QÜESTIIÓ, Vol 11, n° 2, pp. 39–50, 1987

SOBRE LA REPRESENTACION DE UN CONJUNTO MEDIANTE ARBOLES ADITIVOS

ANTONI ARCAS PONS

Universitat de Barcelona

En este trabajo se estudia el problema de la representación de un conjunto mediante árboles aditivos, en el sentido de hallar una formalización que permita abordar el mismo desde la perspectiva general de los métodos geométricos de respresentación del análisis multivariante.

On the additive tree representations

Keywords: Multivariate data analysis, Ultrametric distance, Additive distance, Additive tree, Ultrametric tree.

1. INTRODUCCION

Dentro de los métodos de análisis multivariante cabe resaltar el desarrollo que, en los últimos años, han gozado los métodos geométricos de representación. Este desarrollo ha sido fruto, fundamentalmente, de las necesidades de utilización de tales métodos en las ciencias experimentales. Ejemplos de trabajos teóricos y de aplicación en este sentido son: Alonso [1], Buenman [4], Dallot [10], Torgerson [21], Sattath y Tversky [18], Pruzansky et al. [17].

Los métodos de representación tendrán por objetivo la identificación de los elementos de una colección, en general, heterogénea de objetos, sobre la que tenemos definida una disimilaridad, en un espacio geométrico modelo en el sentido siguiente: Si $S = \{s_1 \ldots, s_n\}$ conjunto de objetos, δ disimilaridad definida sobre S y (X, d) espacio geométrico modelo, se trata de hallar una función

$$\Phi : (S, \delta) \longrightarrow (X, d) \tag{1}$$

–Antoni Arcas Pons.- Dptm. d'Estadística. Universidat de Barcelona.- Av. Diagonal, 645. 08028. Barcelona.

–Article rebut el febrer de 1987.

de modo que $d(\Phi(s_i), \Phi(s_j))$ se aproxime a $\delta(s_i, s_j)$ utilizando algún criterio de ajuste.

En caso de que

$$\delta(s_i, s_j) = d(\Phi(s_i), \Phi(s_j)) \tag{2}$$

tendremos una realización de S sobre X, entendida como una aplicación inyectiva e isométrica.

Se pueden considerar dos grupos de representaciones según el tipo de espacio geométrico modelo: modelos espaciales o continuos y modelos en árbol o discretos. El objetivo de los primeros, es la representación de cada objeto como las coordenadas de un punto en un espacio euclídeo, de modo que las distancias entre los mismos reflejen la disimilaridad observada, mientras que en los modelos discretos se trata de representar cada objeto como un vértice de un grafo conexo, de modo, que las relaciones entre los vértices en el grafo reflejen las relaciones de proximidad observadas.

2. CONCEPTOS PREVIOS Y PLANTEO DEL PROBLEMA

Definiremos a continuación dos tipos de distancias que utilizaremos en el trabajo:

Definición 2.1.

Una distancia $d : S \times S \longrightarrow R$ es ultramétrica si:

$\forall x, y, z \in S$

$$d(x, y) \leq \max \{d(x, z), d(y, z)\} \tag{3}$$

Definición 2.2.

Una distancia $d : S \times S \longrightarrow R$ es aditiva si:

$\forall x, y, z, t \in S$

$$d(x, y) + d(z, t) \leq \max \{d(x, z) + d(y, t), d(x, t) + d(y, z)\} \tag{4}$$

Proposición 2.1.

Si d es distancia ultramétrica, también es distancia aditiva.

Demostración:

Consideremos x, x, z elementos de S y supongamos que

$$d(x,y) \leq \min\{d(x,z), d(x,t), d(x,z), d(y,t), d(z,t)\} \tag{5}$$

Por ser d distancia ultramétrica se verifica que

$$d(x,z) = d(y,z) \quad \text{y} \quad d(x,t) = d(y,t)$$

de donde

$$d(x,z) + d(y,t) = d(x,t) + d(y,z) \tag{6}$$

Por otro lado

$$d(z,t) \leq \max\{d(y,z), d(y,t)\} \tag{7}$$

y, utilizando (5), (6) y (7) se deduce inmediatamente

$$d(x,y) + d(z,t) \leq d(x,z) + d(y,t) = d(x,t) + d(y,z) \tag{8}$$

Dada cualquier cuaterna (x,y,z,t) se ha comprobado que podemos elegir dos pares que cumplen la condición (8), resultando, por tanto, inmediata la verificación del axioma (4), por lo que d es aditiva.

