Modelos de Soporte No Supervisado

Estadísticas de Validación del Cluster

Validación del cluster

- El término validación de conglomerados se usa para diseñar el procedimiento de evaluación de la bondad de los resultados del algoritmo de agrupamiento.

- Esto es importante para evitar encontrar patrones en información aleatoria, así como, en la situación en la que desea comparar dos algoritmos.

- En general, las estadísticas de validación de clustering se pueden clasificar en 3 clases (Theodoridis y Koutroubas, 2008; G. Brock y otros, 2008, Charrad et al., 2014):

1. **Validación interna:** utiliza la información interna de la agrupación para evaluar la bondad de una estructura de agrupamiento sin referencia a información externa. También se puede usar para estimar el número de clusters y el algoritmo de agrupamiento apropiado sin ningún dato externo.

2. **Validación externa**: consiste en comparar los resultados de un análisis de clúster a un resultado conocido. Mide la congruencia de los clústeres con la clasificación suministrada externamente. Dado que conocemos el número de clúster "verdadero" de antemano, este enfoque se usa principalmente para seleccionar el algoritmo de agrupamiento correcto para un conjunto de datos específico.

3. **Validación de conglomerados relativos**: evalúa la estructura de conglomeración variando diferentes valores de parámetros para el mismo algoritmo (por ejemplo,: variando el número de clusters k). Generalmente se usa para determinar la cantidad óptima de clusters.

Validación interna

- Debemos recordar que el objetivo de los algoritmos de agrupamiento es dividir el conjunto de datos en grupos de objetos, tales que:
  - Los objetos en el mismo grupo son similares tanto como sea posible, y los objetos en diferentes clusters son muy distintos
  - Queremos que la distancia promedio dentro del clúster sea lo más pequeña posible; y la distancia promedio entre los clusters es lo más grande posible.
- Las medidas de validación interna reflejan a menudo la compacidad, la conectividad y la separación de las particiones de clúster

Validación interna

1. **Compactación o cohesión del clúster**: mide cuán cerca están los objetos dentro de el mismo grupo. Una menor variación dentro del clúster es un indicador positivo de compacidad (es decir, una buena agrupación).

Los diferentes índices para evaluar el la compacidad de los clusters se basan en medidas de distancia, como distancias promedio/mediana entre observaciones dentro del cluster.

2. Separación: Mide qué tan bien separado está un grupo de otros grupos.

Los índices utilizados como medidas de separación incluyen:

- distancias entre centros de clúster
- las distancias mínimas por pares entre objetos en diferentes clusters

3. Conectividad: corresponde a qué extensión de los elementos se colocan en el mismo clúster así como sus vecinos más cercanos en el espacio de datos. La conectividad tiene un valor entre 0 e infinito y debe ser minimizada.

Validación interna

En general, la mayoría de los índices utilizados para la validación interna de clusters combinan compacidad y medidas de separación de la siguiente manera:

$$Index = \frac{(\alpha \times Separation)}{(\beta \times Compactness)}$$

Donde alfa y beta son los pesos de cada atributo.

Validación interna

Coeficiente Silhouette

Mide qué tan bien se agrupa una observación y estima la distancia promedio entre los grupos. El diagrama de silueta muestra una medida de qué tan cerca está cada punto de un grupo de puntos en los clusters vecinos.

Para cada observación i, el ancho de la silueta se calcula de la siguiente manera:

1. Para cada observación i, calcule la disimilitud promedio (ai) entre i y todos otros puntos del clúster al que (i) pertenece

2. Para todos los demás clusters C, a los que no pertenece, calcule la disimilaridad promedio d (i, C) de i para todas las observaciones de C. El más pequeño de estos d (i, C) se define como bi = minC d (i, C). El valor de bi se puede ver como la disimilaridad entre i y su clúster "vecino", es decir, el más cercano al que no pertenece.

3. Finalmente, el ancho de la silueta de la observación i se define mediante la fórmula:

Si = (bi - ai) / max (ai, bi).

Validación interna

Coeficiente Silhouette

El ancho de la silueta se puede interpretar de la siguiente manera:

- Las observaciones con un **Si** grande (casi 1) están bien agrupadas.

