CONSTRUCCIÓN E INTERPRETACIÓN DEL COEFICIENTE SILUETA EN AGRUPAMIENTOS POR K-MEDIAS

ISADORE NABI

Como se señala en (Datanovia, 2021), el término validación de agrupamiento (cluster validation) se utiliza para diseñar el procedimiento de evaluación de la bondad de los resultados del algoritmo de agrupamiento. Esto es importante para evitar encontrar patrones en datos aleatorios, así como en la situación en la que desea comparar dos algoritmos de agrupamiento. En general, como se señala en el lugar citado, las estadísticas de validación de agrupamiento se pueden clasificar en tres clases [Charrad et al. 2014, Brock et al. (2008), Theodoridis y Koutroumbas (2008)]:

1. Validación interna de clústeres, que utiliza la información interna del proceso de agrupamiento para evaluar la bondad de una estructura de agrupamiento sin referencia a información externa. También se puede utilizar para estimar el número de agrupaciones y el algoritmo de agrupamiento adecuado sin ningún dato externo.
2. Validación de clúster externa, que consiste en comparar los resultados de un análisis de clúster con un resultado conocido externamente, como etiquetas de clase proporcionadas externamente. Mide hasta qué punto las etiquetas de los grupos coinciden con las etiquetas de clase suministradas externamente. Dado que conocemos el número de clúster "verdadero" de antemano, este enfoque se utiliza principalmente para seleccionar el algoritmo de agrupamiento correcto para un conjunto de datos específico.
3. Validación relativa de clústeres, que evalúa la estructura de agrupamiento variando diferentes valores de parámetros para el mismo algoritmo (por ejemplo, variando el número de clústeres k). Generalmente se usa para determinar el número óptimo de clústeres.

El objetivo de los algoritmos de agrupamiento en particiones es dividir el conjunto de datos en grupos de objetos, de modo que:

1. Los objetos en el mismo grupo sean lo más similares posible y los objetos en diferentes los grupos son muy distintos.
2. Las medidas de validación interna reflejan a menudo la compacidad, la conectividad y la separación de las particiones del clúster.

Lo anterior significa que se desea que la distancia promedio dentro del clúster sea lo más pequeña posible y que la distancia media entre agrupaciones sea lo más grande posible. Las medidas de validación interna reflejan a menudo la compacidad, la conectividad y la separación de las particiones del clúster.

1. Compacidad o cohesión del racimo: mide qué tan cerca están los objetos dentro del mismo agrupamiento o conglomerado. Una variación más baja dentro del grupo es un indicador de una buena compacidad (es decir, un buen agrupamiento). Los diferentes índices para evaluar la compacidad de los conglomerados se basan en medidas de distancia, como las distancias por conglomerados dentro de la media/mediana de las distancias entre observaciones.

Matemáticamente hablando señalan (Kolmogórov & Fomin, 1978, pág. 96) que un sistema $\{M_\alpha\}$ de conjuntos se llama *cubrimiento* del espacio topológico T, cuando $\bigcup_\alpha M_\alpha = T$. Un cubrimiento, compuesto por conjuntos abiertos o cerrados, se llama cubrimiento abierto o cerrado, respectivamente. Si una parte de $\{M_{\alpha i}\}$ del cubrimiento M_α también constituye un cubrimiento del espacio T, se dice que $\{M_{\alpha i}\}$ es un subcubrimiento del cubrimiento $\{M_\alpha\}$. Así, si T es un espacio topológico de base numerable (para definir una topología en un espacio T hay que señalar en él el sistema de conjuntos, sin embargo, en muchos casos concretos es más cómodo señalar no la totalidad de subconjuntos abiertos del espacio dado, sino un sistema determinante de subconjuntos que permite definir unívocamente la totalidad de

los subconjuntos abiertos, y es a dicho sistema determinante al que se conoce como base de un espacio topológico; por otro lado, un conjunto es numerable si puede ponerse relación funcional biyectiva con el conjunto de números naturales), de todo cubrimiento suyo abierto se puede extraer un subcubrimiento finito o numerable. Establecido lo anterior, puede comprenderse el concepto de compacidad a través del teorema de Heine-Borel: de cualquier cubrimiento del segmento $[a, b]$ de la recta numérica por medio de intervalos se puede extraer un subcubrimiento finito (Kolmogórov & Fomin, 1978, pág. 104). Este teorema sigue siendo válido cuando en vez de intervalos se consideran cualesquiera conjuntos abiertos: de todo cubrimiento abierto del segmento $[a, b]$ se puede extraer un subcubrimiento finito. Así, un espacio topológico T se llama *compacto*, cuando cualquier cubrimiento abierto suyo contiene un subcubrimiento finito.

