Estudio de algunas propiedades del método de clustering k-modes mediante remuestreo

Diego Araujo , Ramón Álvarez-Vaz

Abstract En este documento se presenta el estudio de la performance de un método de clustering, el kmodes de perfil modal, teniendo en cuenta la variabilidad de algunas métricas que permiten evaluar la homogeneidad de los grupos y decidir una posible solución a la cantidadde grupos. Esa variabilidad se estudió en función del arranque el método, que es aleatorio en la elección de centros, viendo de ese modo la dependencia con la solución encontrada y también en la estructura de los datos por lo cual se trabaja con muestras de aprendizaje, variando el tamaño de la misma mediante remuestreo. Se trabaja con datos que surgen de un estudio en el campo de las finanzas corporativas, desarrollado durante la pandemia en Uruguay, donde se indaga acerca del IMPACTO COVID-19 EN LAS EMPRESAS URUGUAYAS en la toma de decisiones sobre reducción de personal, teletrabajo, capacitación y ambiente laboral entre otras.

Palabras clave: Clustering - K-Modes - Remuestreo

Introducción

El trabajo aquí presentado deriva de una investigación que se realizó sobre una muestra de empresas uruguayas, de forma de conocer el impacto que tuvo la pandemia del Covid-19 sobre las finanzas de las mismas, las medidas tuvieron que implementar para poder dar continuidad a sus negocios, las opciones brindadas por el gobierno que fueron mayormente utilizadas, entre otros aspectos. Para poder recolectar estos datos se utilizó un cuestionario de formato electrónico, en el cuál las respuestas a las preguntas enviadas se pueden codificar como unos y ceros ("Si / No"). Por lo que fue conveniente utilizar algún método de clustering para variables categóricas, de forma de crear el perfil modal de estas empresas y ver como quedaban segmentadas a través de distintos bloques de información.

Mas allá del tema inicial que trataba la investigación, interesa estudiar las propiedades del algoritmo en cuanto a su estabilidad bajo distintos escenarios, evaluando su rendimiento a través de distintas métricas y logrando aproximarnos a una cantidad k óptima de grupos, un problema común en este tipo de análisis. El algoritmo kmodes así como su similar para datos numéricos (k-means), se encuentra afectado por los arranques aleatorios y en este caso también por el tamaño de la muestra.

Es así que se proponen 2 escenarios:

- Escenario 1: Trabajar con una muestra de tamaño fijo y arranques aleatorios.
- Escenario 2: Haciendo remuestreo a distintos niveles (70 %, 80 % y 90 %) de la muestra original, para un k fijo.

Para asegurar la reproducibilidad de los resultados del análisis realizado, se dispuso el código y datos utilizados en un repositorio público al que se puede acceder a través de este link https://gitlab.com/iesta.fcea.udelar/finanzas-de-empresas-iesta.

Metodología

Se utiliza parte de la metodología propuesta por (Tsekouras et al. 2004) para clasificar atributos categóricos, a través del algoritmo mixto *Fuzzy C-modes*, y utilizada por Alvarez-Vaz y Massa para encontrar perfiles de infección parasitaria en escolares de Montevideo (Álvarez-Vaz, Alvarez, and Massa 2012) . En este caso, el algoritmo tiene una lógica de funcionamiento similar a la del algoritmo *k-means* y dada la naturaleza de las variables (binarias), es necesario el uso de otras medidas de disimilaridad, usando un método basado en frecuencias para actualizar los modos (Weihs et al. 2005a). Por lo tanto, del método mixto original planteado por (Tsekouras et al. 2004) se trabaja solamente con el algoritmo *k-modes* (Weihs et al. 2005b).

Métricas de Validación

Se proponen 3 métricas para realizar la validación interna de los cluster y evaluar tanto la compacidad de los grupos (homogeneidad), como la separación entre ellos.

- Diferencias Intracluster Absolutas (DIA): Cantidad total de diferencias que hay dentro de cada cluster, de cada elemento con respecto al perfil modal del grupo, a través de la medida definida en (1). Para un cluster k dado, se tiene la siguiente expresión:

$$DIA(k) = \sum_{i=1}^{n_k} d(x_i, y_k) \tag{1}$$

dónde n_k es la cantidad de elementos en el cluster k y y_k el centroide del grupo (en este caso el modo o perfil modal).

