TÉCNICAS DE AGRUPAMIENTO

José D. Martín Guerrero, Emilio Soria, Antonio J. Serrano

PROCESADO Y ANÁLISIS DE DATOS AMBIENTALES

Curso 2009-2010

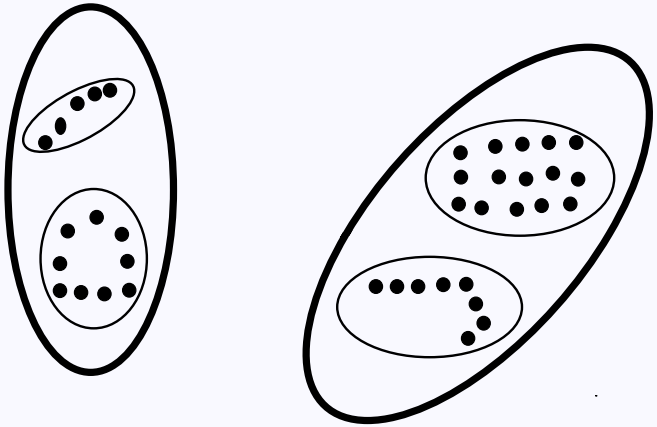
1. Esquema.

- Introducción.
- Algoritmo de las C-Medias.
- Algoritmos de agrupamiento jerárquico.
- Algoritmo *Expectation-Maximization* E-M.
- Red ART2.
- Algunos comentarios finales.

2. Introducción.

GRUPO O CLUSTER

Def.- Región continua del espacio que contiene una densidad relativamente alta de puntos, y que se encuentra a su vez separada de otras regiones de alta densidad por regiones cuya densidad de puntos es relativamente baja (Everitt, 1981).

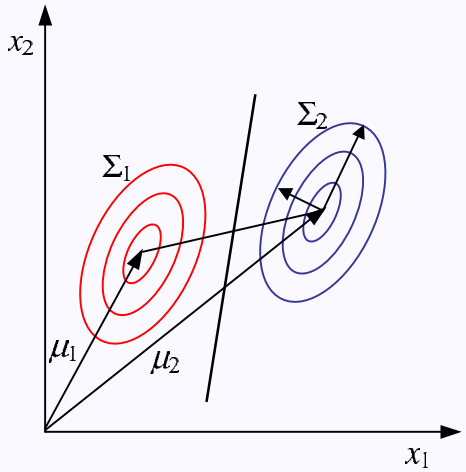


¿Cuántos *clusters* hay? ¿2 ó 4?

2.1. Medidas de proximidad.

- Medidas de similitud: producto escalar, medida de Tanimoto.

- Medidas de disimilitud (distancias).
  - Distancia euclídea.
  - Distancia de Mahalanobis.
  - Distancia de Bhattacharyya.



2.2. Validación del agrupamiento.

¿Número adecuado de grupos *a priori* para una distribución dada?

$$\Downarrow$$

$\nexists$ una respuesta clara.

Tests utilizados para decidir el número correcto de *clusters*:

- Análisis de la bondad del agrupamiento.
- Estudio de la normalidad de los grupos encontrados.
- Índice de Dunn:

$$D_M = \min_{i=1,\ldots,M} \left\{ \min_{j=1,\ldots,M,\ j\neq i} \left(\frac{d(C_i, C_j)}{\max_{k=1,\ldots,M} diam(C_k)} \right) \right\}$$

  para un número M de grupos.

- Índice de Davies-Bouldin:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$R_i = \max_{j=1,\ldots,M, j\neq i} R_{ij},\ i = 1, \ldots, M$$

$$DB_M = \frac{1}{M} \sum_{i=1}^{M} R_i$$

siendo s_i una medida de dispersión para el grupo i.

3. Algoritmo de las C-Medias.

C-Medias busca la minimización de la función de coste:

$$J(\Theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{M} u_{ij} \|x_i - \Theta_j\|^2 \tag{1}$$

$$u_{ij} = \left\{ \begin{array}{ll} 1, & d(x_i, \Theta_j) = \min_{k=1,\ldots,M} d(x_i, \Theta_k) \\ 0, & \textit{en otro caso} \end{array} \right\} \quad i = 1, \ldots, N \tag{2}$$

donde $d(x_i, \Theta_j)$ es la distancia entre el patrón i-ésimo y el prototipo j-ésimo, N el número de patrones y M el número de grupos.

