Revista Colombiana de Estadística
Nº 27

CLASIFICACION JERARQUICA CON VARIABLES BINARIAS Y NOMINALES

CAMPO ELÍAS PARDO T
M. Sc. en Estadística Universidad Nacional

LEONARDO BAUTISTA S.
Profesor Asociado Universidad Nacional

RESUMEN. Se analiza la utilización del algoritmo de clasificación de Ward en conjuntos de datos en los que la información está conformada únicamente por variables binarias o nominales. Se lleva el caso nominal al caso binario mediante una codificación disyuntiva completa y se establecen las matrices de distancias a partir de la distancia ponderada de Manhatan calculada a través de las distancias promedio de Manhatan y de Bray-Curtis. Se estudian todos los casos de dos y tres variables binarias y de dos variables nominales con dos y tres modalidades. Se establece el efecto que sobre los árboles resultantes tiene la asignación apriori de ponderaciones para las variables.

INTRODUCCION

La clasificación no jerárquica de un grupo de individuos caracterizados por p variables binarias conduce en forma natural a pensar en la existencia de 2^p categorías o tipos de individuos los cuales tienen una cantidad determinada de representantes en el grupo de datos. Algunas clases pueden ser vacías mientras otras tienen una amplia cantidad de individuos. La situación se hace muy compleja cuando se pretende que la clasificación sea jerárquica. En primer lugar se hace necesario introducir un concepto adicional de ponderación de variables, en segundo lugar la ausencia de datos en ciertos tipos de datos crea "deformaciones no naturales" en los árboles y tercero no se conocen los efectos del paso binario al caso más general, es decir cuando se está frente a q_j modalidades para la j-ésima variable. Este artículo analiza y presenta estos tres aspectos.

1. Variables y Codificación

Se utilizará la definición que proporciona Lermán (1981) sobre variables binarias y nominales:

Variable binaria o lógica

$$E \longrightarrow \{0,1\}$$
$$x \longrightarrow a(x) = \begin{cases} 1 & \text{si el atributo está presente} \\ 0 & \text{si el atributo está ausente} \end{cases}$$

Esta variable divide a E en dos partes, sea Ea el conjunto de los individuos que poseen el atributo:

$$E_a = a^{-1}(1)$$

Variable nominal

Sea $\{C_1, C_2, \ldots, C_k\}$ el conjunto de modalidades sobre el cual se supone no hay estructura de orden. La variable se define como la aplicación de

$$E \longrightarrow \{C_1, C_2, \ldots, C_k\}$$
$$x \longrightarrow a(x) = C_i \text{ si el objeto } x \text{ posee la modalidad } i$$

Es una variable indicadora de la partición

$$\{E_1, E_2, \ldots, E_k\} \quad \text{donde} \quad E_i = a^{-1}(C_i)$$

Esta forma conocida como de código condensado tiene otra forma de representación conocida como Disyuntiva Completa, la cual establece una variable binaria por cada modalidad así:

$$a_i(x) = \begin{cases} 1 & \text{si el individuo } x \text{ posee la modalidad } i. \\ 0 & \text{si no la posee} \end{cases}$$

La codificación disyuntiva completa (notada en adelante TDC) para variables nominales tiene la ventaja de permitir el uso de los índices de similaridad para variables binarias.

2. Selección de la distancia

Para los desarrollos que aquí se presentan se toma como base la distancia ponderada de Manhatan como índice de disimilaridad:

$$d(i,j) = \sum_{k=1}^{p} p_k |x_{ki} - x_{kj}|$$

donde p_k ($p_k \geq 0$ para todo k; $\sum p_k = 1$) es el peso de la variable k. Esta selección está estrechamente ligada a la decisión de utilizar el algoritmo de WARD de clasificación. Este algoritmo se aplica para variables continuas con la distancias euclidianas. La distancia ponderada de Manhatan equivale, en los casos de variables binarias y nominales al cuadrado de la distancia euclidiana ponderada, por lo que la tabla inicial de distancias será la raíz de las distancias ponderadas de Manhatan.

