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: Data Mining, Clustering, R

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", donde 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, también ha traido 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)

[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.

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 desarrollar 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 característica[2] o Suceso Elemental[3] (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].

Para poder decidir si dos datos deben ser agrupados en el mismo cluster o por el contrario, separados, es necesario introducir los conceptos de similaridad y disparidad; cuanto más similares sean los datos, más tendrán en común y por lo tanto es más probable que terminen bajo el mismo cluster, y por otra parte, si los datos son diferentes entre sí, serán separados en clusters diferentes. Este criterio se basa en la proximidad de dichos objetos. Si midiéramos la similitud entre dos objetos i y j, esta devolvería 0 si ambos fueran totalmente distintos, y cuanto más elevado fuera el valor, más semejantes serían, siendo 1 el valor más alto, indicando que ambos datos son idénticos. De la misma manera, medir la disparidad entre los objetos daría resultados opuestos, con un 0 indicando que son idénticos y con un 1 que no tienen nada en común.

Esta forma de calcular la proximidad no solo

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

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

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

ayuda a la hora de agrupar los datos en clusters y formar subconjuntos a partir de los resultados, sino que con este proceso también somos capaces de detectar outliers, datos anómalos que pueden ser datos erróneos procedentes de errores de medida o fallos que deben ser eliminados o, datos correctos sumamente importantes que son diferentes al esto y deben ser analizados detenidamente. A la hora de agrupar los datos, puede que queden varios objetos sueltos sin pertenecer a ningún cluster, esos son los datos anómalos.

Además, Clustering también sirve como primera aproximación a la hora de procesar los datos, porque aunque de por sí, este proceso puede aportar información útil agrupando datos similares y clasificándolos, permitiendo un análisis de los resultados que puede dar lugar al descubrimiento de nuevos conocimientos, también se usa como método de preprocesamiento de datos para otros algoritmos de Data Mining que trabajarán sobre las clusters generados y los atributos seleccionados como criterio a la hora de crearlos, pues estos se pueden considerar clases nuevas de objetos.

Es gracias a estas características que Clusterización es empleada en muchos ámbitos del mundo actual, siendo la técnica de Aprendizaje No Supervisado más extendida. Campos como la analítica de negocios, el reconocimiento de imágenes, las búsquedas web, seguridad, biología y ciencias de la salud hacen uso habitual de esta técnica para clasificar tipos de consumidores que comparten preferencias, aunar bajo un mismo subconjunto muchas formas diferentes de escribir el mismo caracter para facilitar el reconocimiento de textos escritos a mano, agrupar resultados similares de una consulta en internet y mostrar los más relevantes dentro de cada grupo o el estudio de la taxonomía[5] de las especies. También se puede hacer uso de su capacidad para detectar datos anómalos con la finalidad de revelar posibles fraudes o transacciones financieras sospechosas, incluso ayudar en la disminución de crímenes analizando los resultados obtenidos tras clusterizar datos de detenciones y delitos; incluso aplicada a datos geofísicos[6] para interpretarlos y obtener resultados significativos. Asimismo, se utiliza a la hora de comparar comunidades en redes sociales permitiendo recomendar a los usuarios contenido de su agrado, o dentro del mundo sanitario, ayudando en la identificación y control de diversos tipos de enfermedades e incluso puede usarse como apoyo en la gestión de edificios públicos como bibliotecas, agrupando a los lectores por sus preferencias de manera similar a los consumidores de un negocio [9–13].

