Clasificación de patrones: Métodos no supervisados

Jordi Porta Zamorano

Escuela Politécnica Superior **Dept. de Lingü´ıstica Computacional**

Universidad Autónoma de Madrid **Real Academia Española**

jordi.porta@uam.es **porta@rae.es**

abril de 2005

Agrupamiento (*clustering*)

Los métodos no supervisados (especialmente el *clustering*) se usan cuando:

- recopilar y clasifi car a mano es costoso
- la caracterización de los patrones cambia con el tiempo (o con el corpus)

por otro lado:

- permite encontrar caracterizaciones útiles para construir clasifi cadores
- el descubrimiento de grupos y subgrupos que revele la naturaleza de la estructura del problema

Agrupamiento: distancias

Las técnicas de agrupamiento se basan fundamentalmente en el concepto de similitud (o disimilitud) entre ejemplos y agrupaciones. Muchas veces se utilizan *métricas* (o *distancias*) para medir la similitud entre ejemplos. Las métricas más usadas son:

- distancia de Minkowski: Se trata de una família de métricas con la forma general:

$$L_q(\vec{x},\vec{y}) = (\sum_{i=1}^{n} |x_i - y_i|^q)^{1/q}$$

de entre las que destacan las siguientes:

- Manhattan (o *city block*) ($q = 1$):

$$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

- Euclidea ($q = 2$):

$$L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Agrupamiento: distancias

- de Hamming: normalmente aplicada a vectores binarios, da el número de componentes con valores distintos.

$$H(\vec{x},\vec{y}) = |\{i \,|1 \leq i \leq n \wedge x_i \neq y_i\}|$$

P. ej.: La distancia de Hamming para los vectores $(1, 0, 1, 0, 1)$ y $(0, 1, 1, 1, 0)$ es 4.

- de Tanimoto: tambien aplicada a dos vectores binarios $\vec{x}$ e $\vec{y}$:

$$T(S_1, S_2) = \frac{|S_1| + |S_2| - 2|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|}$$

donde S_1 son las componentes de $\vec{x}$ con valor 1 y S_2 las de $\vec{y}$.

- ...

Agrupamiento: medidad de similitud

- No siempre la medida de similitud es una distancia:
  - *distancia del coseno*:
    $$\cos(\vec{x},\vec{y}) = \sum_i \frac{x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$
    usada en *recuperación de la información* donde un documento se representa por un vector de pesos
  - *Distancia simétrica de Kullback-Leibler*:
    $$\mathrm{KL}(p||q) = \sum_i p_i \times \log_2 \frac{p_i}{q_i}$$
  - *Distancia de* χ^2:
  - *Distancia de edición*: puede ser de aplicación a la comparación de *rankings*
  - ...

Agrupamiento: evaluación

La calidad de un agrupamiento puede medirse con:

- *similitud intragrupo*: normalmente la varianza de las similitudes de los elementos del agrupamiento
- *disimilitud intergrupo*: normalmente la disimilitud entre los centroides

Agrupamiento: técnicas

- Existen muchas técnicas de agrupamiento, las más usadas son las de *agrupamiento jerárquico* y las de *agrupamiento dinámico* que a su vez pueden ser por *reunión* o *separación*.
- Al igual que en los métodos de clasificación supervisados, no existe una técnica de agrupamiento de aplicación universal.

Agrupamiento jerárquico

- agrupamiento:

  crear un grupo con cada ejemplo
  while haya más de un grupo **do**
  buscar los dos grupos más similares y fusionarlos en uno
  end while

El algoritmo de agrupamiento admite tantas variantes como funciones de similitud puedan definirse. Los algoritmos más populares para calcular la similitud entre dos agrupaciones son:

- *single-link*: la mínima de las distancias entre todos los pares de ejemplos de cada agrupación.
- *complete-link*: la máxima de las distancias entre todos los pares de ejemplos de cada agrupación.
- *group average*: la media de las distancias entre todos los pares de ejemplos de cada agrupación.

Agrupamiento dinámico

También conocido como k-means o *Iterative Distance-based Clustering*. Necesita que se le proporcione a priori el número de grupos k. El algoritmo es el siguiente:

- agrupamiento:

  seleccionar al azar k ejemplos como centros iniciales de cada grupo;
  repeat
    asignar cada ejemplo al grupo con menor distancia a su centro;
    recalcular los nuevos centros de cada grupo;
  until los grupos sean estables

- Los centros de cada grupo, también denominados *centroides* pueden corresponderse con ejemplos o no, en ese caso se les denomina *prototipos*.
- Los grupos se consideran *estables* cuando los ejemplos no cambian de grupo respecto la iteración anterior.
- Se pueden obtener agrupaciones jerarquizadas tomando $k = 2$ y aplicando el mismo algoritmo de manera recursiva sobre los ejemplos de cada grupo.

COBWEB/CLASSIT

- son dos sistemas de agrupamiento clásicos
- COWEB funciona con atributos cualitativos y CLASSIT para cuantitativos
- devuelve un árbol en el que:
  - las hojas son ejemplos
  - los nodos son agrupaciones
- es un algoritmo incremental, la adición de un nuevo ejemplo:
  - queda absorvido por un nodo
  - provoca la restructuración del árbol (*split*/*merge*)

  según una función de utilidad que mide cuánto puede un agrupamiento predeci el valor de los atributos de sus ejemplos:

$$CU(C_1, \ldots, C_k) = \frac{\sum_l p(C_l) \sum_i \sum_j (p(a_i = v_{ij}|C_l)^2 - p(a_i = v_{ij})^2)}{k}$$

  en la que C_l son agrupaciones, a_i atributos y v_{ij} los valores del atributo a_i