CIMPA-UCR

Métodos de la Clasificación Automática

Eduardo Piza Volio*

Semana PASI sobre Reducción de Información
Centro de Investigación en Matemática,
Guanajuato, México, abril de 2010

*Centro de Investigación en Matemática Pura y Aplicada (CIMPA) de la Universidad de Costa Rica. Código postal 2060, San José, Costa Rica. Email: epiza@cariari.ucr.ac.cr.

CIMPA

Métodos de la Clasificación Automática

1. Introducción

La naturaleza ofrece una amplia diversidad de poblaciones susceptibles a ser repartidas en categorías. Cada disciplina científica requiere de clasificaciones de sus objetos en estudio, a partir de caracterizaciones adecuadas de los mismos. Más aún, puede afirmarse que la clasificación siempre ha sido una de las actividades inherentes al hombre: es una de las bases de todo conocimiento humano.

Por *Clasificación Automática* se entiende una extensa colección de algoritmos, ideas y técnicas tendientes a resolver racionalmente el problema general de la clasificación de objetos.

Se distinguen dos vastas familias de métodos: las técnicas de *particionamiento* y las técnicas *jerárquicas*.

Las primeras establecen particiones de la colección de objetos en clases disjuntas, con base a criterios de calidad específicos. Las técnicas jerárquicas establecen una tipología completa de los objetos en estudio, conduciendo a clasificaciones de tipo arbóreas o dendogramas.

CIMPA-UCR

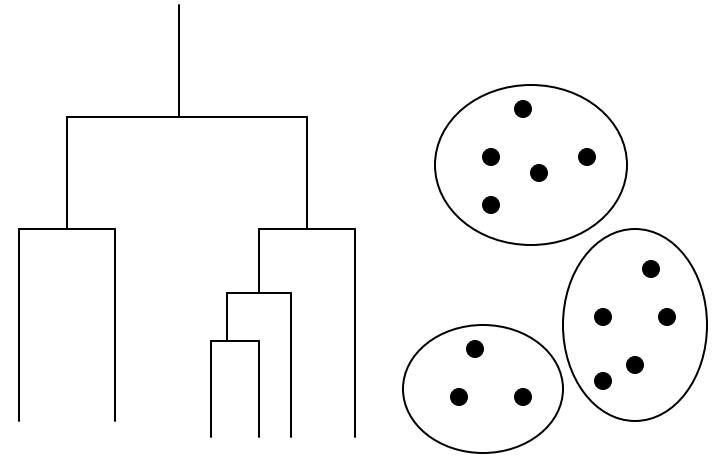
Clasificación Automática

"Cluster analysis"

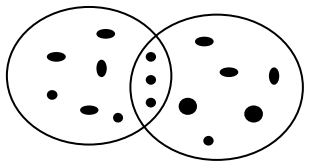
Análisis de conglomerados

Análisis tipológico

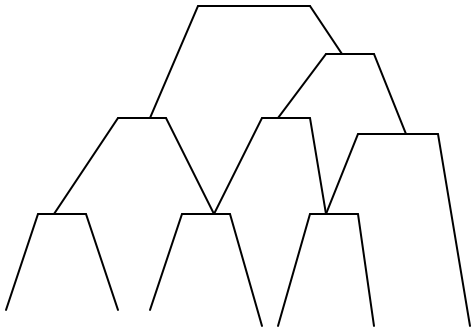
Análisis de grupos



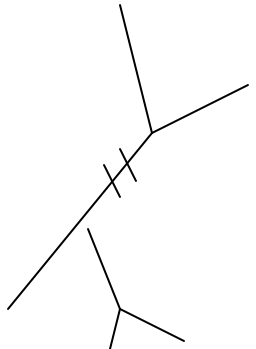
Jerárquicos



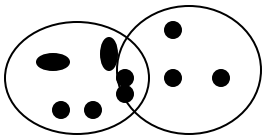
Particionamiento



Piramidales



Arboles aditivos



No disjuntos

Difusos

2. Índices de disimilitud

Las *disimilitudes* generalizan el concepto de las distancias en el conjunto de objetos a clasificar Ω. El primer paso en un análisis de Clasificación Automática generalmente será escoger un índice de disimilitud apropiado.

Específicamente, un *índice de disimilitud* (o simplemente una *disimilitud*) sobre un conjunto Ω es una aplicación $d : \Omega \times \Omega \mapsto \mathbb{R}$ que verifica las siguientes dos propiedades:

(1) *Simetría*: $d(x,y) = d(y,x),\ \forall x, y \in \Omega.$

(2) *No-negatividad*: $d(x,y) \geq 0,\ \forall x, y \in \Omega.$ Además, $d(x,x) = 0,\ \forall x \in \Omega.$

Si adicionalmente d verifica la propiedad

(3) $d(x,y) = 0 \Rightarrow x = y, \quad \forall x, y \in \Omega,$

entonces a d se le acostumbra llamar *desviación* sobre Ω. Cuando d verifica (1), (2), (3) y además verifica la *desigualdad triangular*:

(4) $d(x,y) \leq d(x,z) + d(z,y), \quad \forall x, y, z \in \Omega,$

entonces d es una *distancia* sobre Ω. Finalmente, si d verifica las propiedades (1), (2), (3) y verifica la *desigualdad ultramétrica*:

(5) $d(x,y) \leq \max\{d(x,z), d(z,y)\}, \quad \forall x, y, z \in \Omega,$

entonces a d se le acostumbra llamar *distancia ultramétrica* sobre Ω. La desigualdad ultramétrica es equivalente a la *desigualdad triangular generalizada*:

$$d(x,y) \leq [(d(x,z))^r + (d(z,y))^r]^{1/r}, \quad \forall x, y, z \in \Omega, \forall r > 0.$$

2. Índices de disimilitud

2.1. Disimilitudes para datos de tablas binarias

En este caso cada uno de los n objetos de Ω está caracterizado por p atributos o variables dicotómicas comunes para todos los objetos. Así, $\Omega = \{w_1, \ldots, w_n\}$, donde $w_i \cong (w_{i1}, \ldots, w_{ip})^t$, con $w_{ij} \in \{0, 1\}$, $1 \leq i \leq n$, $1 \leq j \leq p$.

Los valores 1 ó 0 en cada variable dicotómica indican la presencia o ausencia del atributo lógico en cuestión para el objeto específico. Este tipo de datos se presenta con frecuencia en las aplicaciones en biología y antropología, entre otras.

Para $1 \leq i \leq n$ y $1 \leq j \leq n$, con $i \neq j$, sean:

- $a_{ij} = \sum_{k=1}^{p} w_{ik}\, w_{jk}$ el número de atributos presentes simultáneamente en los objetos w_i y w_j.
- $b_{ij} = \sum_{k=1}^{p} (1 - w_{ik})\, w_{jk}$ el número de atributos presentes en el objeto w_j y ausentes en el objeto w_i.
- $c_{ij} = \sum_{k=1}^{p} w_{ik}\, (1 - w_{jk})$ el número de atributos presentes en el objeto w_i y ausentes en el objeto w_j.
- $d_{ij} = \sum_{k=1}^{p} (1 - w_{ik})\, (1 - w_{jk})$ el número de atributos ausentes en ambos objetos w_i y w_j.
- $n_i = \sum_{k=1}^{p} w_{ik}$ el número de atributos presentes en el objeto w_i.

2.1. Disimilitudes para datos de tablas binarias

A continuación se enumeran los índices de disimilitud más importantes para tablas binarias, mencionados en la literatura.

Jaccard–1901:

$$d_1 = 1 - \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

Czekanowki–1913, Dice–1945, Sorensen–1948:

$$d_2 = 1 - \frac{2a_{ij}}{2(a_{ij} + d_{ij} + c_{ij})}$$

Sokal y Michener–1958, Rogers y Tanimoto–1960:

$$d_3 = 1 - \frac{a_{ij} + d_{ij}}{p}$$

Russel y Rao–1940:

$$d_4 = 1 - \frac{a_{ij}}{p}$$

Kulczynski–1927:

$$d_5 = 1 - \frac{a_{ij}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

Kulczynski–1928:

$$d_6 = \frac{b_{ij} + c_{ij}}{a_{ij} + d_{ij}}$$

Ochiaï:

$$d_7 = 1 - \frac{a_{ij}}{\sqrt{n_i n_j}}$$

Sokal y Sneath:

$$d_8 = 1 - \frac{a_{ij} + d_{ij}}{b_{ij} + c_{ij}}$$

Sokay y Sneath:

$$d_9 = 1 - \frac{a_{ij}}{a_{ij} + 2(b_{ij} + c_{ij})}$$

Yule:

$$d_{10} = 1 - \frac{|a_{ij}d_{ij} - b_{ij}c_{ij}|}{a_{ij}d_{ij} + b_{ij}c_{ij}}$$

Distancia euclideana ponderada:

$$d_{11} = \frac{b_{ij} + c_{ij}}{p}$$

Distancia de la correlacion:

$$d_{12} = 1 - \frac{|a_{ij}d_{ij} - b_{ij}c_{ij}|}{\sqrt{(a_{ij} + b_{ij})(c_{ij} + d_{ij})(a_{ij} + c_{ij})(b_{ij} + d_{ij})}}$$

2.2. Disimilitudes para el caso de variables cuantitativas

En este caso cada uno de los n objetos está caracterizado por los valores medidos en p variables cuantitativas $v_1, \ldots, v_p$. Para el cálculo de las disimilitudes entre objetos se cuenta entonces con una tabla rectangular de datos. Así:

$$\Omega \cong \begin{pmatrix} w_{11} & \cdots & w_{1_p} \\ \vdots & & \vdots \\ w_{n1} & \cdots & w_{np} \end{pmatrix},$$

donde $w_{ij} = v_j(w_i)$ es el valor medido del objeto w_i en la variable v_j. A continuación se enumeran los principales índices de disimilitud mencionados en la literatura especializada.

