Una aplicación de la Teoría de la Información al análisis de datos definidos mediante variables cualitativas multi-estado: medidas de similaridad y análisis cluster.

José A. Esquivel Guerrero

Francisco Contreras Cortés

Fernando Molina González

Josefa Capel Martínez

Departamento de Prehistoria.
Universidad de Granada. 18071 Granada.

1. Introducción

El Análisis Cluster constituye una importante técnica de análisis de datos, ampliamente utilizada en distintas áreas de conocimiento (Biología, Psicología, Arqueología, Sociología, etc.) con el propósito de identificar entidades similares a partir de las características que poseen. En el campo de la Arqueología, la necesidad de clasificar los items arqueológicos y construir tipologías precisas conduce a una utilización, cada vez más amplia, de este tipo de técnicas estadísticas, solventando determinados problemas que aparecen en la arqueología tradicional:

- — manejo de grandes cantidades de datos que, debido a su dimensionalidad, son difíciles de estudiar a menos que puedan clasificarse en grupos manejables con la mínima pérdida de información.
- — necesidad de disponer de un método de agrupación útil y nítido, que introduzca un grado de objetividad no obtenible por observación directa.
- — utilización simultánea de varias características a lo largo del proceso para evitar soluciones descriptivas basadas, en general, en una única característica diferenciadora.

El análisis está constituido por dos procesos fundamentales: la obtención de una medida de similaridad adecuada a las características de los objetos, y un algoritmo que consiga la agrupación de los objetos en clusters (grupos) con gran homogeneidad interna y alta heterogeneidad externa.

Las técnicas basadas en variables cuantitativas han sido ampliamente estudiadas, utilizando los métodos y resultados de la Geometría Euclídea para obtener medidas de similaridad (generalmente a partir de distancias) y algoritmos de agrupación (un estudio detallado aparece en SNEATH & SOKAL, 1973; DUDA & HART, 1973; DIDAY & SIMON, 1976, y EVERITT, 1980). Sin embargo, los datos definidos mediante variables binarias o variables cualitativas multiestado presentan mayores problemas, ya que no existen modelos geométricos adecuados. Los modelos binarios más comunes aparecen codificados en tablas presencia/ausencia, 1/0, etc., y se han desarrollado varios coeficientes de similaridad para los mismos (Jacquard-Sneath, Lance y Williams, Sokal y Michener, Rogers-Tanimoto, Yule, etc.) (en SNEATH & SOKAL, 1973, y DUDA & HART, 1973, aparece una revisión completa de este tipo de coeficientes). Las variables multiestado originan, debido a su carácter, mayores problemas que las anteriores, y con frecuencia se

José A. Esquivel, Francisco Contreras Cortés, Fernando Molina González y Josefa Capel Martínez
Complutum, 1 (1991), Madrid (pp. 53-64).

han estudiado considerando cada estado de cada variable como una variable dicotómica (KENDALL, 1975, y ROMESBURG, 1984), aunque el interés del problema en una amplia diversidad de campos (reconocimiento de imágenes, reconocimiento de cadenas de símbolos, secuencias de fonemas, estudio de documentos, organización de bases de datos, etc.) ha suscitado que sea abordado desde distintas teorías (ESQUIVEL, 1988; PAL & MAJUMDER, 1985; BACKER & JAIN, 1981; MICHALSKI & STEPP, 1983; ITO, KODAMA & TOYODA, 1984; RAO, 1984; BEN-BASSAT & ZAIDENBERG, 1984; CHIU & WONG, 1986; WONG & CHIU, 1987).

En este trabajo se proponen varias medidas de similaridad entre objetos definidos mediante variables cualitativas multiestado, a partir de métodos y técnicas de la Teoría de la Información (este tema ha sido objeto de la Tesis Doctoral de uno de nosotros, J.A.E.) (ESQUIVEL, 1988). Además, se desarrolla un algoritmo de clustering basado en dichas medidas de similaridad, aplicándolo a un conjunto de 50 vasos cerámicos extraídos del yacimiento La Cuesta del Negro, Purullena (Granada) de la Edad del Bronce.

2. Incertidumbre y Entropía

El estudio de una distribución de objetos definidos mediante variables cualitativas multiestado exige obtener la máxima información de cada objeto, de cada variable y de cada estado, en función de la frecuencia de aparición de los estados, número de estados de las variables, etc. Términos iguales a rareza, abundancia y otras nociones intuitivas deben reflejarse de forma clara y precisa.

De acuerdo a la teoría de Shannon, en un modelo matemático de comunicación la información vendrá determinada por un parámetro estadístico asociado a un esquema de probabilidad y «debe indicar una medida relativa a la incertidumbre de acuerdo a la ocurrencia de un mensaje particular en el conjunto de mensajes» (REZZA, 1961, y SHANNON, 1948). En la axiomática clásica, la incertidumbre asociada a un suceso E_k perteneciente a un conjunto de sucesos $\Omega = \{E_1, ..., E_n\}$ viene determinada por el valor

$$- \lg_2 p_k$$

siendo p_k la probabilidad de ocurrencia del suceso E_k. Y la media extendida a todos los sucesos de una distribución de sucesos viene determinada por la *entropía* de Shannon y Weaver

$$H(X) = \sum_{1}^{n} - p_i \lg_2 p_i, \qquad \sum_{1}^{n} p_i = 1,$$

que mide la incertidumbre media asociada a un esquema finito y completo de probabilidad, aun cuando varios autores han sugerido otras definiciones de entropía que no verifican algunas de las condiciones de la entropía clásica (REZZA, 1961) y, modernamente, se han realizado diversas generalizaciones de la entropía (RAO, 1984).

Al considerar un «espacio» de unidades definidas mediante variables cualitativas multiestado, la entropía debe tener en cuenta la incertidumbre de los estados de cada variable, el número de estados de las variables y la frecuencia de aparición de los mismos. En este trabajo se propone una medida de entropía enfocada al estudio de este tipo de variables, teniendo en cuenta las consideraciones anteriores según:

- la incertidumbre de un estado muy frecuente debe ser pequeña, ya que la probabilidad de que dicho estado aparezca en una unidad escogida al azar es grande; recíprocamente, si un estado es raro su contribución a la entropía debe ser grande.
- la incertidumbre asociada a una variable será mayor cuanto menor sea el número de sus estados, pues la dicotomía que produce en la distribución es mayor que si tuviese muchos estados.