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 grupos distancias mínimas por pares entre objetos en diferentes grupos.

Matemáticamente, un espacio vectorial separable es aquel espacio provisto de un conjunto numerable siempre denso dentro de sí. Lo anterior significa que el espacio vectorial V debe tener un conjunto ϕ que sea denso dentro del espacio vectorial V. Que un conjunto ϕ sea denso dentro de V implica que para todo $x \in V$ cualquier vecindario N de x de radio δ, escrito como $N(x, \delta)$, contiene elementos de ϕ, es decir, que al menos un elemento de ϕ se encontrará en N. Se puede definir el vecindario $N(x, \delta)$ como $bola$ con radio $\delta > 0$ centrado en x. Nótese aquí algunas cuestiones. Como puede observarse, el vecindario en cuestión viene dado por el radio δ centrado en el vector x y, al ser el radio equidistante a todos los puntos que rodean su punto x de referencia (el vecindario que rodea x), se formará un $bola$, la cual será precisamente el vecindario de x. Evidentemente, el radio no podría ser menor o igual a cero, pues no se formaría ningún entorno al vector x.

3. Conectividad: corresponde a la medida en que los elementos se colocan en el mismo clúster que sus vecinos más cercanos en el espacio de datos. La conectividad tiene un valor entre 0 e infinito y debe minimizarse.

Matemáticamente hablando, la conectividad de un subconjunto[1] dentro de un espacio topológico es la cualidad que establece que tal subconjunto no puede ser descrito como unión disjunta de dos conjuntos abiertos no vacíos del espacio topológico en cuestión. Lo anterior significa que, si se separa el espacio topológico en dos conjuntos abiertos, es decir, que los extremos de cada nuevo conjunto (resultado de la separación) no se incluyan en estos dos conjuntos, al volverlos a unir el resultado es equivalente al conjunto original y, además, esta intersección es vacía (que no tienen elementos comunes entre sí); ambos requisitos no se cumplen simultáneamente, es decir, si se separan en dos subconjuntos, es posible lograr que al intersecarlos (volvernos a unir) el resultado sea un conjunto vacío, pero solo a costa que se omita algún punto de los originales, por lo cual no se vuelve al conjunto original; por otro lado, es posible volver al conjunto original, pero solo a costa de no omitir ninguno de los puntos originales, por lo cual su intersección no sería un conjunto vacío.

Formalmente y recordando lo planteado anteriormente, un conjunto conexo es un subconjunto de $C \subseteq X$ de un espacio topológico (X, T) que no puede ser descrito como una unión disjunta de dos conjuntos abiertos no vacíos de la topología, en donde T es la colección de conjuntos abiertos del espacio topológico. En otras palabras, está formado por una sola pieza y no es divisible.

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

[1] Además de ser también un subespacio, pues dado un espacio vectorial V, se dice que un subconjunto no vacío $U \subseteq V$, es un subespacio vectorial de V cuando al restringir las operaciones de suma y multiplicación por escalares (constantes) para V a U, este es un espacio vectorial.

$$Índice = \frac{(\alpha \times Separación)}{(\beta \times Compacidad)},$$

donde α y β son pesos.

Matemáticamente hablando,

Como señala (Datanovia, 2021), el análisis de silueta mide qué tan bien se agrupa una observación y estima la distancia promedio entre los agrupamientos. El gráfico de silueta muestra una medida de qué tan cerca está cada punto en un grupo de puntos en los grupos vecinos.

La metodología de cálculo para cada observación i, el ancho de la silueta s_i, consiste en que para cada observación i se calcula la disimilitud promedio a_i entre i y todos los demás puntos del grupo al que pertenece i. Para todos los demás conglomerados C, a los que i no pertenece, se calcula la disimilitud promedio $d(i, C)$ de i con todas las observaciones de C. La menor de estas $d(i, C)$ se define como $b_i = \text{mín}_C\ d(i, C)$. El valor de b_i puede verse como la disimilitud entre i y su grupo "vecino", es decir, el más cercano al que no pertenece. Finalmente, el ancho de la silueta de la observación i se define mediante la fórmula: $S_i = (b_i - a_i\ )/max(a_i, b_i)$.

