Métodos no jerárquicos de clusters aplicados a las secuencias

Mag. Virginia Trevignani
Universidad Nacional del Litoral

Abril, 2019

A diferencia de los métodos jerárquicos en los que el objetivo del análisis es hallar un número óptimo de agrupamiento, estos métodos parten de una hipótesis sobre cual sería ese número. En general se suponen más de una solución y el análisis se basa en comparar indicadores computados para cada una.

En los métodos jerárquicos para el agrupamiento, en cada paso del algoritmo, sólo un objeto cambia de grupo (aquel no agrupado aún) y los grupos están conformados por los objetos previamente agrupados. Esto también incluye la situación en que avanzado el análisis, se agrupan ya grupos definidos en los pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo; la clasificación resultante tiene un número creciente de clases anidadas. Los métodos no jerárquicos o de partición comienzan con una solución inicial, un número de grupos fijado de antemano y agrupa los objetos para obtener dichos grupos.

Se ensayaron distintos modelos de construcción de tipologías para las tres universidades, variando el método de agrupamiento (jerárquico y no jerárquico) y las medidas de disimilitud usadas para estimar la matriz de distancias entre secuencias (OM, HAM, DHD).

El modelo de tipología con mejor calidad de ajuste y mejor clasificación de las secuencias (según el análisis de *Silhouette*) para las tres universidades es el que resulta de aplicar un método no jerárquico (PAM) con una matriz de distancias estimada en base a DHD. Describimos a continuación los elementos utilizados.

Las medidas de disimilitud más aplicadas al análisis de secuencias son DHD, HAM y OM; miden la distancia entre secuencias y constituyen el insumo para la construcción de tipologías de trayectorias Esas tres medidas dan como resultado distancias diferentes entre trayectorias interindividuales. DHD es una medida adecuada cuando el orden no es lo relevante, sino el calendario y el ritmo con el cual las relaciones laborales cambian de estados. HAM es una medida que computa distancias en función de las posiciones iguales o diferentes de los estados entre las secuencias. La medida OM estima el costo mínimo necesario de transformar una secuencia en otra (mediante operaciones de inserción, borrado y sustitución de un elemento de la secuencia, con costos diferentes) (Studer, 2013).

Las medidas de calidad de la partición de la técnica de conglomerados sirven para comparar distintos modelos y para guiar la elección de la mejor solución (cuántos grupos de trayectorias seleccionar). El PBC (Point Biserial Correlation); el HG (Hubert's Gamma) y el HGSD (Hubert's Somers' D) miden la capacidad del agrupamiento para reproducir la matriz de distancia original (mientras que la primera medida mide la capacidad de reproducir el valor exacto de las distancias, las otras dos están basadas en la concordancia). Las tres medidas varían entre -1 y 1: a medida que aumenta el índice, mejor es el agrupamiento obtenido. El HC (Hubert's C) mide la brecha entre el agrupamiento obtenido y el mejor posible; varía entre 0 y 1: valores más pequeños indican un mejor agrupamiento. El ASW

(Average Silhouette Width) y el ASWw (Average Silhouette Width weighted) miden la coherencia de la asignación de casos a grupos; varía entre -1 y 1: a medida que aumenta indica mayor coherencia (mucha distancia entre los grupos y alta homogeneidad intra-grupo). El CH y el CHsq (Calinski-Harabasz index) es un Pseudo F computado de las distancias (y el cuadrado de las distancias en el caso del segundo); varían de 0 en adelante. El R2 y el R2sq (Pseudo R2) indican la proporción de la varianza explicada por cada agrupamiento y varían entre 0 y 1. (Studer 2013: 12-15)

La técnica de Silhouettes evalúa la clasificación de los casos dada una solución específica de clúster. Cuando s(i) es más grande (cercano a 1) implica que la disimilitud adentro del grupo es mucho menor que la disimilitud entre grupos, lo cual significa que la solución agrupa bien. Cuando s(i) es cercano a cero, indica que puede haber casos que pueden ser clasificados en diferentes grupos. Cuando s(i) es cercano a - 1, indica que el caso ha sido mal clasificado.

REFERENCIAS.

Abbott, A. (2001). Time matters: on theory and method. Chicago: University of Chicago Press.

ALDENDERFER, Mark & BLASHFIELD, Roger (1984) *Cluster Analysis.* Sage Publications. California.

BAILEY, Kenneth (1994) *Typologies and Taxonomies. An Introduction to Classification Techniques.* Sage Publications. California.

EVERITT, Brian et all (2000) *Cluster Analysis*. Fourth Edition. Published by Edward Arnold. London.

Studer, Mattias (2013) Weighted Cluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. Lausanne. DOI : 10.12682/lives.2296-1658.2013.24