Universidad Autónoma de Baja California

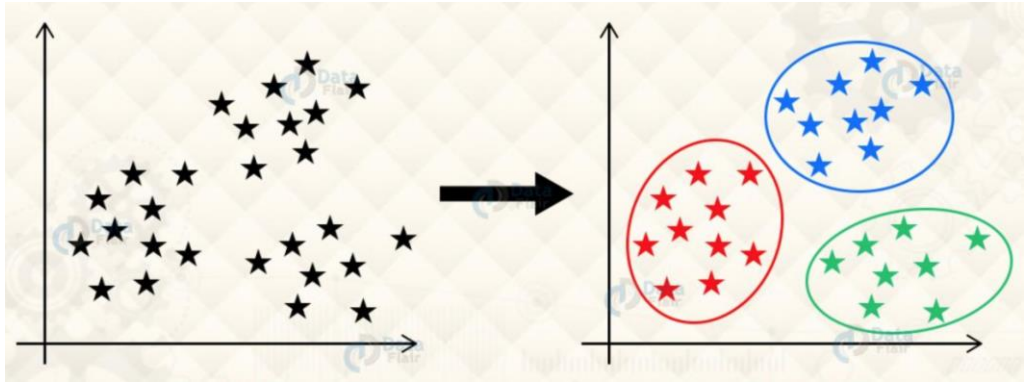
Maestría y Doctorado en Ciencias e Ingeniería
Ingeniería en Computación

6. Agrupación

Minería de Datos

¿Qué es el análisis de grupos?

- Es el proceso de agrupar un conjunto de datos hacia múltiples grupos (clusters) tal que esos objetos dentro un mismo grupo tiene alta similitud, pero son muy disimilares de otros grupos. Las disimilitudes y similitudes son validadas en base en los valores de los atributos que describen a los objetos y frecuentemente involucran medidas de distancias.



Análisis de grupos

- *Agrupación* es el proceso de particionar un conjunto de objetos de datos hacia subconjuntos. Donde cada subconjunto es un grupo, tal que los objetos dentro un grupo son similares el uno al otro, y disimilares con los objetos de otros grupos. Bajo este contexto, diferentes métodos de agrupación pueden generar diferentes grupos a partir del mismo conjunto de datos.
- La agrupación es útil en el hecho de que puede llevar al descubrimiento de grupos desconocidos dentro de los datos.
- En anális de grupos es conocido como *aprendizaje no supervisado* porque las etiquetas de clases no estan presentes. Por esta razón, agrupación es una manera de aprender a partir de observación, en lugar de aprendizaje mediante ejemplos (*aprendizaje supervisado*).

