Curso sobre Modelamiento Estadístico y Sistemas Recomendadores

Clase Presencial 4: Técnicas de Agregació

(Aprendizaje no supervisado)

4.1.- Introducción

Definición

`Clustering` o análisis de conglomerados es el proceso de particionar un conjunto de observaciones en subconjuntos.

Se busca que los objetos dentro de un mismo grupo sean lo más similares posible entre sí y lo más diferentes con los de otro.

Introducción

Gráficamente

(a) Original points.

(b) Two clusters.

(c) Four clusters.

(d) Six clusters.

Introducción

- Las agrupaciones son realizadas por algoritmos, por lo que es útil para detectar agrupamientos previamente desconocidos.
- Diferentes métodos de agrupamiento pueden llevar a diferentes grupos de los mismos datos.
- Un cluster o grupo de objetos pueden tratarse como pertenecientes a una clase implícita $\Longrightarrow$ clasificación automática.
- Aplicaciones: inteligencia de negocios, reconocimiento de imágenes, búsquedas Web.

Introducción

- En algunos contextos, esta técnica se conoce como segmentación de datos.
- También puede utilizarse para detectar *outliers:* fraudes bancarios, actividades criminales en comercio electrónico.

Introducción

Para que un algoritmo sea útil como herramienta necesita cumplir con los siguientes requisitos:

- Escalabilidad: útiles en grandes bases de datos (millones o billones de objetos).
- Capaz de manejar diferentes tipos de atributos: numéricos, binarios, nominales, ordinales o mezclas de ellos.
- Encontrar clusters con cualquier forma: muchos algoritmos se basan en distancias que sólo detectan clusters esféricos.
- Requisitos sobre conocimiento del contexto para determinar los parám de entrada.

Introducción

- Robustos frente a la presencia de datos “ruidosos”.
- Métodos incrementares y no sensibles al orden de llegada de los registros.
- Capaz de agrupar datos con alta dimensionalidad.
- Realizar clustering sujeto a ciertas restricciones.
- Interpretable, comprensible y útil.

Introducción

- Criterio de partición: hay métodos en los que no hay jerarquía entre los clusters y otros que separan los datos jerárquicamente.
- Separación de los clusters: separación de los objetos en grupos mutuamente excluyentes o no.
- Medidas de similaridad: algoritmos basados en distancias vs. algoritmos basados en densidades o métodos continuos.
- Búsqueda de clusters en subespacios de los datos.

4.2.- Métodos de Partición en Número Fijo

Métodos de Partición en Número Fijo

- Es el método más simple y fundamental de análisis de conglomerados. Organiza los datos en varios grupos exclusivos, cuya cantidad es fijada de antemano.

- Dado un conjunto de datos, D, de n objetos y k el número de grupos a formar, el algoritmo distribuye los objetos de D en los grupos, $C_1, \ldots, C_k$, donde $C_i \subset D$ y $C_i \cap C_j = \emptyset$, para $1 \leq i, j \leq k$. De esta manera, cada partición representa un grupo o `cluster`.

Métodos de Partición en Número Fijo

- Los grupos se forman optimizando un criterio de particionamiento, como puede ser la función de disimilaridad basada en distancia, tal que los objetos dentro de un grupo son similares y diferentes a los de otros, en términos de las variables que los describen.

- Se utiliza una función para evaluar la calidad de la partición, que pretende lograr alta similaridad intra-cluster (dentro del grupo) y baja similaridad inter-clusters (entre grupos).

4.2.1.- K-MEANS

K-Means

- K-Means: Utiliza el centroide de un cluster, C_i, para representarlo.
- Conceptualmente, el centroide corresponde es el centro de la distribución del grupo.
- Este centroide puede definirse de varias maneras.
- Se calcula la distancia entre un objeto $\boldsymbol{p} \in C_i$ y el representante del grupo, $\boldsymbol{c}_i \Longrightarrow dist(\boldsymbol{p}, \boldsymbol{c}_i)$.

K-Means

- La calidad del cluster C_i se puede medir usando la *variación dentro del cluster*, que corresponde a la suma de los errores cuadráticos entre los objetos en C_i y el centroide $\boldsymbol{c}_i$ definida como:

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{p} \in C_i} dist(\boldsymbol{p}, \boldsymbol{c}_i)^2$$

- Esta función objetivo trata de asegurarse que los k grupos sean los más compactos y separados posible.

K-Means

Input:

- D: datos con d tuplas clasificadas.
- k: número de grupos o clusters.

Output: un conjunto de k grupos.

Paso 1: Seleccionar aleatoriamente k objetos de D como centros de grupos iniciales.