Para efectos prácticos se puede obtener la distancia ponderada de Manhatan a partir de la conocida distancia promedio de Manhatan en el caso de variables binarias. Basta con reemplazar los unos por el valor del peso de la variable y realizar el cálculo. En el caso de tablas con codificación disyuntiva completa (TDC) la distancia ponderada de Manhatan se hace igual reemplazo y se utiliza la conocida distancia de Bráy-Curtis:

$$d_{ij} = \frac{\sum |x_{ki} - x_{kj}|}{\sum (x_{ki} + x_{kj})} \tag{2}$$

3. Selección del método de afectación

El criterio de afectación del método de Ward tiene como principio unir aquellos grupos para los cuales el incremento de la inercia "Dentro" sea mínima. La definición de distancia de Ward entre grupos disyuntos A y B está dada por:

$$W(A,B) = \frac{f_A f_B}{f_A + f_B} \|\underline{g_A} - \underline{g_B}\|^2$$

y corresponde al incremento en la inercia "Dentro" al unir las clases A y B. En particular para dos grupos cada uno con un único individuo la distancia de Ward está dada por:

$$W(\underline{x_i}, \underline{x_j}) = \frac{1}{2}\|\underline{x_i} - \underline{x_j}\|^2 = \frac{1}{2} d^2(\underline{x_i}, \underline{x_j})$$

Utilizando $d(x_i, x_j)$ como la raíz de la distancia ponderada de Manhatan, la distancia de Ward entre dos individuos está entonces dada por $W(x_i, x_j) = (1/2).d(x_i, x_j)$.

El método de Ward es secuencial y la formula de recurrencia para evaluar la distancia de Ward de un grupo AUB, con respecto a otro C, es:

$$W(A \cap B, C) = \frac{(f_A + f_C)W(A,C) + (f_B + f_C)W(B,C) - f_C W(A,B)}{f_A + f_B + f_C}$$

4. El método de Ward con variables binarias

Antes de presentar estos resultados es necesario precisar el sentido de algunas formas particulares de notación que se van a utilizar. Las variables se nombran V_1, V_2, etc., y se llama V_1 a la variable de mayor peso, V_2 a la que le sigue y así en forma ordenada hasta la última. El peso de las variables se identifica con $P_1 > P_2 > \cdots$, etc. Se utilizan letras mayúsculas para denominar tipos de individuos. El orden $A, B, C, \ldots$ corresponde al sentido descendente en un sistema numérico de base dos. Así por ejemplo en el caso de tres variables binarias la letra A corresponde a los individuos con los valores 1,1,1; y la letra H aquellos con los valores 0,0,0. Los grupos se notan según las letras de los tipos de los individuos que los conforman. Así, ABG es el grupo en el cual hay individuos de tipo A, B y G. Se dice que se está ante una tabla de datos de "caso completo", cuando para cada tipo posible de individuos hay por lo menos un dato, de lo contrario se dirá "caso incompleto".

4.1 Clasificación con DOS variables binarias. Sean las variables V_1 y V_2 que generan los tipos A, B, C y D, con los pesos P_1 y P_2 respectivamente. Los pesos cumplen $P_1 > P_2$ y $P_1 + P_2 = 1$, para lo cual es necesario entonces que $P_1 > 0.5$ y $P_2 < 0.5$.

La tabla de datos para un caso completo y simétrico y la tabla de distancias ponderadas de Manhatan entre tipos están dadas por:

	$V1$	$V2$	$TIPO$
$U1$	1	1	A
$U2$	1	0	B
$U3$	0	1	C
$U4$	0	0	D

	A	B	C	D
A	0	P_2	P_1	1
B	P_2	0	1	P_1
C	P_1	1	0	P_2
D	1	P_1	P_2	0

La distancias de Ward entre individuos es igual a un medio de la matriz de distancias ponderadas de Manhatan.

El menor valor (mayor a cero) en la tabla de distancias de Ward es $\frac{P_2}{2}$ lo que implica la unión a igual altura de los tipos A y B o C y D. El algoritmo se decide por el punto más cercano al borde superior izquierdo de la tabla. Se unen los puntos A y B conformando el grupo AB. Las distancias AB-C y AB-D se obtienen mediante la fórmula de recurrencia.

$$\begin{aligned} W(AB,C) &= \frac{1}{3}(2\frac{P_1}{2} + 2\frac{1}{2} - \frac{P_2}{2}) \\ &= \frac{1}{3}(P_1 + 1 - \frac{P_2}{2}) = \frac{1}{3}(P_1 + P_1 + P_2 - P_2/2) \\ &= (4P_1 + P_2)/6 \end{aligned}$$

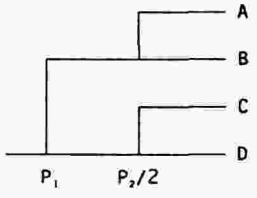
La nueva matriz está dada por:

	AB	C	D
AB	0		
C	$(4P_1 + P_2)/6$	0	
D	$(4P_1 + P_2)/6$	$P_2/2$	0

en la que la siguiente unión es entre C y D, generando la matriz:

	AB	CD
AB	0	
CD	P_1	0

El dendrograma obtenido es:



La inercia total es 1 y para dos clases $I_E/I_T = P_1 > 0.5$. En la matriz original de distancias se presenta un empate, pero se verifica que el resultado es invariante a la decisión sobre la primera unión.

