 5 & 5 & 0 \end{pmatrix}$$

observando que por la proposición 4.1

$\Delta_o(d_1', d_2') \leq \max\{1.5,\ 1.5\} - 1 = 0.5$

y por lo mismo que en el caso anterior, como $\Delta_o(d_1', d_2') = 0.5$, la cota hallada en 4.1 es la más fina posible.

6. CONCLUSIONES.

El estudio de propiedades comunes respecto las imágenes de una misma disimilaridad al considerar la función de clasificación a partir de un método de clasificación fijado como una función multívoca, nos proporciona un camino para ir resolviendo el grave problema que surge de la indeterminación de la aplicación de métodos de clasificación jerárquica. En este trabajo se ha calculado una expresión general de la disimilaridad en un nivel de una clasificación jerárquica por los métodos del máximo y UPGMA, y a partir de una métrica definida en D_X, hallamos una cota de la distancia entre las ultramétricas asociadas a una misma disimilaridad inicial.

7. BIBLIOGRAFIA

/1/ ARCAS, A.: "Contribuciones a la construcción de clasificaciones estratificadas". Tesina, Fac. Matemáticas, Universidad de Barcelona, (inédita) (1983).

/2/ BENZECRI, J.P.: "L'Anal-se des Données. I. La Taxinomie. L'Analyse des Données. II. L'Analyse des correspondances". Dunod. Paris (1976).

/3/ BERGE, C.: "Espaces topologiques. Fonctions multivoques". Dunod. Paris. (1959)

/4/ CUADRAS, C.M.: "Métodos de análisis multivariante". Eunibar. Barcelona. (1981)

/5/ CUADRAS, C.M. y OLLER, J.M.: "Geometría finita aplicada a la Estadística". Actas XIV Reunión Nac. Estad. Inv. Op. e Inform. (1984). Tomo 1, pag. 287-296.

/6/ FARRIS, J.S.: "On the cophenetic correlation coefficient". Syst. Zool., 18(3), 279-285. (1969).

/7/ JARDINE, N. y SIBSON, R.: "Mathematical taxonomy". Jhon-Wiley. New York. (1971).

/8/ JOHNSON, S.C.: "Hierarchical clustering schemes". Psychometrika, 32, 241-254. (1967).

/9/ PREVOSTI, A.: "La distancia genética entre poblaciones". Miscellanea Alcobé, Univ. de Barcelona, 109-118.

/10/ ROHLF, F.J.: "Adaptive hierarchical clustering schemes". Syst. Zool., 19 pp.58-82. (1970).

/11/ SALICRU, M.: "Consideraciones sobre desemejanzas y clasificaciones asociadas". Tesina, Fac. Matemáticas, Univ. de Barcelona, (inédita). (1983).

/12/ SNEATH, P.H.A. y SOKAL, R.R.: "Numerical Taxonomy". W.H. Freeman and Co. S. Francisco. (1973)

/13/ SOKAL, R.R. y ROHLF, F.G.: "The comparation of dendograms by objective methods". Taxon, 11, 33-40. (1962).