UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Clustering: algoritmo C-medias [1]

Alfons Juan
Albert Sanchis
Jorge Civera

DSIC
Departamento de Sistemas
Informáticos y Computación

[1] Para 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

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:

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

- El problema del clustering se aproxima como:

$$\Pi^* = \underset{\Pi=\{X_1,\ldots,X_C\}}{\arg\min} \; J(\Pi)$$

2. Criterio "Suma de Errores Cuadráticos" (SEC)

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

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

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

norma cuadrado

$a \in \mathbb{R}^2$

$$J_c = \sum_{\boldsymbol{x} \in \underline{X_c}} \underline{\|\boldsymbol{x} - \boldsymbol{m}_c\|^2}, \qquad \|a\| = \cancel{\left(\sqrt{a_1^2 + a_2^2}\right)^2}$$

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

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

y n_c su talla.

Ejemplo de cálculo de la SEC

$C = 2$

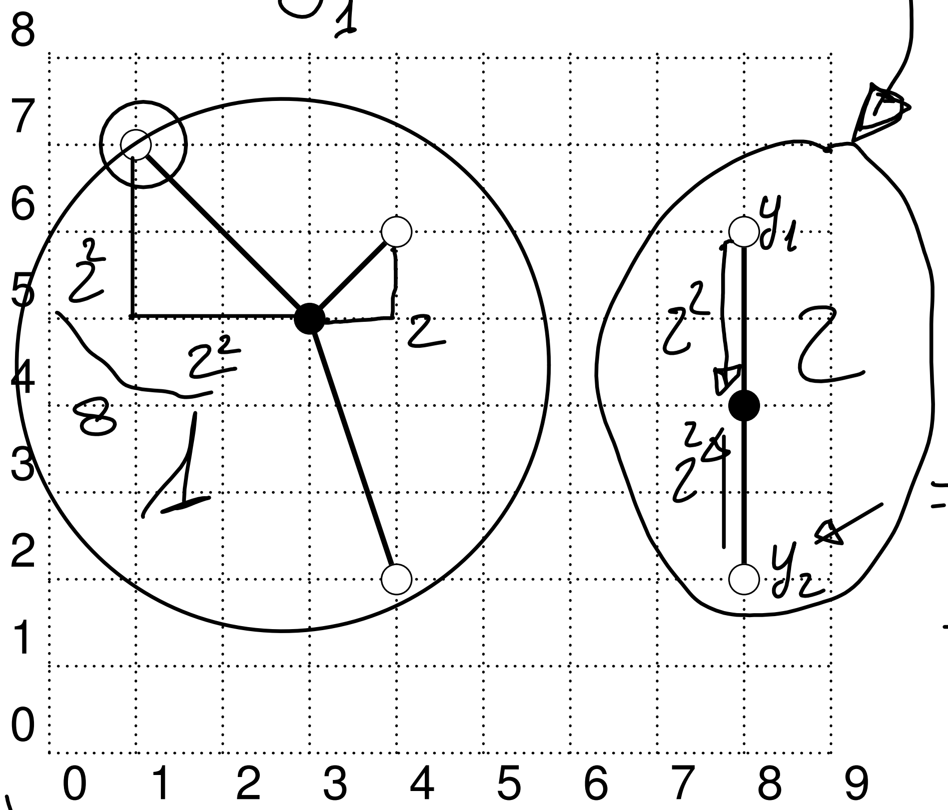
$\Pi = \{X_1, X_2\}$

$X_1 = \left\{ \begin{pmatrix}4\\2\end{pmatrix}, \begin{pmatrix}4\\6\end{pmatrix}, \begin{pmatrix}1\\7\end{pmatrix} \right\}$

$X_2 = \left\{ \begin{pmatrix}8\\2\end{pmatrix}, \begin{pmatrix}8\\6\end{pmatrix} \right\}$

$m_1 = \frac{1}{3}\left(\begin{pmatrix}4\\2\end{pmatrix} + \begin{pmatrix}4\\6\end{pmatrix} + \begin{pmatrix}1\\7\end{pmatrix} \right) = \frac{1}{3}\begin{pmatrix}9\\15\end{pmatrix} = \begin{pmatrix}3\\5\end{pmatrix}$

