INDICE DE DUNN

Miguel Cárdenas-Montes

Las técnicas de clustering permiten agrupar datos en función de su similaridad. Paralelamente, existen índices que evalúan cómo de separables son las agrupaciones producidas por los algoritmos de clustering. El índice de Dunn es uno de ellos.

Objetivos:

- Conocer el uso práctico del índice de Dunn para evaluar la separación de clústeres formados con técnicas como k-means o DBScan, permitiendo obtener el número óptimo de clústeres.

Este documento puede contener imprecisiones o errores. Por favor no lo utilice para citarlo como una fuente fiable.

1 *Indice de Dunn*

El índice de Dunn [1] [2] es una métrica para evaluar el buen funcionamiento de los algoritmos de clustering. El objetivo de este índice es identificar un conjunto de clústeres que sean compactos, con una varianza pequeña entre los miembros del clúster, y que éstos estén bien separados de los miembros de otros clústeres. Un valor más alto del índice de Dunn indica un mejor rendimiento del algoritmo de clustering. Por lo tanto, este índice sirve para encontrar en número óptimo de clústeres en un conjunto de datos.

[1] J. C. Dunn. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104, 1974

[2] Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Anirban Mukhopadhyay. *Multiobjective Genetic Algorithms for Clustering - Applications in Data Mining and Bioinformatics*. Springer, 2011. ISBN 978-3-642-16614-3. 10.1007/978-3-642-16615-0. URL http://dx.doi.org/10.1007/978-3-642-16615-0

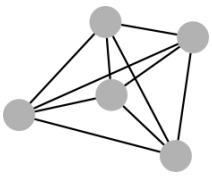

Figura 1: Ejemplo de distancia dentro del cluster. Un valor adecuado del índice de Dunn requiere que este valor sea bajo.

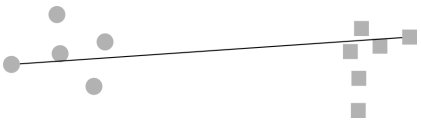

Figura 2: Ejemplo de distancia entre clústeres: la distancia entre los dos puntos más lejanos de diferentes clústeres. Un valor adecuado del índice de Dunn requiere que este valor sea alto.

Figura 3: Ejemplo de distancia entre clústeres: la distancia entre los dos puntos más cercanos de diferentes clústeres. Un valor adecuado del índice de Dunn requiere que este valor sea alto.

El índice de Dunn tiene un valor entre cero y infinito, y debe ser lo más alto posible. Por lo tanto, la distancia entre los miembros de un

clúster debe ser los más baja posible, y la distancia entre los clústeres lo más alta posible. En la ecuación 1 se representa el índice de Dunn.

$$D = \min_{1 \leq i \leq n} \left\{ \min_{1 \leq j \leq n, i \neq j} \left\{ \frac{d(i,j)}{\max_{1 \leq k \leq n} d'(k)} \right\} \right\} \tag{1}$$

donde $d(i,j)$ representa la distancia entre los clústeres i y j, y $d'(k)$ mide la distancia dentro del cluster k. El lector debe tener en cuenta que no queda fijado en la definición qué medida entre clústeres o dentro del cluster se deben utilizar.

Por ejemplo, en el caso de la distancia entre clústeres puede utilizarse la distancia más corta entre dos puntos de diferentes clústeres, o la distancia más larga, o la distancia entre los centroides. También pueden utilizarse diferentes indicadores de la distancia dentro de un clúster.

El índice Davies-Bouldin forma parte del mismo grupo de índices que el índice de Dunn.

Debe observarse que tampoco se especifica que medida de distancia (euclidea, manhattan, etc.) debe usarse, pero sí es obligatorio usar la misma en ambos conceptos.

Referencias

[1] J. C. Dunn. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[2] Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Anirban Mukhopadhyay. *Multiobjective Genetic Algorithms for Clustering - Applications in Data Mining and Bioinformatics*. Springer, 2011. ISBN 978-3-642-16614-3. 10.1007/978-3-642-16615-0. URL http://dx.doi.org/10.1007/978-3-642-16615-0.