Análisis de datos abiertos MEData basados en algoritmos de clustering

Daniel Torres González
daniel.torresg@udea.edu.co

Daniel Santa Rendón
daniel.santar@udea.edu.co

Héctor Mauricio Guerra Londoño
hector.guerra@udea.edu.co

Asesor: Javier Fernando Botia Valderrama
javier.botia@udea.edu.co

Proyecto integrador I, código 2508103, grupo 04

Resumen

En la Alcaldía de Medellín, se impulsó una iniciativa para que cualquier persona accediera a los datos abiertos sobre salud, educación, entre otros, llamado MEData, que es un repositorio de base de datos libres. Uno de los problemas que se presenta con los datos abiertos de la alcaldía es que la mayoría de datos no están etiquetados para hacer tareas posteriores como predicción o análisis de datos. Por esta razón, el proyecto integrador pretende crear soluciones basados en clustering para agrupar los datos en clusters, de tal forma que facilite la etiquetación de los datos. Para encontrar una primera aproximación a la solución de este problema, el proyecto se divide en tres etapas: análisis de la calidad de los datos usando ingeniería de características; aplicación de algunos algoritmos de clustering seleccionados; y el uso de la validación interna de clusters para establecer el número óptimo de clusters.

Palabras Clave

Clustering, Bases de datos, Ingeniería de características, validación de clases, aprendizaje no supervisado.

Planteamiento del problema

La inteligencia artificial (IA) (Pino Díez, Gómez Gómez, & Abajo Martínez, 2001) es uno de los componentes de la llamada cuarta revolución industrial (Schwab, 2017) y se define como la simulación de procesos de inteligencia humana por parte de las máquinas. La IA es ampliamente usada hoy en día tanto para la industria como para uso cotidiano de las

personas. Por ejemplo los bancos usan IA para predecir el valor de las acciones, muchas empresas están implementando chatbots para dar soporte al cliente, también se usa para reconocimiento facial, rellenado de imágenes y ahora con big data es necesaria la IA para análisis de datos.

Dentro de la inteligencia artificial hay un campo llamado machine learning (aprendizaje de máquina) (González, 2014) que es una disciplina científica que crea sistemas que aprenden automáticamente y es fundamental para el análisis de datos. Hay tres tipos:

- Aprendizaje supervisado (supervised learning) cuándo los datos poseen etiqueta, su objetivo es clasificar datos nuevos gracias al entrenamiento recibido con los datos existentes,
- aprendizaje reforzado (reinforcement learning) se basa en que la máquina aprenda por la experiencia y
- aprendizaje no supervisado (unsupervised learning) cuando los datos no poseen etiquetas, su objetivo es agrupar los datos de la mejor manera para darles una etiqueta óptima.

En este proyecto se trabajará el aprendizaje no supervisado (Tello, 2014), viendo los datos como una matriz se puede explicar porqué el use de esta técnica:

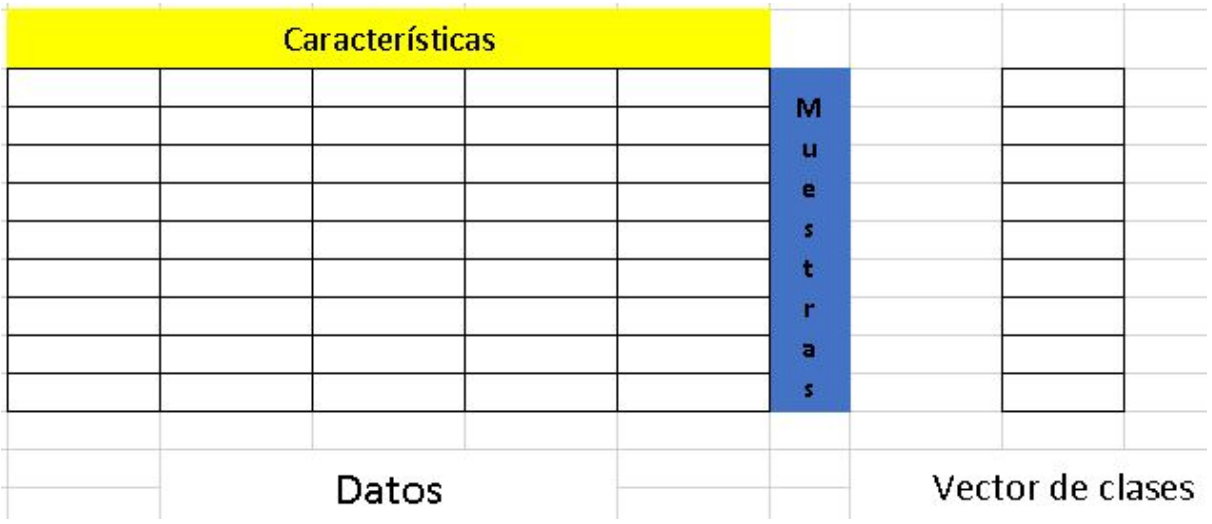


Fig 1. Base de datos representada en una matriz

Las características (columnas) son las variables o atributos de ese conjunto de datos y las muestras son la cantidad de datos obtenidos (filas), ejemplo: Se recoge información sobre 1000 autos de Medellín, las características podrían ser modelo, color, kilometraje, capacidad de pasajeros y las clases son la marca (mazda, toyota, renault, ford). Si ese vector de clases está presente se aplica aprendizaje supervisado, sino aprendizaje no supervisado. En nuestro caso las bases de datos carecen de este vector.

Antes de aplicar cualquier tipo de aprendizaje es necesario asegurar la calidad de los datos, para obtener datos adecuados se realiza ingeniería de características la cual "intenta aumentar la eficacia predictiva de los algoritmos de aprendizaje creando características de los datos sin procesar que facilitan el proceso de aprendizaje"(Tabladillo, 2017).

Para esto, se hace un análisis de outliers (datos atípicos) son datos muy alejados de la media y que pueden causar problemas en el agrupamiento o no ser útiles, por eso se descartan y luego un análisis de los datos faltantes (Ocaña Peinado, n.d.) que son datos que no están debido a un error de tipeo, outliers descartados o campos que no son obligatorios de responder en una encuesta por ejemplo, se pueden completar por medio de la media, la mediana o la frecuencia de los datos existentes. Se realiza un escalamiento de los datos (Solis, n.d.) con el objetivo de llevar los datos a una escala menor para un mejor manejo y comprensión y por último se procede a hacer una reducción de dimensionalidad que consiste en excluir características que no aportan información al modelo, se hará por medio del método PCA (principal components analysis por su siglas en inglés) (IBM Knowledge Center, n.d.).

Con datos óptimos es prudente aplicar los algoritmos de clustering o agrupamiento (Soni Madhulatha, 2012), estos son algoritmos capaces de identificar agrupaciones de elementos de acuerdo a una medida de similitud entre ellos. En la literatura hay varias clases de algoritmos según la similitud pero en este proyecto se trabajarán estos:

- Algoritmos de distancia: K-Means, fuzzy c-means, Gustafson–Kessel means(GK).

Estos algoritmos permiten agrupar los datos mediante una medida de distancia entre los datos y los centros de cada clase. Para estos algoritmos es necesario conocer el número de clases.

- Algoritmos de ruido:DBSCAN y sus variantes.

Son algoritmos que agrupan por la densidad de los datos, por lo tanto si un dato pertenece a un cluster debe estar cerca de un montón de datos de ese cluster. No se requiere conocer el número de clusters.

- Algoritmos de jerarquías (Árboles): agglomerative clustering, DIANA (divisive analysis)
- Algoritmos de espectros: Spectral clustering.

En este proyecto se usarán los algoritmos de k-means, Gustafson-Kessel means y DBSCAN, los demás se dejarán como uso opcional.

Para corroborar la información resultante del clustering se valida la separabilidad (clusters o grupos separados, se busca que sea la máxima) y la compactación (Datos del cluster cercanos, se busca que sea mínima), se usará la validación interna la cual medida mide la calidad del agrupamiento a partir de la información de los datos, la estructura geométrica de cada clase y la dispersión de los datos agrupados en cada clase. Existen varios índices tales como: Bic index, Calinski-Harabasz index, Davies-Bouldin index (DB), Silhouette index, Dunn index. (Rendón, Abundez, Arizmendi, & M. Quiroz, 2011). Se tratará de usar la mayoría. Si se cumplen las dos condiciones anteriores la información ha quedado debidamente agrupada y útil para hacerle análisis.

