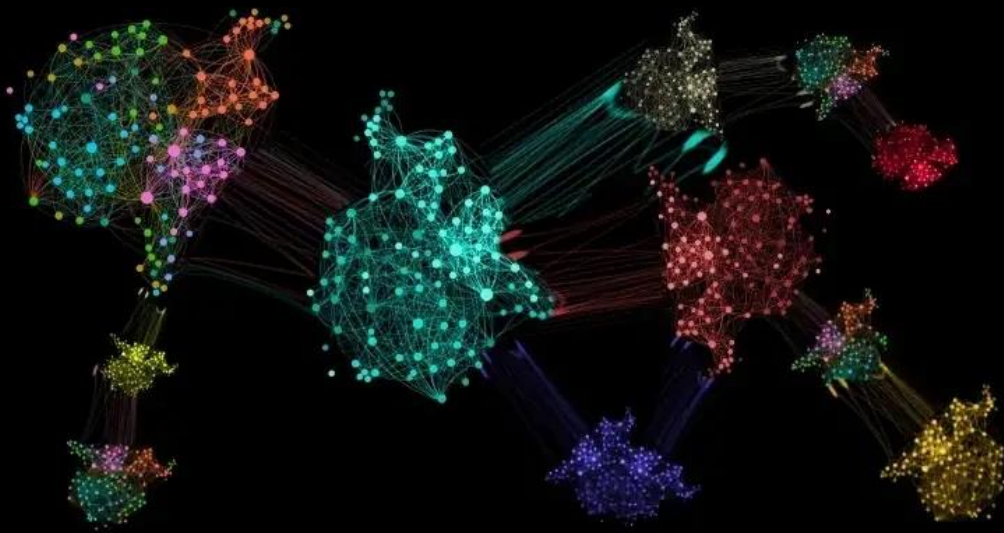


CI Metodología

Una Aproximación Metodológica al Macro Proceso de Clustering



por

Adrian Armando Araneda Toro

Luego del preprocesamiento que toma el 80% del tiempo en un proyecto de analítica de datos, comenzamos el proceso de clusterización:

1. **Clase de las Variables.** Identificar la clase de mis Variables, si son de tipo continuo y cuales no. Aíslo las no continuas. Dejo solo variables continuas. Tipo numérico por ejemplo. Aíslo dummies o dicotómicas. No trabajo con ellas por lo explicado en clases. Documentar esta etapa.
2. **Calidad de los Vectores de Trabajo.** Analizar los vectores desde el punto de vista de la completitud; porcentaje de información, N/As y calidad de la información. Documentar esta etapa.

3. **Eliminación de Variables Correlacionadas.**

Elimino variables correlacionadas por las razones dadas en clases.

- a. Criterio por Poder de la información o poder predictivo (para el caso supervisado).
- b. Completitud, N/A y calidad del vector (para el caso no supervisado y supervisado).
- c. Instrucción de Negocio (para el caso no supervisado y supervisado).

Documentar esta etapa.

4. **Identificación de outliers.** Realizar dos estudios. Uno con Outliers y otro sin Outliers. Y comparar técnicamente ambos al llegar a la etapa 9 particularmente, estudiando la variación de la composición de los clusters.

Si por recurso máquina debo trabajar con Kmeans y no con PAM, una buena alternativa sería aislar los outliers antes de generar los clusters, por las razones dadas en clases.

Se pueden aislar los outliers, en otro clusters por el analista antes de comenzar.

Ojo con la técnica de identificación de outliers. Siempre se deben pensar en "outliers multifactoriales, multidimensionales o multivariantes" y no eliminar outliers o aislarlos solo porque en una o dos de 6 variables, se presentaron los outliers identificados.