Variantes del algoritmo:

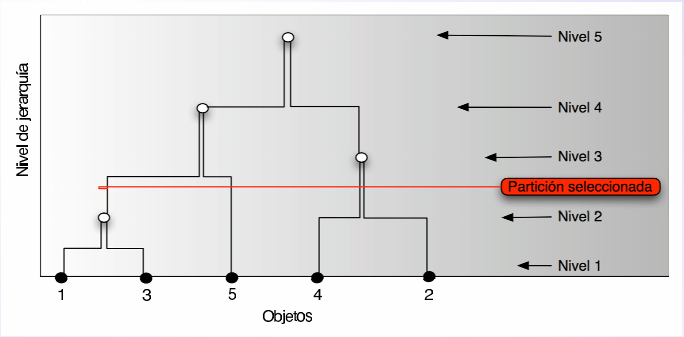
- Isodata: detección automática del número de grupos.
- C-Medias Difuso: Pertenencia difusa del usuario i al grupo j:

$$\mu_{ij} = \frac{\|x_i - \Theta_j\|^{(-\frac{2}{m-1})}}{\sum_{j=1}^{M} \|x_i - \Theta_j\|^{(-\frac{2}{m-1})}} \Rightarrow J(\Theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{M} \mu_{ij}^m \|x_i - \Theta_j\|^2$$

1. Si $m = 1 \Rightarrow$ Agrupamiento no difuso.
2. Si $m > 1 \Rightarrow$ Agrupamiento difuso (tanto más cuanto mayor sea el valor de m).

4. Algoritmos de agrupamiento jerárquico.

Jerarquía de agrupamientos anidados a través de un proceso iterativo, que puede ser de naturaleza **acumulativa** o divisiva.



ALGORITMO ACUMULATIVO:

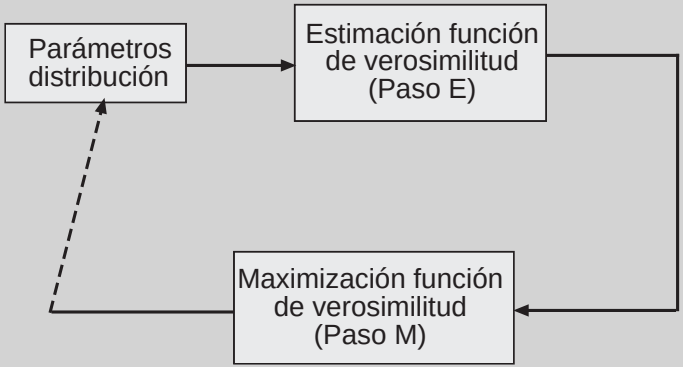
1. Agrupamiento inicial: $R_0 \equiv \{C_i = \{\mathbf{x_i}\},\ i = 1, \ldots, N\}$.
2. Las iteraciones van avanzando $t \rightarrow t+1$.
3. Cálculo de distancias $d(C_r, C_s)$: distancias entre todos los *clusters* posibles C_r, C_s del agrupamiento anterior R_{t-1}, selecccionando la pareja de *clusters* C_i, C_j que presentan la mínima distancia.
4. Se define un nuevo grupo $C_q = C_i \cup C_j$. El nuevo agrupamiento es $R_t \equiv \left\{R^*_{t-1} \cup C_q\right\}$.
5. Vuelve a avanzarse la iteración hasta que queda un único *cluster*.

