Más sobre DBSCAN y los algoritmos basados en densidad

La base del algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) es identificar regiones densas a partir del número de objetos cercanos a un punto en una vecindad específica. Para realizar esta tarea categoriza los datos como núcleo, borde o atípico con base en los valores de dos parámetros: minPtos y épsilon. A partir de esta información, utiliza tres niveles de conectividad para construir los grupos, donde la conectividad implica un enfoque de encadenamiento de puntos:

- Densidad directamente alcanzable (*Direct Density Reachable*): dos puntos, P y Q, se consideran "directamente alcanzables" si Q está dentro del radio de búsqueda de P (definido por épsilon) y P es un punto núcleo. Representa la relación de vecindad directa entre los puntos (Fig. 1.a).

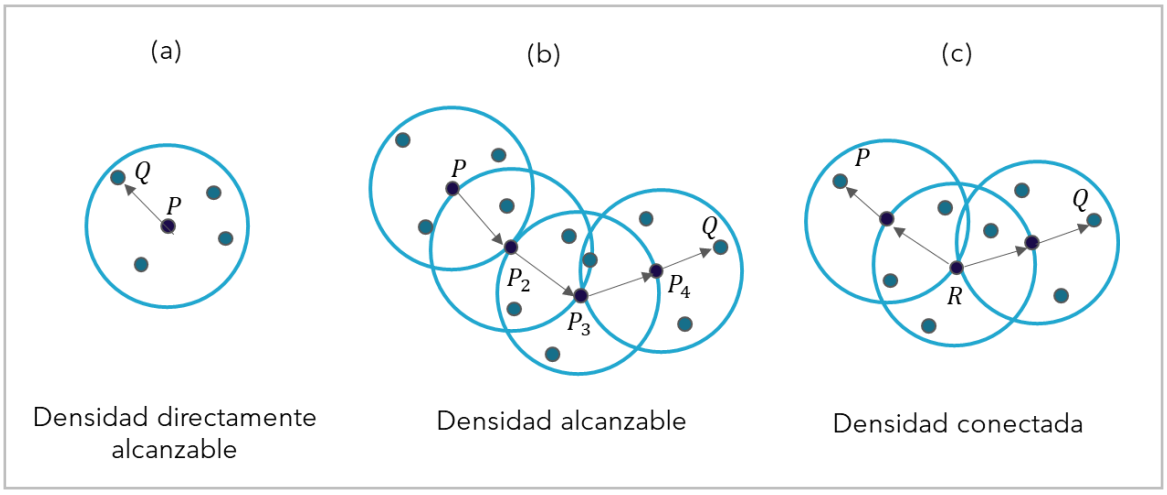

Fig. 1. Niveles de conectividad en DBSCAN.

- Densidad alcanzable (*Density Reachable*): dos puntos, P y Q, se consideran "alcanzables" si existe una secuencia de puntos núcleo $\{P_1, P_2, \ldots, P_n\}$, donde $P_1 = P$ y $P_n = Q$, y cada punto P_{i+1} es directamente alcanzable desde P_i. En otras palabras, hay un camino de puntos núcleo que

conecta a P con Q, lo que permite que los grupos se extiendan a lo largo de una serie de puntos (Fig. 1.b).

- Densidad conectada: dos puntos, P y Q, se consideran "conectados" si ambos son directamente alcanzables desde un mismo punto núcleo R. Esto permite extender el grupo agregando puntos borde desde los puntos núcleo (Fig. 1.c).

Un grupo se define entonces como un conjunto de puntos conectados por densidad. El algoritmo puedes visualizarlo en la Fig. 2.

Algoritmo DBSCAN

Entrada = {D: conjunto de datos, Eps: tamaño o radio de la vecindad, minPtos: número mínimo de datos en la vecindad Eps}

Comienzo_procedimiento

1. *Determinar los puntos núcleo.*
2. *A partir de un punto núcleo no asignado a un grupo, crear un nuevo clúster con los puntos directamente alcanzables por densidad.*
3. *Extender el grupo con los puntos alcanzables por densidad.*
4. *Buscar todos los puntos conectados por densidad y asignarlos al mismo grupo del punto núcleo.*
5. *Repetir los pasos 2 – 4 hasta que no hayan más puntos núcleo.*

Fin-procedimiento

***Salida** = regiones conectadas (grupos), puntos atípicos.*

Fig. 2. Algoritmo DBSCAN.

Otros algoritmos de agrupación basados en densidad

Se han propuesto otros algoritmos para la agrupación basada en densidad y algunos de ellos son variantes de DBSCAN, como HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*). Este algoritmo utiliza una estructura jerárquica para la extracción de grupos a diferentes niveles de densidad. Para esto emplea diversos valores de épsilon por lo que solo requiere que le especifiquemos como parámetro el tamaño mínimo de un grupo. A diferencia de DBSCAN, puede encontrar grupos con densidades variables sin tener que elegir primero un umbral de distancia. Sin embargo, es más complejo computacionalmente. Otros métodos que puedes

explorar son OPTICS (*Ordering Points To Identify the Clustering Structure*) y DENCLUE (*DENsity-based CLUstEring*).

Bibliografía

Aggarwal, Ch., Reddy, Ch. (2013). *Data clustering*. O'Reilly.

Campello, R.J.G.B., Moulavi, D., Sander, J. (2013). *Density-Based Clustering Based on Hierarchical Density Estimates*. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science. Vol 7819: 160–172. Springer.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X. (2017). *DBSCAN Revisited: Why and How You Should (Still) Use DBSCAN*. ACM Transactions on Database Systems. 42(3). Article 19. https://doi.org/10.1145/3068335.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.