3. Medidas de información

Una medida de información que verifique las anteriores consideraciones se define cómo:

La información asociada al estado x_{ik} con probabilidad p_{ik} es

$$I(x_{ik}) = -\frac{1}{n_i} \lg_2 p\,(x_{ik}) \geq 0,$$

siendo n_i el número de estados de la variable X_i. Esta medida se ajusta a la axiomática de Shannon, al ser solución de la ecuación

$$f(1/n) + f(1/m) = f(1/mn).$$

La incertidumbre media (entropía) asociada a la variable Xi viene entonces determinada por

$$H_{X_i} = \overline{I(x_{ik})} = -\frac{1}{n_i} \sum_{k=1}^{n_i} p(x_{ik}) \lg_2 p(x_{ik}) > 0$$

$$\sum_{k=1}^{n_i} p(x_{ik}) = 1,$$

que verifica las condiciones exigidas a las medidas de incertidumbre en la Teoría de la Información.

La entropía así definida tiende a suavizar la influencia de los estados extremos (con frecuencia muy pequeña o muy grande). Sin embargo, la influencia de estos estados es fundamental en el estudio de la asociación que pueda existir entre las unidades ya que la coincidencia de dos unidades en un determinado estado debe valorarse en función de la información completa que aporte dicho estado, esto es:

- la significación de una coincidencia de unidades en un estado poco frecuente debe ser mayor que si coinciden en un estado más frecuente, puesto que «... el acuerdo en estados raros es menos probable que el acuerdo entre estados frecuentes y debe ser más valorado» (SNEATH y SOKAL, 1973).
- es menos significativa una coincidencia en un estado de una variable con muchos estados que si el *número de estados de la variable es escaso.*

Estas consideraciones llevan a la definición de entropía total o «distorsión» de una variable (ESQUIVEL, 1988):

$$D(X_i) = -\frac{1}{n_i}\sum_{k=1}^{n_i} \lg_2 p(x_{ik}),$$

$$\sum_{k=1}^{n_i} p(x_{ik}) = 1,$$

que refleja la influencia que produce cada estado en el espacio de unidades y en qué forma queda afectada la homogeneidad de dicho espacio, en función de la información que aporta cada una de las unidades.

4. Incertidumbre de una unidad

Los elementos del «espacio» aportan su propia incertidumbre en función de las características que los constituyen, modificando la estructura del espacio, puesto que la introducción o eliminación de un elemento trae consigo una modificación en los parámetros que definen las características estructurales del espacio (frecuencia de los estados, desaparición de algún estado, etc.).

Con estas premisas, sea el conjunto de elementos (unidades) $\Gamma = \{A_1, A_2, ..., A_n\}$ definido sobre el conjunto de variables multiestado $V = \{X_1, X_2, ..., X_v\}$, donde cada variable X_i tiene asociado un conjunto de estados $W_i = \{x_{i1}, x_{i2}, ..., x_{i,n(i)}\}$, siendo n(i) (a veces la notación n_i es menos cómoda, como en el caso anterior) el número de estados de la variable X_i. A cada unidad A_i se le asocia el objeto matemático definido por la n-tupla (DUBOIS y PRADE, 1980)

$$m(A_i) = (m_1(A_i), m_2(A_i). ..., m_v(A_i)),$$

siendo m_k el procedimiento de medida asociado a la variable X_k y $m_k(A_i)$ el estado que toma la unidad A_i en la variable X_k, $i = 1, ..., n$ y $k = 1, ..., v$, esto es, $m_k(A_i) = x_{kj}$ si j es el índice del estado de la variable X_k que aparece en A_i. El conjunto de objetos matemáticos correspondiente a una distribución de unidades se denomina espacio de patrones *(pattern space)* S o espacio total y, aunque la diferencia entre una unidad A y su objeto matemático asociado es evidente, por simplicidad se denotarán de igual forma excepto cuando sea necesario llevar a cabo dicha distinción.

Con la anterior notación, $p_i(A)$ es el valor de la probabilidad (frecuencia relativa, o probabilidad en un diseño probabilístico) del estado x_{ij} si la unidad A posee dicho estado en la variable X_i, es decir

$$p_i(A) = p(x_{ij}) \quad \text{si} \quad m_i(A) = x_{ij},\ 1 \le j \le n_i$$

De aquí que la distorsión (denominada campo) *producida por una unidad se define como la incertidumbre* total que dicha unidad produce en el espacio de unidades

$$F(A) = -\sum_{i=1}^{v} \frac{1}{n_i} \lg_2 p_i(A) \ge 0,\ A \varepsilon \Gamma.$$

La distribución de unidades se comporta entonces de forma similar a un campo de fuerzas en equilibrio dinámico, y cualquier modificación (en las unidades, estados o variables) produce un reajuste en los valores de los campos de las unidades, modificando la estructura de la distribución.

A partir de esta medida puede definirse la distorsión o campo producido por un grupo de unidades, que debe reflejar tanto la atracción existente entre unidades semejantes como la repulsión (diversidad) entre unidades no semejantes (o escasamente semejantes), en función de qué variables tienen estados comunes en el grupo (y en qué medida) y cuáles los tienen distintos.

Estas ideas tienen un punto de partida en dos nociones matemáticas que axiomatizan las ideas intuitivas.

5. Unión e intersección de unidades

— *Intersección*

Dadas las unidades A_i, A_j $\varepsilon\Gamma$, i, j = 1, ..., p, la intersección entre ellas está definida por su parte común, es decir,

$$A_i \cap A_j = \{a_{kh}/m_k(A_i) = m_k(A_j) = a_{kh}\},$$

$$k = 1, ..., v, h = 1, ..., n_k.$$

Intuitivamente, la intersecccción está constituida por un objeto matemático (en general no será una unidad, ya que puede no contener todas las variables) caracterizado por los estados de las variables comunes a ambas unidades.

