Evaluación de sensibilidad al ruido en algoritmos de clustering basados en densidad o en particiones

Gabriel Espínola Benítez

Universidad Comuera – Diplomado en Machine Learning

1 INTRODUCCIÓN

El Clustering, es un tipo de aprendizaje no supervisado, radica en agrupar conjuntos de datos (no etiquetados) en subconjuntos de datos llamados Clusters.

Algoritmos Clustering

- Basados en densidad
  - DBSCAN
- Basados en particiones
  - K-MEANS
  - K-MEDOIDS

Métricas de Validación

- Índice de Rand ajustado
- Índice de Fowlkes-Mallows
- Información mutua ajustada

Los datos Ruidosos u Valores atípicos (Outliers) no suelen tener similitud con ningún tipo de dato que está siendo analizado, reflejan características inusuales en la población que está siendo evaluada. Uno de los problemas que pueden causar los datos ruidosos, es que sus propiedades pueden no ser representativos del conjunto de datos en general, pudiendo distorsionar el resultado del análisis al aplicar un algoritmo de clustering.

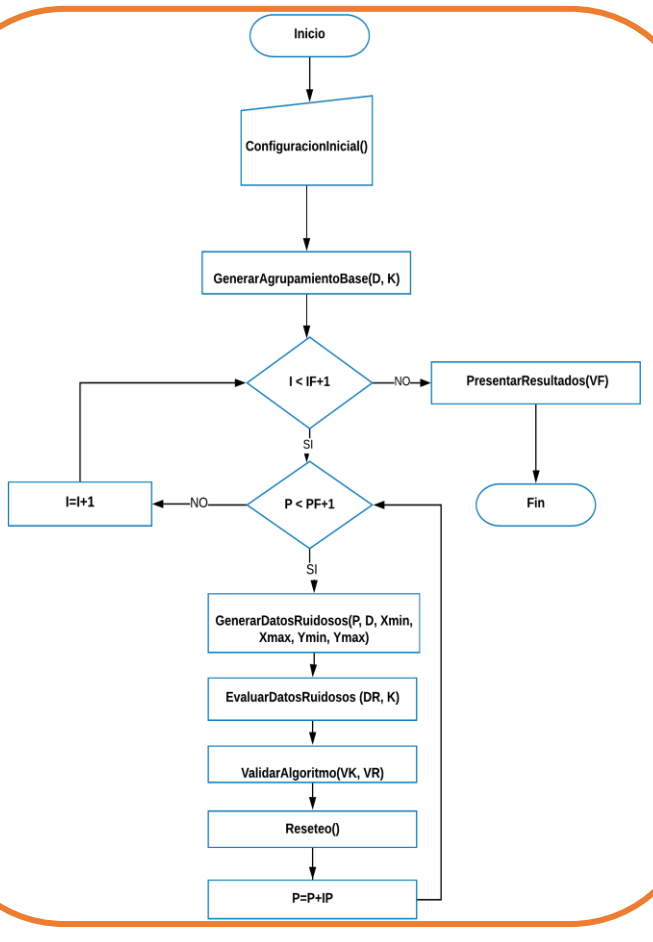
2 OBJETIVOS

General:

- Evaluar la robustez de algunos algoritmos de clustering frente al ruido utilizando métricas de validación existentes.
- **Específicos**:
- Medir la sensibilidad de los algoritmos de agrupamientos K-means, Kmedoides y DBSCAN respecto al ruido creciente usando métricas en problemas donde varia la cantidad y dimensión de datos evaluados, el porcentaje de ruido presente y el valor del parámetro del algoritmo de agrupamiento estudiado
- Determinar la robustez de los algoritmos Clustering aplicados a los conjuntos de datos en base a los resultados conseguidos por la métrica de validación