$m_2 = \frac{1}{2}\left(\begin{pmatrix}8\\2\end{pmatrix} + \begin{pmatrix}8\\6\end{pmatrix} \right) = \begin{pmatrix}8\\4\end{pmatrix}$



$J_2 = \|y_1 - m_2\|^2 + \|y_2 - m_2\|^2$

$= \left\| \begin{pmatrix}8\\2\end{pmatrix} - \begin{pmatrix}8\\4\end{pmatrix} \right\|^2 + \left\| \begin{pmatrix}8\\6\end{pmatrix} - \begin{pmatrix}8\\4\end{pmatrix} \right\|^2$

$= \left\| \begin{pmatrix}0\\-2\end{pmatrix} \right\|^2 + \left\| \begin{pmatrix}0\\2\end{pmatrix} \right\|^2 =$

$= \left(\sqrt{0^2 + (-2)^2}\right)^2 + \left(\sqrt{0^2 + 2^2}\right)^2$

$= 4 + 4 = 8$

SEC $\quad J(\Pi) = 28$

$$J = J_1 + J_2 = 20 + 8 = \underline{\underline{28}}$$

$$J_1 = 8 + 10 + 2 = 20$$

$$J_2 = 4 + 4 = 8$$

3. Algoritmo C-medias de Duda y Hart

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

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

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

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

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, $\Pi = \{X_1, \ldots, X_C\}$
- ***Salida:*** una partición optimizada, $\Pi^* = \{X_1, \ldots, X_C\}$
- ***Método:***

  Calcular medias y J

  repetir

  para todo dato $\boldsymbol{x}$

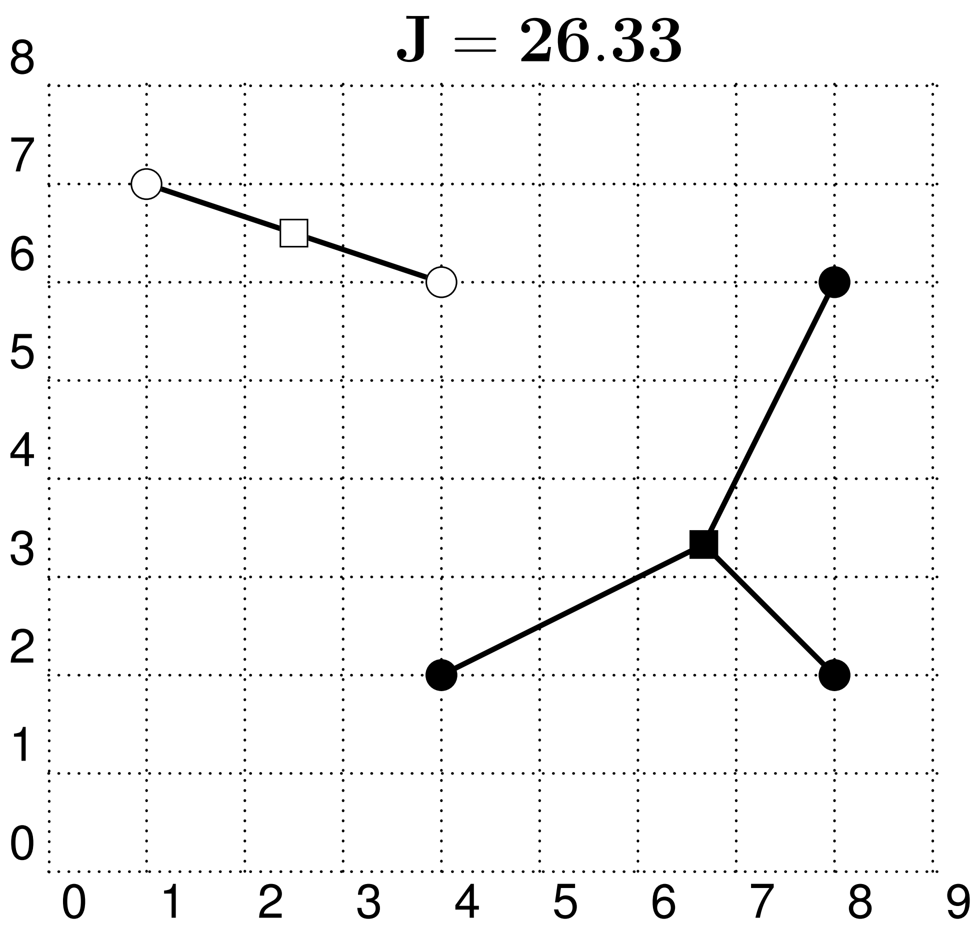
  Sea i el clúster en el que se encuentra $\boldsymbol{x}$

