Algoritmo CURE

Iker Meneses Sales

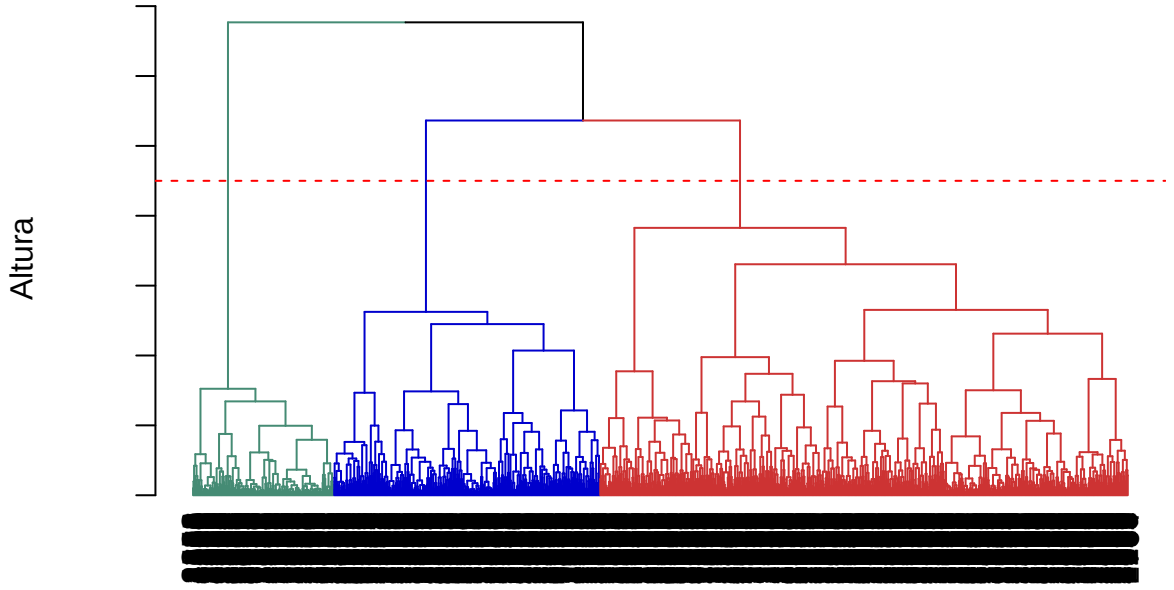
2023-10-29

Algoritmo CURE

Siguiendo con los algoritmos de clusterización para bases de datos grandes, es momento de realizar el CURE. CURE (Clustering Using REpresentatives) es un algoritmo de clustering para base de datos grandes en el cual se gestiona, inicialmente, una muestra de la base de datos a partir de la cual se realiza un clustering jerárquico (usando la distancia euclídea y el método de agregación simple) y se sacan un número pequeño de puntos (representantes) de cada cluster. Entonces, se acercan esos representantes hacia el centroide del cluster un 20% y, a partir de estos representantes acercados, se busca cuál es el que se encuentra más cercano de cada punto de la base de datos restante. Finalmente, una vez se encuentra el representante más cercano a cada individuo, se asigna el individuo al cluster al que pertenece el representante.

En este caso, como la base de datos escogida dispone de datos numéricos y categóricos, se ha decidido modificar las reglas del CURE y usar la distancia de Gower y el método de agregación de Ward en la construcción inicial del clustering. Así pues, realmente se podría afirmar que se está realizando un pseudoCURE en este caso.

Inicialmente, para este caso, se ha decidido escoger una muestra significativa y grande para evitar problemas en la construcción de los clusters iniciales. Así, se ha usado una muestra de $n = 2000$ con el objetivo de realizar el primer cluster a partir del cual se elegirán los representantes. El dendograma resultante reporta la siguiente estructura:



Tras analizar los resultados, se puede apreciar que el número de clusters óptimo es $k = 3$. De esta forma, la partición inicial de la muestra en cada cluster se puede apreciar en la parte inferior:

Table 1: Distribución inicial de individuos por cluster CURE

Cluster	Observaciones
1	569
2	1129
3	302

Ahora, a partir de este clustering jerárquico inicial, se escogerán los representantes. Para ello, se busca aquellos puntos más alejados entre sí y, a la vez, más alejados del centroide de cada cluster. Para este paso, se han elegido exactamente 5 representantes por cluster. Una vez se tienen seleccionados, el siguiente paso es acercarlos al centro. En este caso, se ha decidido aproximarlos un 20% hacia el centroide del cluster al que pertenecen.

Por último, se analiza cada punto y se busca el representante más cercano. Una vez se tiene esa información, se le asigna al individuo el cluster al que pertenece el representante más cercano. Para este paso, se ha procedido a procesar los datos de 500 en 500, para así evitar problemas con la capacidad de gestión de datos del ordenador. Así, el resultado del clustering final se presenta en la taba inferior:

Table 2: Distribución inicial de individuos por cluster CURE

1	1351
2	2275
3	1374