En cada iteración solamente será necesario calcular las distancias entre el nuevo *cluster* formado C_q y el resto de grupos. Algoritmos utilizados:

1. Actualización de centroides (pesados por el número de patrones).
2. Algoritmos basados en la fórmula de Lance y Williams:
   - Algoritmo de enlace simple:
     $d(C_q, C_s) = \min\{d(C_i, C_s), d(C_j, C_s)\}$
   - Algoritmo de enlace completo:
     $d(C_q, C_s) = \max\{d(C_i, C_s), d(C_j, C_s)\}$
   - Algoritmo de promedio no pesado:
     $d(C_q, C_s) = \frac{1}{2}\left[d(C_i, C_s) + d(C_j, C_s)\right]$
   - Algoritmo de promedio pesado:
     $d(C_q, C_s) = \frac{1}{n_i+n_j}\left[n_i \cdot d(C_i, C_s) + n_j \cdot d(C_j, C_s)\right]$
   - Algoritmo de centroide no pesado:
     $d(C_q, C_s) = \frac{1}{2}d(C_i, C_s) + \frac{1}{2}d(C_j, C_s) - \frac{1}{4}d(C_i, C_j)$
   - Algoritmo de centroide pesado:
     $d(C_q, C_s) = \frac{n_i}{n_i+n_j}d(C_i, C_s) + \frac{n_j}{n_i+n_j}d(C_j, C_s) - \frac{n_i n_j}{(n_i+n_j)^2}d(C_i, C_j)$

5. Algoritmo *Expectation-Maximization* E-M.

Maximiza la esperanza de la función de verosimilitud de un determinado vector de parámetros $\Psi(t)$ sobre un conjunto de muestras.



Aplicación en la estimación de mezclas de distribuciones Gaussianas (parámetros a estimar: media μ y desviación estándar σ).

Procedimiento:

1. Estimación inicial de los parámetros (inicialización con C-Medias, por ejemplo): $\Psi(0)$.
2. Cálculo iterativo de los parámetros, concluyéndose cuando $\|\Psi(t+1) - \Psi(t)\| \leq \epsilon$. Las sucesivas estimaciones $\Psi(t)$ nunca hacen decrecer la verosimilitud.

6. Red ART2.

Redes basadas en la Teoría de la Resonancia Adaptativa (*Adaptive Resonance Theory*, ART):

- Ventaja frente a otros algoritmos de agrupamiento: No es necesario conocer *a priori* el número de grupos subyacente en la ditribución.
- Red ART1: Patrones con valores binarios.
- Red ART2: Patrones con valores continuos.

Procedimiento:

1. Dado un patrón de entrada $\mathbf{s_i}$, encontrar el prototipo P más parecido a este patrón.
2. Comprobar si P es una buena representación de $\mathbf{s_i}$:
   $\mathbf{s_i} \cdot \mathbf{P} \geq \alpha \cdot \sum_j \mathbf{s_i^j}$?? (α es la constante de aprendizaje).
3. Si se pasa el anterior test, entonces se pasa el test de vigilancia:
   $\mathbf{s_i} \cdot \mathbf{P} \geq \rho$?? ($\rho$ es el parámetro de vigilancia).
4. Si cualquiera de los dos anteriores tests no se pasa, se inicializa un nuevo prototipo.

7. Algunos comentarios finales.

- C-Medias y C-Medias difuso funcionan bien cuando el conjunto de datos a agrupar no es demasiado complejo, y en particular con grupos compactos y ¡¡¡DEL MISMO TAMAÑO!!!.
- El algoritmo que presenta un mejor comportamiento para conjuntos de dimensionalidad alta suele ser de entre los presentados aquí ART2, destacando que ART2 tiene además la ventaja adicional de que no es necesario conocer *a priori* el número de grupos. Para conjuntos de dimensionalidad media, es recomendable asimismo la utilización de este algoritmo aunque también destaca el rendimiento de E-M y los algoritmos de agrupamiento jerárquico. Para conjuntos de baja dimensionalidad prácticamente todos los algoritmos ofrecen un rendimiento aceptable.
- Realimentación incorporando nuevos comportamientos que puedan aparecer ⇒ Modelos sencillos basados en *Learning Vector Quantization* (LVQ), obteniéndose un esquema final semi-supervisado.
- Red neuronal supervisada con parte del algoritmo de aprendizaje no supervisado ⇒ Red RBF (Radial Basis Function), que realiza mapeos locales de un conjunto de entrada en un conjunto de salida.
- Red neuronal que asocia patrones de entrada similares a neuronas cercanas simulando comportamiento cerebral ⇒ SOM (Self-Organizing Map), que preserva las relaciones topológicas con los patrones de entrada. Más pensada para visualización que para agrupamiento de patrones.