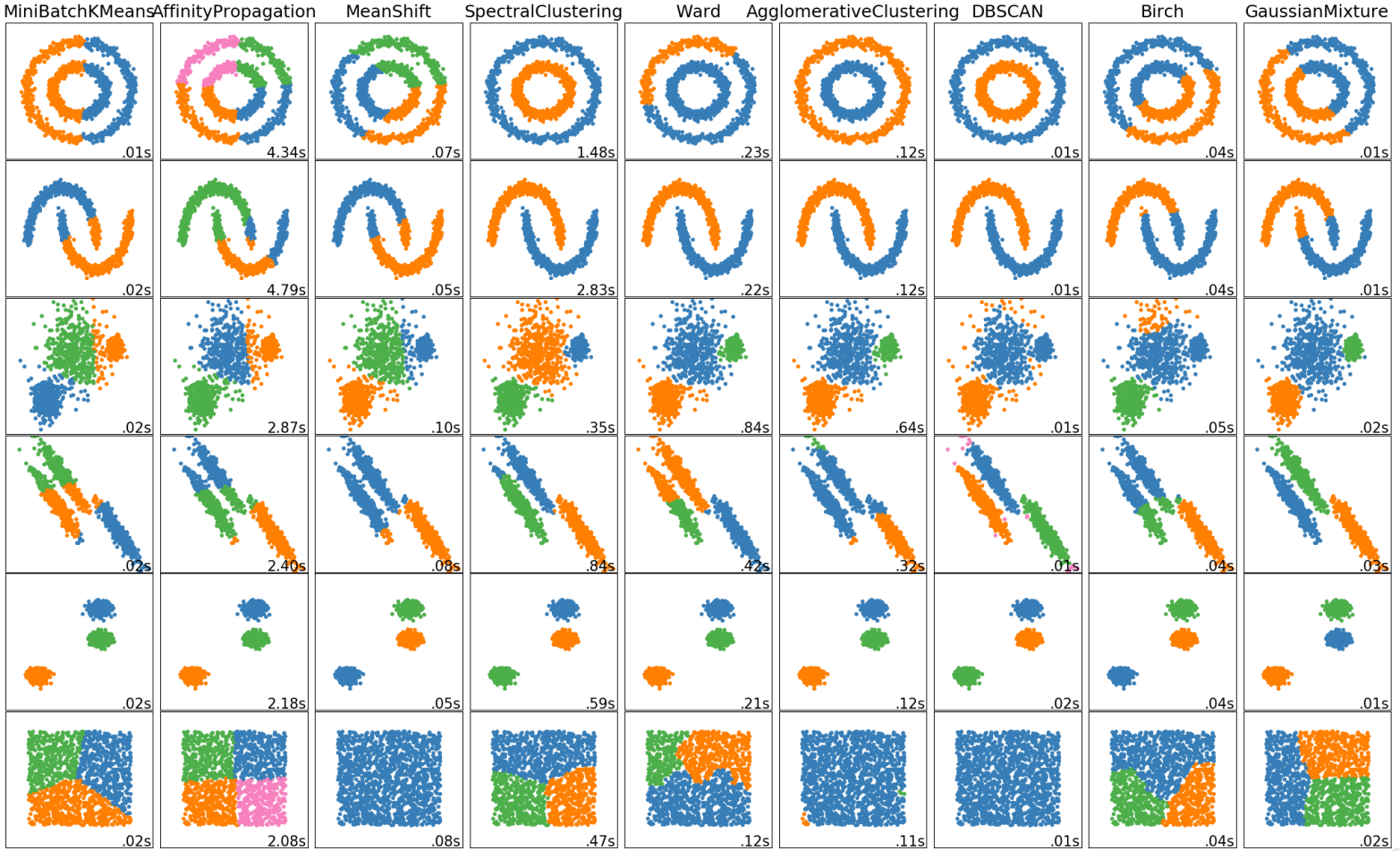
Técnicas típicas de agrupación

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Ejemplos de tipos de algoritmos existentes

- Métodos basados en partición
  - K-Means
  - K-Medoids
- Métodos jerárquicos
  - BIRCH
  - Jerárquico probabilístico
- Métodos basados en densidad
  - DBSCAN
  - OPTICS
  - DENCLUE
- Métodos basados en grid
  - STING
  - CLIQUE

Ejemplos de tipos de algoritmos existentes



Evaluación de la agrupación

- Las principales tareas en la evaluación de las agrupaciones incluyen lo siguiente:
  - **Evaluación de la tendencia de los grupos.** La tarea de agrupación sólo es significativa si no existen datos aleatorios en la estructura.
  - **Determinación del número de grupos en un conjunto de datos.** Es importante previo al proceso de agrupación tener un estimado estadístico de la cantidad de grupos que se desea y pueden encontrar.
  - **Medición de la calidad de la agrupación.** Hay varias técnicas para evaluar la calidad de los grupos encontrados.

Evaluación de la tendencia de los grupos

- **Hopkins Statistic.** Es una estadísitca espacial que prueba la aleatoriedad espacial de una variable como este se distribuye dentro de un espacio. Dado un conjunto de datos *D*, que es considerado como una muestra de una variable aleatoria *o* queremos determinar que tan lejos está *o* de estar uniformemente distribuido del espacio de datos.
- Otros.

