UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Clustering: algorisme C-mitjanes

Alfons Juan
Albert Sanchis
Jorge Civera

DSIC

Departament de Sistemes Informàtics i Computació

Objectius formatius

- Analitzar el problema del clustering particional sota el ***criteri Suma d'Errors Quadràtics***
- Aplicar ***l'algorisme C-mitjanes de Duda i Hart***
- Aplicar ***l'algorisme C-mitjanes convencional***

Índex

1 Clustering particional

L'aprenentatge no supervisat o ***clustering*** és un problema clàssic de ***l'aprenentatge automàtic***

Una aproximació usual és el ***clustering particional:***

- Assumim disponible una ***funció criteri*** J per a avaluar la qualitat de qualsevol partició de N dades en C clústers:

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

- El problema del clustering s'aproxima com:

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

2 Criteri "Suma d'Errors Quadràtics" (SEQ)

La SEQ d'una partició $\Pi = \{X_1, \ldots, X_C\}$ es defineix com:

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

on J_c és la ***distorsió*** del clúster c,

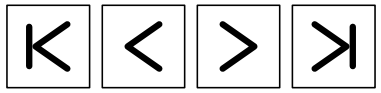
$$J_c = \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{m}_c\|^2,$$

sent $\boldsymbol{m}_c$ la ***mitjana*** o ***centroide*** del clúster c,

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

i n_c la seua talla.

Exemple de càlcul de la SEQ



3 Algorisme C-mitjanes de Duda i Hart

Donada una partició $\Pi = \{X_1, \ldots, X_C\}$, l'increment de la SEQ a causa de la transferència d'una dada $\boldsymbol{x}$ del clúster i al j és:

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

La transferència serà profitosa si $\triangle J < 0$, açò és, si:

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

Donada una partició inicial, ***l'algorisme C-mitjanes de Duda i Hart [1, 2]*** aplica transferències profitoses successives ...

Algorisme C-mitjanes de Duda i Hart (cont.)

- ***Entrada:*** una partició inicial, $\Pi = \{X_1, \ldots, X_C\}$
- ***Eixida:*** una partició optimitzada, $\Pi^* = \{X_1, \ldots, X_C\}$
- ***Mètode:***

  Calcular mitjanes i J

  repetir

    per a tota dada $\boldsymbol{x}$

      Siga i el clúster en el qual es troba $\boldsymbol{x}$

      Trobar un $j^* \neq i$ que minimitze $\triangle J$ en transferir $\boldsymbol{x}$ d'i a j^*

      Si $\triangle J < 0$, transferir $\boldsymbol{x}$ d'i a j^* i actualitzar mitjanes i J

  fins a no trobar transferències profitoses

Exemple: aplicació del C-mitjanes de Duda i Hart

4 Algorisme C-mitjanes convencional

La condició de Duda i Hart es compleix si es compleix la condició:

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

Aquesta condició és la base del C-mitjanes convencional:

- ***Entrada:*** una partició inicial
- ***Salida:*** una partició optimitzada
- ***Mètode:***

  repetir

    Calcular les mitjanes dels clústers

    Reclasificar les dades segons les mitjanes més properes

  fins que no es reclasifique cap dada

Exemple: aplicació del C-mitjanes convencional

Referències

[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.