Proposición 2.2.

Si u distancia ultrmétrica sobre S y $\varphi : S \longrightarrow R$ función de S en R verificando: $: -(\varphi(s_i) + \varphi(s_j)) < u(s_i, s_j), i \neq j$ la distancia $d : S \times S \longrightarrow R$ definida como

$$d(s_i, s_j) = u(s_i, s_j) + \varphi(s_i) + \varphi(s_j) \quad \text{si} \quad i \neq j$$

$$d(s_i, s_i) = 0$$

es distancia aditiva.

La demostración es inmediata.

El problema que se plantea en este trabajo consiste en demostrar que existe una realización entre (S, d), siendo d distancia aditiva, y un determinado modelo de árbol que denominaremos árbol aditivo. De este modo se obtendrá una formalización de las representaciones mediante árboles aditivos. También se hallarán, en base a esta formalización, las relaciones con las representaciones mediante árboles ultramétricos.

Es de resaltar la existencia de otros tipos de formalizaciones como la desarrollada por Buneman [4] en donde se estudia la representación mediante árboles aditivos utilizando modelos conjuntistas, en una línea parecida al estudio de árboles ultramétricos mediante jerarquías indexadas. La principal ventaja que conlleva este estudio respecto al trabajo anteriormente citado es que se ajusta mejor a los planteamientos intuitivos con relación al problema de la representación en el sentido indicado en (1).

3. REPRESENTACIONES MEDIANTE ARBOLES ADITIVOS

Definiremos previamente el concepto de árbol aditivo como modelo de representación.

Definición 3.1.

Dado (S, d) donde S es un conjunto a representar, d una distancia definida en S, supongamos que existe un conjunto finito T tal que $S \cap T = \phi$, y consideremos $V = S \cup T$ al que denominaremos conjunto de vértices; si existe $G_V \subset V \times V$ verificando

a) G_V grafo conexo sin ciclos.

b) Si $(x, y) \in G_V$ entonces $(y, x) \in G_V$.

c) Si $(x, y) \in G_V, (x, z) \notin G_V, \forall z \in V, z \neq y$, entonces $x \in S$ (x vértice terminal).

d) Existe $d^* : V \times V \longrightarrow R$, distancia sobre V, verificando:

d1) $d^*(x,y) = \sum_{i=1}^{m-1} d^*(x_i, x_{i+1}) | x_1 = x, x_m = y, (x_i, x_{i+1}) \in G_V$ para $i = 1, \ldots, m-1$

d2) $d^*_{|S} = d$

diremos que (G_V, d^*) es una representación en árbol aditivo asociada a (S, d)

Los resultados que se exponen a continuación muestran que dada d distancia aditiva sobre S, existe una realización entre (S, d) y un determinado árbol aditivo (G_V, d^*).

Se debe demostrar previamente el siguiente lema:

Lema 3.1.

Dado S conjunto finito, d distancia aditiva definida sobre S existe $(x_0, y_0) \in S \times S$ tal que

$$(11) \qquad d(x_0, y_0) + d(z,t) \leq d(x_0, z) + d(y_0, t) = d(x_0, t) + d(y_0, z)$$

$\forall (z,t) \in S \times S$ de modo que $\{x_0, y_0\} \cap \{z,t\} = \phi$

La demostración se efectúa por inducción sobre el cardinal $S(|S| = n)$, resultando inmediata en el caso en que el mismo sea $n = 4$. Si se considera cierto el lema para un conjunto de cardinal $n-1$, para un conjunto S de cardinal n, consideramos

$$S' = S - \{y\}$$

siendo y un elemento cualquiera de S. Por hipótesis de inducción podemos encontrar $(x_0, y_0) \in S' \times S'$ tales que

$$\forall (z,t) \in S' \times S' \text{ verificando } \{x_0, y_0\} \cap \{z,t\} = \phi$$

$$d(x_0, y_0) + d(z,t) \leq d(x_0, z) + d(y_0, t) = d(x_0, t) + d(y_0, z)$$

$$\text{Si } \; d(x_0, y_0) + d(y,t) \leq d(x_0, y) + d(y_0, t) = d(z_0, t) + d(y_0, y)$$