Los casos incompletos están conformados por conjuntos de dos y tres tipos de individuos. Para el análisis se estudian todos los casos posibles y se encuentra que de acuerdo a la forma de los árboles resultantes los casos posibles conforman grupos de idénticos resultados. Estos grupos se nombran comenzando por los caracteres B2 (de binarias con dos variables), seguidos del número de individuos y un consecutivo de grupo. B2I2G1 es el primer grupo en el caso de dos individuos con dos variables binarias.

Casos incompletos con dos tipos de individuos.

De los cuatro tipos de individuos (A, B, C, D), se pueden tomar seis subconjuntos de dos $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{B, C\}$, $\{B, D\}$ y $\{C, D\}$. De estos seis casos se obtienen tres grupos :

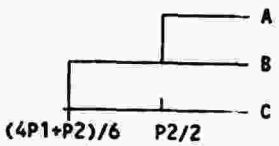
B2I2G1 : $\{A, B\}$ y $\{C, D\}$

B2I2G2 : $\{A, C\}$ y $\{B, D\}$

B2I2G3 : $\{A, D\}$ y $\{B, C\}$

Casos incompletos con tres tipos de individuos.

Hay cuatro casos posibles $\{A, B, C\}$, $\{B, A, D\}$, $\{C, D, A\}$ y $\{D, C, B\}$ los cuales concluyen siempre en igual clasificación, dando así origen al grupo B2I3G1 cuyo árbol es:



4.2 Clasificación con TRES variables binarias.. Con tres variables se pueden tener $2^3 = 8$ tipos de individuos posibles, que se identifican con las letras $A, B, \ldots, H$. Los pesos de las variables son P_1, P_2 y P_3 con $P_1 > P_2 > P_3$ y $P_1 + P_2 + P_3 = 1$. En el plano P_3 vs. P_2 los puntos que cumplen las condiciones anteriores son:

$$\{(P_2, P_3) : P_3 < P_2, P_3 < 1 - 2P_2, P_3 > 0\}$$

Estos puntos son interiores al triángulo de vértices: $A(0,0)$, $E(1/3, 1/3)$ y $H(1/2, 0)$ de la figura 1.

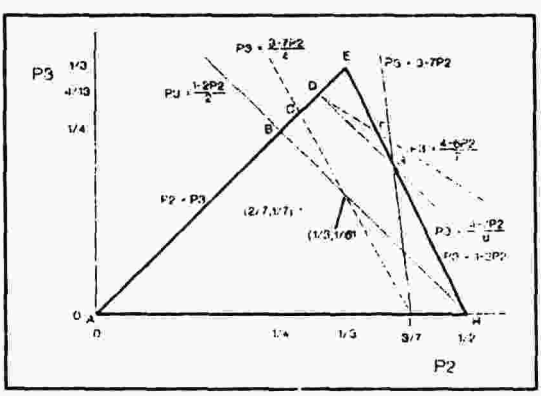
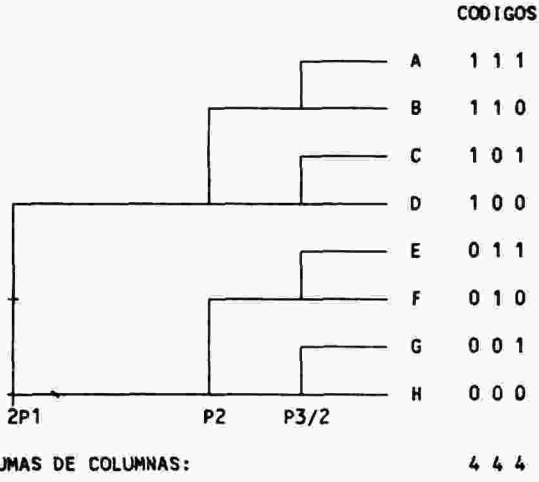