  Hallar un $j^* \neq i$ que minimice $\triangle J$ al transferir $\boldsymbol{x}$ de i a j^*

  Si $\triangle J < 0$, transferir $\boldsymbol{x}$ de i a j^* y actualizar medias y J

  hasta no encontrar transferencias provechosas

Ejemplo: aplicación del C-medias de Duda y Hart



Partición optimizada

4. Algoritmo C-medias convencional

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

$$\|\boldsymbol{x} - \boldsymbol{m}_j\|^2 < \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

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

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

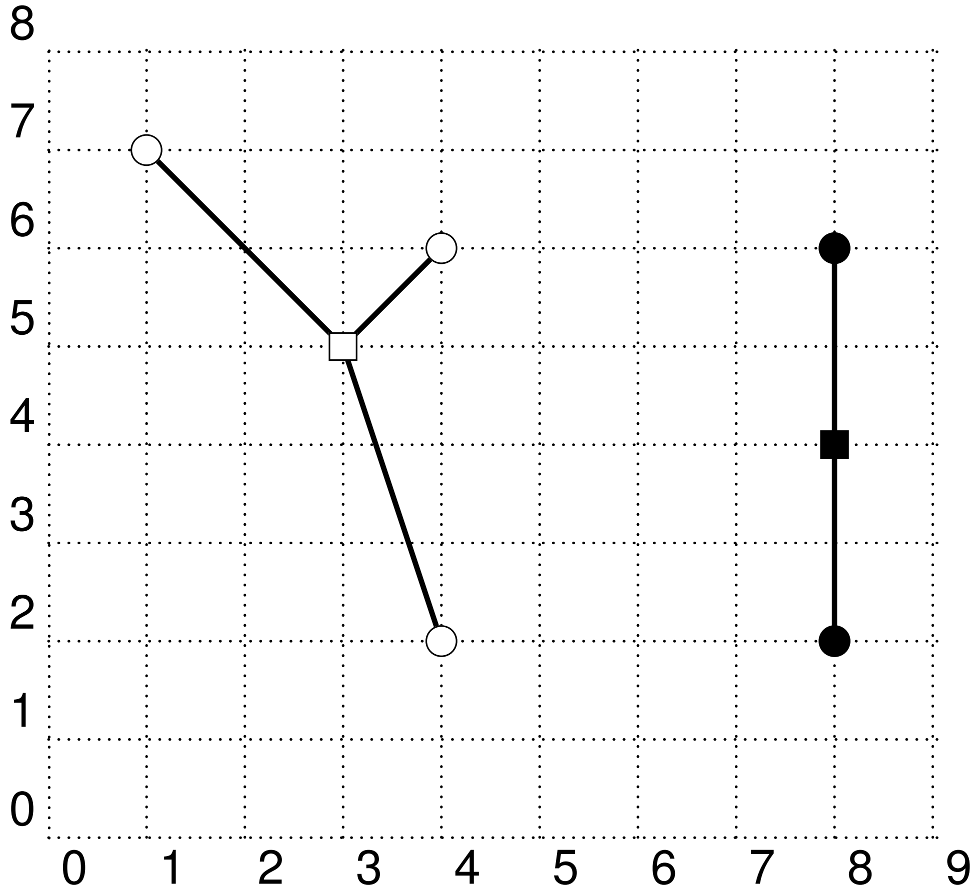
  repetir

  Calcular las medias de los clústers

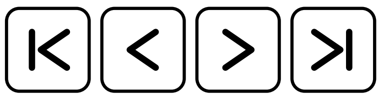
  Reclasificar los datos según sus medias más cercanas

  hasta que no se reclasifique ningún dato

Ejemplo: aplicación del C-medias convencional



Partición optimizada



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

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.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA