Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Clusterización de grandes bases de datos: un ejemplo práctico.

Alar Urruticoechea[a], Elena Vernazza[b], Diana del Callejo Canal[c] , Margarita Canal Martínez[d], Ramón Álvarez Vaz[e]

[a] Universidad Católica del Uruguay (UCU), Uruguay.,

[b e] Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República (UdelaR), Uruguay.,

[c d] Instituto de Investigación de Estudios Superiores, Económicos y Sociales de la Universidad Veracruzana (IIESES-UV). Xalapa, Veracruz, México.,

[a] *alar.urruticoechea@ucu.edu.uy,* [b] *elenavernazza@gmail.com*, [c] *ddelcallejo@uv.mx*,
[d] *mcanal@uv.mx,* [e] *ramon@iesta.edu.uy*

Resumen

Realizar clusterización de grandes bases de datos categóricos presenta con regularidad un reto en los análisis estadísticos. El objetivo de este trabajo es proponer una alternativa viable para el tratamiento de grandes bases de datos. **Metodología:** Se utilizó la base de datos PISA 2018 con la muestra específica de México (n = 5874) y se estudiaron las variables tecnológicas. Se empleó el análisis de correspondencias múltiple (ACM) para las variables categóricas (tecnología) y el algoritmo Clustering Large Applications (CLARA) para la clusterización de grandes bases de datos. **Resultados:** Se encontraron 3 cluster, caracterizándose: Cluster 1: entre el 57% a 75% de los estudiantes no tienen acceso a las variables tecnológicas estudiadas exceptuando a las variables a) para uso en casa tienes internet (92% tienen y usan) y, b) tienen acceso a un celular con acceso a internet (95% tiene y usa). Cluster 2: entre el 70%-95% de los estudiantes tienen acceso y usan las variables tecnológicas estudiadas. Cluster 3: entre el 60%-90% de los estudiantes no tienen acceso a las variables tecnológicas estudiadas. **Discusión:** El uso de esta metodología parece útil a la hora de clusterizar grandes bases de datos.

Palabras claves: Big Data, Educación, Tecnología, CLARA.

Introducción

En la actualidad, analizar grandes cantidades de datos (Big Data) se ha vuelto una tarea indispensable en la investigación y desarrollo en todas las áreas desde la industria, hasta la academia (Nayar & Puri, 2017).

La naturaleza de Big Data es diferente a la de muestreos o diseños de experimentos que utilizan cantidades de datos manejables. El Big Data es una palabra relativamente nueva y existen varias definiciones al respecto. Sin embargo, Ganz & Rainsel la definen como una

nueva generación de tecnologías y arquitecturas, diseñadas para extraer de una forma económica y veloz la información de valor de grandes volúmenes de datos (2011).

El tratamiento de Big Data implica cuatro 'v': Volumen, variedad, velocidad y valor. Dentro de las herramientas estadísticas que pueden ser útiles para cumplir con las cuatro 'v' se encuentra la clusterización (Gantz & Rensel, 2011).

La clusterización consiste en descubrir grupos homogéneos al interior de una tabla de datos y heterogéneos entre ellos. Esta tarea se puede volver complicada cuando se trata de grandes volúmenes de datos, varios expertos en el área han propuesto diferentes algoritmos para solventar el desafío que implica clusterizar grandes volúmenes de datos. Los algoritmos del clúster se pueden clasificar en: Partition-based; Hierarchical-based; Density-based; Grid-Based and Model-Based (Nayar & Puri, 2017).

CLARA (Clustering Large Applications) es una extensión del algoritmo PAM (Partitioning Around Medoids) propuesto por Kaufman & Rousseeuw (1990). El algoritmo CLARA (Maechler, Rousseeuw, Struyf, Hubert & Hornik) en R-Studio (R Core Team, 2019) separa muestras de la base completa y aplica el algoritmo PAM sobre cada una de ellas. El principal motivo para crear CLARA es la deficiencia de PAM para trabajar con grandes volúmenes de información (Leiva-Valdebenito y Torres-Avilés, 2010).