Figura 1

La tabla de datos no difiere conceptualmente del caso de dos variables, como tampoco lo hace la tabla de distancias de Ward entre individuos. El árbol de clasificación para el caso completo está dado por:



Inercia total = 2

Al hacer un corte a una altura entre P_2 y $2P_1$ se obtiene las clases $\{A, B, C, D\}$ y $\{E, F, G, H\}$, que corresponden a la partición dada por la variable 1, la de mayor peso. Si se hace un corte a una altura comprendida entre $P_3/2$ y P_2 se obtienen las cuatro clases: $\{A, B\}$, $\{C, D\}$, $\{E, F\}$ y $\{G, H\}$, que corresponden a las particiones según la variable 2 de cada una de las dos clases anteriores. Como se nota de la codificación, los resultados se corresponden a un ordenamiento de números de un sistema numérico binario de tres dígitos.

El interés está obviamente en los casos incompletos. Algunos tablas incompletas conducen a una partición que sigue una lógica similar a la obtenida en el caso completo, en cuyo caso se dirá que el resultado es un "árbol parecido al completo (A.P.C)".

Los casos incompletos con tres variables binarias.

Los subconjuntos de dos individuos son todos reducibles a una variable, por tener variables con el mismo valor o correlacionadas. De los 56 subconjuntos de tres individuos solo ocho son no reducibles a dos variables y corresponde precisamente a aquellos cuya tabla de datos tiene la característica de conformar una matriz no singular, todos ellos concluyen en el grupo B3I3G7.

De los 70 subconjuntos de cuatro individuos, doce son reducibles a dos variables, seis porque una variable tiene el mismo valor para los cuatro individuos y concluyen en tres árboles (B3I4G1 a B3I4G3) y seis por existir correlación entre dos variables (árboles B3I4G4 a B3I4G6).

Los conjuntos de dos individuos son complementarios a los de seis y los de cinco

a los de tres. Los grupos que se forman son también "complementarios", es decir los complementos de los conjuntos que pertenecen a un grupo de dos individuos, por ejemplo, pertenecen a un grupo de seis individuos.

En el cuadro 1 se puede observar que la mayoría de los grupos se pueden identificar por el valor de la inercia total y las sumas de las columnas de las variables en la tabla de datos. Otros grupos, como B3I5G4 a B3I5G7 tienen la misma inercia y las mismas posibilidades para la suma de las columnas, se pueden diferenciar entonces por la fracción de inercia entre los grupos sobre inercia total. En estos grupos la inercia total no depende de los pesos.

Para algunos conjuntos se obtiene un solo árbol con tal que se cumpla la condición $P_1 > P_2 > P_3$. Dicho árbol es "parecido al completo". Para otros se puede obtener uno de dos árboles, dependiendo de una condición adicional en los pesos. Uno de estos dos árboles es el "parecido la completo". En estos grupos hay un enfrentamiento entre el peso de las variables y el "peso de las ramas".

En el conjunto $\{A, B, C, G\}$, por ejemplo se tienen las dos posibilidades. En un primer árbol hay primero una partición según la variable uno formando las ramas $\{A, B, C\}$ y $\{G\}$. Luego el conjunto $\{A, B, C\}$ se divide según la variable dos, en las ramas $\{A, B\}$ y $\{C\}$ y finalmente se separan A y B que tienen diferente valor para la variable tres. Esto es análogo a lo que sucede en el caso completo. En este sentido se dice que el primer árbol es el "parecido al completo". En el segundo árbol al obtener dos clases se tiene una partición según la variable dos, que no es la de mayor peso. $\{A, B\}$ se divide luego según la variable tres y finalmente $\{C, D\}$ lo hace según la variable uno.

En el primer caso se tienen las ramas $\{A, B, C\}$ y $\{G\}$, pero como el individuo G está solo, tiende a unirse al individuo C de la otra rama, el cual a su vez está solo dentro de esa rama. La distancia de Ward entre los dos individuos "solos" es, $(W(C, G) = P_1/2)$. Con pesos de las variables lo suficientemente diferentes se puede contrarrestar el efecto del "peso de la rama" $(P_3 < (3 - 7P_2)/4)$.