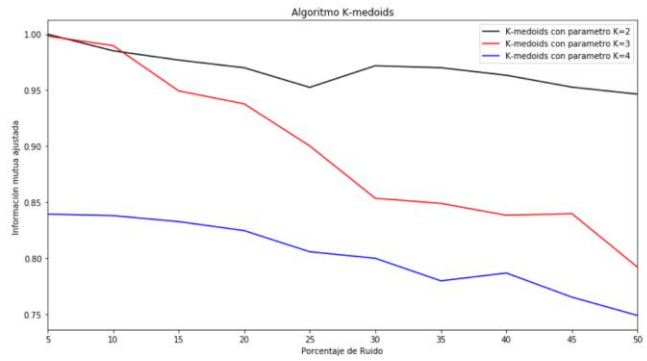
3 MARCO METODOLÓGICO

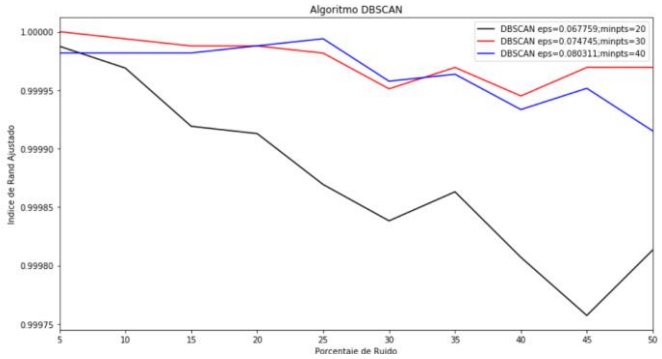
(i) Se genera un conjunto de datos (dataset) de manera aleatoria o se define un conjunto de datos existentes para ser evaluado y se establece un parámetro inicial n que denota el número de veces que se correrá un algoritmo sobre una instancia para promediar resultados obtenidos por un mismo algoritmo.
(ii) Se crean datasets ruidosos a partir del dataset original, donde el ruido se genera a partir de una distribución uniforme de forma que el ruido sea un porcentaje 10i del número de instancias original, donde i es entero y 0<i<11. Finalmente se promedian las n aplicaciones de un mismo algoritmo sobre cada porcentaje de ruido, usando los valores obtenidos de la métrica de validación seleccionadas.

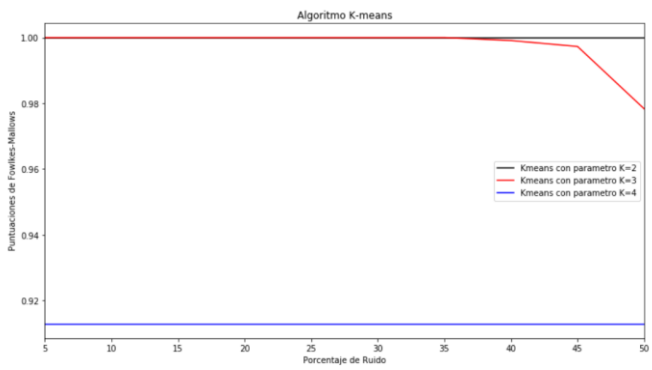


4 RESULTADOS

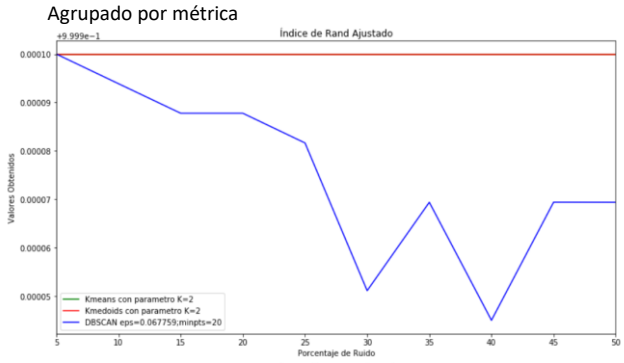
Conjunto de datos de 4 dimensiones

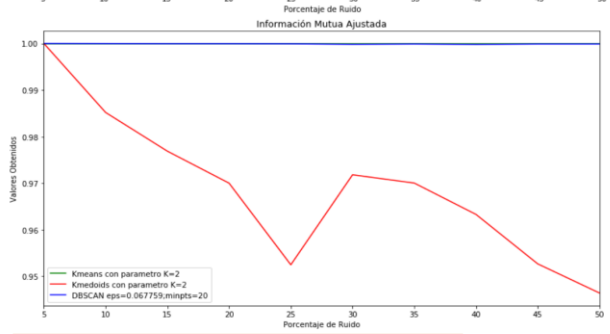






Agrupado por métrica





5 CONCLUSIÓN

- La metodología aplicada expone que el exceso de datos ruidosos en un dataset afecta al resultado final de un algoritmo de agrupamiento, pero de forma relativa a cada caso.
- Se observó que un mismo algoritmo puede presentar distinto comportamiento respecto al ruido dependiendo de los parámetros elegidos.

6 REFERENCIAS

- Campello, R.J., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. PAKDD.
- Rahmah, N., & Sitanggang, I.S. (2016). Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra.
- V Estivill-Castro, V., & Yang, J. (2000). Fast and Robust General Purpose Clustering Algorithms. PRICAI.
- Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data.