En este estudio se utilizaron datos de la prueba PISA de la para el 2018 (PISA, 2018) Organización para la Cooperación y el Desarrollo Económicos (OCDE) originalmente 1118 variables y 612004 casos. Se filtraron variables correspondientes al acceso y uso de tecnología en México, se utilizó CLARA con la finalidad de identificar agrupaciones que ayudarán a retratar la realidad mexicana en acceso y uso de tecnología en jóvenes de 15 años para el año mencionado.

Se encontraron 3 clústeres, el primero caracterizado por jóvenes que no tienen acceso a las variables tecnológicas estudiadas, pero con acceso a internet en casa y en el celular. El segundo está caracterizado por estudiantes que tienen acceso y usan las variables tecnológicas estudiadas y el tercero que clasifica a estudiantes no tienen acceso a las variables tecnológicas estudiadas. El uso de CLARA parece útil a la hora de clusterizar grandes bases de datos.

Metodología

Aspectos generales

Los datos analizados se obtuvieron de la prueba PISA de la Organización para la Cooperación y el Desarrollo Económicos (OCDE). La base de datos filtrada utilizada contiene 10 variables de corte categórico referentes al uso de la tecnología para 5874 estudiantes de 15 años en México que tomaron la prueba en el año 2018.

Metodología Estadística

Después de la limpieza de la base de datos, se realizó un análisis exploratorio donde se obtuvieron las estadísticas descriptivas para cada una de las variables estudiadas. Posteriormente se trasladaron las variables originalmente categóricas a su dimensión numérica a través de un Análisis de Correspondencias Múltiple (ACM).

A partir de estas 7 dimensiones se realizó, utilizando el algoritmo CLARA, un análisis clúster. Se utilizaron 3, 4 y 5 clúster para poder comparar a través del gráfico de siluetas cual era la mejor agrupación para los datos. Se eligieron 3 clústeres, que posteriormente se caracterizan de acuerdo a tablas cruzadas entre los clústeres y las categorías originales.

Resultados y Discusión

Análisis descriptivo de la muestra

Del total de la muestra (n = 5874, México) el 52.5% son niñas y el 47.5% niños.

Análisis descriptivo de las variables de interés

Como se observa en la tabla 1, una gran cantidad de jóvenes mexicanos no cuenta con acceso a internet desde su casa, USB y celular sin internet. El resto de las variables, presentan alrededor del 50% de uso.

Tabla 1.- Frecuencia de uso - variables interés

En su casa cuenta con:	No tiene	Tiene no Usa	Tiene y usa
Computadora escritorio	**40.13**	**9.21**	**50.66**
Laptop	**46.31**	**8.95**	**44.74**
Tablet	**32.8**	**14.3**	**52.89**
Internet	**71.76**	**1.97**	**26.27**
Videojuegos	**35.10**	**12.46**	**52.43**
Celular con internet	**33.40**	**14.98**	**51.62**
Celular sin internet	**79.49**	**5.69**	**14.83**
Reproductor de Música	**47.62**	**11.32**	**41.06**
Impresora	**34.51**	**8.82**	**56.67**
USB	**81.65**	**10.38**	**7.97**

ACM y Clusterización

Los resultados del análisis de correspondencias múltiples indican que con 7 dimensiones se explica el 54.78% de la inercia (ver tabla 2). Estas 7 dimensiones son las que se usaron en el análisis clúster.

Tabla 2.- Inercias explicadas por dimensión

Dimensión	Inercia explicada	Inercia acumulada
1	**17.95**	**17.95**
2	**10.00**	**27.95**
3	**6.67**	**34.62**
4	**5.72**	**40.34**
5	**4.99**	**45.33**
6	**4.73**	**50.06**
7	**4.71**	**54.78**

Tal como se observa en la Figura 1. al seleccionar 3 clústeres los grupos presentan mayor homogeneidad interna que al seleccionar 4 clústeres, por lo que se decide quedarse con 3 clústeres.