La información común a ambas unidades es entonces

$$F(A_i \cap A_j) = -\sum_{k=1}^{v} \frac{1}{n_k} \lg_2 p_k(A_i) \geq$$

$$\geq 0, \text{ si } m_k(A_i) = m_k(A_j)$$

Naturalmente, esta formulación es equivalente a

$$F(A_i \cap A_j) = -\sum_{k=1}^{v} \frac{1}{n_k} \lg_2 p_k(A_j) \geq$$

$$\geq 0, \text{ si } m_k(A_i) = m_k(A_j)$$

— *Unión*

Dadas las unidades A_i, A_j $\varepsilon\Gamma$, i, j = 1, ..., p, la intersección entre ellas está definida por los estados que aportan algunas de las unidades, es decir,

$$A_i \cup A_j = \{a_{kh}/a_{kh} = m_k(A_i) \text{ o } a_{kh} = m_k(A_j)\},$$

$$k = 1, ..., v, h = 1, ..., n_k.$$

Intuitivamente, la unión de dos unidades está constituida por un objeto matemático (en general no será una unidad, ya que en cada variable puede tomar más de un estado) caracterizado por los estados que aparecen en alguna de las unidades.

La definición de unión de dos unidades permite obtener la información conjunta a ambas unidades según:

$$F(A_i \cup A_j) = -\sum_{k=1}^{v} \frac{1}{n_k} \lg_2 p_k(A_i) -$$

$$-\sum_{k=1}^{v} \frac{1}{n_k} \lg_2 p_k(A_j) \; (m_k(A_i) = m_k(A_j))$$

que mide la distorsión aportada por los estados que aparecen en algunas de ambas unidades, eliminando los estados repetidos (ESQUIVEL, 1988).

Estas dos medidas verifican la relación fundamental $F(A \cup B) = F(A) + F(B) - F(A \cap B)$, enunciada por Pal y Majumder, (PAL & DUTTA MAJUMDER, 1985) en el contexto de medir el grado de ambigüedad en un conjunto.

Esta propiedad puede generalizarse al cálculo del campo conjunto de varias unidades en función de los campos individuales y de las intersecciones múltiples entre ellas (dos a dos, tres a tres, etc.). La computación de la información conjunta proporcionada por los elementos de un grupo G_n formado por los elementos $\{A_1, A_2, ..., A_n\}$ será entonces

$$F(G_n) = F\left(\bigcup_{k=1}^{n} A_i\right)$$

que incluye tanto la similitud entre los elementos del grupo como las diferencias existentes entre ellos (en ESQUIVEL, 1988, se establece una axiomática completa).

6. Afinidad entre grupos

La noción de información conjunta asociada a un grupo contiene tanto la similitud como la disimilitud entre sus elementos, reflejando la estructura subyacente al grupo en base a los estados que los configuran. Estas propiedades estructurales del grupo pueden reflejarse a partir de los valores de la afinidad entre un elemento y el grupo o entre dos grupos, y debe verificar (BACKER & JAIN, 1981; PAL & MAJUMDER, 1985):

(i) La afinidad entre un elemento y un grupo no debe ser menor si el elemento es un miembro del grupo que si no está contenido en el grupo.
(ii) La afinidad será aproximadamente 0 si el elemento es muy extraño respecto al grupo («si el elemento está distante del grupo o fuera de la región de interés», sic).
(iii) La afinidad será igual a un máximo absoluto si el grupo consiste en un único elemento que tenga la misma localización que el elemento bajo consideración.

Una caracterización intuitiva de la noción de afinidad entre dos elementos es:

$$A_t(A_1, A_2) = F(A_1 \cap A_2),$$

pero la extensión a afinidad unidad-grupo o grupo-grupo debe tener en cuenta, respectivamente, la atracción que se ejerce entre dicha unidad y los elementos del grupo, y la atracción mutua ejercida por los elementos de los dos grupos:

1. La afinidad elemento-grupo se define como

$$A_t(A,G) = F(G \cap A), \text{ o } A_t(A, G) = F\left(\bigcup_{k=1}^{n} A_i\right)$$

si $G = \{A_1, A_2, ..., A_n\}$.

2. La afinidad grupo-grupo se define como

$$A_t(G, G') = F(G \cap G')$$

que intuitivamente es una medida de la información común a G y G', computando las conexiones entre G y G'.

Ambas definiciones verifican las propiedades de Backer y Jain (ESQUIVEL, 1988).

7. Medidas de similaridad

Las técnicas de la Teoría de Conjuntos inducen una medida conjuntista de similaridad en la forma

$$r(A_i\, A_j) = \frac{F(A_i) + F(A_j) - F(A_i \cup A_j)}{F(A_i \cup A_j)} \quad A_i, A_j \in \Gamma,$$

que equivale a

$$r(A_i, A_j) = \frac{F(A_i \cap A_j)}{F(A_i \cup A_j)} = \frac{A_t(A_i, A_j)}{F(A_i \cup A_j)}$$

basada en la medida de similaridad establecida por ITO, KODAMA & TOYODA (1984), que, basada en la teoría de conjuntos, combina unión e intersección para variables no-independientes.

Esta medida está referida a dos unidades pero, al extenderla a similaridad entre grupos, surgen dos posibilidades en función de la intersección que se prefiera, puesto que con solamente dos unidades se tiene que

$$A_t(A_i, A_j) = F(A_i \cap G_j),$$

Exísten dos posibles extensiones:

- — extensión de $A_i \cap A_j$ como intersección de todos los elementos del grupo, denominada intersección fuerte y denotada por $A_i \wedge A_j$.
- — extensión de $A_i \cap A_j$ en la forma $A_t(A_i, A_j)$.

De acuerdo a la posibilidad que se adopte, resultarán dos medidas básicas de similaridad (ESQUIVEL, 1988).

Similaridad fuerte S_1

Esta medida (fig. 1) considera la intersección fuerte de los elementos de los grupos en la forma: dados los grupos $G_1 = \{A_1, A_2, ..., A_n\}$ y $G_2 = \{B_1, B_2, ..., B_p\}$, la intersección fuerte de G_1 y G_2 se define como

$$G_1 \wedge G_2 = \left(\bigcap_{i=1}^{n} A_i\right) \cap \left(\bigcap_{j=1}^{p} B_j\right)$$

La similaridad S_1 es entonces:

$$S_1(G_1, G_2) = \frac{F(G_1 \cap G_2)}{F(G_1 \cup G_2)}, \quad 0 \le S_1 \le 1.$$

Esta medida verifica de forma estricta las propiedades de Backer y Jain, y Pal y Majumder.

