UNIVERSIDAD DE GRANADA

MÁSTER DE CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

CURSO ACADÉMICO 2019-2020

MINERÍA DE DATOS: AGRUPAMIENTO NO SUPERVISADO Y DETECCIÓN DE ANOMALÍAS

Análisis Clúster de datasets

Aplicación de diversos algoritmos de clustering sobre un dataset y análisis de resultados.

Nicolás Cubero Torres

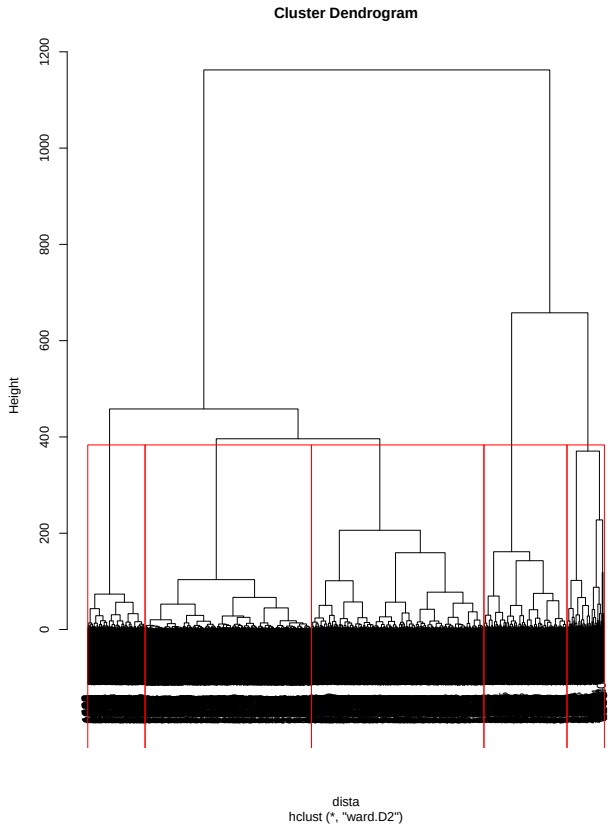
2 de Diciembre de 2019

En este documento, se decide llevar a cabo un análisis jerárquico sobre el dataset **demo_cs** mediante el método de Ward.

Dicho análisis se realizará considerando tomando como medida de distancia un promedio de las distancias medidas sobre un subconjunto de los atributos numéricos (para los cuales se ha empleado la distancia euclídea como medida de distancia) y las distancias medidas sobre un subconjunto de atributos binarios.

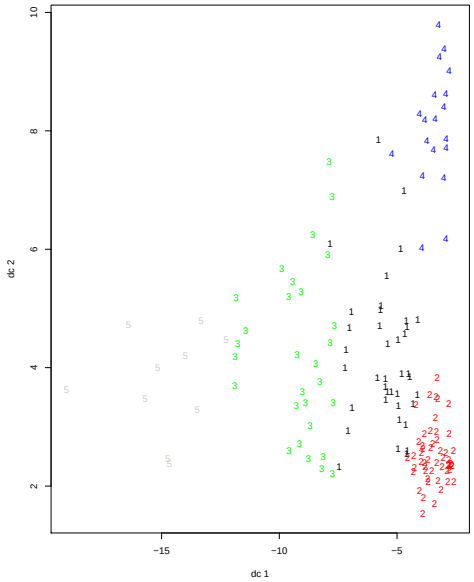
Se ejecuta el *script distancias-jerarquico.R* para la aplicación de un clustering jerárquico usando el método de Ward.

Tras la ejecución del algoritmo de clustering, se selecciona realizar la agrupación en 5 grupos.