La caracterización de los 3 clúster indica que:
Cluster 1: entre el 57% a 75% de los estudiantes no tienen acceso a las variables tecnológicas estudiadas exceptuando a las variables a) para uso en casa tienes internet (92% tienen y usan) y, b) tienen acceso a un celular con acceso a internet (95% tiene y usa).
Cluster 2: entre el 70%-95% de los estudiantes tienen acceso y usan las variables tecnológicas estudiadas.
Cluster 3: entre el 60%-90% de los estudiantes no tienen acceso a las variables tecnológicas estudiadas.

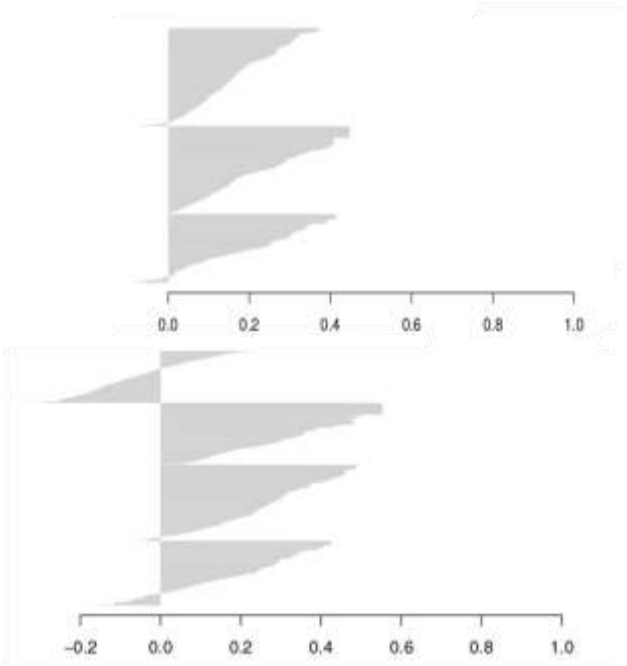

Figura 1. Gráfico de siluetas con 3 (superior) y 4 clústeres (inferior).

Conclusiones

Los resultados muestran que el algoritmo CLARA resulta de utilidad al momento de clusterizar grandes cantidades de datos.

La elección de tres clústeres es la mejor representación que se encontró hasta el momento para poder describir el acceso y uso de la tecnología para jóvenes de 15 años en México.

Dos de las dimensiones utilizadas en el ACM, siguen un comportamiento distinto al resto. En futuros estudios valdría la pena estudiar estos comportamientos con mayor detalle y proponer alternativas al ACM.

En un próximo estudio derivado de este, sería relevante explorar las diferencias y coincidencias entre el algoritmo CLARA y k-means.

Referencias

Gantz, J. & Reinsel, D. (2011). Extracting value from chaos. **IDC iview**, vol. 1142, pp. 1-12.

Nayar A.& Puri, V. (2017). Comprehensive Analysis & Performance Comparison Of Clustering Algorithms For Big Data. **Review of Computer Engineering Research**, Vol. 4, No 2, pp. 54-80.

Kaufman, L. & Rousseeuw, P. (1990). **Finding groups in data: An introduction to cluster analysis.** John Wiley and Sons, New York.

Leiva-Valdebenito, S. y Torres-Avilés, F. (2010). Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados. Un estudio comparativo. **Revista colombiana de estadística**, Vol. 33, No.2, pp. 321-339.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2019). **Cluster: Cluster Analysis Basics and Extensions.** R package version 2.1.0.

R Core Team (2019). **R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.** URL https://www.R-project.org/.

Sebastien Le, Julie Josse, Francois Husson (2008). **FactoMineR: An R Package for Multivariate Analysis.** Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

PISA (2018). Pisa Database. https://www.oecd.org/pisa/data/2018database/