Instituto Nacional de Astrofísica Óptica y Electrónica

Coordinación de Ciencias Computacionales

Cross-Clustering: A partial clustering algorithm with automatic estimation of the number of clusters

Cruz Paz Daniel

Melendez Melendez Gabriel

Septiembre 2016

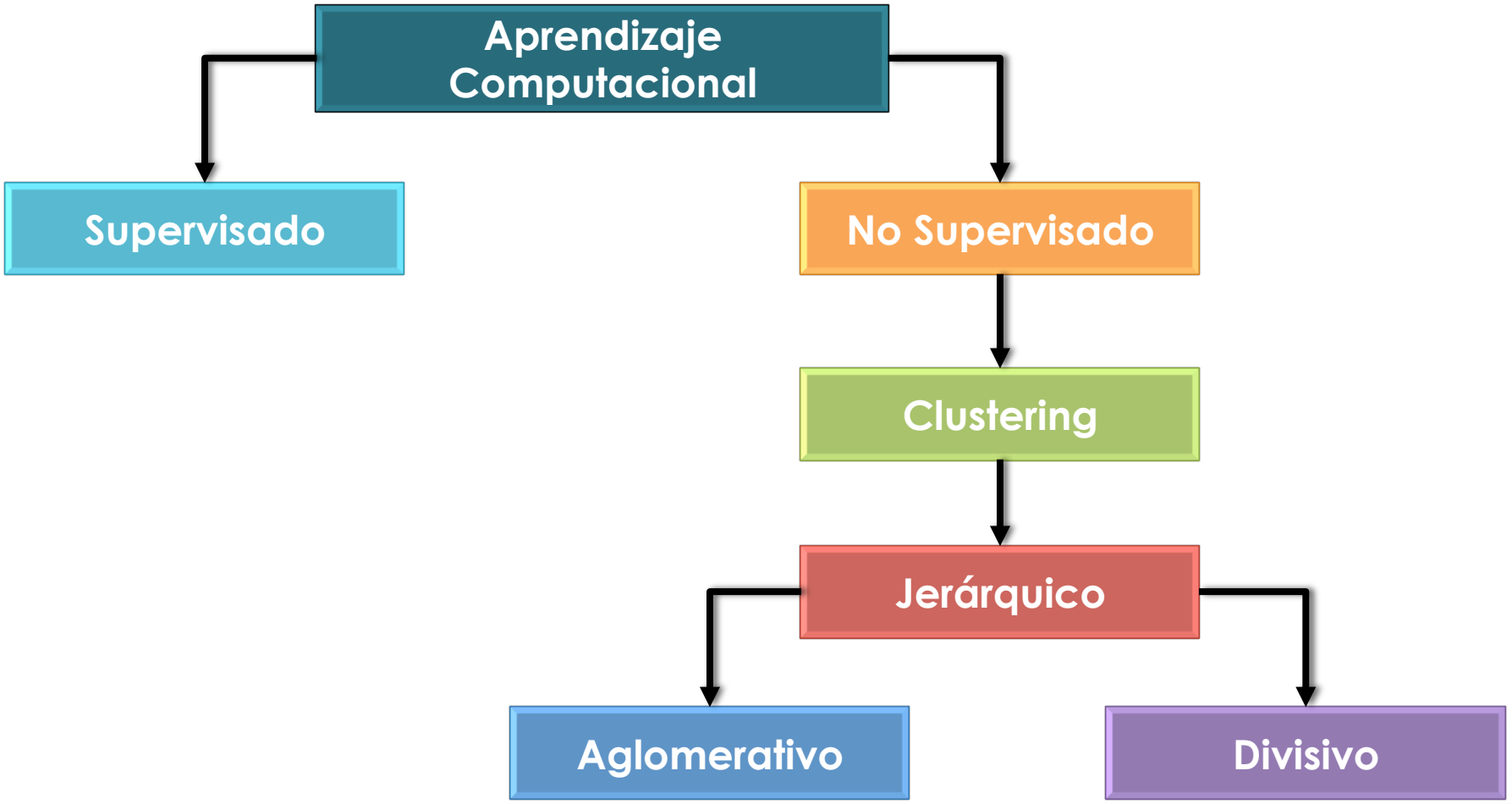
Puntos

- Descripción del problema abordado en el artículo seleccionado.
- Breve descripción del trabajo relacionado que se comenta en el artículo.
- Breve descripción de la solución propuesta en el artículo.
- Descripción de los experimentos que soportan dicha solución, incluyendo tablas y/o gráficas tomadas del artículo.
- Conclusiones a las que llegan los autores.
- Análisis de debilidades y/o modificaciones posibles para mejorar la solución propuesta por los autores. (Puede basarse en el trabajo futuro propuesto).