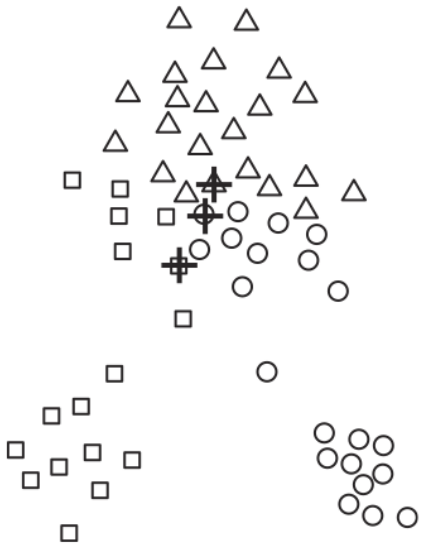
Paso 2: Asignar cada objeto al grupo al cual es más similar, basado en la distancia Euclidiana entre él y el promedio del grupo.

Paso 3: Actualizar las medias de los grupos y volver al paso 2.

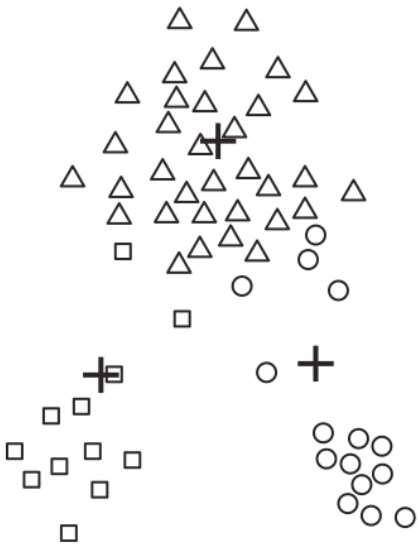
Paso 4: Se para cuando la asignación es estable.

K-Means

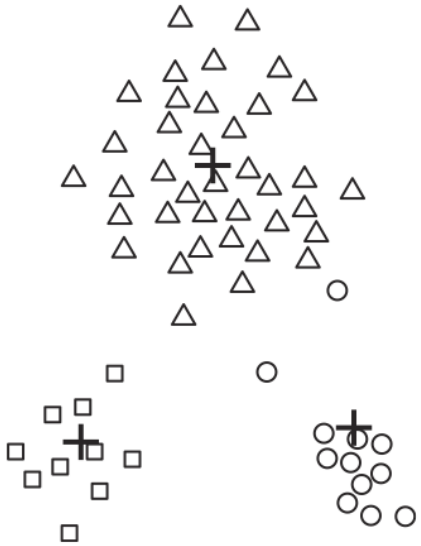
Gráficamente



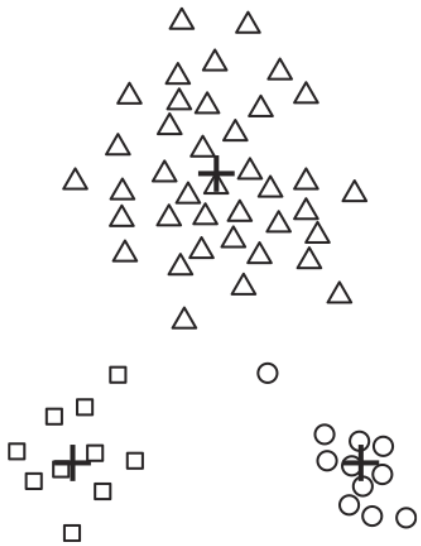
(a) Iteration 1.



(b) Iteration 2.



(c) Iteration 3.



(d) Iteration 4.

K-Means

- El algoritmo K-Means no está garantizado que logre convergencia a un óptimo global y usualmente termina en un óptimo local.
- Esto depende de la selección aleatoria inicial de los centros iniciales.
- Se recomienda realizar varias veces el algoritmo con diferentes centros de grupos iniciales.
- El método es relativamente escalable y eficiente en el procesamiento de grandes bases de datos.

K-Means

- Este método es útil sólo cuando es posible calcular la media de un grupo de objetos.
- En presencia de variables nominales $\Longrightarrow$ k-modas.
- k-medias + k-modas $\Longrightarrow$ datos con atributos mixtos.

K-Means: Desventajas

- Necesidad de especificar k anticipadamente.
  Propuesta: rango de valores de k y comparar resultados para determinar el mejor k.
- No sirve cuando se quieren encontrar grupos de forma no convexa o de tamaños muy diferentes.
- Sensible a observaciones extremas o outliers.
- No funciona bien con bases de datos muy grandes (no es muy escalable).