(a) *Distancias de Minkowsy*:

$$d_{13}(w_i, w_j) = \left(\sum_{k=1}^{p} |w_{ik} - w_{jk}|^r \right)^{1/r}, \text{ con } r > 0.$$

(b) *Distancia de Chebychev*:

$$d_{14}(w_i, w_j) = \max_{k=1,\ldots,p} |w_{ik} - w_{jk}|.$$

(c) *Distancias euclídeas*:

Las disimilitudes mayormente utilizadas en las aplicaciones son las distancias euclídeas. Estas son de la forma $d^2(w_i, w_j) = (w_i - w_j)^t \, Q \, (w_i - w_j)$, donde Q es una matriz simétrica y definida positiva. Diferentes selecciones de la métrica Q conducen a diferentes disimilitudes. Las más empleadas son:

2.2. Disimilitudes para el caso de variables cuantitativas

(c1) *Distancia euclídea clásica*:

$$d_{15}(w_i, w_j) = \sum_{k=1}^{p} (w_{ik} - w_{jk})^2.$$

(c2) *Distancia de las varianzas*:

$$d_{16}(w_i, w_j) = \sum_{k=1}^{p} \frac{1}{\sigma_k^2} (w_{ik} - w_{jk})^2,$$

(c3) *Distancia de Mahalanobis*:

$$d_{17}(w_i, w_j) = (w_i - w_j)^t \, V^{-1} \, (w_i - w_j),$$

donde V es la matriz de varianzas y covarianzas de $v_1, \ldots, v_p$. La disimilitud de Mahalanobis aparece con frecuencia en la teoría del Análisis de Datos, en especial en el Análisis Discriminante. Su empleo en el contexto de clasificación es apropiado en situaciones en las cuales las variables $v_1, \ldots, v_p$ no son independientes. Obsérvese que d_{17} generaliza a d_{16}, que a su vez generaliza a d_{15}.

(c4) *Distancia del* χ^2:

$$d_{18}(w_i, w_j) = \sum_{k=1}^{p} \frac{1}{w_{\cdot k}} \left(\frac{w_{ik}}{w_{i\cdot}} - \frac{w_{jk}}{w_{j\cdot}} \right)^2,$$

donde $w_{\cdot k} = \sum_{r=1}^{n} w_{rk}$, y para $s \in \{i, j\}$, $w_{s\cdot} = \sum_{k=1}^{p} w_{sk}$. La distancia del χ^2 es la base del Análisis de Correspondencias. Su empleo es particularmente apropiado para tablas de datos de frecuencias.

(c5) *Distancia del ACP*:

$$d_{19}(w_i, w_j) = \sum_{k=1}^{p} \lambda_k \, (w_{ik}^* - w_{jk}^*)^2.$$

2.2. Disimilitudes para el caso de variables cuantitativas

(d) *Índices de intensidad de presencia de atributos.*

Para las situaciones en que las variables $v_1, \ldots, v_p$ son atributos y w_{ik} es un indicador continuo de la intensidad de la presencia del atributo v_k en el objeto w_i, otros índices de disimilitud han sido propuestos. En este caso $w_{ik} \geq 0$, $1 \leq i \leq n$, $1 \leq k \leq p$. Por ejemplo, los objetos a clasificar pueden ser regiones geográficas adyacentes, mientras los atributos pueden ser distintas especies biológicas, de manera que w_{ik} es un indicador de la abundancia de la especie k-ésima en la región i-ésima. Las disimilitudes más importantes en este contexto son las siguientes:

(d1) *Czekanowski–1932, Halternorth–1937, Cain y Harrison–1958*:

$$d_{20}(w_i, w_j) = \frac{1}{p}\sum_{k=1}^{p} |w_{ik} - w_{jk}|.$$

(d2) *Lance y Williams–1966*:

$$d_{21}(w_i, w_j) = \frac{1}{p}\sum_{k=1}^{p} \frac{|w_{ik} - w_{jk}|}{w_{ik} + w_{jk}}.$$

(d3) *Clark–1952*:

$$d_{22}(w_i, w_j) = \left\{ \frac{1}{p}\sum_{k=1}^{p} \left(\frac{w_{ik} - w_{jk}}{w_{ik} + w_{jk}} \right)^2 \right\}^{1/2}.$$

2.2. Disimilitudes para el caso de variables cuantitativas

(d) *Índices de intensidad de presencia de atributos.*

(d4) *Odum–1950, Bray y Curtis–1957*:

$$d_{23}(w_i, w_j) = \frac{\sum_{k=1}^{p} |w_{ik} - w_{jk}|}{\sum_{k=1}^{p} (w_{ik} + w_{jk})}.$$

(d5) *Kulczynski*:

$$d_{24}(w_i, w_j) = \frac{\sum_{k=1}^{p} \text{mín}\{w_{ik}, w_{jk}\}}{\sum_{k=1}^{p} (w_{ik} + w_{jk})}.$$

(d6) *Marczewski y Steinhaus*:

$$d_{25}(w_i, w_j) = \frac{\sum_{k=1}^{p} |w_{ik} - w_{jk}|}{\sum_{k=1}^{p} \text{máx}\{w_{ik}, w_{jk}\}}.$$

3. Índices de agregación

Los índices de agregación generalizan el concepto de distancia entre grupos de objetos. Un segundo paso en Clasificación Automática consiste generalmente en seleccionar un índice de agregación apropiado para los fines del estudio.

Formalmente, un *índice de agregación* (o simplemente una *agregación*) sobre $\mathcal{P}(\Omega)$ es una aplicación $\delta: \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \mapsto \mathbb{R}$ (donde $\mathcal{P}(\Omega)$ es la colección de subconjuntos de Ω) que verifica las siguientes dos propiedades:

(6) *Positividad*: $\delta(h_1, h_2) \geq 0, \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$

(7) *Simetría*: $\delta(h_1, h_2) = \delta(h_2, h_1), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$

(8) *Preordenanza*: δ establece la misma preordenanza que d sobre $\Omega \times \Omega$, esto es,

$$d(w_p, w_q) \geq d(w_r, w_s) \Rightarrow \delta(\{w_p\}, \{w_q\}) \geq \delta(\{w_r\}, \{w_s\}), \quad \forall w_p, w_q, w_r, w_s \in \Omega.$$

Algunas de las agregaciones más conocidas en la literatura especializada son extensiones de una disimilitud, en el sentido que verifican además la propiedad siguiente:

(9) *Extensionalidad*: $\delta(\{w_i\}, \{w_j\}) = d(w_i, w_j), \quad \forall w_i, w_j \in \Omega.$

La propiedad (9) es un caso particular de la (8). Los ocho índices de agregación $\delta_1, \ldots, \delta_8$ que se enumeran a continuación cumplen la propiedad (8), aunque no todos cumplen la propiedad (9).

3. Índices de agregación

(a) *Vecino más cercano, o "single linkage", Jardine y Sibson–1962*:

$$\delta_1(h_1, h_2) = \text{mín}\{d(w_i, w_j) : w_i \in h_1, w_j \in h_2\}, \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

(b) *Vecino más lejano, o "complete linkage", Sorensen–1948*:

$$\delta_2(h_1, h_2) = \text{máx}\{d(w_i, w_j) : w_i \in h_1, w_j \in h_2\}, \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

(c) *Promedio de las distancias, Sokal y Michener–1958*:

$$\delta_3(h_1, h_2) = \frac{1}{|h_1|\,|h_2|} \sum_{w_i \in h_1, w_j \in h_2} d(w_i, w_j), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

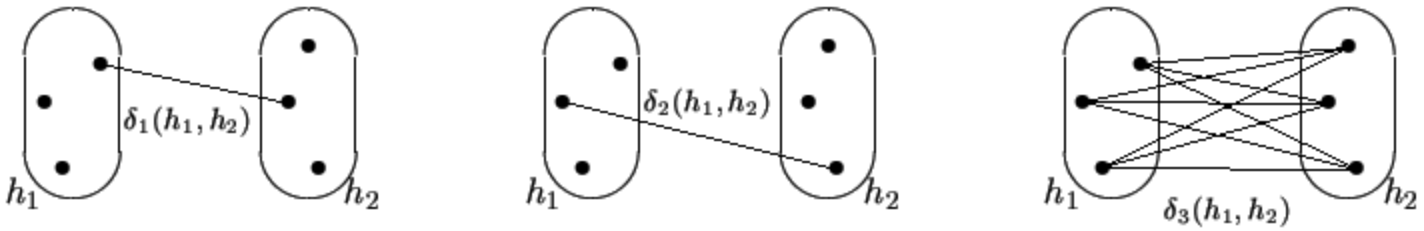


Figura 1: Representación esquemática de las agregaciones δ_1, δ_2, δ_3.

3. Índices de agregación

(d) *Agregaciones euclídeas*:

En los restantes índices de agregación δ_4, δ_5, δ_6, δ_7, δ_8, se considera a Ω como un subconjunto de un espacio euclídeo $\mathbb{R}^p$. Algunas nociones fundamentales deben definirse:

- A cada objeto $w \in \Omega$ se le asocia un peso, $p(w)$, indicador de la importancia del mismo dentro del análisis. Los pesos de los objetos están caracterizados por las propiedades

$$p(w) > 0, \quad \forall w \in \Omega \qquad ; \qquad \sum_{w \in \Omega} p(w) = 1.$$

  En la mayoría de las aplicaciones todos los objetos tienen idéntica importancia en el análisis, de modo que $p(w) = 1/|\Omega|$, $\forall w \in \Omega$ (*pesos canónicos*). Adicionalmente, se define el peso $p(h)$ de un grupo $h \subseteq \Omega$ de objetos, mediante $p(h) = \sum_{w \in h} p(w)$.

- El *centro de gravedad* o *baricentro* de un grupo de objetos $h \subseteq \Omega$ se denota por $G(h)$ y se define mediante:

$$G(h) = \frac{1}{p(h)} \sum_{w \in h} p(w) \cdot w.$$

  Al baricentro $G(h)$ se le asigna el peso $p(h)$.

- La *inercia* de un grupo de objetos $h \subseteq \Omega$ es una medida de cohesión/dispersión de los objetos del grupo con respecto a su centro de gravedad. Se denota como $I(h)$ y se define mediante

$$I(h) = \sum_{w \in h} p(w) \cdot \|w - G(h)\|^2.$$

3. Índices de agregación

- La *inercia* de un grupo de objetos $h \subseteq \Omega$ es una medida de cohesión/dispersión de los objetos del grupo con respecto a su centro de gravedad. Se denota como $I(h)$ y se define mediante

$$I(h) = \sum_{w \in h} p(w) \cdot \|w - G(h)\|^2.$$

Todas las disimilitudes empleadas en la definición de las agregaciones euclídeas $\delta_4, \ldots, \delta_8$ serán distancias euclídeas cuadráticas. A continuación se enumeran estas agregaciones.

