



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Clustering: algoritmo C-medias ¹

Alfons Juan
Albert Sanchis
Jorge Civera

DSIC

Departamento de Sistemas
Informáticos y Computación

¹Para una correcta visualización, se requiere Acrobat Reader v. 7.0 o superior

Objetivos formativos

- Analizar el problema del clustering particional bajo el ***criterio Suma de Errores Cuadráticos***
- Aplicar el ***algoritmo C -medias de Duda y Hart***
- Aplicar el ***algoritmo C -medias convencional***

Índice

1	Clustering particional	3
2	Criterio “Suma de Errores Cuadráticos” (SEC)	4
3	Algoritmo C -medias de Duda y Hart	6
4	Algoritmo C -medias convencional	9
5	Optimalidad de los algoritmos C -medias	11

1. Clustering particional

El *aprendizaje no supervisado* o *clustering* es un problema clásico del *aprendizaje automático*

El *clustering particional* es una de sus aproximaciones usuales:

- Asumimos disponible una *función criterio* J para evaluar la calidad de cualquier partición de N datos en C clústers:

$$J(\Pi) \quad : \quad \Pi = \{X_1, \dots, X_C\}$$

- El problema del clustering se aproxima como:

$$\Pi^* = \arg \min_{\Pi = \{X_1, \dots, X_C\}} J(\Pi)$$

2. Criterio “Suma de Errores Cuadráticos” (SEC)

La SEC de una partición $\Pi = \{X_1, \dots, X_C\}$ se define como:

$$J(\Pi) = \sum_{c=1}^C J_c$$

donde J_c es la **distorsión** del clúster c ,

$$J_c = \sum_{\mathbf{x} \in X_c} \|\mathbf{x} - \mathbf{m}_c\|^2,$$

siendo \mathbf{m}_c la **media** o **centroide** del clúster c ,

$$\mathbf{m}_c = \frac{1}{n_c} \sum_{\mathbf{x} \in X_c} \mathbf{x}$$

y n_c su talla.

Ejemplo de cálculo de la SEC

3. Algoritmo *C*-medias de Duda y Hart

Dada una partición $\Pi = \{X_1, \dots, X_C\}$, el incremento de la SEC debido a la transferencia de un dato \mathbf{x} del clúster i al j es:

$$\Delta J = \frac{n_j}{n_j + 1} \|\mathbf{x} - \mathbf{m}_j\|^2 - \frac{n_i}{n_i - 1} \|\mathbf{x} - \mathbf{m}_i\|^2$$

La transferencia será provechosa si $\Delta J < 0$, esto es, si:

$$\frac{n_j}{n_j + 1} \|\mathbf{x} - \mathbf{m}_j\|^2 < \frac{n_i}{n_i - 1} \|\mathbf{x} - \mathbf{m}_i\|^2$$

Dada una partición inicial, el **algoritmo *C*-medias de Duda y Hart [1, 2]** aplica transferencias provechosas sucesivas ...

Algoritmo C -medias de Duda y Hart (cont.)

- **Entrada:** una partición inicial, $\Pi = \{X_1, \dots, X_C\}$
- **Salida:** una partición optimizada, $\Pi^* = \{X_1, \dots, X_C\}$

- **Método:**

Calcular medias y J

repetir

para todo dato x

Sea i el clúster en el que se encuentra x

Hallar un $j^* \neq i$ que minimice ΔJ al transferir x de i a j^*

Si $\Delta J < 0$, transferir x de i a j^* y actualizar medias y J

hasta no encontrar transferencias provechosas

Ejemplo: aplicación del C -medias de Duda y Hart

4. Algoritmo C -medias convencional

La condición de Duda y Hart se cumple si se cumple la condición:

$$\|x - m_j\|^2 < \|x - m_i\|^2$$

Esta condición es la base del algoritmo C -medias convencional:

- **Entrada:** una partición inicial
- **Salida:** una partición optimizada
- **Método:**

repetir

Calcular las medias de los clústers

Reclasificar los datos según sus medias más cercanas

hasta que no se reclasifique ningún dato

Ejemplo: aplicación del C -medias convencional

5. Optimalidad de los algoritmos *C*-medias

- Ninguna de las versiones del algoritmo *C*-medias garantiza la obtención de un mínimo global de la SEC
- La versión de Duda & Hart obtiene un *mínimo local*
- La versión “popular” no garantiza la minimización local en algunos casos

Referencias

- [1] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001.