El clústering espectral. Grafos más reducción de la dimensionalidad

El clústering espectral busca agrupar elementos basándose en las relaciones entre los datos a partir de un grafo de conectividad. El conjunto de datos se proyecta en un espacio de características de menor dimensión, en el cual los grupos se pueden construir mediante la aplicación de un algoritmo de agrupación, como K-medias. A continuación, se describen los pasos que realiza el algoritmo:

1. Construcción de la matriz de afinidad. Primero, se construye un grafo de afinidad en la forma de una matriz de adyacencia, la cual mide la similitud o "relación" entre cada par de datos. Para este propósito se pueden emplear métricas como el Kernel Gaussiano y métodos basados grafos, como el de vecindad Épsilon (*Épsilon-neighborhood graph*) y K-vecinos más cercanos (*k-nearest neighbor graph*).

2. Proyección en un nuevo espacio. El siguiente paso es calcular los vectores y los valores propios de la matriz Laplaciana que se obtiene a partir de la matriz de afinidad. Los vectores propios se utilizan entonces para proyectar los datos en un espacio de dimensión reducida.

3. Determinación de los grupos. A continuación, se aplica un algoritmo de agrupación, como K-medias, en el nuevo espacio de características. Aunque este algoritmo suele ser el más común en este contexto, es posible utilizar otros métodos de clústering, como el jerárquico o el basado en densidad.

¿Cuáles son sus hiperparámetros?

- Número de grupos. Como se suele utilizar el algoritmo K-medias es necesario determinar el valor de K que indica el número de grupos a formar en el espacio de características.
- Matriz de afinidad. El rendimiento de la agrupación espectral depende en gran medida de la calidad de esta matriz, la cual se utiliza para modelar las relaciones entre los datos. Existen diferentes métodos para construirla los cuales, a su vez, tienen hiperparámetros que también deben ser ajustados.

Como para otros métodos, es posible hacer una búsqueda sobre una "grilla" de valores de los hiperparámetros y hacer la selección de la mejor combinación con base en una métrica de evaluación.

En resumen.

- El clústering espectral es capaz de identificar grupos de formas y tamaños irregulares en el espacio original de los datos.
- Al contar con un paso de reducción de la dimensionalidad puede manejar datos con muchos atributos.
- Además, al basarse en un grafo de las relaciones entre los datos puede identificar grupos o comunidades bien conectadas.

Bibliografía

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

Bonaccorso, G. (2020). *Mastering Machine Learning Algorithms. Second Edition*. Packt Publishing.

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.