- Un pequeño **Si** (alrededor de 0) significa que la observación se encuentra entre dos grupos.

- Las observaciones con un **Si** negativo probablemente estén ubicadas en el grupo incorrecto.

Validación interna

Dunn Index

El índice de Dunn es otra medida interna de validación de clusters que puede ser calculado de la siguiente manera:

1. Para cada grupo, calcule la distancia entre cada uno de los objetos en el grupo y los objetos en los otros grupos.
2. Use el mínimo de esta distancia por pares como la separación entre clusters (min.separation)
3. Para cada cluster, calcule la distancia entre los objetos en el mismo grupo.
4. Use la distancia máxima dentro del clúster (es decir, el diámetro máximo) como compacidad intracluster
5. Calcule el índice de Dunn (D) de la siguiente manera:

$$D = \frac{min.separation}{max.diameter}$$

Si el conjunto de datos contiene grupos compactos y bien separados, el diámetro esperado de los clusters debe ser pequeño y se espera que la distancia entre los clusters sea grande.

Por lo tanto, el índice de Dunn debe maximizarse.

Validación interna

Es posible cuantificar las coincidencias entre los clusters generados y una referencia externa usando el índice de Rand corregido y el índice de variación VI de Meila.

Ambos se implementan con la función R cluster.stats () [fpc package].

El índice de Rand corregido varía de **-1** (sin coincidencias) **a 1** (coincidencia perfecta).

La validación externa de clusters se puede usar para seleccionar el algoritmo de agrupamiento adecuado para un conjunto de datos dado.

Validación interna

Los siguientes paquetes R son requeridos:

- factoextra para visualización de datos
- fpc para calcular estadísticas de validación de clustering
- NbClust para determinar el número óptimo de clústeres en el conjunto de datos

Validación interna

Utilizaremos la función eclust () [enhanced clustering, en factoextra] que proporciona varias ventajas:

- Simplifica el flujo de trabajo del análisis de clustering
- Se puede usar para calcular clusters jerárquicos y particionales en una sola línea
- Comparar las funciones de particionamiento estándar (kmeans, pam, clara y fanny)

requiere que el usuario especifique la cantidad óptima de clusters

- La función eclust () calcula automáticamente la estadística de ***gap*** para estimar el número correcto de clusters
- Proporciona el estadístico ***Silhouette*** para todos los métodos de particionamiento y jerárquicos
- Gráficos usando ggplot2

Validación interna

El formato simplificado de la función eclust () es el siguiente:

eclust (x, FUNcluster = "kmeans", hc_metric = "euclidiano", ...)

• x: vector numérico, matriz de datos o data frame

• FUNcluster: una función de clúster que incluye "kmeans", "pam", "clara", "fanny", "Hclust", "agnes" y "diana"

• hc_metric: especifica la métrica que se utilizará para calcular las disimilaridades entre las observaciones.

Los valores permitidos son aquellos aceptados por la función dist () [incluyendo "euclidiana", "manhattan", "máxima", "canberra”, "Binary", "minkowski"] y medidas de distancia basadas en la correlación ["pearson”, "Spearman" o "kendall”]

Validación interna

La función eclust () devuelve un objeto de clase eclust que contiene el resultado de la Función utilizada (por ejemplo, kmeans, pam, hclust, agnes, diana, etc.).

Incluye también:

- cluster: la asignación de cluster de observaciones
- nbclust: la cantidad de clusters
- silinfo: la información de la silueta de las observaciones (silhouette)
- tamaño: el tamaño de los clusters
- data: una matriz que contiene los datos originales o estandarizados (si stand =CIERTO)
- gap_stat: que contiene estadísticas de brechas (gap)

Validación interna

Validación de cluster: Gráfico de Silhouette

Recuerde que el coeficiente de silueta ***(Si)*** mide cuán similar es un objeto i a los otros objetos en su propio clúster frente a los del clúster vecino. Valores de ***Si*** rango de **(1 a – 1)**:

- Un valor de Si cercano a 1 indica que el objeto está bien agrupado. El objeto i es similar a los otros objetos en su grupo.