(d1) *Distancia entre baricentros, Sokal y Michener–1958*:

$$\delta_4(h_1, h_2) = d(G(h_1), G(h_2)), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

(d2) *Inercia de la unión, Jambu–1978*:

$$\delta_5(h_1, h_2) = I(h_1 \cup h_2), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

(d3) *Varianza de la unión, Jambu–1978*:

$$\delta_6(h_1, h_2) = \frac{1}{p(h_1 \cup h_2)} I(h_1 \cup h_2), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

(d4) *Incremento de la inercia, Ward–1963*:

$$\delta_7(h_1, h_2) = I(h_1 \cup h_2) - I(h_1) - I(h_2), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

(d5) *Incremento de la varianza, Ward–1963*:

$$\delta_8(h_1, h_2) = \frac{p(h_1)\, p(h_2)}{[p(h_1) + p(h_2)]^2} d(G(h_1), G(h_2)), \quad \forall h_1, h_2 \in \mathcal{P}(\Omega).$$

4. Jerarquías indexadas

En Clasificación Automática las jerarquías indexadas brindan una estratificación completa de la colección de objetos en estudio, de acuerdo a las relaciones de similitud/disimilitud entre los mismos. Una *jerarquía* sobre Ω es cualquier subconjunto H de $\mathcal{P}(\Omega)$ que verifica las siguientes tres propiedades:

1. $\Omega \in H, \quad \phi \notin H.$
2. $\{w\} \in H, \quad \forall w \in \Omega.$
3. $h_1 \cap h_2 \in \{\phi, h_1, h_2\}, \quad \forall h_1, h_2 \in H.$

Si adicionalmente H verifica la siguiente propiedad:

4. $h \in H \Rightarrow$ (o bien h es unitario, o bien $h = h_1 \cup h_2$, con $h_1, h_2 \in H$),

entonces se le llama *jerarquía binaria* sobre Ω. Una *indexación* f para la jerarquía H es una aplicación $f: H \mapsto \mathbb{R}$ que verifica las siguientes dos propiedades:

5. $f(\{w\}) = 0, \quad \forall w \in \Omega.$
6. $h_1 \subseteq h_2 \Rightarrow f(h_1) \leq f(h_2), \quad \forall h_1, h_2 \in H.$

Al par (H, f) —donde H es una jerarquía sobre Ω y f es una indexación para H— se le llama *jerarquía indexada* sobre Ω.

Las jerarquías indexadas suelen representarse gráficamente por medio de árboles o dendogramas.

4. Jerarquías indexadas

Por ejemplo, supóngase que Ω está compuesto por 7 objetos: $\Omega = \{w_1, w_2, \ldots, w_7\}$ y la jerarquía H está formada por los 12 grupos de objetos: $H = \{h_1, h_2, \ldots, h_{12}\}$, donde:

$$\begin{array}{lcllcl} h_i & = & \{w_i\},\ 1 \leq i \leq 7 & h_8 & = & \{w_4, w_5\}, \\ h_9 & = & \{w_1, w_2, w_3\}, & h_{10} & = & \{w_4, w_5, w_6\}, \\ h_{11} & = & \{w_4, w_5, w_6, w_7\} & h_{12} & = & \Omega. \end{array}$$

Supóngase además que H está indexada mediante la función f definida por $f(h_i) = 0$, cuando $1 \leq i \leq 7$ y $f(h_i) = i - 7$, cuando $8 \leq i \leq 12$. Entonces la jerarquía indexada (H, f) se representa gráficamente con un dendograma como se ilustra en la Figura 2.

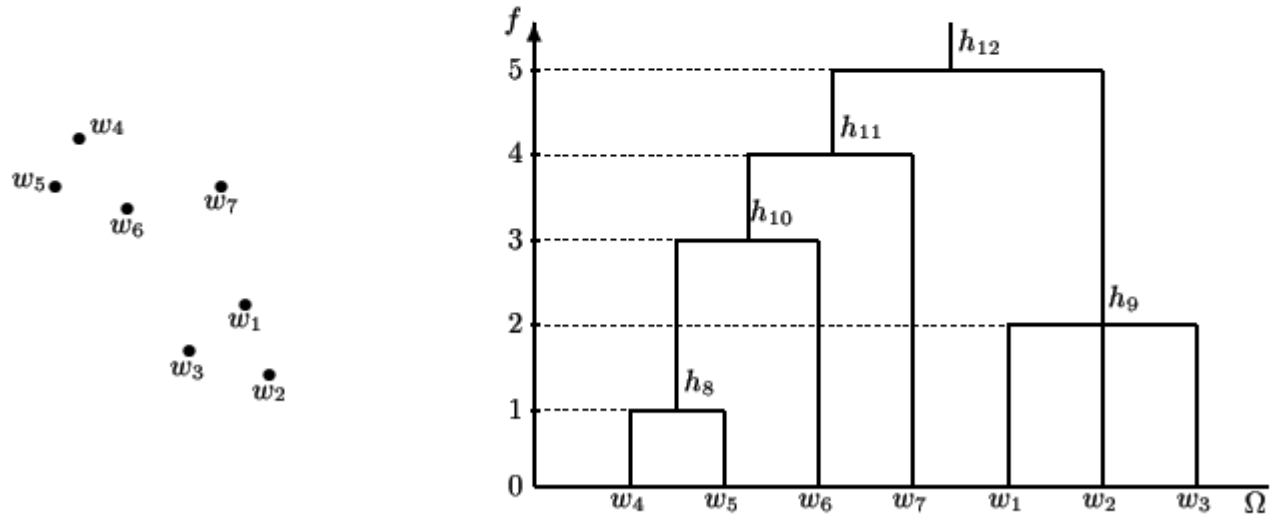


Figura 2: Ejemplo de una jerarquía indexada, a partir de 7 objetos $w_1, \ldots w_7$. A la izquierda se ilustra cómo podría ser la posición relativa de los 7 puntos en el plano.

4. Jerarquías indexadas

4.1. Construcción del índice de la jerarquía

Cada jerarquía H sobre Ω puede indexarse de una infinidad de maneras, conservando la misma preordenanza de los grupos de la jerarquía. Interesa en las aplicaciones aquellas indexaciones que permitan una fácil interpretación de los resultados.

Cuando la jerarquía H es binaria, es posible construir una indexación f para H a partir de un índice de agregación δ sobre $\mathcal{P}(\Omega)$, empleando el siguiente esquema recursivo:

$$\begin{aligned} f(\{w\}) &= 0, \quad \forall w \in \Omega. \\ f(h) &= \delta(h_1, h_2), \quad \text{cuando } h = h_1 \cup h_2, \text{ con } h_1, h_2 \in H. \end{aligned}$$

No obstante, este esquema no siempre define una indexación válida sobre H, pues podría conducir a *inversiones* (esto es, existencia de grupos h_1, $h_2 \in H$ para los cuales $h_1 \subset h_2$ pero $f(h_1) > f(h_2)$).

Para evitar la producción de inversiones en el esquema anterior basta con modificar la definición de

$$f(h_1 \cup h_2) := \text{máx}\{\delta(h_1, h_2), f(h_1), f(h_2)\}.$$

5. Algoritmo general de la clasificación jerárquica aglomerativa

El objetivo terminal en un estudio de clasificación automática jerárquica es producir una jerarquía indexada (H, f) a partir de la caracterización que se posea de los objetos a clasificar. Una vez seleccionados los índices de disimilitud d sobre Ω y agregación δ sobre $\mathcal{P}(\Omega)$, el algoritmo básicamente consiste en:

- Se construyen los grupos unitarios $\{w\}, \quad \forall w \in \Omega$.

- Se construyen paso a paso los grupos restantes que conforman la jerarquía indexada (H, f). En cada paso se buscan los grupos h_p y h_q que estén más próximos (de acuerdo a la agregación δ) y se unen en un nuevo grupo de la jerarquía.

La indexación construida por este algoritmo es precisamente la inducida por δ, evitando las inversiones.