$\forall t \in S'$ tal que $t \notin \{x_0, y_0\}$, la proposición quedará probada, siendo (x_0, y_0) la pareja buscada satisfaciendo las condiciones del lema. Si por el contrario existe $z_0 \in S'$ tal que

$$d(x_0, y) + d(y_0, z_0) < d(x_0, y_0) + d(y, z_0) = d(x_0, z_0) + d(y, y_0)$$

se demuestra primero que

$$d(x_0, y) + d(y_0, z) \leq d(x_0, y_0) + d(y, z) = d(x_0, z) + d(y, y_0)$$

si $z \notin \{x_0, y\}$ y, a partir del mismo, el resultado más general

$$d(x_0, y) + d(z_1, z_2) \leq d(x_0, z_1) + d(y, z_2) = d(x_0, z_2) + d(y, z_1)$$

$\forall (z_1, z_2) \in SxS$ tal que $\{z_1, z_2\} \cap \{x_0, y\} = \phi$

con lo cual quedará probado el elma, siendo (x_0, y) la pareja que cumple las condiciones del mismo.

Teorema 3.1.

Dado (S, d) con d distancia aditiva asociada a S, existe una realización de (S, d) en un árbol aditivo, y recíprocamente, dado (G, d^*) árbol aditivo $d^*_{|S}$ es distancia aditiva.

Demostración:

Sea (S, d) con $|S| = n$, y d distancia aditiva asociada a S. La demostración se realiza por inducción sobre n

a) Caso $n = 4$

Sea $S = \{x, y, z, t\}$ y

$$d(x, y) + d(z, t) \leq d(x, t) + d(y, z) = d(x, z) + d(y, t) \tag{12}$$

Veamos que existe un árbol aditivo (Fig. 1) asociado a (S, d)

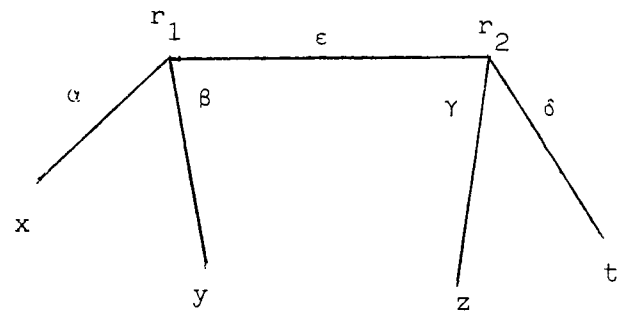

FIGURA 1

Planteado el sistema de ecuaciones, en base a imponer $d^*_{|S} = d$, se obtienen como soluciones:

$$\begin{aligned}
\alpha &= 1/2.(d(x,y) + d(x,t) - d(y,t)) \\
\beta &= 1/2.(d(x,y) + d(y,t) - d(x,t)) \\
\gamma &= 1/2.(d(y,z) + d(z,t) - d(y,t)) \\
\delta &= 1/2.(d(z,t) + d(y,t) - d(y,z)) \\
\varepsilon &= 1/2.(d(y,z) + d(x,t) - d(x,y) - d(z,t))
\end{aligned}$$

que, a partir de (12), son positivas. La construcción del árbol aditivo resulta, ahora, inmediata.

b) Caso $n > 4$

Veamos que si es cierto para $|S| = n - 1$, también es cierto para n.

Sea (x, y) verificando las condiciones del lema. Consideremos $S' = S - \{y\}$, y la distancia $d_1 = d_{|S'}$ que verifica, evidentemente, el axioma aditivo. A partir de la hipótesis de inducción se puede asegurar la existencia de un árbol aditivo que representa al par (S', d_1). Se trata de efectuar una "ampliación" del árbol (Fig. 2) introduciendo "y", al igual que un nuevo vértice del conjunto T en el

camino entre "x" y "r_1", con (x, r_1) elemento del árbol aditivo asociado a S'. Se elige $z \epsilon S'$ y se hallan valores α, β, β' (Fig. 2) que verifiquen $d^*_{|S} = d$.

(13)

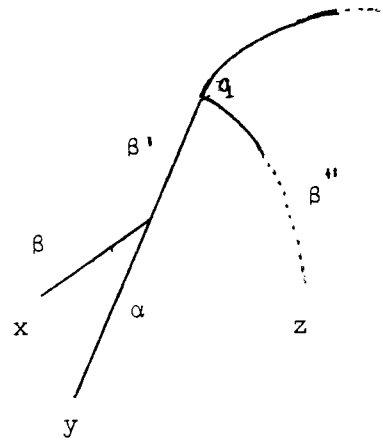

FIGURA 2

Para ello se plantea el sistema

$$\text{(14)} \qquad \begin{cases} d(x,y) = \alpha + \beta \\ d(y,z) = \alpha + \beta' + \beta'' \\ d(x,z) = \beta + \beta' + \beta'' \end{cases}$$

