# Métodos no supervisados y detección de anomalías: agrupamiento jerárquico

UCO ONLINE

# Agrupamiento jerárquico

## 1. Introducción

Hasta el momento, los métodos de agrupamiento que hemos visto se engloban en la categoría de métodos particionales, pues dividen el conjunto de puntos directamente en un número predefinido de grupos. En contraposición, el agrupamiento jerárquico supone construir una secuencia de grupos anidados, en base igualmente a una matriz de proximidad. En esta lección se estudiarán las dos formas habituales de crear estas jerarquías: el modo aglomerativo (de abajo a arriba) o el divisivo (de arriba abajo).

Los resultados de un agrupamiento jerárquico pueden representarse como un árbol, al que llamamos dendrograma, como el que se muestra en la Figura 1. El nodo raíz es la representación de todo el conjunto de datos, mientras que los nodos hoja son cada uno de los puntos o muestras que componen el conjunto de datos. Los niveles intermedios establecen las relaciones de proximidad entre nodos, y la altura del dendrograma expresa la distancia entre pares de puntos o grupos. De esta forma, un agrupamiento en n grupos se obtiene "partiendo" el dendrograma en un nivel determinado. Es decir, el agrupamiento jerárquico nos da la posibilidad de, en una única ejecución del método, obtener varios agrupamientos a posteriori con distinto número de grupos. Esta representación proporciona una descripción muy informativa y una visualización sencilla de las estructuras grupales que existen en el conjunto de datos. Esto es relevante si realmente existen relaciones de jerarquía entre los datos (por ejemplo, especies de organismos que efectivamente componen un árbol biológico).

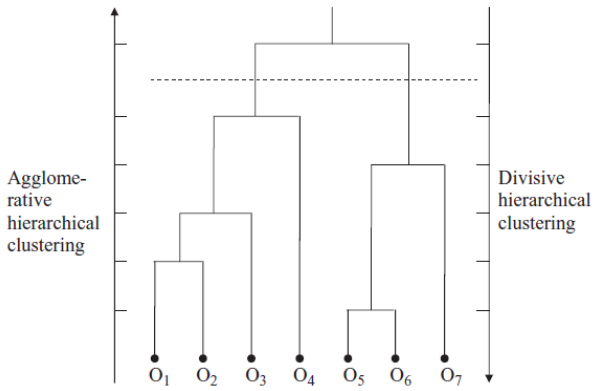


*Figura 1. Representación gráfica del agrupamiento jerárquico (Xu & Wunsh, 2009)*

## 2. Agrupamiento jerárquico aglomerativo

El agrupamiento jerárquico aglomerativo inicia el proceso de análisis de grupos con N grupos, donde N es igual al número de objetos a agrupar. Esto es, al comienzo del proceso cada grupo está formado por un único objeto. A continuación, se lleva a cabo una secuencia de operaciones que irán mezclando los grupos, hasta llegar al extremo de que todos los objetos estén en un único grupo. El proceso consiste en los siguientes pasos:

1. Asignar un grupo a cada objeto que existe en el conjunto.
2. Calcular la matriz de proximidad (normalmente basada en distancia) entre los grupos.
3. Determinar la menor distancia entre pares de grupos, en base a una función de distancia concreta (más adelante se verán las distintas opciones).
4. Unir los dos grupos que están más próximos, que forman un nuevo grupo a partir de ahora.
5. Recalcular la distancia entre grupos con la nueva distribución de grupos.
6. Repetir los pasos 3 a 5 hasta que solo quede un grupo.

El proceso de unir dos grupos para conformar uno nuevo es claramente dependiente de la función de distancia que se define. Hay que destacar que estamos hablando de una distancia entre grupos, no entre objetos y, por tanto, dicha distancia puede interpretarse de varias maneras. Este concepto es el que se conoce como *clustering linkage*. Las principales formas de enlace que se utilizan son:

1. ***Single linkage***. La distancia entre dos grupos se determina como la distancia mínima entre dos objetos que pertenecen a cada uno de los grupos. Otra forma de conocerlo es como el método del vecino más cercano. Funciona bien si los grupos están muy separados entre sí. Por el contrario, si hay ruido (algún objeto muy alejado del resto de su grupo), puede provocar la unión de grupos poco relacionados.
2. ***Complete linkage***. Es el caso contrario al anterior donde la distancia se establece como la máxima distancia entre dos objetos, uno asignado en cada grupo. Este mecanismo es efectivo para localizar grupos pequeños y compactos.
3. ***Group average linkage***. La distancia entre dos grupos se define como la media entre todas las distancias que se pueden calcular entre pares de objetos asignados a grupos diferentes. A este método también se le conoce como *unweighted pair group method average* (UPGMA).
4. ***Weight average linkage***. Se basa en el anterior, pero las distancias se ponderan en base al número de objetos en cada grupo. Su acrónimo es WPGMA (*weighted pair group method average).*

5. ***Centroid linkage.*** También conocido como UPGMC (*unweighted pair group method centroid*), que considera la distancia euclídea entre los centroides (calculada como la media de las coordenadas de todos los puntos en cada grupo).
6. ***Ward's method.*** Tiene en cuenta la suma de errores que resultaría de la unión de los dos grupos, de forma que busca minimizar ese valor.

