Modelos de Soporte No Supervisado

Eligiendo el mejor algoritmo para cluster

Eligiendo el mejor algoritmo

- Elegir el mejor método de agrupación para un dato dado puede ser una tarea difícil para el analista
- Revisaremos el paquete R clValid (G. Brock et al., 2008)
- Dicho paquete que se puede usar para comparar simultáneamente múltiples algoritmos de agrupamiento
- Y para identificar el mejor enfoque de agrupación y el número óptimo de clusters.

Medidas para comparar algoritmos de agrupamiento

El paquete clValid compara los algoritmos de agrupamiento utilizando dos medidas de validaciones de clúster:

1. **Medidas internas**: que utiliza información intrínseca en los datos para evaluar la calidad de la agrupación. Las medidas internas incluyen la conectividad, coeficiente Silohuette y el índice de Dunn.

2. **Medidas de estabilidad**: una versión especial de medidas internas, que evalúa la consistencia de un resultado de agrupamiento al compararlo con los grupos obtenidos después de que cada columna se elimine, una a la vez.

Medidas para comparar algoritmos de agrupamiento

Las medidas de estabilidad del clúster incluyen:

- La proporción promedio de no superposición (APN- average proportion of non-overlap)
- La distancia promedio (AD- average distance)
- La distancia promedio entre medias (ADM- average distance between means)
- La figura del mérito (FOM- figure of merit)

El APN, AD y ADM se basan en la tabla de clasificación cruzada del agrupamiento original con la información completa con la agrupación basada en la eliminación de una columna.

- La APN mide la proporción promedio de observaciones no colocadas en el mismo clúster en función de los datos completos y la agrupación en función de la datos con una sola columna eliminada.
- AD mide la distancia promedio entre las observaciones colocadas en el mismo clúster en ambos casos (conjunto completo de datos y eliminación de una columna).
- El ADM mide la distancia promedio entre los centros del grupo para las observaciones colocadas en el mismo grupo en ambos casos.
- El FOM mide la varianza media dentro del clúster de la columna eliminada, donde la agrupación se basa en las columnas restantes (no eliminadas).

Medidas para comparar algoritmos de agrupamiento

Los valores de ***APN, ADM y FOM van de 0 a 1***, con un valor menor correspondiente con resultados de agrupamiento altamente consistentes.

AD tiene un valor ***entre 0 y infinito***, y valores más pequeños también son preferidos.

Medidas para comparar algoritmos de agrupamiento

Usaremos la función clValid () [en el paquete clValid], cuyo formato simplificado es :

clValid (obj, nClust, clMethods = "hierarchical", validation = "estabilidad", maxitems = 600, metric = "euclidean", method = "average")

• obj: una matriz numérica o marco de datos. Las filas son los elementos que se agruparán y las columnas son muestras.

• nClust: un vector numérico que especifica el número de clusters que se evaluarán. Por ejemplo, 2:10

• clMethods: el método de agrupamiento a ser utilizado. Las opciones disponibles son "jerárquicas", "Kmeans", "diana", "fanny", "som", "modelo", "sota", "pam", "clara", y "agnes", con múltiples opciones permitidas.

• validation: el tipo de medidas de validación que se utilizarán. Los valores permitidos son "Interno", "estabilidad" y "biológico", con múltiples opciones permitidas.

• maxitems: la cantidad máxima de elementos (filas en la matriz) que pueden ser agrupado.

• metric: la métrica utilizada para determinar la matriz de distancia. Las posibles opciones son "Euclidiano", "correlación" y "manhattan".

• método: para la agrupación jerárquica (hclust y agnes), método de aglomeración a ser utilizado. Las opciones disponibles son "ward", "single", "complete" y "promedio".

Medidas para comparar algoritmos de agrupamiento

Por ejemplo, considere el conjunto de datos del iris, la función clValid () se puede usar de la siguiente manera.

Comenzamos por medidas internas del clúster, que incluyen la conectividad, el ancho de el coeficiente Silohuette y el índice de Dunn.

Es posible calcular simultáneamente estas medidas internas para múltiples algoritmos de agrupamiento en combinación con un rango de números de grupo.

Medidas para comparar algoritmos de agrupamiento

Se puede ver que la agrupación jerárquica con dos clusters ofrece el mejor rendimiento en cada caso (es decir, para conectividad, medidas de Dunn y Silhouette).

A pesar de algoritmo de agrupamiento, el número óptimo de clusters parece ser dos utilizando las tres medidas.

Calculando las medidas de estabilidad tenemos que para las medidas APN y ADM, agrupación jerárquica con dos clústeres nuevamente da la mejor puntuación.

Para las otras medidas, PAM con seis clusters tiene la mejor Puntuación.