Similaridad-afinidad S_2

La medida S_1 computa solamente la incertidumbre proporcionada por aquellos estados de las variables que aparecen en todos y cada uno de los elementos de los grupos, y basta que un estado no aparezca en un elemento para que sea tomado en cuenta, es decir, es bastante estricta. Tomando como base la afinidad y, por tanto, incluyendo todas las relaciones entre los miembros de los grupos, se define la medida S_2 (fig. 2) en la forma:

$$S_2(G_1, G_2) = \frac{A_t(G_1, G_2)}{F(G_1 \cup G_2)}$$

que se inspira en la medida de similaridad de ITO, KODAMA Y TOYODA (1984), puesto que equivale a

$$S_2(G_1, G_2) = \frac{F(G_1) + F(G_2) - F(G_1 \cup G_2)}{F(G_1 \cup G_2)}$$

Por construcción, la medida S_2 no verifica las propiedades de Backer y Jain, y Pal y Majumder, debido a que toma en consideración todas las coocurrencias entre elementos de G_1 y G_2, y estas coocurrencias añaden su efecto a la incertidumbre. Estos problemas inducen a considerar otras medidas de similaridad derivadas de S_2, que se inspiran en distintos contextos y toman en cuenta el número de elementos coincidentes en los grupos.

Similaridad S_3

Inspirada en la definición de entropía de Kauffman (en PAL y MAJUMDER, 1985), considera la

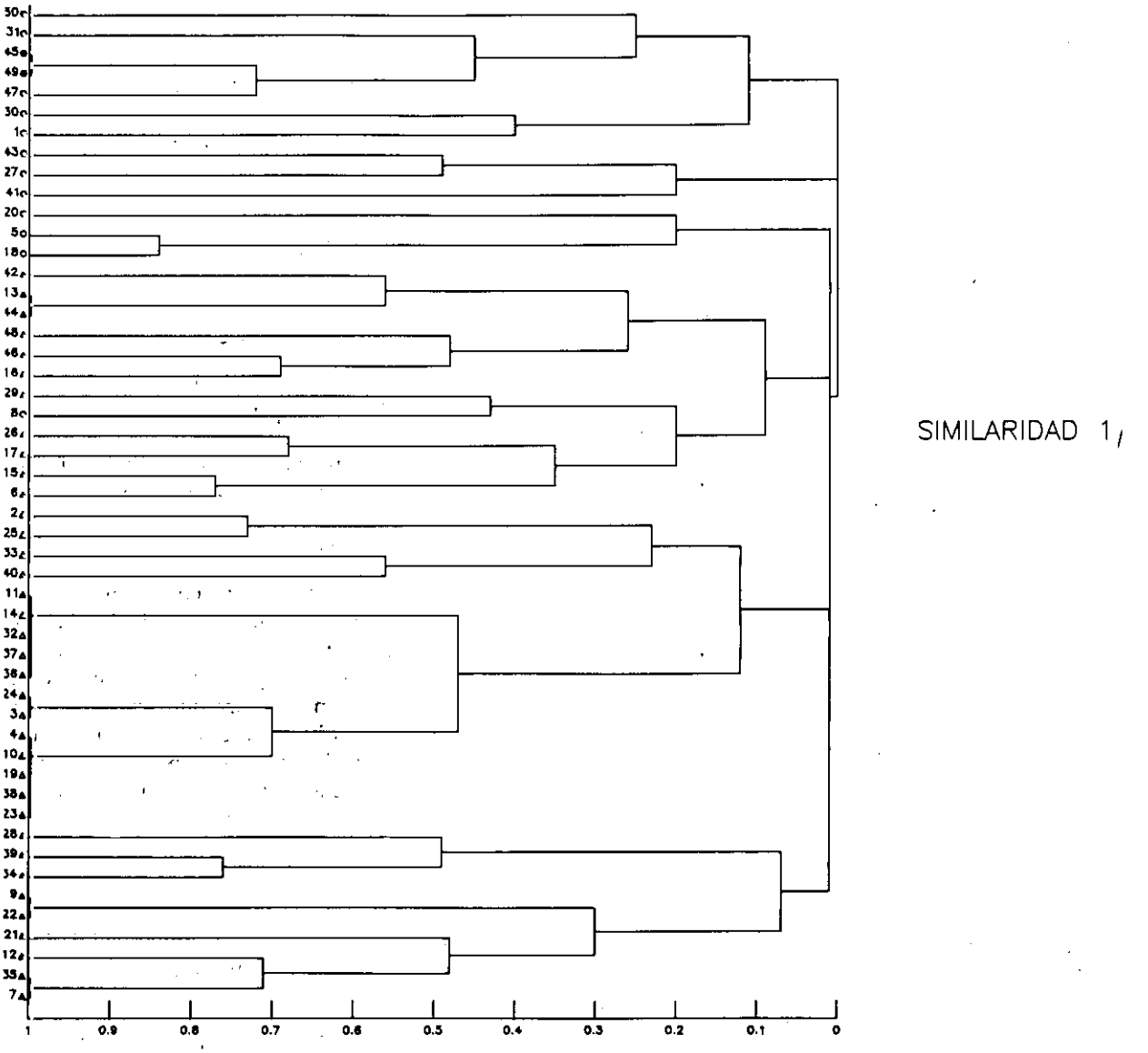

Fig. 1.—*Dendograma obtenido con la medida de similaridad* S_1.

media logaritmica de la afinidad relativa al campo conjunto de los grupos según:

$$S_3 (G_1, G_2) = \frac{\lg_2 n_e}{\lg_2 n} \frac{A_t (G_1, G_2)}{F (G_1 \cup G_2)}, \text{ siendo}$$

$$n_e = \begin{cases} 1, \text{ si } Vx, Vy \varepsilon G_1 \cup G_2, x = y \\ n, \text{ si } x, y \varepsilon G_1 \cup G_2, x = y, \ n = N (G_1 \cup G_2) \end{cases}$$

El término $\lg_2 n$ mantiene el valor máximo de S_3 comprendido entre 0 y 1, y evita una influencia excesiva del número de elementos en la similaridad (fig. 3).

