XLVIII Coloquio Argentino de Estadística

VI JORNADA DE EDUCACIÓN ESTADÍSTICA "MARTHA DE ALIAGA"

27 al 30 oct 2020

Poster:

Estudio comparativo de métodos de clasificación no supervisada en contextos de grandes bases de datos

Emanuel Ciardullo, Marta Quaglino

FACULTAD DE CIENCIAS ECONÓMICAS

UNC Universidad Nacional de Córdoba

SAE

ESTUDIO COMPARATIVO DE MÉTODOS DE CLASIFICACIÓN NO SUPERVISADA EN CONTEXTOS DE GRANDES BASES DE DATOS

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

UNR

Emanuel Ciardullo. Licenciatura en Estadística, Universidad Nacional de Rosario
Tutora: Dra. Marta Quaglino. Licenciatura en Estadística, Universidad Nacional de Rosario

INTRODUCCIÓN

En estadística se conoce como análisis *cluster* al estudio formal de los métodos para el agrupamiento de objetos según las características intrínsecas de los mismos. Estos métodos, tienen por objetivo obtener grupos dentro de los cuales los individuos, que a priori conforman un grupo heterogéneo, sean homogéneos entre si y distintos de los pertenecientes a otro grupo. Se pueden encontrar cientos de algoritmos de *clustering* propuestos a través de las distintas disciplinas científicas, además de las modificaciones y adaptaciones de estos a casos particulares.

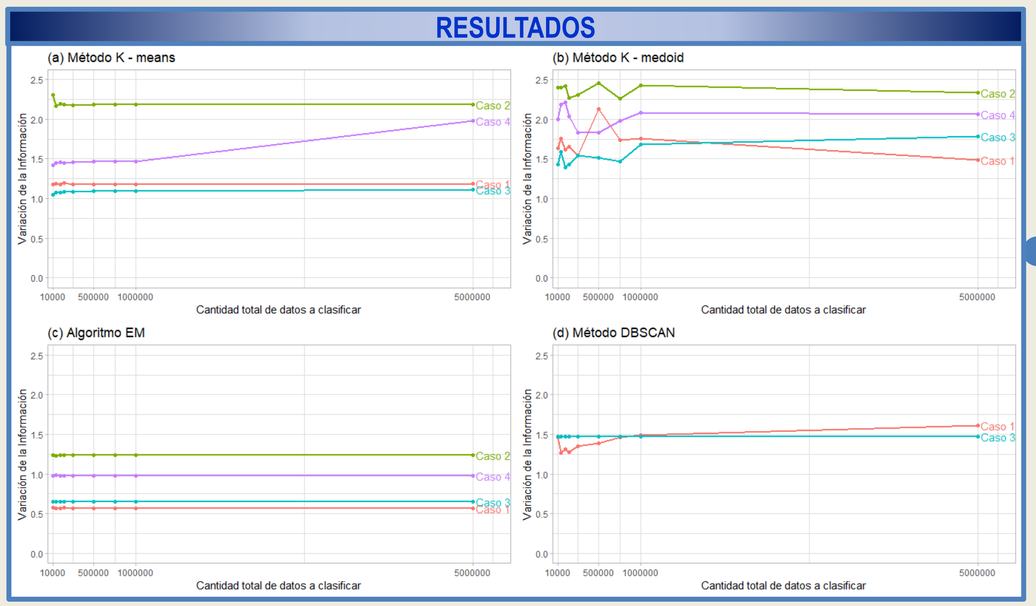
OBJETIVO

Evaluar las diferencias en los agrupamientos obtenidos a partir de un conjunto de algoritmos de clasificación no supervisada, aplicados a grandes bases de datos mediante un estudio por simulación.

VALIDACIÓN

El criterio, denominado Variación de la Información (VDI) fue propuesto por Marina Meila (2003) y mide la cantidad de información que se gana (o se pierde) al considerar agrupamientos diferentes. Mientras mayor sea el valor VDI mas grande será la diferencia entre los agrupamientos.

RESULTADOS



ALGORITMOS DE CLUSTERING

- ***K – means*:** Proceso iterativo que comienza con K centroides ubicados aleatoriamente, y asigna cada observación al centroide más cercano. Después de asignarlos, los centroides se mueven a la ubicación promedio de todos los datos asignados a él, y se vuelven a reasignar los puntos.
- ***K – medoid*:** A diferencia de *K – means* los centroides son puntos observados. Es más robusto ante el ruido y outliers que *K – means*.
- **DBSCAN:** Una observación forma parte de un grupo si hay una cierta cantidad mínima de observaciones dentro de un radio de proximidad. Los clústeres deben estar separados por regiones de baja densidad.
- **Algoritmo EM**: Trata de identificar clústeres entre los individuos asumiendo que los datos provienen de un modelo de probabilidad identificado.

SIMULACIÓN

Se simularon cuatro casos considerando 10 variables y observaciones provenientes de poblaciones gaussianas mixtas El primer caso fue simulado con un nivel de solapamiento entre 0.05 y 0.15, y el segundo caso con un solapamiento mayor entre las poblaciones, de entre 0.15 y 0.30. Los casos 3 y 4 repitieron el solapamiento de los casos 1 y 2 respectivamente pero se contaminaron los datos agregando dos variables de confusión y 5% de outliers.

CONCLUSIONES

El método DBSCAN fue incapaz de identificar los agrupamientos en los casos con mayor superposición de los grupos. Tanto *K – means* como EM demostraron ser estables con la calidad de los resultados obtenidos al aumentar el tamaño de los grupos a clasificar. El algoritmo EM es el que mejor resultados arroja en todos los escenarios.

BILIOGRAFÍA