5. Algoritmo general de la clasificación jerárquica aglomerativa

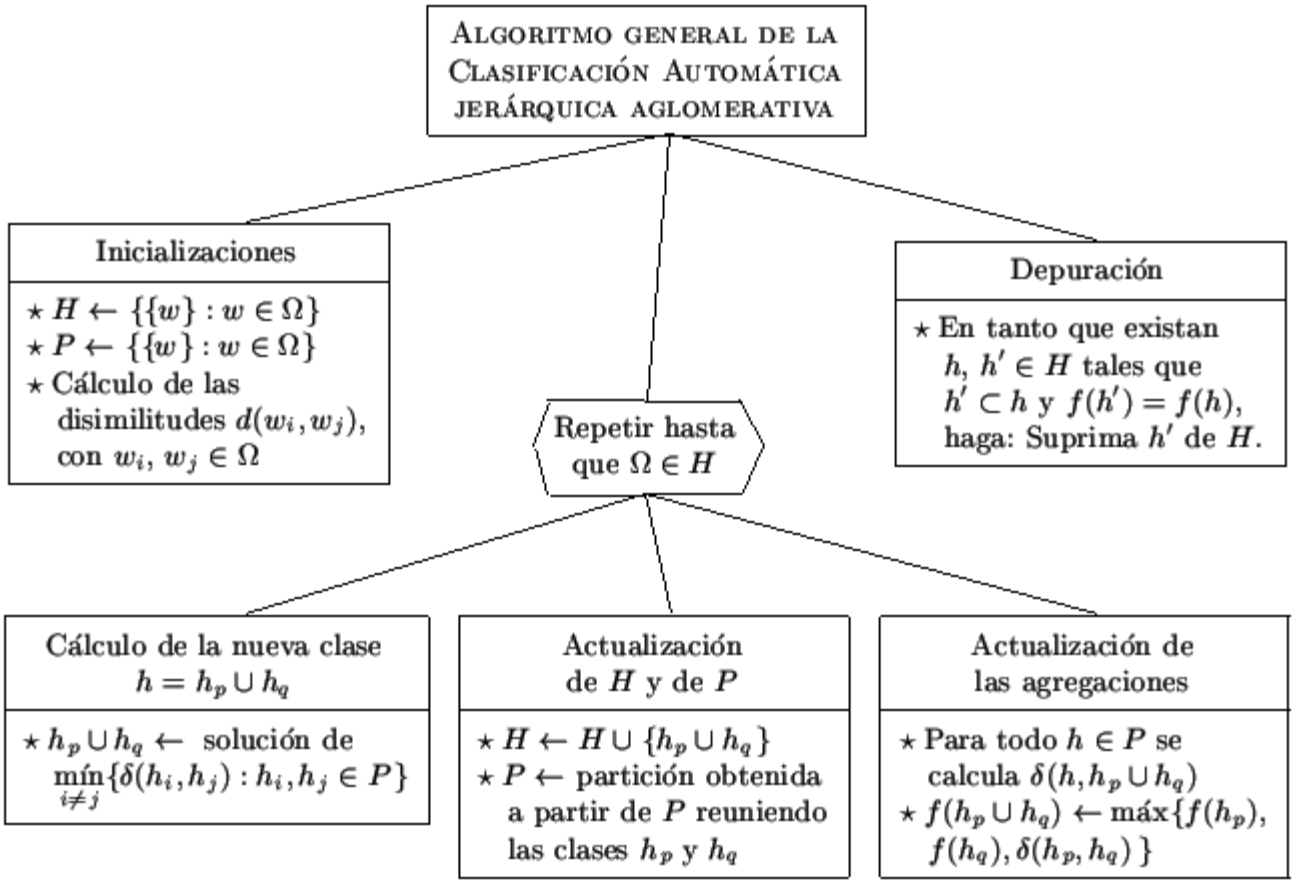


Figura 3: Algoritmo general de la clasificación automática jerárquica aglomerativa, partiendo de una disimilitud d y una agregación δ.

5. Algoritmo general de la clasificación jerárquica aglomerativa

5.1. Ventajas y desventajas del algoritmo general de clasificación jerárquica

- **Ventajas**:
  - Complejidad $O(n^2)$, si se aplican las fórmulas de recurrencia de Lance, Williams, Jambu.
  - Dadas d y δ, hay una única solución.
  - Fácil lectura
- **Desventajas**:
  - Resultado depende de δ.
  - Resultado depende de cómo el programador resuelve las igualdades.
  - Cargar en memoria tabla de n^2 disimilitudes.
  - Una jerarquía impone restricciones de inclusión.
  - Se aproxima d con una ultramétrica.
  - Es un algoritmo voraz ("greedy")

5. Algoritmo general de la clasificación jerárquica aglomerativa

5.1. Implementación del algoritmo: recurrencia de Lance-Williams & Jambu

En la implementación del algoritmo es de fundamental importancia poder calcular las nuevas agregaciones $\delta(h, h_p \cup h_q)$ mediante una fórmula de recurrencia, en términos de agregaciones e índices previamente calculados. Ello lleva a un ahorro sensible de tiempo de computación. Jambu [16] brinda una fórmula general para el cálculo de las nuevas agregaciones $\delta(h, h_p \cup h_q)$. Esta es:

$$\begin{aligned} \delta(h, h_1 \cup h_2) \;=\; & a_1\,\delta(h, h_1) + a_2\,\delta(h, h_2) + a_3\,\delta(h_1, h_2) + \\ & a_4\,f(h) + a_5\,f(h_1) + a_6\,f(h_2) + \\ & a_7\,|\delta(h, h_1) - \delta(h, h_2)| \end{aligned} \tag{1}$$

La recurrencia anterior generaliza una antigua fórmula de recurrencia de Lance y Williams [6] en la cual solamente se consideran los términos asociados a los parámetros a_1, a_2, a_3 y a_7. En la recurrencia (1) de Jambu los valores de los parámetros $a_1, a_2, \ldots, a_7$ fueron calculados por varios autores (ver [16, 6, 27]) y son presentados en la Figura 4, para cada una de las agregaciones clásicas $\delta_1, \delta_2, \ldots, \delta_8$ descritas en la sección 2.

CIMPA-UCR

5. Algoritmo general de la clasificación jerárquica aglomerativa

5.1. Implementación del algoritmo: recurrencia de Lance-Williams & Jambu

Parámetros a_i $\rightarrow$	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Agregación δ $\downarrow$	$\delta(h,h_1)$	$\delta(h,h_2)$	$\delta(h_1,h_2)$	$f(h)$	$f(h_1)$	$f(h_2)$	$\lvert\delta(h,h_1)-\delta(h,h_2)\rvert$
$\delta_1(h,h_1\cup h_2)$	$1/2$	$1/2$	0	0	0	0	$-1/2$
$\delta_2(h,h_1\cup h_2)$	$1/2$	$1/2$	0	0	0	0	$1/2$
$\delta_3(h,h_1\cup h_2)$	$\frac{p_1}{p_1+p_2}$	$\frac{p_2}{p_1+p_2}$	0	0	0	0	0
$\delta_4(h,h_1\cup h_2)$	$\frac{p_1}{p_1+p_2}$	$\frac{p_2}{p_1+p_2}$	$\frac{-p_1\,p_2}{(p_1+p_2)^2}$	0	0	0	0
$\delta_5(h,h_1\cup h_2)$	$\frac{p+p_1}{p_T}$	$\frac{p+p_2}{p_T}$	$\frac{p_1+p_2}{p_T}$	$\frac{-p}{p_T}$	$\frac{-p_1}{p_T}$	$\frac{-p_2}{p_T}$	0
$\delta_6(h,h_1\cup h_2)$	$\frac{(p+p_1)^2}{p_T^2}$	$\frac{(p+p_2)^2}{p_T^2}$	$\frac{(p_1+p_2)^2}{p_T^2}$	$\frac{-p^2}{p_T^2}$	$\frac{-p_1^2}{p_T^2}$	$\frac{-p_2^2}{p_T^2}$	0
$\delta_7(h,h_1\cup h_2)$	$\frac{p+p_1}{p_T}$	$\frac{p+p_2}{p_T}$	$\frac{-p}{p_T}$	0	0	0	0
$\delta_8(h,h_1\cup h_2)$	$\frac{(p+p_1)^2}{p_T^2}$	$\frac{(p+p_2)^2}{p_T^2}$	$\frac{-p\,(p_1+p_2)}{p_T^2}$	0	0	0	0

Figura 4: Valores de los parámetros a_i en la fórmula de recurrencia de Jambu para las diferentes agregaciones. Aquí $p=p(h)$, $p_1=p(h_1)$, $p_2=p(h_2)$, $p_T=p+p_1+p_2$.

5. Algoritmo general de la clasificación jerárquica aglomerativa

5.1. Implementación del algoritmo: recurrencia de Lance-Williams & Jambu

La fórmula de recurrencia general de Jambu (1) caracteriza por completo el tipo de agregación δ utilizado en el algoritmo general. Se podría utilizar esta recurrencia, con alguna selección específica de los parámetros $a_1, \ldots, a_7$ (no necesariamente alguna de las citadas en la Figura 4), para construir una nueva agregación δ. No obstante es de esperar que algunas selecciones de los parámetros $a_1, \ldots, a_7$ en la fórmula de recurrencia (1) conduzcan a inversiones en la agregación δ. Diday y Lebart [9] estudiaron este problema y establecieron condiciones suficientes sobre los valores de los parámetros $a_1, \ldots, a_7$ para que las inversiones no se produzcan. Estas son:

Teorema 1 *Supóngase que se cumplen simultáneamente las condiciones: i)* $a_1 + a_2 + a_3 \geq 1$, *ii)* $a_j \geq 0$, *para* $j \in \{1, 2, 4, 5, 6\}$, *iii)* $a_7 \geq -\min\{a_1, a_2\}$. *Entonces, la agregación* δ *construida a partir de la fórmula de recurrencia de Jambu (1) no contiene inversiones.*

De las agregaciones clásicas $\delta_1, \ldots, \delta_8$ solamente δ_1, δ_2, δ_3 y δ_7 verifican estas condiciones. La agregación δ_4 (distancia entre centros de gravedad) puede conducir a inversiones, al igual que la agregación δ_8 (incremento de las varianzas). Para evitarlas se emplea la definición alternativa arriba señalada para la construcción del índice f de la jerarquía: $f(h_p \cup h_q) = \max\{\delta(h_p, h_q), f(h_p), f(h_q)\}$.

CIMPA-UCR

6. Un ejemplo: clasificación de pintores del Renacimiento

Núm	Pintor	Color	Expresividad	Composición	Diseño
01	Bassano	06	08	17	00
02	Bellini	04	06	14	00
03	Cortona	16	14	12	06
04	Da Udine	10	08	16	03
05	Da Vinci	15	16	04	14
06	Del Piombo	08	13	16	07
07	Del Sarto	12	16	09	08
08	Durer	08	10	10	08
09	Guercino	18	10	10	04
10	Holbein	09	10	16	13
11	Jordaens L.	13	12	09	06
12	Michelangelo	08	17	04	08
13	Murillo	06	08	15	04
14	Perugino	04	12	10	04
15	Pordenone	08	14	17	05
16	Poussin	15	17	06	15
17	Raphael	17	18	12	18
18	Rembrandt	15	06	17	12
19	Romano G.	15	16	04	14
20	Rubens	18	13	17	17
21	Teniers	15	12	13	06
22	Tintoretto	15	14	16	04
23	Van Dyck	15	10	17	13
24	Veronese	15	10	16	03

Figura 5: Los 24 pintores más conocidos del Renacimiento, calificados según el criterio de Roger de Piles. Las calificaciones van de 0 (mínima) a 20 (máxima).