- Un valor de Si cercano a -1 indica que el objeto está pobremente agrupado, y la asignación a algún otro grupo probablemente mejore los resultados generales.

Es posible dibujar coeficientes Silhouette de observaciones usando la función fviz_silhouette () [factoextra package], que también imprimirá un resumen del análsisi Silhouette.

Validación externa

Cálculo del índice de Dunn y otras estadísticas de validación del clúster

La función cluster.stats () [fpc package] y la función NbClust () [en NbClust paquete] se puede utilizar para calcular el índice de ***Dunn*** y muchos otros índices.

El formato simplificado es:

cluster.stats (d = NULL, clustering, al.clustering = NULL)

- d: un objeto de distancia entre casos generado por la función dist ()
- clustering: vector que contiene el número de clúster de cada observación
- alt.clustering: vector que indica una agrupación alternativa

Validación externa

La función cluster.stats () devuelve una lista que contiene muchos componentes útiles para analizar las características intrínsecas de una agrupación:

• cluster.number: cantidad de clusters

• cluster.size: vector que contiene el número de puntos en cada grupo

• average.distance, median.distance: vector que contiene distancias promedio / mediana dentro del cluster

• average.between: distancia promedio entre clusters. Queremos que sea tan grande

como sea posible

average.within: distancia promedio dentro de los clusters. Queremos que sea tan pequeño como sea posible

• clus.avg.silwidths: vector de anchos promedio de silueta del clúster. Recordemos que, el el ancho de la silueta también es una estimación de la distancia promedio entre los grupos. Su valor está comprendido entre 1 y -1 con un valor de 1 que indica buena agrupación.

• within.cluster.ss: una generalización de la suma de cuadrados dentro de los clusters (kmeans función objetivo), que se obtiene si d es una matriz de distancia euclidiana.

• dunn, dunn2: índice Dunn

• corrected.rand, vi: dos índices para evaluar la similitud de dos agrupamientos: el índice de Rand corregido y el índice VI de Meila

Todos los elementos anteriores se pueden usar para evaluar la calidad interna de la agrupación.

Validación externa

Entre los valores devueltos por la función cluster.stats (), hay dos índices para evaluar la similitud de dos agrupaciones, el índice de Rand corregido y el de Meila VI.

- Sabemos que los datos del iris contienen exactamente 3 grupos de especies.
- ¿El agrupamiento K-means coincide con la verdadera estructura de los datos?
- Podemos usar la función cluster.stats () para responder a esta pregunta.

Comencemos por calcular una tabulación cruzada entre clusters k-means y la referencia variable especies.

Puede observarse que:

• Todas las especies de setosa (n = 50) se han clasificado en el grupo 1

• Se ha clasificado un gran número de especies versicor (n = 39) en el grupo 3. Algunos de ellos (n = 11) se han clasificado en el grupo 2.

• Una gran cantidad de especies virginica (n = 36) ha sido clasificada en el grupo 2. Algunos de ellos (n = 14) se han clasificado en el grupo 3.

Es posible cuantificar la coincidencia entre especies y clusters k-means usando ya sea el índice de Rand corregido y el VI de Meila

Validación externa

El índice de Rand corregido proporciona una medida para evaluar la similitud entre dos particiones, ajustadas por casualidad.

Su rango es -1 (sin coincidencia) a 1 (perfecta coincidencia). La coincidencia entre especies y la solución de clúster es 0.62 usando el índice Rand y 0.748 usando el VI de Meila.

El mismo análisis se puede calcular tanto para la agrupación PAM como para la agrupación jerárquica.

La validación de clústeres externos se puede usar para seleccionar el algoritmo de agrupamiento adecuado para un conjunto de datos dado.

Validación externa

Resumen

Describimos cómo validar los resultados de agrupamiento utilizando el método Silhouette y el Índice de Dunn.

Esta tarea se facilita utilizando la combinación de dos funciones R: eclust () y fviz_silhouette en el paquete factoextra.

También demostramos cómo evaluar coincidencias entre un resultado de agrupamiento y una referencia externa.