En los grupos B3I4G10, B3I4G13, B3I4G14, B3I5G4, B3I6G3 y B3I6G4, el árbol "parecido al completo" se obtiene cuando los pesos caen en la región opuesta al vértice de pesos iguales, $E(1/3, 1/3)$ (ver cuadro y figura 1). En el grupo B3I4G10, por ejemplo, el árbol "parecido al completo" forma dos ramas, una con tres individuos y otra con uno. En los grupos B3I4G11 y B3I6G5 el árbol "parecido al completo" se obtiene en la región que toca el vértice de pesos iguales. En el grupo B3I4G11 el árbol "parecido al completo" tiene la misma forma que el del grupo B3I4G10 (descrita en el párrafo anterior). En el conjunto $\{A, B, C, H\}$, por ejemplo, las dos ramas son

CUADRO 1. Inercia, inercia explicada, suma de columnas y condición para ser árbol parecido al completo en casos incompletos con dos y tres variables binarias

GRUPO	INERCIA	SUMA DE COLUMNAS	CONDICION PARA AP
B2I3G1	2/3		
B2I4G1	1		
B3I3G7	2/3	SC1,SC2,SC3 = 1 o 2	
B3I4G7	1	SC1,SC2,SC3 = 2	
B3I4G8	3/4	SC1,SC2,SC3 = 1 o 3	
B3I4G9	$3/4 + P_3/4$	SC3 = 2	
B3I4G10	$3/4 + P_2/4$	SC2 = 2	$P_3 < (3 - 7P_2)/4$
B3I4G11	$3/4 + (P_2 + P_3)/4$	SC2,SC3 = 2	$P_3 < 3 - 7P_2$
B3I4G12	$3/4 + P_1/4$	SC1 = 2	
B3I4G13	$3/4 + (P_1 + P_3)/4$	SC1,SC3 = 2	$P_3 < (4 - 7P_2)/6$
B3I4G14	$3/4 + (P_1 + P_2)/4$	SC1,SC2 = 2	$P_3 < (4 - 6P_2)/7$
B3I5G1	$\frac{4}{5} + \frac{2(P_2+P_3)}{5}$	SC1 = 4 o 1	
B3I5G2	$\frac{4}{5} + \frac{2(P_1+P_3)}{5}$	SC2 = 4 o 1	
B3I5G3	$\frac{4}{5} + \frac{2(P_1+P_2)}{5}$	SC3 = 4 o 1	
B3I5G4 a	6/5	SC1,SC2,SC3 = 2 o 3	
B3I5G7			
B3I6G1	$4/3 + P_3/6$	SC3 = 3	
B3I6G2	$4/3 + P_2/6$	SC2 = 3	
B3I6G3	$4/3 + P_1/6$	SC1 = 3	$P_3 < (3 - 7P_2)/4$
B3I6G4	$4/3 + (P_2 + P_3)/6$	SC2,SC3 = 3	$P_3 < (4 - 7P_2)/6$
B3I6G5	$4/3 + (P_1 + P_3)/6$	SC1,SC3 = 3	$P_3 < 3 - 7P_2$
B3I6G6	$4/3 + (P_1 + P_2)/6$	SC1,SC2 = 3	
B3I6G7	3/2	SC1,SC2,SC3 = 3	
B3I7G1	12/7		
B3I8G1	2		

GRUPO	I_E/I_T	CONDICION PARA APC
B3I5G4	$P_2 + (P_1 + 16P_3)/36$	$P_1 < (1 - 2P_2)/2$
B3I5G5	$P_1 + (P_2 + 16P_3)/36$	
B3I5G6	$P_1 + (16P_2 + P_3)/36$	
B3I5G7	$P_1 + (P_2 + P_3)/36$	

$\{A, B, C\}$ y $\{H\}$, C y H son los individuos "solos" y la distancia de Ward entre ellos es, $W(C, H) = (P_1 + P_3)/2$. En este caso la tendencia es hacia el "árbol parecido al completo" debido a que la distancia de Ward entre C y H es grande. Sin embargo C y H se pueden unir si $P_3 > 3 - 7P_2$.

Si se le asignan a las tres variables los pesos en una relación igual a la del sistema numérico binario (el primer peso es el doble del segundo y el segundo es el doble del tercero), es decir 4/7, 2/7 y 1/7, siempre se obtiene el árbol "parecido al completo".