Observación1: Una buena medida para la identificación de outliers que se suele utilizar en ciencias es la "la ley de los 5 sigmas" (desviación estándar). Para esto consultar los siguientes links:

https://www.linkedin.com/posts/adrianaranedat_cienciaabrdeabrdatosabrparaabrfaed_sicaabrdeabrpartaedculasabrsubataejimicasabrbosaejnabrdeabrhiggsabr-activity-7002070557550592000-exkX?utm_source=share&utm_medium=member_desktop

<https://lapirangablogs.wordpress.com/2021/05/19/5-sigma-y-el-sueno-del-cientifico-de-datos/>

<https://francis.naukas.com/2013/08/14/la-historia-de-las-cinco-sigmas-en-fisica-de-particulas/>

<https://web.archive.org/web/20080513015956/http://www.onesixsigma.com/node/7630>

5. **Escalar.** Escalamos mi Master con el objeto de dejar las variables en una única unidad de medida para poder compararlas y preprocesarlas. Luego, no tiene sentido aplicar log para normalizar, dado que no estoy trabajando en el mundo supervisado (predecir, explicar una variable observada), ergo, no necesito forzar a las distribuciones a que se parezcan o se asimilen a una normal o gaussiana (para el caso supervisado esto sirve para disminuir o suavizar la variabilidad o ruido para ayudar con la precisión de las predicciones). No lo necesito. Este es el mundo No Supervisado. No es necesario aplicar log.
6. **Visualizar.** Procedo a una primera aproximación sólo visual de los datos para observar la dispersión natural de los datos. PLOTEAR en 2D o en 3D (utilizar alguna técnica de reducción de variables, por ejemplo, la técnica PCA enseñada en clases). Esto es deseable por las razones dadas y demostradas en clases.

Ojo, no se pueden tomar decisiones a partir de sólo esta aproximación. Es sólo, un primer acercamiento al hallazgo de agrupaciones (clústers) naturales.

7. **Testear Algoritmos de Clustering.** Estudiamos con qué algoritmo obtendremos los mejores clústeres (en virtud de los 7 índices enseñados en clases) con la librería `clValid` por ejemplo y complementarlo con otras 4 técnicas enseñadas también en clases.
8. **Testear Óptimo de K.** Estudiamos con cuantos K obtendremos los mejores clústeres (en virtud de los 7 índices enseñados en clases) con la librería `clValid` de R por ejemplo y complementarlo con otras 4 técnicas enseñadas en clases.

Hay que recordar que nunca se debe elegir como óptimo el mínimo de clústeres ni el máximo de clústeres testeados. Dado que es lógico obtener los mejores resultados desde el punto de vista de la maximización de la inercia entre-clústers con el menor número de clústers. De la misma manera, es lógico también obtener la mayor minimización de la inercia intra-clústers al aumentar al máximo el número de clústers (ya que las observaciones dentro de un clústers comenzarían a compactarse). Entonces los parámetros que miden por separado cada distancia siempre sugerirán en primer lugar, con mínimo y máximo de clústers.

9. **Valorización de la técnica de Clustering y de los óptimos obtenidos, a partir de los indicadores o parámetros enseñados.** En esta etapa es recomendable, una primera valorización preliminar comercial de los clústers en unidad \$ en términos del beneficio aportado y reducción de costos, según la técnica y número de clústers seleccionado en las técnicas del paso anterior. No se debe profundizar. Es solo enunciar. Generar preguntas. Es conducir ya desde una perspectiva comercial los parámetros técnicos obtenidos en virtud de la naturaleza del problema del fenómeno que se desea clusterizar.
10. Generación de Clústers.
11. Inspección de la composición de los clústers a través de a lo menos, las medidas de tendencia central.
12. **Valorización comercial de la inspección de la composición de los clústers.** Valorizar las medidas de tendencia central en \$. Y fundamentar en términos de costos y beneficios en \$, qué técnica y con qué número de K obtengo la mejor rentabilidad (comercial, social, etc.). En esta etapa se debe vender la solución analítica. Tanto en términos técnicos, pero sobre todo, comercial (reitero. Por comercial entiéndase tanto en rentabilidad social como económica monetaria).