Clustering

Curso 2019/2020

Sofía Almeida Bruno
Daniel Bolaños Martínez
José María Borrás Serrano
Fernando de la Hoz Moreno
Pedro Manuel Flores Crespo
María Victoria Granados Pozo

Índice

1. Introducción

El clustering consiste en agrupar objetos similares. Dos objetos se consideran similares si, considerando alguna medida de error, las observaciones podrían ser del mismo objeto. Esta clasificación ocurre constantemente en nuestra vida diaria, damos el mismo nombre a objetos que difieren en detalles insignificantes.

Si tenemos una serie de observaciones sin clasificar, el objetivo del clustering es agrupar los datos en clases o clusters. Por ejemplo, cuando en ámbitos biológicos se quiere determinar las especies de una planta concreta. Este tipo de problema aparece cuando queremos no solo identificar especies nuevas, sino también cuando queremos establecer las relaciones entre ellas. El término clustering se considera sinónimo a taxonomía numérica o clasificación. En el ámbito de la ciencia de datos se considera problemas diferentes clasificación y agrupamiento. En el primero conocemos de antemano las clases de los objetos, mientras que en el segundo, a partir de los grupos creados se inferirán las características principales de los grupos.

Utilizando el lenguaje matemático, podemos definir el problema como sigue:

> Dadas $x_1, \ldots, x_n$ medidas de p variables en n objetos considerados *heterogéneos*. El objetivo del análisis cluster es agrupar estos objetos en k clases *homogéneas*, donde k es también desconocido (aunque habitualmente se asume que es mucho menor que n).

Decimos que un grupo es *homogéneo* si sus miembros están cerca unos de otros pero los miembros de otros grupos son muy diferentes a estos. Esto lleva a definir dos métricas entre los puntos para indicar el grado de alejamiento y el de asociación o similaridad. Se pueden tomar distintas distancias, creando aproximaciones diferentes al problema (trataremos este tema en más profundidad en la Sección 2).

El análisis cluster se aplica en numerosos campos como las ciencias naturales, médicas, económicas, *marketing*, ... En *marketing*, por ejemplo, es útil dividir a los clientes y conocer las necesidades de cada segmento de mercado para lograr alcanzar a los clientes potenciales. En psicología puede ser útil encontrar tipos de personalidad a partir de los cuestionarios realizados. En arqueología se puede aplicar esta técnica para clasificar objetos en diferentes periodos.

Para llevar a cabo un análisis cluster hay que realizar principalmente dos pasos:

1. Elegir una medida de proximidad.
2. Elegir un algoritmo para construir los grupos. Hay dos grupos principales: de particionamiento y jerárquicos (que a su vez se dividen entre divisivos y aglomerativos)

2. Medidas

3. Métodos de agrupamiento

3.1. Jerárquicos

3.2. No jerárquicos

4. Número de clústeres

5. Parte práctica

6. Bibliografía

Referencias