cuyas soluciones son:

$$\text{(15)} \qquad \begin{aligned} \beta' &= 1/2 \cdot (d(y,z) + d(x,z) - d(x,y) - 2_{\beta''}) \\ \beta &= 1/2 \cdot (d(x,z) + d(x,y) - d(y,z)) \\ \alpha &= 1/2 \cdot (d(y,z) + d(x,y) - d(x,z)) \end{aligned}$$

comprobándose fácilmente, a partir del lema, que las soluciones son no negativas. La distancia a través del árbol entre "y" y otro elemento $z' \epsilon S'$ es

$$\alpha + \beta' + d(x, z') - (\beta + \beta')$$

sustituyendo en (15) resulta

$$d(x, z') + d(y, z) - d(x, z)$$

y, por la elección del par (x, y), se verifica

$$d(x, y) + d(z, z') \leq d(x, z) + d(y, z') = d(x, z') + d(y, z)$$

resultando finalmente

$$d(x, z') + d(y, z) - d(x, z) = d(y, z')$$

quedando comprobado (13).

Para demostrar el recíproco, dados $x, y, z, t \in S$ y suponiendo

(16) $$d(x, y) + d(z, t) \leq d(x, z) + d(y, t) \leq d(x, t) + d(y, z)$$

se consideran

$$V_1 = \{v \in V \mid \quad v \text{ pertenece al camino entre } \quad x \; e \; y \; (v \neq x, y)\}$$
$$V_2 = \{v \in V \mid \quad v \text{ pertenece al camino entre } \quad z \; y \; t \; (v \neq z, t)\}$$

De este modo es sencillo probar: a) x e y no pueden pertenecer ambos a $V_2 b)$. Si $x \epsilon V_2$ e $y \notin V_2$ entonces las desigualdades (16) son igualdades c) Si $x, y \notin V_2$ entonces $d(x, t) + d(y, z) = d(x, z) + d(y, t)$, quedando con ello demostrado el teorema.

Se puede comprobar, asimismo, que el árbol aditivo asociado a (S, d) tendrá una configuración única.

Dado un árbol aditivo, es fácil demostrar (Arcas [2]) que se puede definir una función.

(17) $$\Phi : V \longrightarrow R^2$$

inyectiva, de manera que para cada par de puntos $(u, v) \in G_V$ con $\Phi(u) = (\Phi_1(u), \Phi_2(u))$ y $\Phi(v) = (\Phi_1(v), \Phi_2(v))$ (Supongamos por ejemplo $\Phi_1(u) \leq \Phi_1(v)$ y $\Phi_2(u) \leq \Phi_2(v)$) existe un conjunto P_{uv} tal que

$$P_{uv} = \{(\Phi_1(u), \alpha) | \alpha \in [\Phi_2(u), \Phi_2(v)]\} \cup \{(\alpha, \Phi_2(v) | \alpha \in [\Phi_1(u), \Phi_1(v)]\}$$

verificándose

$$d^*(u,v) = |\Phi_2(u) - \Phi_2(v)| \tag{18}$$

y de modo que dado $(x,u) \in G_V$ con x vértice terminal,

$$\Phi_2(x) \leq \Phi_2(u)$$

se puede comprobar entonces

Teorema 3.2.

Dado (S,u) con u distancia sobre S, existe un árbol aditivo tal que el conjunto de vértices terminales coincide con los elementos de S y $\Phi_2(s_1) = ... = \Phi_2(s_n)$ si y sólo si u es distancia ultramétrica

Si u es ultramétrica, la construcción del árbol aditivo se realizará utilizando el algoritmo fundamental de clasificación (Cuadras,[6]). El recíproco se demostrará directamente a través de las propiedades del árbol. El árbol resultante se denominará árbol ultramétrico o dendrograma.

Como consecuencia de los resultados anteriores, es inmediato comprobar entonces

Teorema 3.3.

Dada d distancia aditiva, existe una distancia ultramétrica u y una función $\varphi : S \longrightarrow R$ de modo que

$$d(s_i, s_j) = u((s_i, s_j) + \varphi(s_i) + \varphi(s_j) \text{ si } i \neq j \tag{20}$$

Quedarán de este modo explicadas las relaciones entre las representaciones asociadas a distancias aditivas y distancias ultramétricas.