Similaridad S_4

Tomando como punto de partida la entropía de De Luca y Termini (DE LUCA y TERMINI, 1972), que es una extensión de la información de Shannon considerando n fuentes binarias, la medida S_4 (fig. 4) se define como:

$$S_4 (G_1, G_2) = \frac{n_e}{n} \frac{A_t (G_1, G_2)}{F (G_1 \cup G_2)}$$

Varias medidas derivadas han sido utilizadas, sin el término n_e, en distintos contextos: XIE y BEDROSIAN (1984) aplican estos conceptos en los

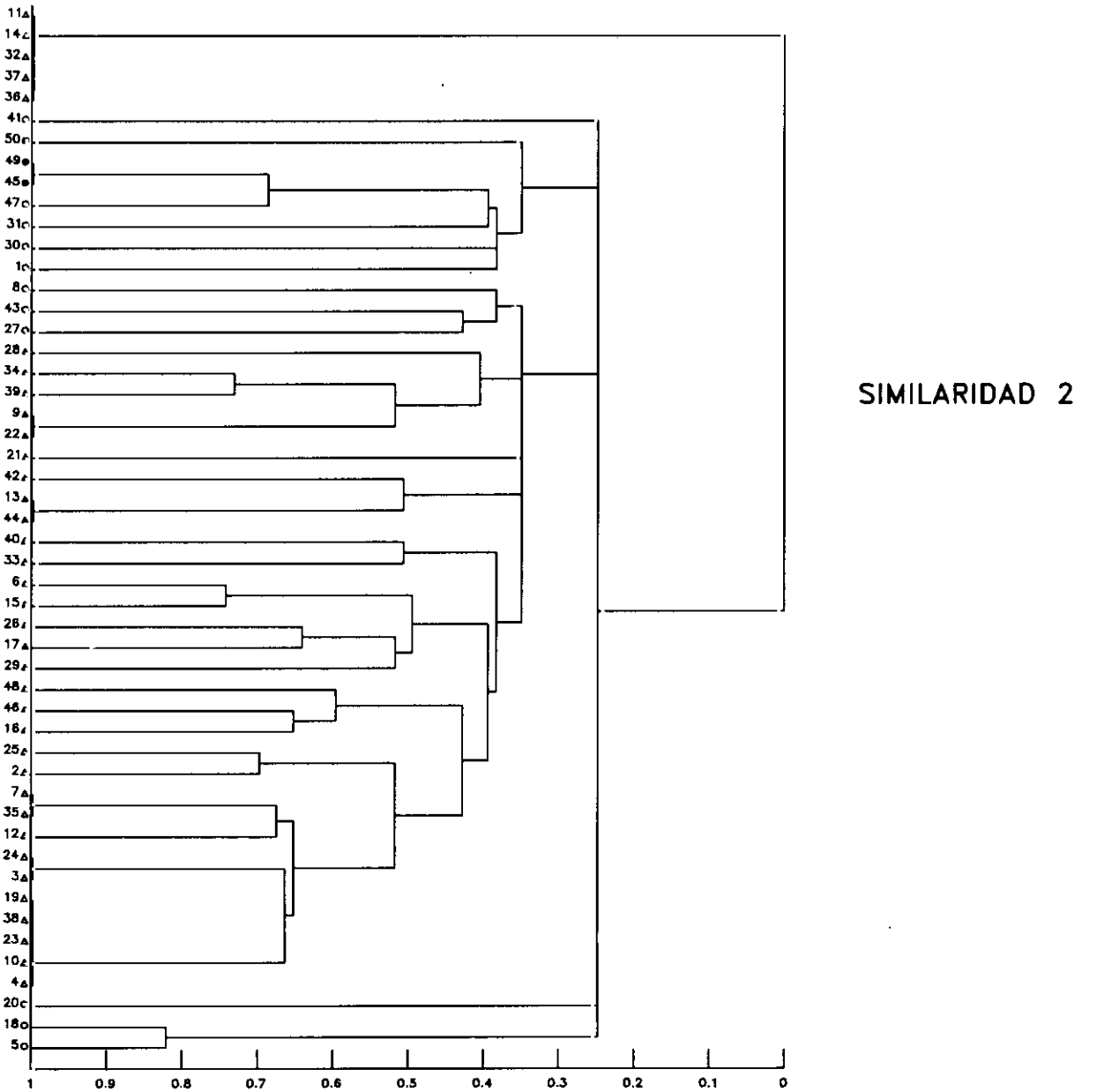

Fig. 2.—*Dendograma obtenido con la medida de similaridad S2.*

campos de tratamiento de imágenes y Termodinámica Estadística; PAL y CHAKARBORTY (1986) definen un índice de evaluación de patrones mediante medidas interclases e intraclases.

8. Algoritmo de agrupación

Los índices $S_1 - S_4$ permiten desarrollar un algoritmo de agrupación jerárquico y aglomerativo en el que, en cada nivel, se fusionan los grupos con mayor similaridad para, en un segundo paso, actualizar las similaridades teniendo en cuenta los parámetros del nuevo grupo creado.

El esquema del algoritmo consiste en (ESQUIVEL, 1988):

1. Cálculo de las probabilidades (frecuencias) p(x) para todos los estados x de todas las variables.
2. N= p (el número inicial de clusters N coincide con el número de elementos p).

