Clustering

Christopher Expósito Izquierdo
Airam Expósito Márquez
Israel López Plata
Belén Melián Batista
J. Marcos Moreno Vega

{**cexposit, aexposim, ilopezpl, mbmelian, jmmoreno**}**@ull.edu.es**

Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna

Índice

INTRODUCCIÓN

Propósito

- Agrupar los elementos de un conjunto en grupos (*cluster*) que tengan significado o sean útiles.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar [1]

- La utilidad de los grupos de objetos depende del objetivo del análisis y del dominio de aplicación.
- En general, los grupos están formados por elementos similares o relacionados entre sí.

Dificultades

- En aplicaciones prácticas es habitual que no exista una noción bien definida de cluster.
- En general, el número de clusters es desconocido, lo que supone una dificultad añadida.
- La noción de cluster está íntimamente relacionada con el concepto de distancia o similitud.
- La amplia variedad de aplicaciones y tipos de datos hace que existan diferentes modelos y formas de agrupar un conjunto de objetos.

¿Cuántos grupos?

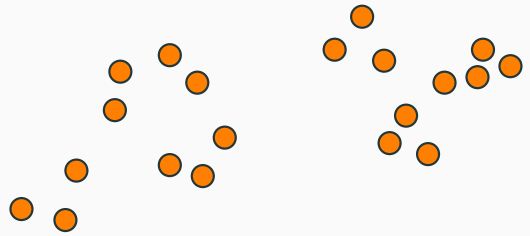

Figura 1: Datos originales.

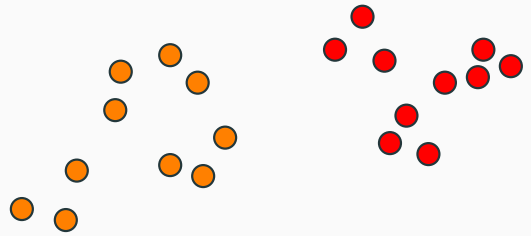

Figura 2: Dos clusters.

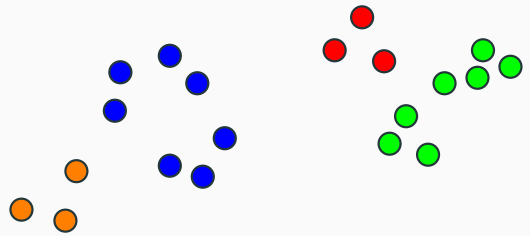

Figura 3: Cuatro clusters.

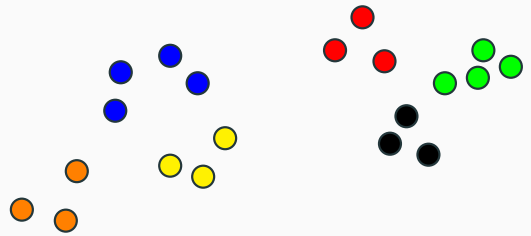

Figura 4: Seis clusters.

Tipos de clusters

- **Bien separados.** Los objetos de cada cluster están más cercanos (son más similares) a los objetos de su propio cluster que a los objetos de cualquier otro cluster.
- **Basados en prototipos.** Los objetos de cada cluster están más cercanos (son más similares) al prototipo que define al cluster que a los prototipos del resto de clusters.
- **Basados en densidad.** Un cluster es una región densa del espacio de objetos rodeada por una región de baja densidad.

Tipos de clusters

Figura 5: Bien separados.

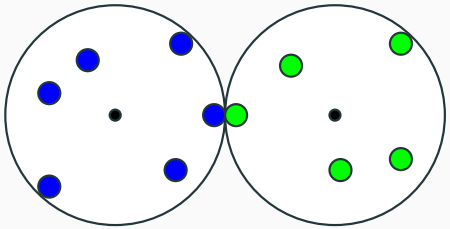

Figura 6: Basados en prototipos.

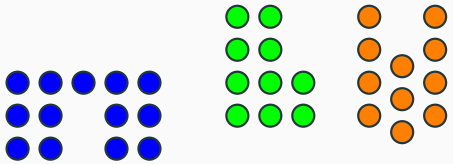

Figura 7: Basados en densidad.

Funciones de distancia o similitud (i)

- **Atributos numéricos (i)**

Dados $x = (x_1, x_2, \ldots, x_n)$ e $y = (y_1, y_2, \ldots, y_n)$

- Distancia euclídea

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Distancia rectilínea (o de Manhattan)

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Funciones de distancia o similitud (ii)

- **Atributos numéricos (ii)**
  - Distancia de Chebychev

$$d(x, y) = \max_{1 \leq i \leq n} |x_i - y_i|$$

  - Distancia del coseno

$$d(x, y) = \arccos \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Funciones de distancia o similitud (iii)

- **Atributos numéricos (iii)**
  - Suele ser conveniente normalizar el valor de los atributos antes de calcular las distancias.
  - La normalización pretende evitar que la distancia se vea altamente influenciada por los atributos con valores altos.
  - Es conveniente también detectar previamente los valores anómalos o atípicos ya que estos suelen afectar gravemente a la normalización empleada.

Funciones de distancia o similitud (iv)

- **Atributos nominales**
  - Distancia

$$d(x, y) = \omega \sum_{i=1}^{n} \delta(x_i, y_i)$$

con

$$\delta(a, b) = \begin{cases} 0 & \text{si } a = b \\ 1 & \text{si } a \neq b \end{cases}$$

Funciones de distancia o similitud (v)

- **Tiras de caracteres**
  - Distancia de edición

  Se ponderan las inserciones, borrados y sustitucines de caracteres necesarios para obtener una tira de carecteres desde la otra.

- **Conjuntos**
  - Índice Jaccard

$$d(x, y) = \frac{|x \cup y| - |x \cap y|}{|x \cup y|}$$

BIBLIOGRAFÍA

Bibliografía

Tan, P.-N., Steinbach, M., and Kumar, V.
Introduction to Data Mining.
Addison-Wesley, 2006.

Licencia