4. CONCLUSIONES

– Se ha desarrollado una formalización sobre las relaciones entre una distancia aditiva (verificando el axioma del cuarto punto) y un árbol aditivo, en el sentido de demostrar que la obtención de una distancia aditiva sobre un conjunto permite obtener una realización en un espacio geométrico modelo constituido por un árbol aditivo.

– Se han hallado en términos de la misma formalización las relaciones entre representaciones mediante árboles aditivos y árboles ultramétricos.

– Las relaciones (20) permiten pensar en la posibilidad de abordar el problema de la representación espacial de distancias aditivas de forma paralela al caso de distancias ultramétricas, en el mismo sentido que los trabajos de Cuadras [7] y Ohsumi y Nakamura [16].

5. NOTAS BIBLIOGRAFICAS

[1] **Alonso, G.** (1975) "Estudio de la distribución geográfica del polimorfismo crosmosómico en Drosophila Subobscura". Tesina, Fac. Biología Univ. Barcelona

[2] **Arcas, A** (1986). "Contribuciones a la representación de datos multidimensionales mediante árboles aditivos". Tesis, Fac. de Matemáticas, Univ. Barcelona.

[3] **Berge, C.** (1973). "Graphes et Hipergraphes". Dunod, Paris.

[4] **Buneman, P.** (1971). "The recovery of trees from measures of dissimilarity". F. R. Hodson, D. G. Kendall, P. Tautu (Edit). "Mathematics in the Archeological and historical sciences". Edinburgh University Press.

[5] **Cuadras, C. M.** (1980) "Mètodes de representació de dades y la seva aplicació en Biologia". Col. Soc. Catalana Biologia, 13,95-133.

[6] **Cuadras, C. M.** (1981). "Métodos de análisis multivariante". Eunibar, Barcelona.

[7] **Cuadras, C. M.** (1983), "Análisis algebraico sobre distancias ultramétricas". Actas 44 Per. de Sesiones del Inst. Intern. de Estad., Madrid, V. 2,554-557.

[8] **Cuadras, C. M., Oller J. M., Arcas A. y Rios M.** (1986). "Métodos geométricos de la estadística".Questió, V.9, nº4, pp. 219-250.

[9] **Cunningham, J. P.** (1978). "Free trees and bidirectional trees as representations of psychological distance". Journal of Mathematical Psychology, 17, 165-188.

[10] **Dallot S., e Ibañez F.** (1972) "Étude préliminaire de la morphologie et de l'évolution chez les chaetognates". Inv. Pesq. 36 (1), 31-41.

[11] **De Soete; G.** (1983). "A least square algorithm for fitting aditive trees to proximity data". psychometrika, 48 (4), 621-626.

[12] **Fitch, W. y Margoliash, E.** (1967). "Construction of Phylogenetic trees". Science, V. 155, 279-284.

[13] **Gower, J. C,** (1967) "Euclidean distance geometry". Math. Scientist. 7-14.

[14] **Jardine, N. y Sibson, R.** (1971) "Mathematical Taxonomy". John Wiley,New York.

[15] **Kruskal, J. B. y Wish, M.** (1978). "Multidimensional Scaling". Sage Publication, Beverly Hills.

[16] **Ohsumi, N. y Nakamura, T.** (1981). "Some properties of monotone hierchical dendrogram in numerical classification". Proc. Inst. Statist. Mathem., 28 (1), 117-133.

[17] **Pruzanski S., Tversky A. y Carroll J. D.** (1982). "Spatial versus tree representations of proximity data". Psychometrika, V. 47, 1,3-24.

[18] **Sattath, S. y Tversky , A.** (1977). "Additive similarity trees". Psychometrika, 42(3), 319-345.

[19] **Sneath, P. H. A. y Sokal R. S.** (1973) "Numerical Taxonomy". W. H. Freeman and Co., San Francisco.

[20] **Sokal, R. R. y Sneath P. H. A.** (1963). "Principles of numerical taxonomy". W. H. Freman and Co., San Francisco.

[22] **Torgerson, M. S.** (1958). "Theory and methods of scaling". John Wiley, New York.

[23] **Waterman, M. S.; Smith T. F.; Singh M. y Beyer N. A.** (1977). "Additive evolutionary trees". J. Theor. Biol., 64, 199-213.

[24] **Wish, M. y Carroll J. D.** (1982) "Multidimensional Scaling and its Applications". P. R. Krishnaiah and L. N. Kernel Eds. Handboock of Statistics, V.2, 317-345.