6. Un ejemplo: clasificación de pintores del Renacimiento

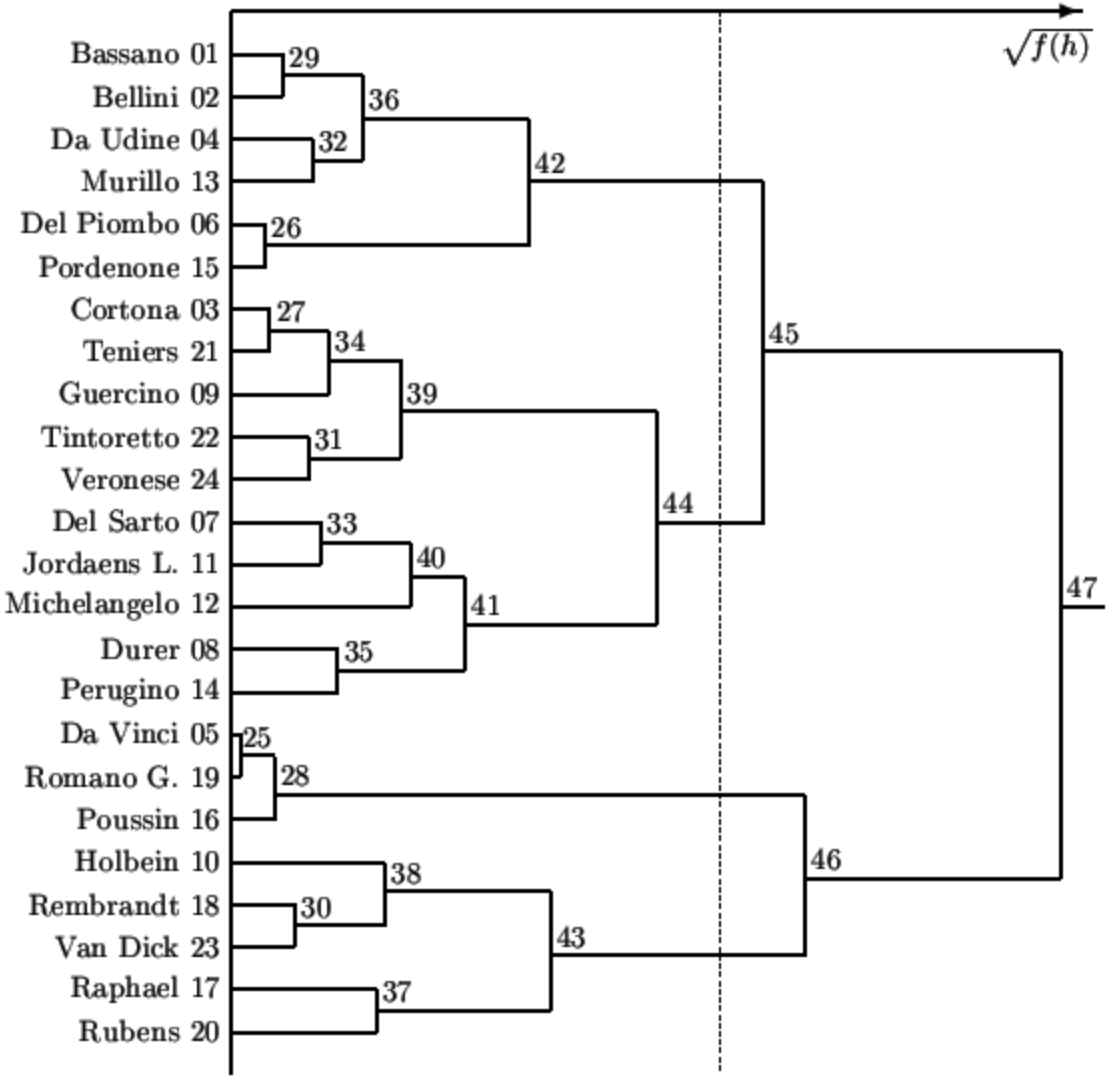


Figura 6: Dendograma de clasificación de los 24 pintores, empleando la distancias euclídeas (d_{15}) entre los pintores y la agregación de Ward (δ_7) entre los grupos de pintores.

6. Un ejemplo: clasificación de pintores del Renacimiento

Grupo	Pintores	Caracterización del grupo	
1	Bassano Bellini Da Udine Murillo Del Piombo Pordenone	Color deficiente Regular expresividad Buena composición Pésimo diseño	(8.6) (9.6) (15.0) (3.3)
2	Cortona Teniers Guercino Tintoretto Veronese Del Sarto Jordaens L. Michelangelo Durer Perugino	Buen color Buena expresividad Composición normal Mal diseño	(11.8) (13.0) (11.0) (5.9)
3	Da Vinci Romano G. Poussin	Excelente color Excelente expresividad Pésima composición Excelente diseño	(15.0) (16.3) (4.7) (14.3)
4	Holbein Rembrandt Van Dyck Raphael Rubens	Excelente color Expresividad normal Excelente composición Excelente diseño	(15.0) (11.4) (15.8) (14.6)

Figura 7: Interpretación de la clasificación de los pintores en 4 grupos. Entre paréntesis se ha anotado el promedio de cada atributo en cada grupo.

7. Las pirámides

Formalmente, una *pirámide* P sobre Ω es un conjunto de partes no vacías de Ω (llamados "grupos" de la pirámide), que satisface las siguientes 4 condiciones:

1. $\Omega \in P, \quad \phi \notin P.$

2. $\{w\} \in P$, para todo $w \in \Omega$, esto es, P contiene a todos los grupos unitarios.

3. Si $h_1, h_2 \in P$, entonces tendremos que, o bien $h_1 \cap h_2 = \phi$, o bien $h_1 \cap h_2 \in P$.

4. Existe un ordenamiento θ de los objetos de Ω bajo el cual todo grupo $h \in P$ está asociado a un "intervalo" de θ, esto es, no se presentan "cruzamientos" en el gráfico de P luego del ordenamiento θ de los objetos.

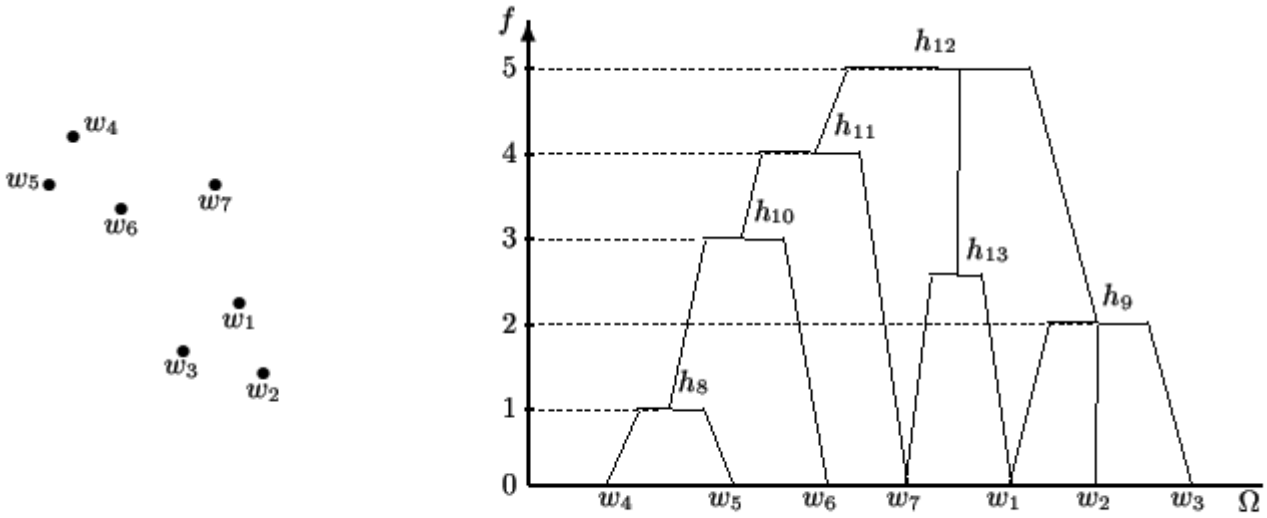


Figura 8: Ejemplo de una pirámide P para mismos 7 objetos $w_1, \ldots, w_7$ de la Figura 2.

8. Optimización en clasificación jerárquica

El estudio de las ultramétricas es de fundamental importancia en clasificación automática jerárquica, debido entre otras cosas al siguiente resultado atribuido a Benzecri.

Teorema 2 (Benzecri, 1973) *Existe una biyección constructiva entre el conjunto $\mathcal{H}$ de todas las jerarquías indexadas sobre Ω y el conjunto $\mathcal{U}$ de todas las ultramétricas sobre Ω.*

Indicamos cómo construir una ultramétrica μ a partir de una jerarquía indexada (H, f) sobre Ω, y viceversa:

- A partir de una jerarquía indexada (H, f) sobre Ω, podemos construir la ultramétrica μ mediante
$$\mu(w, w') = \text{mín}\{f(h) : h \in H \text{ y } w, w' \in h\}.$$
- A partir de una ultramétrica μ sobre Ω, podemos construir una jerarquía indexada (H, f) de la siguiente manera: a) Primeramente, para cada $\alpha \in \mathbb{R}^+$ definimos la relación R_α sobre Ω mediante:
$$w \; R_\alpha \; w' \Leftrightarrow \mu(w, w') \leq \alpha.$$
Es fácil probar que R_α es una relación de equivalencia sobre Ω . b) A continuación definimos la jerarquía indexada (H, f) sobre Ω como el conjunto H de todas las clases de equivalencia de las relaciones R_α, con $\alpha > 0$, en donde a cada grupo $h \in H$ le asociamos el índice $f(h) = \text{máx}\{\mu(w, w') : w, w' \in h\}$, cuando h no es unitario, mientras que $f(h) = 0$, cuando h es unitario.

Se demuestra que estos esquemas de construcción son inversos el uno del otro: aplicados en pares uno después del otro produce que retornemos a la misma jerarquía indexada o misma ultramétrica de partida.

8. Optimización en clasificación jerárquica

Gracias a la equivalencia entre las jerarquías indexadas y las ultramétricas, es natural plantearse el problema de la optimalidad en clasificación automática jerárquica como sigue: a partir de una disimilitud d sobre Ω dada, se busca una ultramétrica $\mu^\star \in \mathcal{U}$ que sea solución al problema

$$\text{minimizar}\,\{\Delta(d,\mu) : \mu \in \mathcal{U}\}, \tag{2}$$

donde $\Delta(d,\mu)$ es alguna *medida de adecuación* entre d y μ que debe ser definida con exactitud. Algunas de las medidas de adecuación Δ que han sido propuestas en la literatura son:

(a) *Benzecri–1973*:

$\Delta_1(d,\mu) =$ correlación entre d y μ, medida sobre todos los objetos $w \in \Omega$.