La utilidad de este método y la flexibilidad que ofrece con las diversas técnicas y formas de aplicarlo de las que dispone e es también parte fundamental de que sea tan comúnmente utilizado y con objetivos tan dispares en gran parte de las áreas del conocimiento. Sin embargo, también es un método frágil, pues depende en gran medida de los datos con los que se trabaje y el criterio escogido para determinar si dos objetos deben agruparse bajo el mismo cluster o separarlos, Por lo que es necesario seguir una serie de pasos a la hora de realizar un Análisis de Grupos de manera correcta [14]:

1. Primero, hay que seleccionar cuidadosamente los datos con los que se va a trabajar, puesto que tanto el tipo de dato como la cantidad de los mismos influyen directamente en los resultados. Si se utilizan demasiados datos el procesamiento puede tener un coste computacional alto y es difícil ofrecer una visualización elegante de los resultados. Pero si se escogen pocos datos se pierde información útil y puede dar lugar a equivocaciones. Es por ello que este paso suele realizarlo un experto en el sector o con su ayuda.
2. Segundo, se debe elegir la forma en la que se va a calcular la similaridad entre los diferentes datos, que analizaremos más adelante cuando analicemos las diferentes técnicas de Clustering.
3. Tercero, debe definirse el criterio con el que se va a tomar la decisión de agrupar en el mismo cluster dos datos. Es decir, escoger el umbral de similaridad a partir del cual dos datos pasan a formar parte del mismo cluster y qué característica o conjunto de ellas van a ser utilizadas para medir la similitud. La elección del umbral suele ser complicada, y es necesaria mucha experiencia o varias iteraciones de ensayo y error para encontrar un valor correcto.
4. Cuarto, hay que optar por uno de los diversos algoritmos de Clusterización que existen, pues los resultados pueden cambiar considerablemente al variar la estrategia con la que se

[5] Ciencia que trata de los principios, métodos y fines de la clasificación. Se aplica en particular, dentro de la biología, para la ordenación jerarquizada y sistemática, con sus nombres, de los grupos de animales y de vegetales.

[6] La Geofísica es la ciencia que estudia la Tierra desde el punto de vista de la física.

realiza la agrupación. Estos se verán a continuación.

5. Por último, una vez obtenidos los resultados hay que validarlos e interpretarlos. La resolución de este paso depende del objetivo inicial por el que se haya decidido realizar el análisis.

Como puede verse, tanto el paso inicial como el último requieren de la ayuda de personas cualificadas si pretenden realizarse correctamente, por lo que es en los pasos intermedios donde la intervención de computadoras y programas informáticos son más útiles y están más desarrollados, así como el siguiente tema a abordar para comprender el funcionamiento del método de Clustering. A continuación se explorarán las diferentes técnicas de Clusterización y de las que hablaremos posteriormnte dentro del apartado práctico con la ayuda del lenguaje de programación R.

2.1. Técnicas de Clustering

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] Definición Data Mining libro 100

[6] The R Project for Statistical Computing. (n.d.). Retrieved from https://www.r-project.org/

[7] Moreno, A. (1994). Aprendizaje automático. Llibre, Edicions UPC.

[8] Apuntes JJ clustering

[9] Alkhaibari, A. A., & Chung, P. (2017). Cluster analysis for reducing city crime rates. 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT). doi:10.1109/lisat.2017.8001983

[10] Song, Y., Meng, H., & Zhang, Y. (2010). Clustering analysis and its applications. 2010 Second IITA International Conference on Geoscience and Remote Sensing. doi:10.1109/iita-grs.2010.5602787

[11] Prabhu, J., Sudharshan, M., Saravanan, M., & Prasad, G. (2010). Augmenting Rapid Clustering Method for Social Network Analysis. 2010 International Conference on Advances in Social Networks Analysis and Mining. doi:10.1109/asonam.2010.55

[12] Baron, J. N., Aznar, M. N., Monterubbianesi, M., & Martínez-López, B. (2020). Application of network analysis and cluster analysis for better prevention and control of swine diseases in Argentina. PLoS ONE, 15(6), 1–26. https://doi.org/10.1371/journal.pone.0234489

[13] Li, J., & Chen, P. (2008). The application of Cluster analysis in Library system. 2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop. doi:10.1109/kamw.2008.4810639

[14] Carugo, O. & Eisenhaber, F. (2010). Data mining techniques for the life sciences (Vol. 609). HTotowa, NJ: Humana Press.

[15] 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.