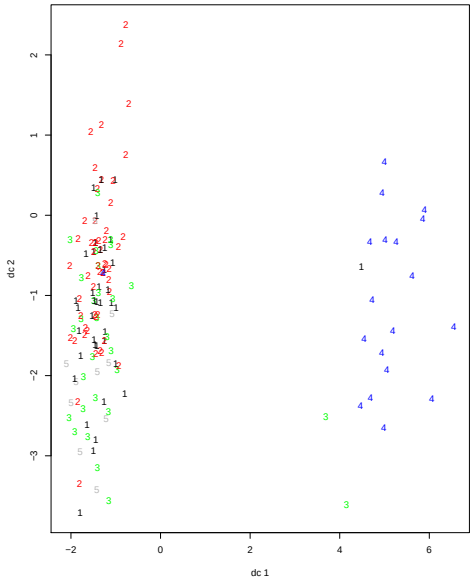
Se representan 150 patrones aleatorios del dataset en dos gráfica: En la primera gráfica, en base a los atributos numéricos, se representan la segunda

distancia factorial de los atributos frente a la primera:



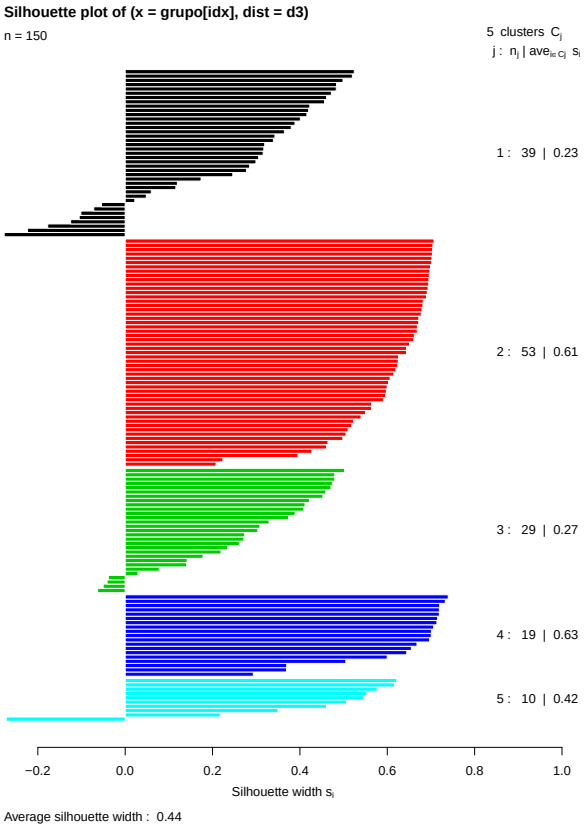
En la anterior gráfica, se pueda apreciar una mayor facilidad para separar el clúster 5 y 4 del resto, mientras que para el resto, puede existir una mayor dificultad para realizar esta separación.

Por su parte, en la segunda gráfica, basándose en los atributos binarios, se representa también la segunda distancia factorial frente a la primera:



En la anterior gráfica, a diferencia de la anterior, se puede apreciar un mayor solapamiento entre las instancias de diferentes clústeres.

Por su parte el diagrama de Silhouette revela valores negativos del coeficiente de Silhouette para alguno de las instancias del clúster nº 1, el clúster nº 3 y el clúster nº 5, asimismo, se puede observar que la puntuación dada a los diferentes clústeres no es relativamente alta (valores de 0.61 y 0.63 respectivamente para los clústeres 2 y 5, mientras que para el resto, los valores se sitúan por debajo de 0.5), dando un valor promedio de 0.38, lo cual nos da lleva a plantear que el agrupamiento de los patrones realizado no es especialmente óptimo considerando 5 clústeres.



Se decidió tratar de modificar el número de clústeres con el objetivo de verificar si el coeficiente de Silhoutte garantizaba una mejor bondad en la agrupación de los patrones. Los resultados se recogen a continuación:

Nº de Clústeres	2	5	7	10
coeficiente de Silhouette promedio	0.59	0.38	0.43	0.36

Los resultados recogidos en la anterior tabla no nos parecen indicar ninguna tendencia clara de que el aumento o decremento del número de grupos lleve a mejorar o empeorar la bondad del agrupamiento efectuado.