La Figura 2 ilustra gráficamente alguno de estos métodos de enlace.

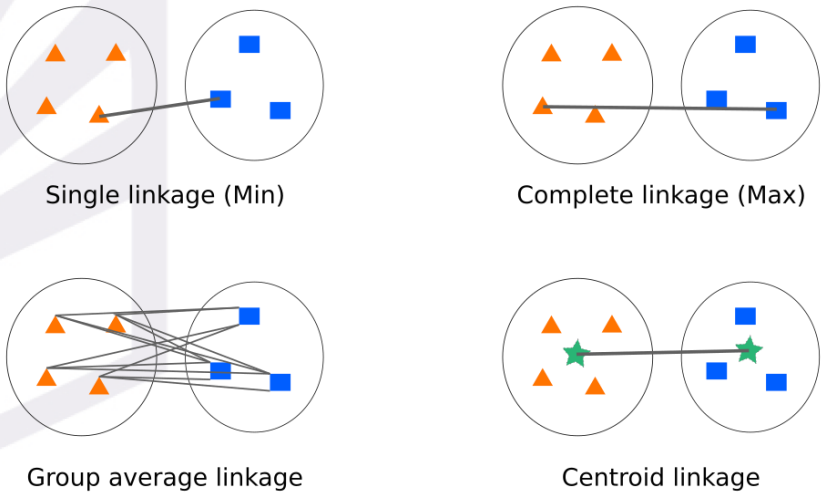


*Figura 2. Formas de calcular la distancia entre grupos*

## 3. Agrupamiento jerárquico divisivo

El agrupamiento jerárquico divisivo actúa de forma inversa al agrupamiento jerárquico aglomerativo. Al comienzo, solo existe un único grupo formado por todos los objetos, que va descomponiéndose en grupos más pequeños hasta que cada grupo contiene un único objeto. Si el conjunto de datos tiene N objetos, existen $2^{N-1}-1$ formas de dividir el grupo inicial en dos grupos. El número de combinaciones es por tanto muy elevado incluso para conjuntos de datos pequeños. Por este motivo, el agrupamiento jerárquico divisivo es más costoso que el aglomerativo, y se usa menos en la práctica. Por otro lado, su proceso parece ser más intuitivo respecto a la estructura de los datos, ya que los grupos grandes que pueden no ser muy cohesionados al principio se van refinando paulatinamente.

Existen diferentes algoritmos que consideran solo algunas de las posibles combinaciones. Uno de ellos es DIANA, que toma como decisión dividir el grupo con el mayor diámetro. En concreto, el proceso que sigue DIANA es el siguiente:

1. Inicializar el grupo $C_i$ con todos los objetos del conjunto de datos. Inicializar otro grupo $C_j$ vacío.
2. Para cada objeto en $C_i$, calcular la distancia media al resto de objetos.
3. Mover el objeto con mayor distancia al grupo $C_j$.
4. Escoger el grupo con el mayor diámetro como aquel a ser dividido.
5. Para cada objeto en $C_i$, calcular la diferencia entre la distancia media al resto de objetos en $C_i$ y la distancia media a los objetos en $C_j$.
6. Si la mayor diferencia en el paso 5 es mayor que 0, mover el objeto con mayor diferencia a $C_j$. Si la diferencia es menor a 0, terminar. En caso contrario, repetir los pasos 4-6.
7. Una vez construido el dendrograma a partir de las separaciones realizadas, seleccionar un agrupamiento "cortando" el dendrograma a un nivel concreto.

## Referencias

C. C. Aggarwal, C. K. Reddy (eds.). "Data Clustering: Algorithms and Applications". Chapman & Hall / CRC Press, 1ª edición, 652 páginas. 2014.

T. Hastie, R. Tibshirani, J. Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer Series in Statistics, 2ª edición, 745 páginas. 2017.

G. James, D. Witten, R. Tibshirani, T. Hastie. "An Introduction to Statistical Learning with Applications in R". Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/

A. Kassambara. "Practical Guide to Cluster Analysis in R". STHDA, 187 páginas. 2017.

S. Shalev-Shwartz, S. Ben-David. "Understanding Machine Learning: From Theory to Algorithms". Cambridge University Press, 1ª edición, 449 páginas. 2014.

R. Xu, D.C. Wunsch II. "Clustering". Wiley/IEEE Press, 1ª edición, 363 páginas. 2009.