4.3 Clasificación con p variables binarias.

Si se tienen variables binarias $V_1, V_2, \ldots, V_p$ con pesos $P_1 > P_2 > \cdots > P_p$, se tienen $n = 2^p$ tipos de individuos posibles, identificados $A, B, D, \ldots,$. La condición de los pesos están dentro de un hipertriángulo de dimensión $P - 1$. El caso completo constituye una generalización de los caso completos de 2 y 3 variables, es decir que hay una partición en dos por la primera variable (la de mayor peso), cada una de las dos particiones se dividen a su vez en dos por la segunda variable y así sucesivamente hasta llegar a las ramas terminales de todos los individuos. Las alturas de los nodos del árbol son $P_p/2, P_{p-1}, 2P_{p-2}, 4P_{p-3}, 8P_{p-4}, \ldots, 2^{p-2}P_1$

Clasificación con tablas incompletas de p variables binarias

El conjunto de individuos a clasificar corresponde a un subconjunto de n elementos del conjunto de los 2^p tipos de individuos posibles, es decir de los 2^p combinado n subconjuntos posibles. Las particiones del conjunto, como se vio en el caso de dos y tres variables, no se pueden considerar como derivaciones directas del caso completo.

El análisis debe comenzar con la determinación del conjunto mínimo y completo de información de la tabla de datos. Primero se debe detectar y eliminar la presencia de variables que no aportan a la clasificación porque son constantes, en estos casos la suma de las columnas en la tabla de datos es 0 o n. De otra parte hay variables que se corresponden para lo cual parece suficiente un estudio por parejas. Si para todos los individuos la suma de las dos columnas es 1, es porque las dos variables se corresponden complementariamente, es decir si una toma el valor 0 la otra toma el valor 1 y viceversa. Si las sumas son siempre 0 o siempre 2 es porque las dos variables se corresponden en forma que una es la repetición de la otra. En los dos casos el aporte a la clasificación de una de las variables dada la presencia de la otra es cero. En casos donde se tienen más variables que individuos $p > n$ se pueden eliminar al menos $p - n$ variables, ya que en tal caso, al menos $p - n$ columnas son linealmente dependientes y no aportan información útil en el proceso de clasificación.

Sin comprobación analítica y como generalización de los casos analizados para dos y tres variables es de esperar que si la asignación de pesos es $P_i = 2^{i-1}/Q$, con $Q = 2^0 + 2^1 + \cdots + 2^{i-1}$, se obtengan siempre clasificaciones APC.

5. Clasificación con variables nominales

Para la identificación de los tipos de individuos se usan las mismas normas que se utilizaron en el estudio con variables binarias. De igual manera, se notan los grupos de resultados de la clasificación comenzando por la letra N (por nominal) seguida del numero de modalidades de cada una de las variables, comenzando por la primera variable y separándolas con la letra x. Finalmente le sigue la cantidad de individuos y el consecutivo del grupo.

5.1 Los casos completos $N2 \times 3$, $N3 \times 2$ y $N3 \times 3$.

Caso $N2 \times 3$

La primera variable, es decir aquella con mayor peso, tiene dos modalidades y la segunda tiene tres. Se tienen seis tipos de individuos posibles, los cuales se notan en forma disyuntiva completa así:

	$V1$	$V2$
A	10	100
B	10	010
C	10	001
D	01	100
E	01	010
F	01	001

La distancia de Bray-Curtis es para este caso:

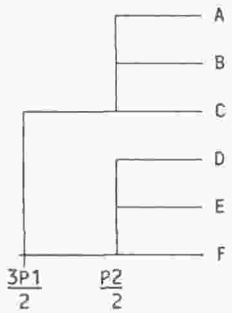
$$d(i,j) = (P_1|X_{i1}-X_{j1}+P_1|X_{i2}-X_{j2}|+P_2|X_{i3}-X_{j3}+P_2|X_{i4}-X_{j4}|+P_2|X_{i5}-X_{j5}|)/2$$

donde el segundo subíndice de X indica el número de la columna.

La matriz de distancias de Bray-Curtis es:

	A	B	C	D	E	F
A	0					
B	P_2	0				
C	P_2	P_2	0			
D	P_1	1	1	0		
E	1	P_1	1	P_2	0	
F	1	1	P_1	P_2	P_2	0

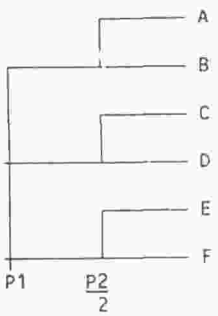
El árbol para el caso completo ($N2 \times 3I6G1$) es:



Inercia total $=3/2 + P_2/2$

Caso $N3 \times 2$

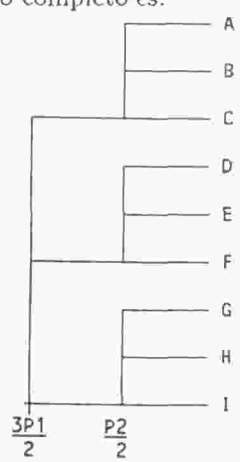
En este caso la primera variable, la de mayor peso, tiene tres modalidades y la segunda tiene dos. El árbol resultante difieren del caso anterior y es el siguiente:



Inercia total $=3/2 + P_1/2$

Caso $N3 \times 3$

Aquí las dos variables tienen cada una tres modalidades y se generan nueve tipos de individuos. El árbol del caso completo es:



Inercia total $= 3$

5.2 Los casos incompletos.

En el cuadro 2 se presenta un resumen de los casos posibles con dos variables nominales de dos y tres modalidades, que no son reducibles.

Con tres tipos de individuos en los casos $N2 \times 3$ y $N3 \times 2$ de los 20 subconjuntos posibles solo los seis de los grupos $N2 \times 3I3G3$ y $N3 \times 2I3G3$ son nuevos. Los dos casos de los grupos $N2 \times 3I3G1$ y $N3 \times 2I3G1$ son reducibles al caso de una variable de tres modalidades y los 12 de los grupos $N2 \times 3I3G2$ y $N3 \times 2I3G2$ a casos de dos variables binarias.

Los casos de tres individuos con dos variables de tres modalidades son todos reducibles, es decir que con tres individuos no hay casos nuevos para dos variables nominales con tres modalidades frente a los casos de una variable, dos variables binarias y $N2 \times 3$ y $N3 \times 2$.

En los casos $N3 \times 3I4$ solo los grupos $N3 \times 3I4G5$ y $N3 \times 3I4G6$ no son reducibles y tienen la inercia común de 5/4.

Para la mayoría de conjuntos se obtiene la misma forma de árbol para la condición de $P_1 > P_2$, pero para algunos se pueden obtener árboles que conducen a clasificaciones diferentes dependiendo de una condición adicional de los pesos. (Grupos: $N3 \times 2I4G3$, $N3 \times 3I6G5$).

Se presentan sin embargo dos situaciones nuevas respecto a los casos con variables binarias. En el resultado notado $N3 \times 3I5G7$ no existe una condición de pesos que permita obtener un árbol APC. Segundo, hay casos en los que lo consecución de árboles APC se logra mediante un criterio diferente a los utilizados hasta ahora: el orden en que se encuentran los individuos en la tabla de datos. La influencia del orden se debe a la presencia de empates tanto en la matriz original de distancias como en las matrices intermedias, los cuales se resuelven teniendo en cuenta su posición. Si se desea que esos empates se resuelven siempre de la misma manera se deben ordenar los individuos en la tabla de datos. Para orientar, además los resultados en el sentido de tener siempre la influencia de las variables de menor peso o importancia en las últimas ramificaciones, dicho orden debe ser de menor a mayor en la codificación de individuos, es decir para culminar con el tipo A. En un problema de clasificación con n individuos, el incremento del número de variables, aumenta también el número posible de valores diferentes para las distancias entre individuos, de esta manera la posibilidad de empates disminuye y con ello el efecto del orden de los individuos.

Para el caso general de p variables nominales $V_1, V_2, \ldots, Vp$, con $q_1, q_2, \ldots, q_p$ modalidades y con pesos $P_1 > P_2 > \cdots > P_p$. El número de tipos de individuos posibles es $n = q_1 \times q_2 \times \cdots \times q_p$. La tabla de datos en la forma TDC tiene n filas

CUADRO 2. Inercia total, cantidad de conjuntos por grupo, condición para obtener árboles APC y presencia de empates.