Descripción del problema

- Limitaciones comunes en la mayoría de métodos de agrupamiento
  - Agrupación de objetos alejados (Falta de buena estrategia)
    - Detectar elementos aislados
  - Estimación a priori del número de clusters
  - Falta de un método capaz de detectar cuando la partición de un dataset no es apropiada
  - Inicialización (Métodos no deterministas)

Descripción del trabajo relacionado

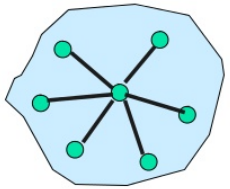


Descripción del trabajo relacionado

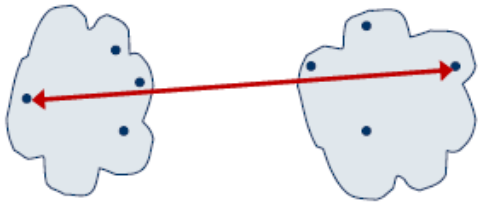
- Métodos clásicos de agrupamiento
  - Método de Ward
  - Vinculación completa
  - K-means
  - DBSCAN
  - SOM

Descripción de la solución propuesta

- Cross clustering
  - Método de Ward
  - Vinculación completa



Ward



Vinculación completa

- Parametros
- Ward
  - Rango para el número de clusters $I^W = [n_{W_{min}}, \ldots, n_{W_{max}}]$

- Vinculación completa
  - Rango para el número de clusters $I^C = [n_{W_{min}} + 1, \ldots, n_{W_{max}}]$

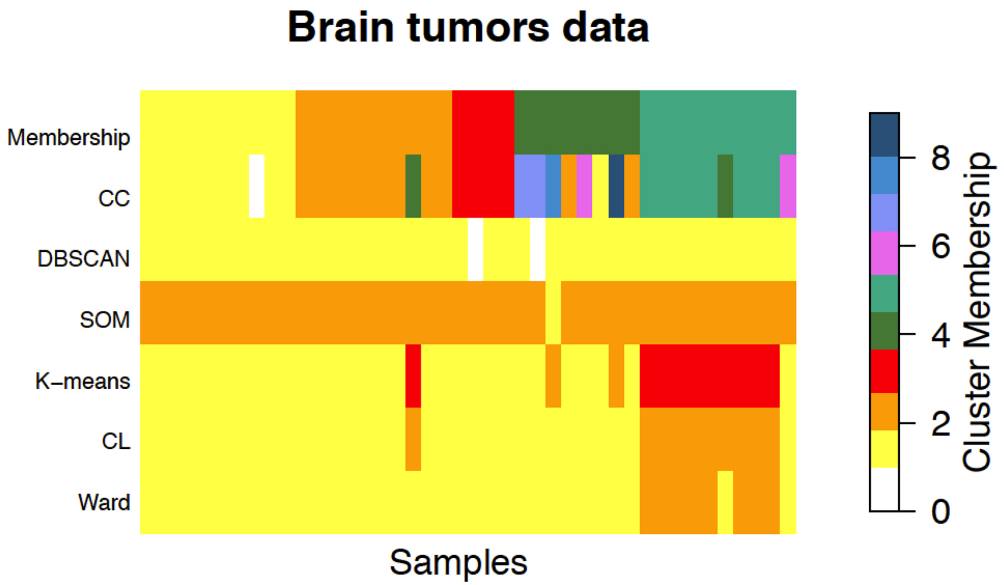
Experimentos (Gráficas, tablas, etc)