(b) *Hartingan–1967, Carrol y Pruzansky–1975, Chanon, Lamaire y Pouget–1980*:

$$\Delta_2(d,\mu) = \sum_{w,w' \in \Omega} p(w)\,p(w')\,[d(w,w') - \mu(w,w')]^2.$$

(c) *Defays–1975*:

$$\Delta_3(d,\mu) = \sum_{w,w' \in \Omega} |d(w,w') - \mu(w,w')|.$$

(d) *Lebart–1982*:

$$\Delta_4(d,\mu) = \left[\sum_{w,w' \in \Omega} |[d(w,w') - \mu(w,w')|^r \right]^{1/r}, \quad \text{con } r > 0.$$

8. Optimización en clasificación jerárquica

8.1. Resultados para ultramétricas dominadas

Una ultramétrica $\mu \in \mathcal{U}$ se dice que es *dominada* por la disimilitud d si se verifica

$$\mu(w, w') \leq d(w, w'), \qquad \forall w, w' \in \Omega.$$

Describimos esta situación escribiendo simplemente $\mu \preceq d$.

Teorema 3 *Dada la disimilitud d sobre Ω, el problema*

$$\textit{minimizar}\{\, \Delta(d, \mu) : \mu \in \mathcal{U},\ \textit{con}\ \mu \preceq d \,\} \tag{3}$$

posee solución única para las tres medidas de adecuación Δ_2, Δ_3 y Δ_4. Además, la solución es la misma en los tres casos: la llamada "ultramétrica bajo-dominación" μ_d, definida por

$$\mu_d(w, w') = \sup\{\, \mu(w, w') : \mu \in \mathcal{U},\ \mu \preceq d \,\}. \tag{4}$$

CIMPA-UCR

8. Optimización en clasificación jerárquica

Comentarios

- μ_d es en efecto es una ultramétrica sobre Ω.

- El problema análogo para ultramétricas dominadoras $\mu \succeq d$, esto es,

$$\text{minimizar}\,\{\,\Delta(d,\mu) : \mu \in \mathcal{U},\ \text{con}\ \mu \succeq d\,\}, \tag{5}$$

  no tiene una solución útil, pues entre otros cosas se demuestra que la cantidad análoga a μ_d definida por

$$\inf\{\,\mu(w,w') : \mu \in \mathcal{U},\ \mu \succeq d\,\} \tag{6}$$

  no define una ultramétrica sobre Ω.

- La ultramétrica bajo-dominación μ_d es equivalente —de acuerdo al esquema constructivo del teorema de Benzecri— a la jerarquía indexada (H, f) que se obtiene al aplicar el algoritmo general de la clasificación jerárquica aglomerativa con la agregación δ_1 del *vecino más cercano*.

- Por esta razón a veces se afirma que esta agregación δ_1 produce resultados óptimos en clasificación jerárquica, aunque en realidad la optimalidad es relativa a los criterios de adecuación y dominación anteriormente mencionados.

9. Clasificación automática mediante particionamiento

- Sea $\Omega = \{w_1, \ldots, w_n\}$ el conjunto finito de n objetos que deseamos clasificar y sea $k < n$ el número de clases en los cuales deseamos clasificar a los objetos.

- Una partición $P = (C_1, \ldots, C_k)$ de Ω en k clases $C_1, \ldots, C_k$, está caracterizada por las siguientes dos condiciones:

$$\text{(a)} \quad \Omega = \bigcup_{i=1}^{k} C_i \qquad\qquad \text{(b)} \quad C_i \cap C_j = \phi, \quad \forall i \neq j.$$

- En el curso de nuestros algoritmos permitiremos eventualmente que algunas de las clases C_i sea vacía, de manera que en realidad las particiones $P = (C_1, \ldots, C_k)$ que estaremos considerando son particiones de Ω en k *o menos* clases.

- Sea $\mathcal{P}_k$ el conjunto de todas las particiones $P = (C_1, \ldots, C_k)$ de Ω en k o menos clases. Andamos interesados en encontrar "buenas particiones", esto es, aquellas particiones que reflejen las relaciones de similitud existentes entre los objetos $w_i \in \Omega$.

- Cada objeto $w_i \in \Omega$ estará caracterizado por p distintos atributos o variables, medidos en una escala numérica, de donde cada objeto w_i será visto como un vector del espacio euclídeo $\mathbb{R}^p$.

9. Clasificación automática mediante particionamiento

- En este espacio de representación contamos con una métrica euclídea M, que nos permite definir el producto interno $\langle w_i \,|\, w_j \rangle_M = w_i^t M w_j$ entre los objetos, así como la norma $\|w\|_M^2 = w^t M w$. En la programación de los algoritmos se ha supuesto, sin pérdida de generalidad para efectos de convergencia, que $M = \mathrm{Id}$ (métrica euclídea clásica). En efecto, en el caso general la métrica M se descompone como $U^t U$, y la transformación $z_i = U w_i$ nos lleva a la métrica euclídea clásica, con los nuevos datos z_i.

- Asociado con cada objeto $w_i \in \Omega$ tendremos el peso de w_i, denotado por p_i, que refleja la importancia relativa del objeto w_i en el estudio. Los pesos p_i son todos positivos y su suma es la unidad: $\sum_{i=1}^n p_i = 1$.

- La calidad de una partición $P = (C_1, \ldots, C_k)$ se mide a través de la *inercia inter-clases* $B(P)$, índice que refleja la intensidad de la separación entre los centros de gravedad de las diversas clases C_i:

$$B(P) := \sum_{i=1}^{k} p(C_i) \cdot \|G(\Omega) - G(C_i)\|,$$

donde $p(C_i)$ es el peso relativo de la clase C_i mientras que $G(\Omega)$ y $G(C_i)$ son los vectores centros de gravedad de Ω y C_i respectivamente, calculados mediante las fórmulas siguientes:

$$p(C_i) = \sum_{w_s \in C_i} p_s, \qquad G(\Omega) = \sum_{s=1}^{n} p_s \, w_s, \qquad G(C_i) = \frac{1}{p(C_i)} \sum_{w_s \in C_i} p_s \, w_s.$$

9. Clasificación automática mediante particionamiento

9.1. Sobre la complejidad de los algoritmos de particionamiento

- El problema de la clasificación automática por métodos de particionamiento consiste entonces en hallar la partición $P \in \mathcal{P}_k$ que maximiza la inercia inter-clases $B(P)$. La complejidad computacional de este problema es del tipo *NP-completo*, pues se trata de una generalización del conocido problema *NPP*, el cual originó la terminología de este grado de complejidad para problemas análogos [21].

- Para tener una idea del tamaño combinatorio de este problema, si denotamos por $S(n,k)$ y B_n el número de particiones de Ω en k clases no vacías y el número total de particiones de Ω respectivamente, entonces,

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^n \qquad , \qquad B_n = e^{-1}\sum_{i=0}^{\infty}\frac{(i+1)^n}{i!}.$$

- Por ejemplo $S(60,2) \approx 0{,}58 \times 10^{18}$, $S(60,5) \approx 0{,}72 \times 10^{40}$, $S(100,5) \approx 0{,}66 \times 10^{68}$, mientras que $B_{10} = 115975$, $B_{15} \approx 0{,}14 \times 10^{10}$ y $B_{40} \approx 0{,}16 \times 10^{36}$. En un problema de clasificación automática de tamaño "regular", en el cual Ω tenga 100 objetos y el número de clases sea $k = 5$, si existiera sobre la Tierra un computador tan veloz que fuese capaz de calcular $B(P)$ para cada una de las particiones P de Ω en un tiempo de 10^{-10} segundos en busca de un máximo global, ¡a este veloz computador le tomaría algo más de 2×10^{48} siglos en completar el análisis de todas las particiones del problema!

9. Clasificación automática mediante particionamiento

9.1. Sobre la complejidad de los algoritmos de particionamiento

- En virtud del monstruoso tamaño del espacio de configuraciones de este problema de optimización discreta, no cabe esperarse la existencia de un algoritmo *eficiente* para la obtención del óptimo global. *Eficiente* en el sentido de la complejidad computacional: que realice los cálculos en *tiempo polinómico*.

- De hecho, si la *Conjetura Fundamental de las Ciencias de la Computación* denominada

$$\mathcal{P} \neq \mathcal{NP}$$

fuese realmente verdadera (la prueba o refutación tiene un valor de $1.000.000, pagaderos por el Instituto Clay de E.E.U.U.), entonces del todo no existiría ningún *algoritmo eficiente* para el problema del particionamiento óptimo.

9. Clasificación automática mediante particionamiento

9.2. El teorema de Huygens

En los algoritmos que desarrollamos utilizamos algunas veces el criterio de la inercia inter-clases $B(P)$ y en otras el criterio de la inercia intra-clases $W(P)$, motivados por el resultado del teorema de Huygens que sigue.

Teorema 4 (Huygens) *Para cada partición $P \in \mathcal{P}_k$ en k o menos clases, se tiene que la suma $B(P) + W(P)$ es siempre igual a la constante I_Ω, llamada* inercia total de Ω, *cuya fórmula es*

$$B(P) + W(P) = I_\Omega := \sum_{i=1}^{n} p_i \, \|w_i - G(\Omega)\|_M^2.$$

Corolario 5 *Los siguientes dos problemas de optimización combinatoria son equivalentes:*

$$(1) \begin{cases} \text{Maximizar } B(P) \\ \text{sujeto a } P \in \mathcal{P}_k \end{cases} \qquad (2) \begin{cases} \text{Minimizar } W(P) \\ \text{sujeto a } P \in \mathcal{P}_k \end{cases}$$

10. Los métodos tradicionales de particionamiento

Los algoritmos tradicionales de clasificación automática mediante particionamiento más conocidos son los siguientes:

- El algoritmo de "k-means" de MacQueen-Forgy, o la generalización de Diday en la misma dirección, denominada "nubes dinámicas".
- Los "algoritmos de transferencias", de Régnier.
- El "Análisis de Particiones Principales".

Ventajas

Son "algoritmos codiciosos", muy rápidos.

Desventajas

Converjen a óptimos locales y no globales.