Determinación el número de grupos en un conjunto de datos

- Un método simple para identificar un número inicial de grupos es mediante $\sqrt{\frac{n}{2}}$ para un conjunto de datos con *n* puntos.
- El método del codo, realiza una búsqueda de $K>0<K^{max}$ mediante el cual la suma de todas las varianzas dentro del grupo *var(K)*, y posteriormente se grafica el comportamiento de *K*/*var(K)* y donde haya un cambio pronunciado se asume es el *K* óptimo.
- Otros.

Medición de la calidad de la agrupación

- Hay dos categorías de éstas medidas de calidad:
  - **Extrínsecas.** Si el *ground truth* está disponible, se pueden comparar los grupos contra los grupos verdaderos y medir.
    - Se usan para validar técnicas de agrupación y los datos provienen de datasets de clasificación donde las etiquetas son conocidas.
  - **Intrínsecas.** Si el *ground truth* no está disponible, se usan éstos métodos donde se evalúan que tan bien se realizó la agrupación considerando que tan bien los grupos están separados.
    - Se usan cuando se tiene la intención real de agrupar para analizar los datos.

Métodos extrínsecos

- **Cluster homogeneity.** Esto requiere que entre más puros sean los grupos, mejor será la agrupación.
- **Cluster completeness.** Es la contraparte del anterior. Requiere que una agrupación, si cualquiera de dos objetos pertenecen a la misma categoría de acuerdo al ground truth, entonces estos deberían ser asignados al mismo grupo. Esto es, la agrupación debería asignar objetos pertenecientes a la misma categoría (de acuerdo al *ground truth*) al mismo grupo.
- **Rag bag.** En muchos escenarios, la categoría *rag bag* contiene objetos que no puedes unificarse con otros objetos. Este criterio establece que poner un objeto heterogéneo dentro de un grupo puro debe de ser penalizado más que el ponerlo dentro de un *rag bag*.
- **Small cluster preservation.** Si una categoría se parte en pequeños pedazos durante una agrupación, esos pedazos pequeños pudieran volverse ruido y por lo tanto la pequeña categoría no pudiera ser encontrada durante la agrupación. Este criterio establece que partir una categoría pequeña en pedazos es más dañino que partir una categoría grande en pedazos.

Métodos extrínsecos

- Varias medidas de calidad de agrupación satisfacen los cuatro criterios anteriormente mencionados. Tal como las métricas *Bcubed*, *precision* y *recall*.
  - Bcubed, evalua la precision y recall por cada objeto dentro de una agrupación en un dado conjunto de datos acorde al *ground truth*.
  - La precisión, de un objeto indica que tantos objetos dentro del mismo grupo pertenecen a la misma categoría que el objeto.
  - El recall, de un objeto refleja que tantos objetos de la misma categoría son asignados al mismo cluster.

Métodos intrínsecos

- En general, estos métodos evaluan la agrupación mediante la examinación de que tan bien los grupos se encuentran separados y que tan compactos los grupos estan.
- Un ejemplo de estos métodos es el **silhouette coefficient**.

Actividad

- Conseguir 1 dataset de agrupación y 1 dataset de clasificación (con etiquetas)
- Usar 2 técnicas diferentes para agrupar (de naturaleza diferente)
  - Ponerse de acuerdo en el *Foro de Bb* para no repetir técnicas
- Utilizar mínimo 3 técnicas intrínsecas y 3 extrínsecas, dependiendo del caso, para determinar la calidad de los grupos
  - Para el caso de la dataset de clasificación, para usarlo como dataset de agrupación hay eliminar las etiquetas pero usarlas como el *ground truth* (extrínsecas)
- De los experimentos realizados, exponer:
  - Cómo funcionan las técnicas elegidas, poner atención a cómo se eligen los parámetros de la técnica
  - Explicar el procedimiento para la experimentación completa
  - Explicar brevemente qué buscan los indices de validación que eligieron
  - Identificar cuál algoritmo fue mejor para resolver cada dataset
  - Resultados obtenidos