Lo anterior, expresado algorítmicamente, puede establecerse como en (Kumar, 2020):

1. Calcular a_i: la distancia promedio de ese punto con todos los demás puntos en los mismos grupos.
2. Calcular b_i: La distancia promedio de ese punto con todos los puntos en el grupo más cercano a su grupo.
3. Calcular s_i, el coeficiente de silueta o $i - ésimo$ punto usando la ecuación $S_i = (b_i - a_i\ )/max(a_i, b_i)$ antes expuesta.

La interpretación del coeficiente silueta debe hacerse de la siguiente forma:

1. Las observaciones con un S_i grande (casi 1) están muy bien agrupadas.
2. Un S_i pequeño (alrededor de 0) significa que la observación se encuentra entre dos grupos.
3. Las observaciones con un S_i negativo probablemente estén ubicadas en el grupo equivocado.

Adicionalmente, señala (Kumar, 2020) que el valor de la silueta es una medida de cuán similar es un objeto a su propio grupo (su cohesión frente a su grupo) en comparación con otros grupos (su separación frente a otros grupos). El valor del coeficiente silueta oscila en el intervalo $[-1, 1]$, donde un valor alto indica que el objeto se corresponde bien con su propio grupo y no se corresponde con los grupos vecinos. Si la mayoría de los objetos tienen un valor alto, la configuración de agrupación en clústeres es adecuada. Si muchos puntos tienen un valor bajo o negativo, es posible que la configuración de agrupación tenga demasiados o muy pocos clústeres.

Como señala (Kumar, 2020), para el agrupamiento por $K - medias$, existen tres hiperparámetros fundamentales para definir la mejor configuración del modelo:

1. Valores iniciales de clústeres
2. Medidas de distancia
3. Número de clústeres

¿Pero qué son los hiperparámetros? Los hiperparámetros son propiedades de configuración del modelo que definen el modelo y permanecen constantes durante el entrenamiento del modelo. El diseño del modelo se puede cambiar ajustando los hiperparámetros. Como señala (Nabi, 2020, pág. 88), la mejor manera de pensar en los hiperparámetros es en términos de la configuración de un algoritmo que se puede ajustar para optimizar el rendimiento, del mismo modo que puede girar las perillas de una radio AM para obtener una señal clara.

Los parámetros del modelo son las propiedades de los datos de entrenamiento que el clasificador u otro modelo ML (Machine Learning) aprende durante el entrenamiento. Los parámetros del modelo difieren para cada experimento y dependen del tipo de datos y la tarea en cuestión. Los hiperparámetros son aquellos que proporcionamos al modelo, por ejemplo: número de nodos y capas ocultos, características de entrada, tasa de aprendizaje, función de activación, etc., en la red neuronal, mientras que los parámetros son aquellos que aprendería la máquina, como pesos y sesgos (Nabi, 2020, pág. 89).

Como señala (Kumar, 2020), los valores iniciales de los clústeres tienen un gran impacto en el modelo de clústeres, existen varios algoritmos para inicializar los valores. Las medidas de distancia se utilizan para encontrar puntos en grupos al centro del grupo, diferentes medidas de distancia producen grupos diferentes.

El número de conglomerados k es el hiperparámetro más importante en el agrupamiento de $K - medias$. Si se sabemos de antemano el número de grupos en los que agrupar los datos, entonces no es de interés ajustar el valor de k. Por ejemplo, $k = 10$ para el conjunto de datos de clasificación de dígitos MNIST.

Si no se tiene siquiera una idea sobre el valor óptimo de k, existen varios métodos para encontrar el valor óptimo/mejor de k, entre los que se encuentra el método de la silueta o silhouette antes expuesto, que según (Kumar, 2020) es superior a otros métodos como el método Elbow o método del codo.

REFERENCIAS

Datanovia. (18 de Mayo de 2021). *Cluster Validation Statistics: Must Know Methods*. Obtenido de CLUSTER VALIDATION ESSENTIALS: https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods

Kolmogórov, A. N., & Fomin, S. V. (1978). *Elementos de la Teoría de Funciones y del Análisis Funcional* (Tercera ed.). (q. e.-m. Traducido del ruso por Carlos Vega, Trad.) Moscú: MIR.

Kumar, S. (18 de Octubre de 2020). *Silhouette Method — Better than Elbow Method to find Optimal Clusters*. Obtenido de Towards Data Science: https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

Nabi, I. (2020). *Sobre los Estimadores de Bayes, el Análisis de Grupos y las Mixturas Gaussianas.* Documento inédito.