Método de nubes dinámicas

Forgy (1965)

Mc Queen (1967) *"k - means"*

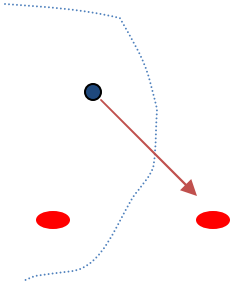
Diday (1969) → MND

- Da una partición inicial al azar: *P*.
- calcula los centros

ciclos

- Asigna los individuos al centro más cercano:

  forma $c_1,\ldots,c_k$ nueva
- Recalcula los centros $g_1,\ldots,g_k$



Hasta alcanzar una estabilización.

Método de nubes dinámicas

Forgy (1965) : esquema básico

Diday (1969) : esquema general

- Una clase se representa por un núcleo o prototipo

$C_l \rightsquigarrow N_l$

- A partir de una representación inicial en núcleos, se iteran:

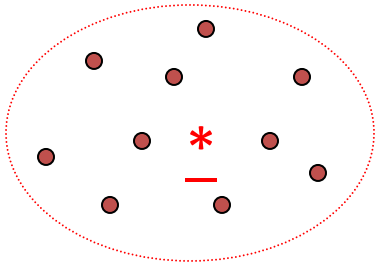
  - se hacen clasificaciones por asignación de los objetos al

    al núcleo más cercano

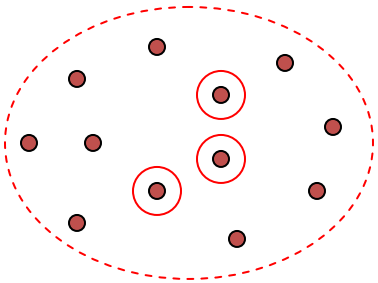
  - se representan las clases mediante el cálculo de los núcleos

Ejemplos de núcleos

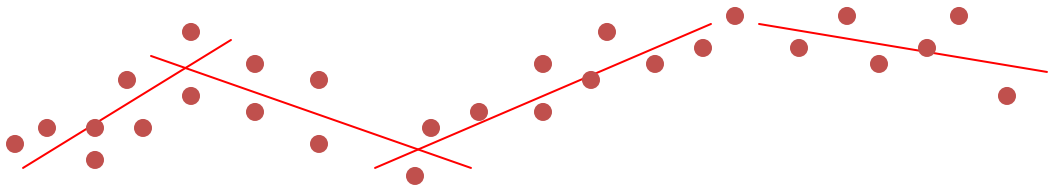
- Caso euclídeo: centro de gravedad
  (punto u objeto promedio)



- Caso no euclideano: una muestra
  (objetos más representativos)



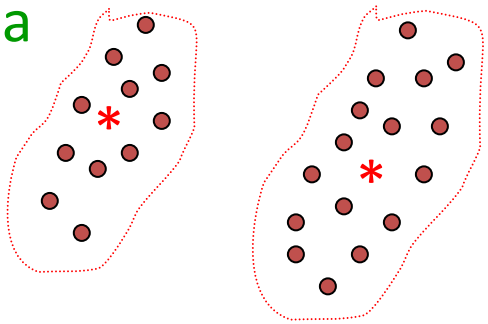
- Caso explicativo: rectas de regresión



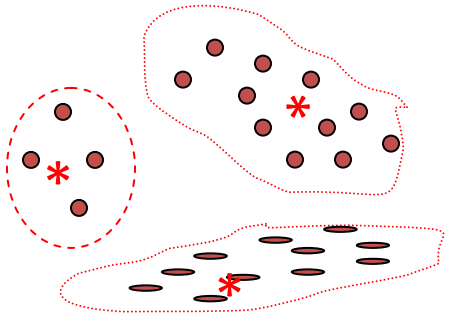
Ejemplos de núcleos

- Reconocimiento de formas: métricas o distancias adaptativas

Una sola métrica



Una métrica por clase:



Etapas en el MND

Asignación

$$x_i \rightarrow C_l \text{ si}$$

$$d(x_i, N_l) \leq d(x_i, N_h) \quad \text{para } h \in \{1,\ldots,k\}$$

ie: $$d(x_i, N_l) \leq \min_h d(x_i, N_h)$$

En caso de igualdad, se asigna x_i a la clase de índice menor

Representación

N_l es núcleo de c_l si el criterio W es mínimo con N_l

Caso euclídeo: $N_l = g_l$, el centro de gravedad

Teorema de Huygens

$$I_a(C_l) = I_{gl}(C_l) + \mu \|g_l - a\|^2$$

$$I_a(C_l) = \sum_{x_i \in C_l} p_i \|x_i - a\|^2$$

MND: núcleos son centros de gravedad

Forgy 1965, Diday 1967, Mac Queen 1967

d : distancia Euclídea clásica (cuadrática)

x^j : cuantitativas

$$W(P,L)=\sum_{l=1}^{k}\sum_{x_i\in C_l} p_i \left\|x_i - N_l\right\|^2$$

$$=\sum_{l=1}^{k}\sum_{x_i\in C_l} p_i \left\|x_i - g_l\right\|^2 + \left|C_l\right| \left\|g_l - N_l\right\|^2$$

➡ Núcleo que minimiza: centro de gravedad g_l

➡ $W(P,L)=W=\sum_{l=1}^{k}\sum_{x_i\in C_l} p_i \left\|x_i - g_l\right\|^2$: Inercia intra-clases

MND: núcleos son centros de gravedad

Forgy 1965, Diday 1967, Mac Queen 1967

Algoritmo:

1. Escoger *k* individuos: (al azar o con experticia) $g_1^{(0)}, g_2^{(0)}, \ldots, g_k^{(0)}$

2. Para *i* = 1 hasta *n*: asignar x_i al centro $g_l^{(t)}$ tal que:

$$\left\| x_i - g_l^{(t)} \right\| = \underset{l=1\ldots k}{\text{Mín}} \left\{ \left\| x_i - g_l^{(t)} \right\| \right\}$$

(caso de igualdad: menor índice)

Se forman clases $C_1^{(t)}, C_2^{(t)}, \ldots, C_k^{(t)}$

3. Calcular núcleos: para *l* = 1 hasta *k*

$$g_l^{(t)} = \frac{1}{\mu_l^{(t-1)}} \sum_{x_i \in C_l^{(t-1)}} p_i x_i \qquad \text{con} \qquad \mu_l^{(t-1)} = \sum_{x_i \in C_l^{(t-1)}} p_i \qquad (t = t+1)$$

4. Hasta que ningún individuo cambia de clase

CIMPA-UCR

MND: convergencia

W decrece en cada iteración del MND

ASIG: *i*) Sean $P=(C_1,\ldots,C_k)$, $L=(g_1,\ldots,g_k)$, $f(L)$: partición alrededor de los g_l

$$W(L,P)=\sum_{l=1}^{k}\sum_{x_i\in C_l} p_i\left\|x_i-g_l\right\|^2$$

$$W(L,f(L))=\sum_{l=1}^{k}\sum_{x_i\in D_l} p_i\left\|x_i-g_l\right\|^2, \qquad \text{con } f(L)=(D_1,\ldots,D_k)$$

Sea $z\in\Omega$: $z\in C_j$ $z\in D_h$

por definición de D_h: $\left\|z-g_h\right\|<\left\|z-g_j\right\|$

$$\Rightarrow p_z\left\|z-g_h\right\|^2\le p_z\left\|z-g_j\right\|^2$$

Razonando $\forall\, z\in\Omega$: $W(L,f(L))\le W(L,P)$

DESCRIPCIÓN DE UNA PARTICIÓN (1)

M: daigonal

Indice global: $R = \frac{B}{I}$ si $R \approx 1 \Rightarrow$ Buena Clasificación

si $R \approx 0 \Rightarrow$ Mala Clasificación

Contribución de las variables:

$$x^j \rightarrow cor(j) = \frac{\mathrm{var}\left(\bar{x}^j\right)}{\mathrm{var}\left(x^j\right)}$$ con : $\bar{x}$ medidas de x^j en cada clase

Descripción de las clases:

$$B(l) = \frac{\mathrm{var}_l\left(\bar{x}^j\right)}{B}:$$ posición de C_l respecto a g $B(I) \uparrow \Rightarrow C_l$ es excéntrico

$$W(l) = \sum_{j=1}^{p} \sum_{x_i \in C_l} p_i \left(x_i^j - g_l^j\right)^2 :$$ concentración de la clase $W(I) \downarrow$ $\Rightarrow C_l$ está concentrado

DESCRIPCIÓN DE UNA PARTICIÓN (2)

Descripción de las clases por variable:

$$x^j \approx C_l : cor(j,l) = \frac{\mu_l \left(\bar{x}_l^j - \bar{x}^j \right)^2}{\mathrm{var}\left(x^j\right)}$$