- **Pruebas con datasets simulados**
  - Se generaron 100 datasets simulados
    - 2000 perfiles de expresión de genes para 5 ejemplos con valor:
      - Positivo (500)
      - Creciente (250)
      - Decreciente (700)
      - Oscilante (300)
      - Convexo (100)
      - Además se introdujo ruido en la información
  - Aplicación de Ward y CL – Distancia euclediana

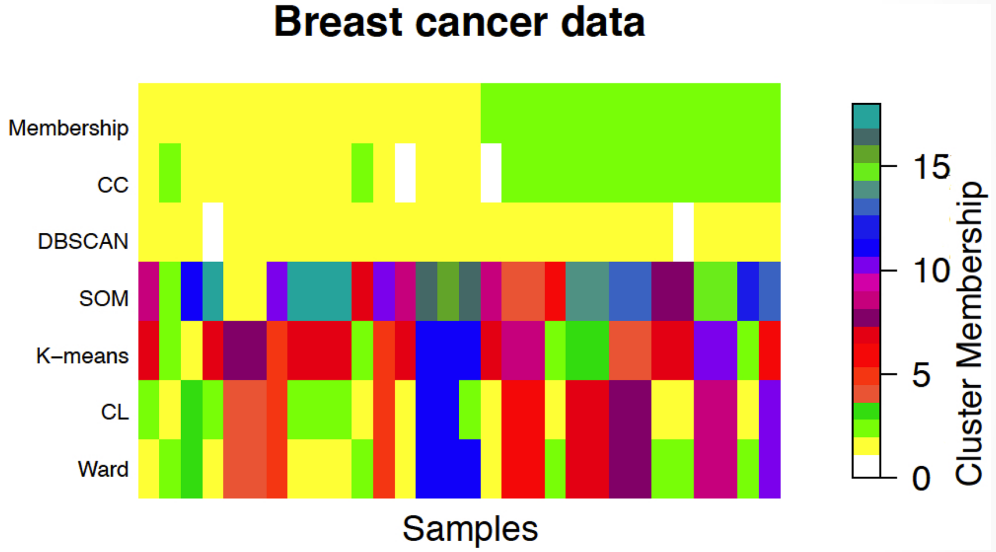
Experimentos (Gráficas, tablas, etc)

- Pruebas con datasets reales
  - **Tumores del cerebro**
    - 5299 perfiles de expresión genética
    - 42 ejemplos
    - ASW para seleccionar el # de clusters
      - Se aplicaron los CC, Ward, CL, DBSCAN, K-means, SOM
  - **Cancer de seno**
    - 719,690 pruebas
    - 30 ejemplos
    - ASW para seleccionar el # de clusters
      - Se aplicaron los métodos CC, Ward, CL, DBSCAN, k-means y SOM

Experimentos (Gráficas, tablas, etc)



	Results	
Method	# Clusters	ARI
Ward	2	0.14
CL	2	0.18
K-means	3	0.19
SOM	2	0.003
CC	$9 + 1$	**0.64**



	Results	
Method	# Clusters	ARI
Ward	11	0.09
CL	11	0.10
K-means	11	0.04
SOM	18	0.07
CC	$2 + 1$	**0.63**

Experimentos (Gráficas, tablas, etc)

- **Aceite de oliva**
  - 572 diferentes tipos de aceite de oliva
    - 8 mediciones
    - Se aplicaron los métodos CC, Ward, CL, DBSCAN, k-means y SOM

Conclusiones(autores)

- CC ofrece grandes ventajas con respecto a algoritmos clásicos de agrupamiento como:
  - CC no necesita un conocimiento a priori del # de clusters
  - Quita elementos aislados del agrupamiento
  - Es capaz de sugerir cuándo un dataset no debe ser agrupado
  - Se hicieron comparaciones con los siguientes métodos:
    - DBSCAN, Ward, CL, NMF, SPARCoC, K-means y SOM
    - CC siempre tuvo mejor rendimiento

Análisis de CC

- Complejidad del orden O(n^2)
  - n = número de elementos a ser agrupados
- Crear un umbral que sirva como referencia para el valor del parámetro ASW y así funcione como una condición de paro.