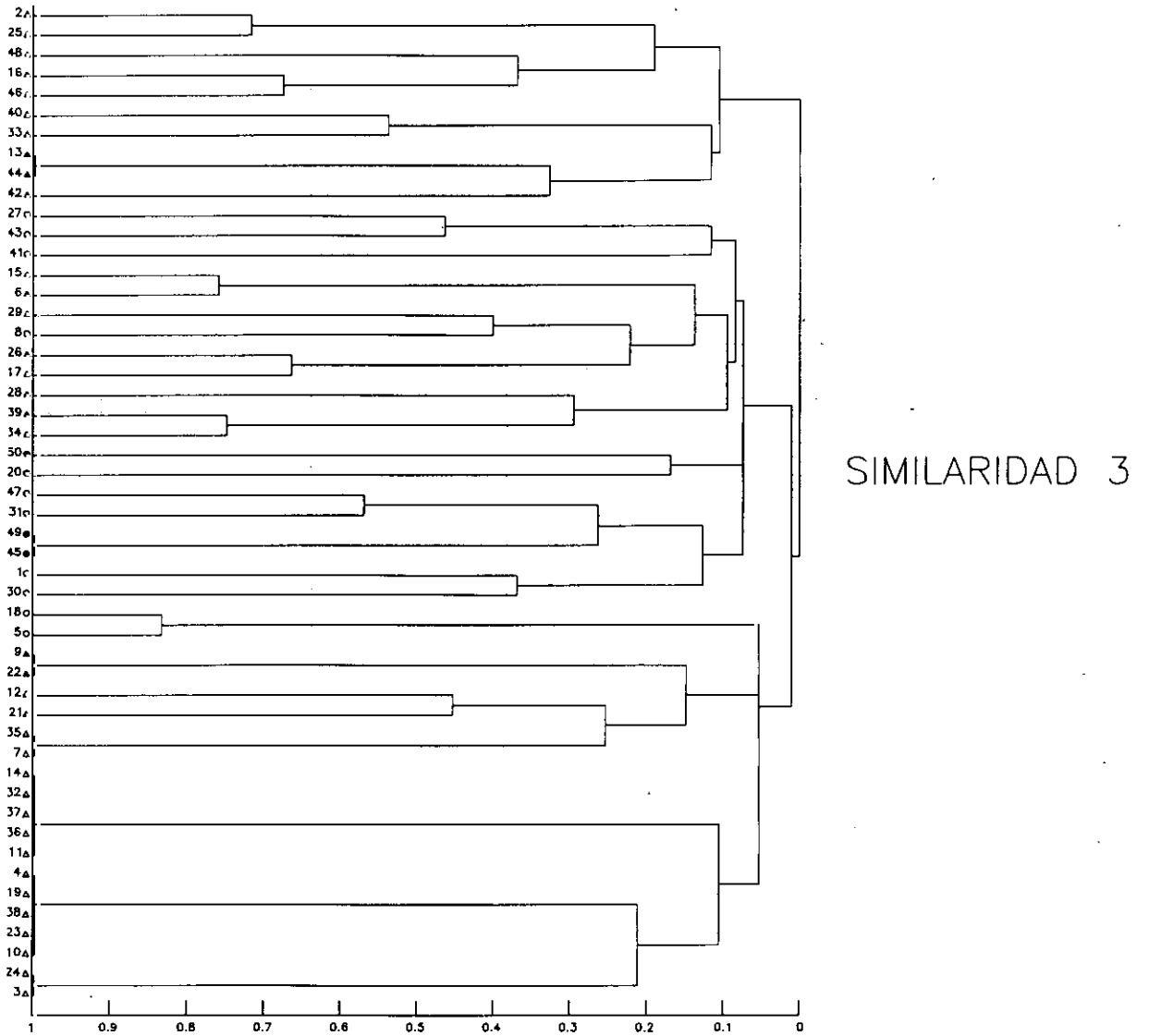

Fig. 3.—*Dendograma obtenido con la medida de similaridad* S_3.

3. $C_i = \{i\}$ (cada cluster C_i solamente contiene al elemento i al comienzo del proceso).

4. Cálculo de $S(C_i, C_j)$ para todos los grupos C_i, C_j, mediante $S_1 - S_4$.

5. Cálculo de los valores i, j tales que $S^* = S(C_i, C_j)$ sea máxima.

6. Si $S^* = 0$ ó $N = 1$ ó $N = N_0$ entonces finalizar. N_0 es el número prefijado de clusters (opcional).

7. $C_i = C_i \cup C_j$; $N = N - 1$; $C_j = \{0\}$.

8. Para todo $k = i$, si $C_k = \{0\}$ calcular $S(C_i, C_k)$.

9. Calcular $F(C_i)$.

10. Repetir el proceso desde el paso 5.

El algoritmo requiere actualizar, en cada paso, las afinidades existentes entre el nuevo grupo y los grupos restantes, incluyendo el campo del nuevo grupo. Este proceso se realiza en los pasos 7-9, y es complicado computacionalmente debido al gran número de intersecciones múltiples que aparecen en su desarrollo. Sin embargo, existen métodos numéricos alternativos que computan directamente A_t a partir de las probabilidades originales de los estados de la distribución (ESQUIVEL, 1988).