- Diferencias Intracluster Relativas (DIR): Cantidad total de diferencias dentro de cada cluster, pero condicionadas a la cantidad de elementos del grupo. Para un cluster k dado, se define como:

$$DIR(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} d(x_i, y_k) \tag{2}$$

- Coeficiente de silueta (CS): Métrica utilizada para calcular que tan similares son las observaciones o elementos en un mismo grupo en comparación con las observaciones de otros grupos(Maechler et al. 2022). Para una observación i se define como:

$$s(i) = \frac{b(i) - a(i)}{max\,(a(i), b(i))} \tag{3}$$

donde $a(i)$ es el promedio de disimilaridades de i con el resto de observaciones que pertenecen al mismo cluster que i, que se define como:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C, i \neq j} d(i, j) \tag{4}$$

dónde $|C_I|$ es la cantidad de elementos en el cluster al cual pertenece i y $b(i)$ es el promedio de disimilaridades de i con los elementos del cluster más cercano al cúal i no pertenece (cluster "vecino' ').

Esta última métrica en comparación con las primeras dos, nos da una información más íntegra del agrupamiento global, ya que no tiene en cuenta únicamente lo que sucede dentro del grupo, sino la separación a otros.

Discusión

Los resultados que se visualizan con *ggplot2* (Wickham 2016) encontrados con lo hecho hasta ahora permiten ver 2 aspectos del algoritmo de clusterización utilizado. Se puede ver que a medida que se va desagregando en más grupos, la heterogeneidad va aumentando, medida a través de 3 métricas que muestra que se llega a un k 'óptimo' en 2, decidiendo en base a la silueta (CS), ya que no hay empresas con valores negativos.

- Escenario 1: Para el escenario donde se prueban diferentes arranques (ya que es aleatorio) iterando 1000 veces se ve que la densidad media de pares discordantes (medida local) es mayor para $k = 2$, mostrando mayor dispersión relativa y mayor discontinuidad, mientras que la silueta es mayor y presenta un comportamiento con un cambio marcado al llegar a valores de 0.28.
- Escenario 2: Teniendo en cuenta que los datos utilizados son una muestra que tiene una tasa de respuesta baja (por lo tanto el número de observaciones es reducido), toma especial relevancia ver que tanta dependencia hay en los datos y de ahí el probar cambiando el tamaño de muestra de aprendizaje. Para $k = 2$ para la métrica (DIA), se observa que para un tamaño menor de muestra de aprendizaje la distribución está corrida a valores más bajos pero con mayor dispersión, mientras que a mayor tamaño de muestra de aprendizaje la variabilidad de (CS) es un poco menor y con menor dispersión en las iteraciones, resultando casi bimodal.

10 Álvarez-Vaz, Ramón, Federico Alvarez, and Fernando Massa. 2012. "Determinación de Tipologías de Infecciones Parasitarias Intestinales, En Escolares Mediante, Técnicas de Clustering Sobre Datos Binarios." In *III Jornadas Académicas de Facultad de Ciencias Económicas y de Administración-Universidad de La República*.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2022. *Cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster.

Tsekouras, George E., Dimitris Papageorgiou, Sotiris Kotsiantis, Christos Kalloniatis, and Panagiotis Pintelas. 2004. "Fuzzy Clustering of Categorical Attributes and Its Use in Analyzing Cultural Data." *International Journal of Computational Intelligence* 1: 147–51.

Weihs, Claus, Uwe Ligges, Karsten Luebke, and Nils Raabe. 2005b. "klaR Analyzing German Business Cycles." In *Data Analysis and Decision Support*, edited by D. Baier, R. Decker, and L. Schmidt-Thieme, 335–43. Berlin: Springer-Verlag.

———. 2005a. "klaR Analyzing German Business Cycles." In *Data Analysis and Decision Support*, edited by Daniel Baier, Reinhold Decker, and Lars Schmidt-Thieme.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.