$cor(j,l) \uparrow : x^j$ es homogénea sobre C_l

Ej: R = 94%

cor (1) = 96.7% → discrimina

cor (2) = 89.8%

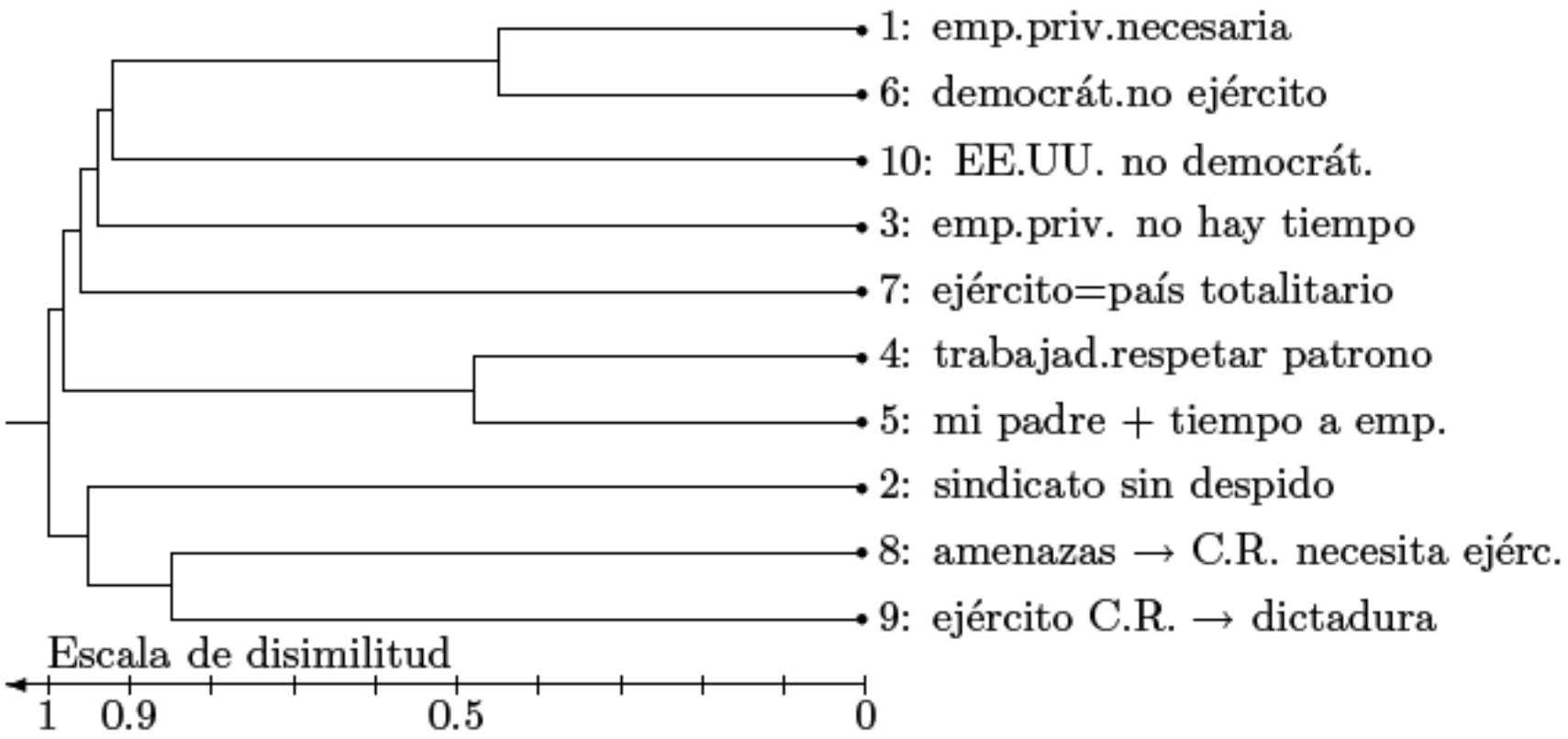


Figura 6.9: Arbol de clasificación para las variables de opinión.

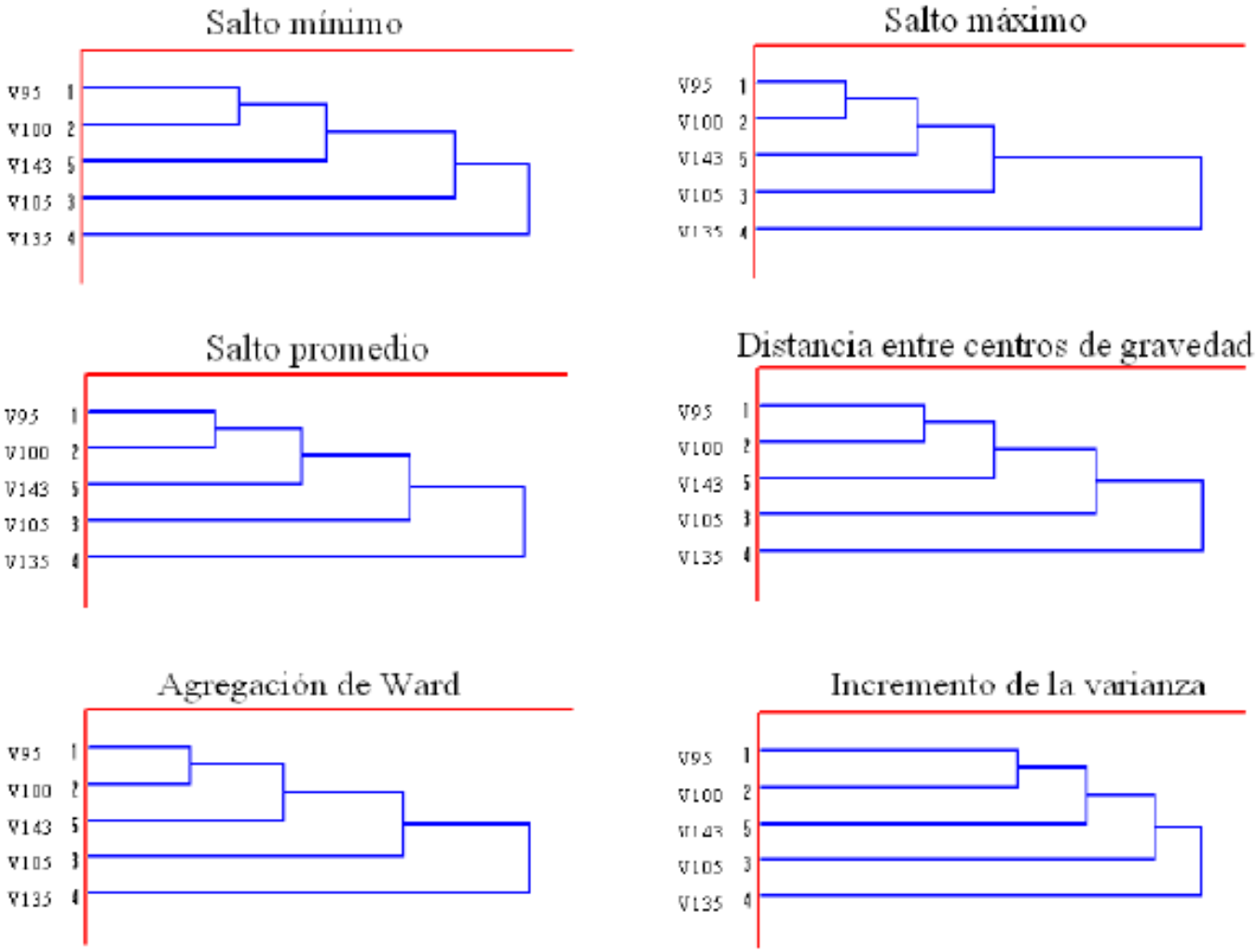


Figura 6.10: Arbol jerárquico de las variedades de fabes asturianas usando seis criterios de agregación.

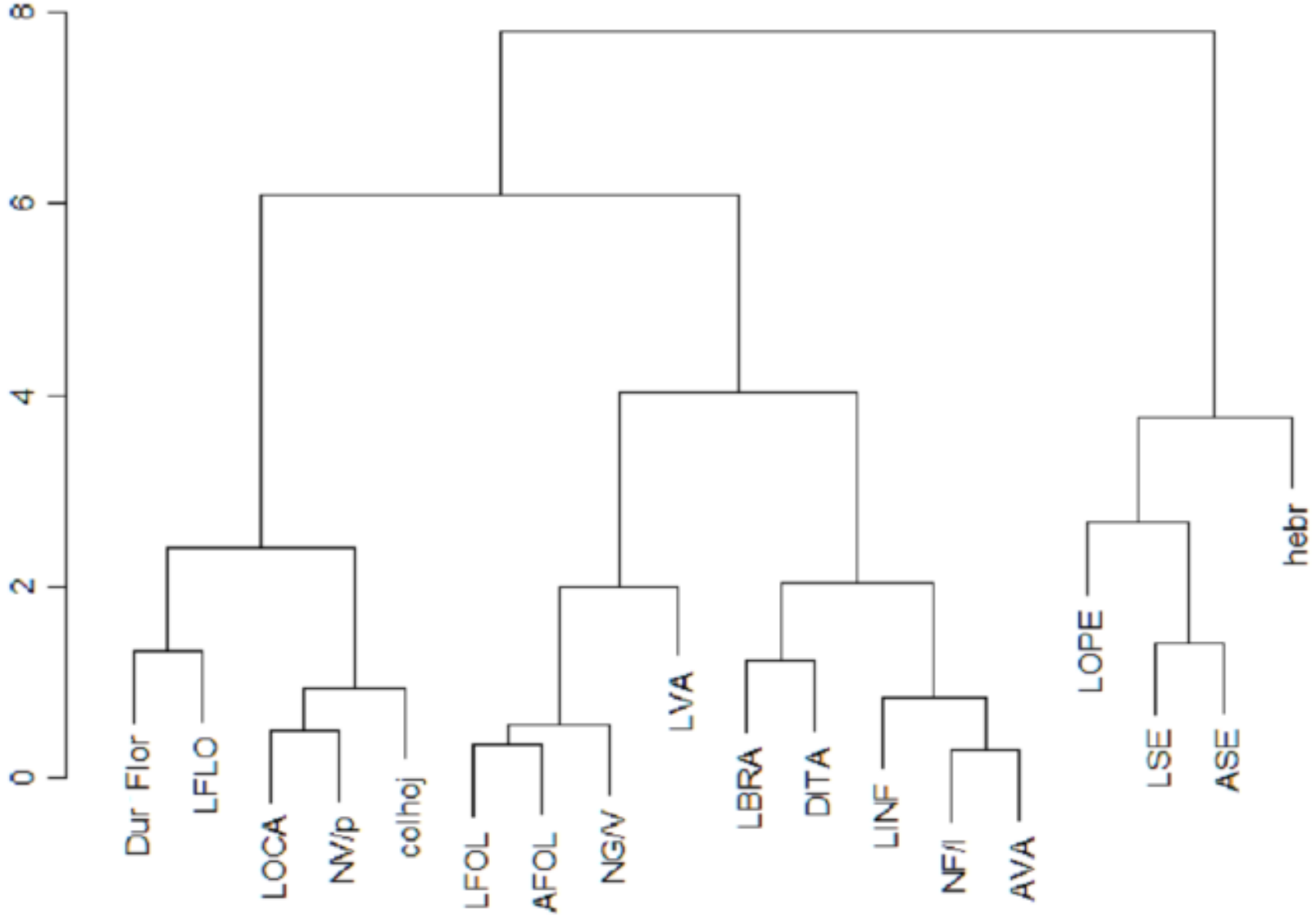


Figura 6.11: Fabes asturianas: árbol jerárquico de las variables.