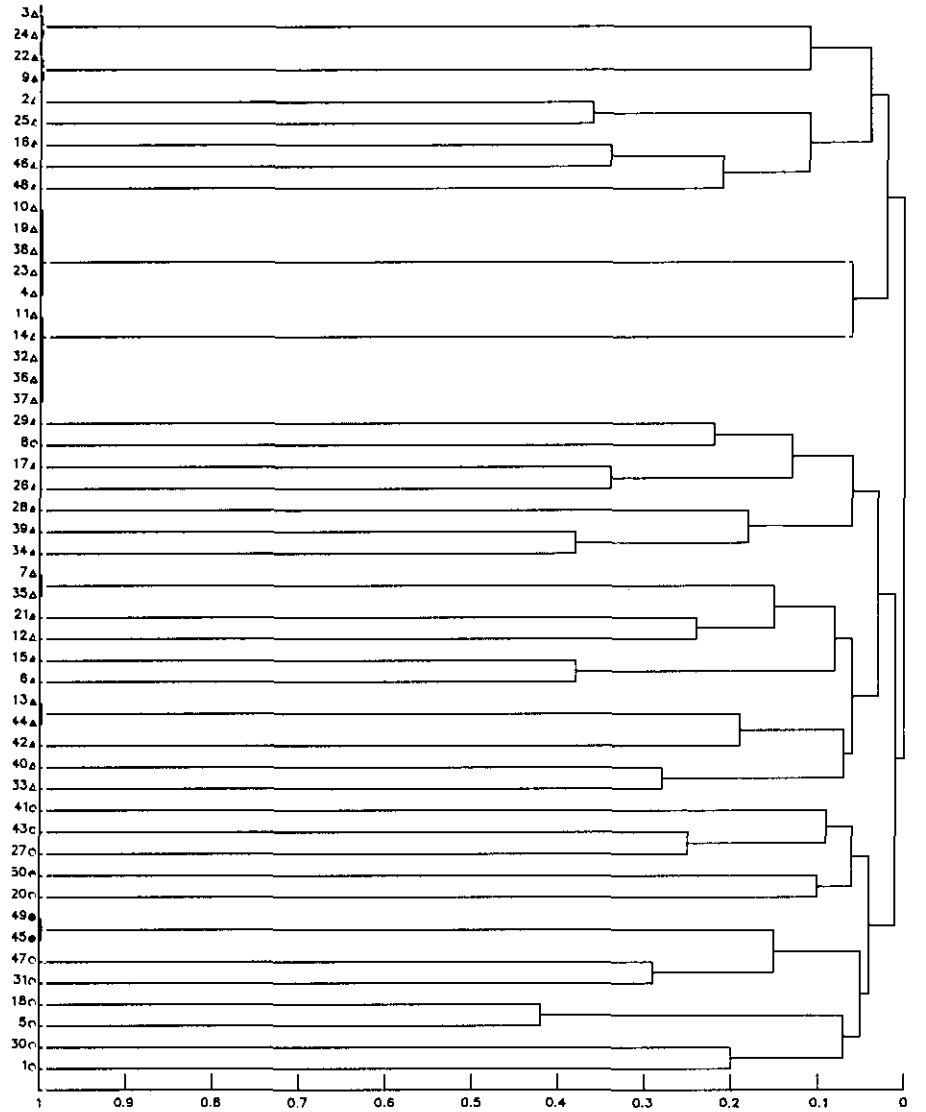

Fig. 4.—*Dendograma obtenido con la medida de similaridad* S_4.

9. Resultados experimentales

Los experimentos computacionales para evaluar las medidas de similaridad y el algoritmo de agrupación se han aplicado a un conjunto de 50 vasos cerámicos extraídos del yacimiento La Cuesta del Negro, en Purullena (Granada), de la Edad del Bronce. Los datos se han definido a partir de 8 variables nominales de tipo tecnológico, arqueológicamente relevantes:

1. Tratamiento de la superficie.
2. Color de la superficie.
3. Color de la pasta.
4. Temperatura de cocción.
5. Matriz.
6. mineralogía.
7. densidad.
8. Tipo de desgrasante.

Para definir los atributos tecnológicos más apropiados para el análisis multivariante hemos utilizado dos tipos de métodos. Por un lado, métodos de observación directa de los artefactos (lupa binocular, tablas de colores...). Con este tipo de procedimiento hemos analizado el tratamiento de las superficies, el color de la pasta y el color de las paredes. Por otro lado, hemos recurrido a métodos ana-

	VARIABLES							
	1	2	3	4	5	6	7	8
Número de estado de las variables	3	4	4	3	3	3	3	3

ESTADOS	VARIABLES							
	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3
4		4	4					

ESTADOS	VARIABLES							
	1	2	3	4	5	6	7	8
1	alisado	beige marrón	gris claro	500°	compacta	+ filo silíc.	1.35 1.52	fino
2	pulido	gris medio	marrón grisáceo	600°	magra	valor medio	1.53 1.69	medio
3	bruñido	marrón grisáceo	rojizo	650°	muy magra	+ cuarz.	1.70 1.86	grueso
4		gris oscuro	gris oscuro					

Tabla 1.—*Definición de las variables y sus estados.*

líticos de laboratorio más complejos para poder computar una serie de atributos que requieren un examen físico-químico o petrológico (difracción de Rayos X, estudio óptico, etc.) que nos han permitido analizar la composición mineralógica de la arcilla, su matriz y contenido en desgrasante, la temperatura de cocción y la densidad. Estos análisis se han realizado en la Estación Experimental del Zaidín (CSIC) de Granada, bajo la dirección de J. Capel, J. Linares y F. Huertas.

Las variables tienen consideración nominal, es decir, no han sido categorizadas, por lo que los símbolos asignados a cada uno de los estados en cada una de ellas no tienen significación alguna (tablas 1 y 2).

En la tabla de los elementos (tabla 2) aparecen los estados correspondientes a cada uno de ellos en las distintas variables. La segunda columna contiene un símbolo, que no se tiene en cuenta en el análisis aunque aparece en el dendrograma, y que proporciona una clave previa introducida por el investigador con un determinado fin: una clasificación previa dictada por la experiencia del investigador, una clave que indique alguna característica de las unidades (cuenco carenado, olla, etc.). Esta clave puede omitirse puesto que el análisis no la toma en cuenta para realizar la agrupación.

Cuando se aplica el algoritmo de agrupación utilizando las distintas medidas se encuentran algunas diferencias debidas a la distinta naturaleza de dichas medidas, pero los resultados son bastante consistentes. Si bien, una vez analizadas las cuatro medidas de similaridad podemos concluir que la medida 2 ofrece unos mejores resultados arqueológicos en este caso concreto, distinguiéndose cuatro grandes grupos de vasos cerámicos. Cada uno de ellos presenta características tecnológicas distintas. Su aparición como ajuar funerario no es arbitraria, sino que cada grupo aparece asociado a distintos tipos de contextos funerarios. Para la discusión ar-

		VARIABLES							
N.º	Si.	1	2	3	4	5	6	7	8
1	C	2	1	1	3	2	3	2	2
2	A	3	4	2	1	1	2	3	1
3	A	3	3	3	1	1	3	3	1
4	A	3	3	2	1	1	3	3	1
5	C	3	2	2	3	3	2	2	2
6	B	3	3	3	2	2	2	3	2
7	A	3	3	2	1	1	2	3	1
8	C	3	3	2	2	2	3	1	2
9	B	3	3	3	3	1	2	3	1
10	A	3	3	2	1	1	3	3	1
11	A	3	3	2	1	1	3	2	1
12	A	3	3	3	1	1	2	3	1
13	B	3	3	2	2	1	1	3	1
14	A	3	3	2	1	1	3	2	1
15	B	3	3	3	2	2	3	3	2
16	B	3	3	2	2	1	2	3	1
17	B	3	3	3	2	2	1	2	1
18	C	3	2	2	3	3	3	2	2
19	A	3	3	2	1	1	3	3	1
20	C	3	3	2	1	3	3	3	3
21	B	3	3	2	1	1	2	3	2
22	B	3	3	3	3	1	2	3	1
23	A	3	3	2	1	1	3	3	1
24	A	3	3	3	1	1	3	3	1
25	A	3	4	2	1	1	3	3	1

