Clustering: Un paquete R para facilitar el análisis de algoritmos de agrupamiento.

1st Luis Alfonso Pérez Martos
Departamento de Informática
Universidad de Jaén, España
lapm0001@gmail.com

2nd Pedro González
DaSCI Andalusian Research Institute
Universidad de Jaén, España
pglez@ujaen

3rd Cristóbal J. Carmona
DaSCI Andalusian Research Institute
Universidad de Jaén, España
ccarmona@ujaen.es

***Abstract*—El agrupamiento es una técnica de aprendizaje no supervisado frecuentemente estudiada y aplicada cuyo objetivo es la división de los datos en grupos de objetos similares. Es bastante común entre los investigadores, ya que permite extraer conocimientos de forma rápida y sencilla. Además, su uso es adecuado para la clasificación automática de datos con el fin de revelar cierta similitud entre ellos. En este trabajo presentamos el paquete `Clustering` desarrollado íntegramente en R, que contiene un conjunto de algoritmos de agrupamiento y que busca dos objetivos: primero, agrupar los datos de forma homogénea estableciendo diferencias entre grupos; y segundo, generar un ranking de algoritmos y atributos a partir del conjunto de datos. Este paquete incorpora una GUI que facilita su uso.**

***Index Terms*—Técnicas de aprendizaje no supervisado, Agrupamiento, R, Clustering.**

I. Introducción

Explorar las propiedades de la información para generar grupos es una técnica de aprendizaje no supervisado conocida como agrupamiento [1] [2]. Esta técnica es un modelo de datos conciso en el que un conjunto de datos debe ser particionado e introducido en grupos. Estos grupos deben cumplir dos condiciones: que los grupos sean lo más dispares posible y que los elementos contenidos sean lo más parecidos. Por regla general los algoritmos de agrupamiento se basan en la optimización de una función objeto, que suele ser la suma ponderada de las distancias a los centros. En la literatura podemos agrupar los datos de múltiples formas, entre las que destacamos [3]: particionales, jerárquicos, basados en densidad, en cuadrículas y en modelos. Uno de los algoritmos más conocidos que afronta el problema del agrupamiento es el k-means [4].

Normalmente, la tarea de comparar algoritmos de agrupamientos es tediosa, ya que debe realizarse manualmente. Esto requiere tiempo y en algunos casos podemos tener problemas en la transmisión de los resultados. Por tanto, para evaluar la distribución de los datos en grupos, es necesario indicar una variable categórica, por lo que la elección de una u otra variable de un conjunto de datos puede variar los resultados de la evaluación.
En este artículo presentamos el paquete `Clustering` [1]. Se trata de un paquete que permite comparar múltiples algoritmos de agrupamiento de forma simultánea y evaluar la precisión de sus resultados. El objetivo es permitir la evaluación de un conjunto de datos para determinar qué atributos ofrecen los mejores resultados para el agrupamiento. De esta forma, evaluamos los grupos creados, su distribucción y la categorización de los datos. La estructura de esta contribución es la siguiente. En primer lugar, en la sección II se presenta el concepto de agrupamiento y sus tipos. En la sección III tenemos la definición de las medidas de validación para evaluar la distribución de los datos en los grupos. Finalmente, la sección IV describe la estructura del paquete y presenta un ejemplo completo del funcionamiento del paquete.

[1]El paquete está disponible para su descarga en el siguiente https://CRAN.R-project.org/package=Clustering

II. Agrupamiento

El análisis de los grupos es un método de aprendizaje no supervisado que constituye la piedra angular de un proceso de análisis inteligente de datos. Se utiliza para la exploración de las interrelaciones entre una colección de patrones, organizándolos en grupos homogéneos. Se denomina aprendizaje no supervisado porque, a diferencia de la clasificación (conocida como aprendizaje supervisado), no se dispone de un etiquetado a priori que pueda utilizarse en la categorización de los datos [5]. El concepto básico de agrupamiento puede expresarse de la siguiente manera: agrupar es el proceso de identificación de grupos naturales dentro de los datos multidimensionales basados en alguna medida de similitud (Euclidean, Manhattan, etc.) [6]. Esta es una definición básica de agrupamiento, por lo que las variaciones en la definición del problema pueden ser significativas, dependiendo sobre todo del modelo especificado. Por ejemplo, un modelo generativo debería definir la similitud basándose en un mecanismo generativo probabilístico, mientras que un enfoque basado en la distancia utilizará una función de distancia tradicional para cuantificarla. Además, los tipos de datos especificados tienen un impacto significativo en la definición del problema.

A. Tipos de Agrupamiento

Existe una gran variedad de algoritmos de agrupamiento que se pueden clasificar en: jerárquicos, particionales, basados en densidad, en cuadrículas y en modelos [7]:

- Jerárquico: crea un desglose jerárquico de los datos en un dendograma que divide recursivamente el conjunto de datos en grupos cada vez más pequeños. Algunos