	GRUPO	INERCIA	No.CONJ.	CON.PARA APC
EMPATE				
	$N2 \times 3I3G3$	$2/3 + P_2/3$	6	
	$N3 \times 2I3G3$	$2/3 + P_1/3$	6	
	$N2 \times 3I4G1$	$3/4 + P_2/2$	6	
	$N2 \times 3I4G3$	$1 + P_2/4$	6	
	$N3 \times 2I4G2$	$3/4 + P_1/2$	6	
	$N3 \times 2I4G3$	$1 + P_1/4$	6	$P_1 < 2/3$
	$N3 \times 3I4G6$	$5/4$	9	
	$N3 \times 3I4G7$	$5/4$	36	
	$N2 \times 3I5G1$	$6/5 + 2P_2/5$	6	ORDEN
SI				
	$N3 \times 2I5G1$	$6/5 + 2P_1/5$	6	
	$N3 \times 3I5G2$	$7/5$	9	
	$N3 \times 3I5G3$	$7/5 + P_2/5$	18	
	$N3 \times 3I5G5$	$7/5 + P_1/5$	18	ORDEN
SI				
	$N3 \times 3I5G6$	$8/5$	9	
	$N3 \times 3I5G7$	$8/5 + P_1/5$	36	NO HAY
	$N2 \times 3I6G1$	$3/2 + P_2/2$	1	
	$N3 \times 2I6G2$	$3/2 + P_1/2$	1	
	$N3 \times 3I6G3$	2	6	
	$N3 \times 3I6G4$	$7/6 + P_2/3$	36	
	$N3 \times 3I6G5$	$11/6 + P_2/6$	18	$P_1 > 3/4$
	$N3 \times 3I6G6$	$11/6 + P_1/6$	18	
	$N3 \times 3I7G1$	$9/7$	9	ORDEN
SI				
	$N3 \times 3I7G2$	$15/7 + P_1/7$	9	
	$N3 \times 3I7G3$	$57/28 + P_1/4$	12	
	$N3 \times 3I8G1$	$21/8$	9	ORDEN
SI				
	$N3 \times 3I9G1$	3	1	

por Q columnas (número total de modalidades). Los árboles completos producen particiones que se obtienen directamente con ordenamiento numérico. Para los casos incompletos no es posible generalizar resultados que permitan señalar los casos en los que es posible derivar clasificaciones a partir de árboles completos. Como se anota en el cuadro 2 en el caso $N2 \times 3$ para obtener un APC se exige que P_1 sea menor que $2/3$, pero en el caso $N3 \times 3$ para obtener lo mismo se tiene como condición que p_1 sea mayor que $3/4$.

Bibliografía

1. Anderberg, M, *CLUSTER ANALYSIS FOR APPLICATIONS*, Academic Press, Londres, 1973.
2. Bautista, L. y Ramos, J., *ANALISIS DE DATOS DE ENCUESTAS Y TABULADOS*, Universidad Nacional, Bogotá, 1988.
3. Cailliéz F. y Pagés, J.P., *INTRODUCTION A L'ANALYSE DES DONEES*, Smash, Paris, 1976.
4. Diday et al., *DATA ANALYSIS AND INFORMATICS.*, Proceedings of the Second International Symposium on Data Analysis and informatics. Versailles, octubre 17-19 1979, North Holland, Amsterdan, 1980.
5. Diday et al., *DATA ANALYSIS AND INFORMATICS III*, Proceedings of the Third International Symposium on Data Analysis and Informatics. Versailles, octubre 4-7 1983, North Holland, Amsterdam, 1984.
6. Hartigan, John, *CLUSTERING ALGORITHMS*, Wiley, New York, 1975.
7. Hayashi, C. et al., *RECENT DEVELOPMENTS IN CLUSTER AND DATA ANALYSIS*, Proceedings of the Japanese-French Scientific Seminar. March 24-26 1987, Academic Press, San Diego, 1987.
8. Jambu, M. y Lebeaux, M., *CLUSTER ANALYSIS AND DATA ANALYSIS*, North Holland, Amsterdam, 1983.
9. Lebart, L., Morineau, A. y Warwick, K., *MULTIVARIATE DESCRIPTIVE STATISTICAL ANALYSIS*, John Wiley, New York, 1984.
10. Lerman, I.C., *CLASSIFICATION ET ANALYSE ORDINALE DES DONEES*, Dunod, Paris, 1981.
11. Rohlf, J., *NTSYS-PC. Numerical taxonomy and multivariate analysis system. Version 1.4.*, Exter Publishing, New York, 1988.
12. Sneath, P. y Sokal, R., *NUMERICAL TAXONOMY. The principles and practice of numerical classification*, 1973.
13. Sokal, R., *PHENETIC TAXONOMY. Theory and methods*, Ann. Rev.Ecol. 17 (1986), 423-42.
14. Volle, Michel, *ANALYSE DES DONEES. 3 Ed. Económica*, París, 1985.
15. Ward, Joe, *HIERARCHICAL GROUPING TO OPTIMIZE AN OBJETIVE FUNCTION*, American Statistical Association Journal (1963).
16. Wishart, David, *AN ALGORITHM FOR HIERARCHICAL CLASSIFICATIONS*, Biometrics (1969).