TFG

Aarón Casado Monge[1], Juan José Cuadrado Gallego[1]

University of Alcala, Polytechnic School, Computer Science Department, Scientific and Technological Campus, Politechnic Building. Office: O243, 28805, Alcala de Henares, Madrid, Spain

Abstract

Clusterización (qué es y para qué sirve) - Técnicas (para qué sirven) - Paquetes

Keywords: BigData, Statistics, Bioestatistics, Data Science, Clustering

1. Introducción

Desde finales del sigo XX se ha considerado que vivimos en la "era de la información", una etapa caracterizada por el incremento, desarrollo y propagación de emergentes tecnologías de la información y comunicación que han permitido al ser humano romper las barreras de la distancia, el tiempo y el lugar a la hora de comunicarse y compartir información; actividades que han sido decisivas en nuestra historia [1]. Sin embargo, la era en la que realmente vivimos es la "era de los datos"; cada día se generan más de dos mil quinientos millones de petabytes[1] de datos provenientes de comercios, ciencias, Internet y casi cualquier actividad del día a día [2] que acaban volcados en redes de ordenadores, sitios web, bases de datos y otros medios de almacenaje.

Esta explosión de datos, a la que se ha denominado Big Data, se debe al alto grado de computarización de la sociedad y el avance de herramientas de recolección y almacenamiento de datos. Negocios en todo el mundo generan grandes cantidades de datos derivados de transacciones, stock de productos, platillas de empleados, etc. Las ramas de la ciencia producen datos de manera constante frutos de experimentos, observaciones, recogida de muestras, etc. Y más recientemente, Internet y las redes sociales han sido las principales responsables del aumento excesivo de datos, siendo usadas por millones de personas simultáneamente.

Y, aunque esto ha supuesto una considerable mejora para la humanidad pues la información nunca había sido tan accesible, tammbíen ha traido consecuencías negativas y problemas, como el almacenamiento y organización de los datos, datos no estructurados que entorpecen su acceso y procesamiento, dificultades a la hora de analizar los datos apropiadamente pudiendo generar desinformación y complicaciones para mostrar los resultados de forma apropiada y aplicarlos de manera eficiente y útil en el mundo real [3].

Como resultado, ha surgido una nueva ciencia que se ha posicionado rápidamente como una de las disciplinas más influyentes de esta era: Data Science (Ciencia de los Datos), que debido a su reciente aparición, carece de una definición consensuada, pero podríamos concretarla como "*Ciencia que usa Estadística, Inteligencia Artificial (IA), Programación y Bases de Datos para posibilitar la extracción de conocimiento a partir de datos*" [4]. A su vez, dentro de esta cienca se han desarrollado otras tres ramas: Data Warehousing, Data Mining y Visualization; cada una de ellas enfocada a resolver o afrontar uno o varios de los problemas mencionados previamente: organización y agrupación de datos, análisis de los datos y presentación de los resultados, respectivamente.

Email addresses: aaron.casado@uah.es (Aarón Casado Monge), jcg@uah.es (Juan José Cuadrado Gallego)

[1]Un Petabyte es una unidad de información o almacienamiento de datos equivalente a un cuadrillon de bytes, mil terabytes o un millón de gigabytes. En este caso, es el equivalente a 2.5 quintillones de bytes.

2. Clustering

Clustering o Cluster Analysis, adaptado al español como **Clusterización**, Agrupamiento o Análisis de Grupos es un método de clasificación no supervisada perteneciente a Data Mining, que busca definir, para una característica determinada o Suceso Elemental (SE) [2], un conjunto de grupos de observaciones (suceso) [3] con valores cercanos. Estos grupos son los denominados clusters o grupos y permiten a partir de los diferentes sucesos elementales que configuran dicho suceso, asignar dicho SE al mismo. Clustering nos permite definir los valores de cada cluster durante el proceso de clasificación [5].

3. Referencias

[1] Alberts, D. S., & Papp, D. S. (1997). The information age: An anthology on its impact and consequences. Office of the Assistant Secretary of Defense Washington DC Command and Control Research Program (CCRP).

[2] Becoming A Data-Driven CEO — Domo. (2018). Data never sleeps 6.0 https://www.domo.com/solution/data-never-sleeps-6

[3] Xu, Z., & Shi, Y. (2015). Exploring big data analysis: fundamental scientific problems´. Annals of Data Science, 2(4), 363-372.

[4] Definición Data Science apuntes FCD

[5] Apuntes JJ clustering

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed., p. 740). 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann Publishers, Elsevier.

https://normas-apa.org/referencias/citar-diccionario/

https://www.scribbr.es/detector-de-plagio/

https://tablesgenerator.com/

elsevier dos páginas latex

[2] Definir suceso elemental

[3] Definir Suceso