		VARIABLES							
N.º	Si.	1	2	3	4	5	6	7	8
26	B	3	3	3	2	2	3	2	1
27	C	2	3	4	2	3	1	1	2
28	B	3	3	2	3	1	1	1	1
29	B	3	3	2	2	2	1	2	2
30	C	3	3	1	3	2	2	2	1
31	C	3	3	2	3	2	3	2	3
32	A	3	3	2	1	1	3	2	1
33	A	3	3	2	1	2	2	3	1
34	B	3	3	2	3	1	1	3	1
35	A	3	3	2	1	1	2	3	1
36	A	3	3	2	1	1	3	2	1
37	A	3	3	2	1	1	3	2	1
38	A	3	3	2	1	1	3	3	1
39	B	3	3	3	3	1	1	3	1
40	A	3	3	2	1	2	1	3	1
41	C	1	3	4	2	1	2	1	1
42	B	3	3	2	2	1	3	3	1
43	C	2	3	4	3	2	3	1	2
44	B	3	3	2	2	1	1	3	1
45	D	2	3	2	3	2	2	3	3
46	B	3	3	2	2	1	2	2	1
47	C	3	3	2	3	2	2	3	3
48	B	3	4	2	2	1	2	2	1
49	D	2	3	2	3	2	2	3	3
50	D	1	1	3	3	2	3	3	3

Tabla 2.—*Estados de las variables para las cincuenta vasijas de La Cuesta del Negro (Purullena, Granada)*

queológica de estos resultados nos remitimos a CONTRERAS, MOLINA, CAPEL y ESQUIVEL, 1988.

BIBLIOGRAFIA

BACKER, E., y A.K. JAIN
1981 «A Clustering Performance Measure Based on Fuzzy Set Decomposition», *IEEE Trans. Pattern Anal. and Machine Intell.*, Vol. PAMI-3, January, pp. 66-75.

BEN-BASSAT, M., y L. ZAINDENBERG
1984 «Contextual Template Matching: A Distance Measure for Patterns with Hierarchically Dependent Features», *IEEE Trans. Pattern Anal. and Machine Intell.*, Vol. PAMI-6, March, pp. 201-211.

CHIU, D.K.Y., y A.K.C. WONG
1986 «Synthesizing Knowledge: A Cluster Analysis Approach Using Event Covering», *IEEE Trans. Syst.,* Man and Cyberns., Vol. SMC-16, March/April, pp. 251-259.

CONTRERAS, F.; F. MOLINA, J. CAPEL y J.A. ESQUIVEL
1988 «Los ajuares cerámicos de la necrópolis argárica de la Cuesta del Negro (Purullena, Granada). Avance al estudio analítico y estadístico», *I Curso de Ciencia en Arqueología,* La Laguna, Universidad de La Laguna (En prensa).

DE LUCA, A., y S. TERMINI
1972 «A Definition of Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory», *Inform. and Control,* Vol. 20, pp. 301-312.

DIDAY, E., y J.C. SIMON
1976 *Clustering Analysis: Communication and Cybernetics,* Vol.10, Springer Verlag, New York.

DUBOIS, D., y H. PRADE
1980 *Fuzzy Sets and Systems: Theory and Applications,* Academic Press, New York.

DUDA, R. O., y P.E. HART
1973 *Pattern Classification and Scene Analysis,* John Wiley, New York.

EVERITT, B.
1980 *Cluster Analysis,* Hubted Press, New York.

ESQUIVEL GUERRERO, J. A.
1988 *Una aplicación de la entropía al Análisis Cluster mediante Variables Cualitativas Multiestado: Afinidad, Similaridad y Agrupación,* Tesis Doctoral, Departamento de Estadística, Universidad de Granada.

ITO, T.; Y. KODAMA y J. TOYODA
1984 «A Similarity Measure Bettwen Patterns with Nonindependent Attributes», *IEEE Trans. Pattern Anal. and Machine Intell.,* Vol. PAMI-6, January, pp. 111-115.

KENDALL, M. G.
1975 *Multivariate Analysis,* Charles Griffin, London.

MICHALSKI, R.S., y R.E. STEPP
1983 «Automated Construction of Classifications: Conceptual Clustering versus Numerical Taxonomy», *IEEE Trans. Pattern Anal. and Machine Intell.,* Vol. PAMI-5, July, pp. 396-410.

MIYAMOTO, S., y K. NAKAYAMA
1986 «Similarity Measures Based on a Fuzzy Set Model and Application to Hierarchical Clustering», *IEEE Trans. Syst., Man and Cyberns.,* Vol. SMC-16, May/June, pp. 479-482.

PAL, S.K., y B. CHAKRABORTY
1986 «Fuzzy Set Theoretic Measure for Automatic Feature Evaluation», *IEEE Trans. Syst., Man and Cyberns.,* Vol. SMC-16, October, pp. 754-760.

PAL, S.K., y D. DUTTA MAJUMDER
1985 *Fuzzy Mathematical Approach to Pattern Recognition,* Wiley Eastern, New Delhi.

REZZA, F.M.
1961 *An Introduction to the Information Theory,* McGraw-Hill, New York.

RAO, C.R.
1984 «Use of Diversity and Distance Measures in the Analysis of Qualitative Data», en N. van Wark & W. W. Howell (eds.): *Multivariate Statistical Methods in Physical Antropology,* Reidel Publishing Co., Dordrecht, Holland, pp. 49-67.

ROMESBURG, H.C.
1984 *Cluster Analysis for Researchers,* Lifetime Learning Publications, Belmont C. A.

SHANNON, C.E.
1948 «A Mathematical Theory of Communication», *Bell System Tech. Journal,* Vol. 27, pp. 379-423, 623-656.

SNEATH, P.H.A., y R.R. SOKAL
1973 *Numerical Taxonomy: The Principles and Practice of Numerical Classification,* W. H. Freeman, San Francisco.

WONG, A.K.C., y D.K.Y. CHIU
1987 «Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data», *IEEE Trans. Pattern Anal. and Machine Intell.,* Vol. PAMI-9, November.

XIE, W.X., y S.D. BEDROSIAN
1984 «A Information Measure for Fuzzy Sets», IEEE Trans. Syst., Man and Cyberns., Vol. SMC-14, January/February, pp. 151-156.