
Abstract

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

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

1. Introducción

Desde finales del siglo 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”, donde cada día se generan más de dos mil quinientos millones de petabytes¹ 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, también ha traído consecuencias 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 la actualidad: **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, Programación y Bases de Datos para posibilitar la extracción de conocimiento a partir de datos*” [4]. A su vez, dentro de esta ciencia 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 mismos y presentación de los resultados, respectivamente.

De entre estas nuevas disciplinas, **Data Mining** es la que se centra en el procesamiento de los datos, procedimiento por el cual se obtiene la información, y se podría definir como “*Proceso de descubrimiento de patrones interesantes y conocimiento a*

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

¹Un Petabyte es una unidad de información o almacenamiento 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.

partir de grandes volúmenes de datos” [5], donde dentro de la misma podemos encontrar diferentes técnicas para encontrar patrones y relaciones, y dependiendo de cuál se aplique se puede obtener un resultado totalmente diferente incluso con el mismo conjunto de datos, por lo que es fundamental emplear la técnica apropiada en función del del objetivo a conseguir, los datos con los que se pretende trabajar y el ámbito de aplicación.

Y en algunos de los campos más importantes del mundo moderno como la analítica de negocios, el reconocimiento de imágenes, las búsquedas web, seguridad, biología y ciencias de la salud; existen dificultades a la hora de clasificar o agrupar ciertos datos porque estos no disponen de una etiqueta o valor conocido por el que se pueda hacerlo, pues este no existe o no ha sido definido. Para poder afrontar este problema se utiliza la técnica de **Clustering**, que permite exactamente generar valores y etiquetas para un conjunto de datos realizando agrupaciones denominadas *clusters* o grupos, donde los datos de un mismo cluster sean muy similares entre ellos y a su vez tengan diferencias claras con datos de otros clusters, permitiendo que cada grupo resultante puede ser etiquetado y tratado como una clase propia.

De esta manera, el objetivo de este Trabajo de Fin de Grado (TFG) es realizar un estudio sobre Clustering, exponiendo de manera teórica qué es este método y qué tipo de utilidades tiene, así como el desarrollo de las diferentes técnicas que existen y las aplicaciones que estas ofrecen. Posteriormente se explorará este método dentro de Bioinformática, un campo centrado en desarrollar técnicas y programas software para analizar datos biológicos, viendo qué aporta a dicha disciplina y cuáles de las técnicas expuestas previamente se emplean y por qué.

Una vez realizado el marco teórico previo, se pretende hacer una parte práctica los diferentes paquetes que ofrecen técnicas de Clustering tanto de carácter general como dentro de Bioinformática para el lenguaje de programación R [6], uno de los más relevantes dentro de Data Science.

2. Clustering

Clustering o Cluster Analysis, adaptado al español como **Clusterización**, Agrupamiento o Análisis de Grupos, es un método de Data Mining que se basa en el Aprendizaje Automático (Machine Learning), una rama de la Inteligencia Artificial (Artificial Intelligence) que pretende desarrol-

lar sistemas que resuelvan problemas basándose en los resultados de experiencias previas, aprendiendo de sus errores. Concretamente, se fundamenta en uno de los métodos de aprendizaje explorados en esta disciplina: Aprendizaje No Supervisado, enfocado en el desarrollo y descubrimiento de nuevos conocimientos. Este, a diferencia de otros métodos de aprendizaje, no dispone de conocimiento previo sobre lo que poder aprender, por lo que su objetivo principal es discernir patrones y relaciones entre los datos para poder separarlos [7].

En esencia, el proceso de Clustering es el mismo, hasta el punto en que básicamente podríamos considerarlos sinónimos, puesto que Clusterización se aplica principalmente sobre conjuntos de datos que se desean organizar en diferentes grupos pero los valores que delimitan cada cluster son desconocidos y por lo tanto se definen en el propio proceso de clasificación.

De una manera más técnica, podríamos decir que Clustering busca definir para una determinada característica² o Suceso Elemental³ (SE), un conjunto de grupos de observaciones (suceso)⁴ con valores cercanos; donde los clusters permiten, dados los diferentes sucesos elementales que configuran un suceso, asignar cada SE al mismo cluster [8].

²Dicho de una cualidad que da carácter o sirve para distinguir a algo o alguien de sus semejante.

³Cada uno de los resultados más simples que se pueen obtener de la realización de un experimento. Obtener el número 3 al tirar un dado.

⁴Se denomina así al subconjunto total de resultados posibles al realizar un experimento. Obtener un 3 o sacar par al tirar un dado.

3. Referencias

- 140 [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).
- 145 [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.
- 150 [4] Definición Data Science apuntes FCD
- [5] Definición Data Mining libro cita100
- [6] The R Project for Statistical Computing. (n.d.). Retrieved from <https://www.r-project.org/>
- 155 [7] Moreno, A. (1994). [Aprendizaje automático](#). Llibre, Edicions UPC.
- [8] Apuntes JJ clustering
- [9] 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.
- 160