Este proyecto se basa en la iniciativa que impulsó la alcaldía de Medellín: MEData, que consta de 237 conjuntos de datos sobre salud, población, seguridad, etc. Con el fin de estar un paso más cerca de convertirse en Smart City; pero alguna información allí almacenada no está en condiciones óptimas (datos con ruido, datos faltantes, información no relevante) para que los expertos en bases de datos las analicen y tomen las decisiones más

convenientes. Para entregar datos óptimos se necesita realizar un proceso de aprendizaje no supervisado (debido a que no hay vector de clases) y todo lo que este conlleva y fue explicado anteriormente.

Revisión Bibliográfica

González, A. (2014). ¿Qué es Machine Learning? Retrieved June 17, 2019, from https://cleverdata.io/que-es-machine-learning-big-data/

IBM Knowledge Center. (n.d.). Escalamiento multidimensional: Crear la medida a partir de los datos. Retrieved June 15, 2019, from https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_alsc_mea.html

Ocaña Peinado, F. M. (n.d.). *Técnicas estadísticas en Nutrición y Salud*. Retrieved from https://www.ugr.es/~fmocan/MATERIALES DOCTORADO/Tratamiento de outliers y missing.pdf

Pino Díez, R., Gómez Gómez, A., & Abajo Martínez, N. de. (2001). *Introducción a la inteligencia artificial : sistemas expertos, redes neuronales artificiales y computación evolutiva*. Retrieved from https://books.google.es/books?hl=es&lr=&id=RKqLMCw3IUkC&oi=fnd&pg=PA10&dq=inteligencia+artificial&ots=iGJynYC_dR&sig=ExDG3G2IZjZUOmDFcFpq_wfd5TU#v=onepage&q=inteligencia artificial&f=false

Rendón, E., Abundez, I., Arizmendi, A., & M. Quiroz, E. (2011). *Internal versus External cluster validation indexes*. Retrieved from https://pdfs.semanticscholar.org/2054/29d63883b041436fdf2be8170a1f98fa90da.pdf

Schwab, K. (2017). *The fourth industrial revolution*. Retrieved from https://books.google.es/books?id=ST_FDAAAQBAJ&dq=fourth+industrial+revolution+%2B+overview&lr=&hl=es&source=gbs_navlinks_s

Solis, D. (n.d.). Ejemplo de Reducción de Dimensionalidad con la Base de Datos Iris. Retrieved June 15, 2019, from https://rpubs.com/dsolis/iris-pca-lda

Soni Madhulatha, T. (2012). *AN OVERVIEW ON CLUSTERING METHODS*. *2*(4), 719–725. Retrieved from https://arxiv.org/ftp/arxiv/papers/1205/1205.1117.pdf

Tabladillo, M. (2017). Ingeniería de características en ciencia de datos: proceso de ciencia de datos. Retrieved June 17, 2019, from https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/create-features

Tello, J. C. (2014). Reconocimiento de patrones y el aprendizaje no supervisado. In *Escuela Técnica Superior de Informática Universidad de Alcalá*. Retrieved from https://www.researchgate.net/profile/Jesus_Tello/publication/228857048_Reconocimiento_de_patrones_y_el_aprendizaje_no_supervisado/links/0c960517e7e677b522000000.pdf

Objetivos

Objetivo General

Implementar un proceso de aprendizaje no supervisado con las bases de datos de MEData obteniendo datos agrupados (clusters) para facilitar el análisis y las decisiones a los expertos en bases de datos.

Objetivos Específicos

- Implementar ingeniería de características para obtener datos limpios antes de aplicar clustering.

- Proponer el desarrollo de un entorno de varios algoritmos de clustering a cada base de datos con el fin de tener diferentes resultados.

- Evaluar la calidad de los clusters por medio de algoritmos de validación interna.

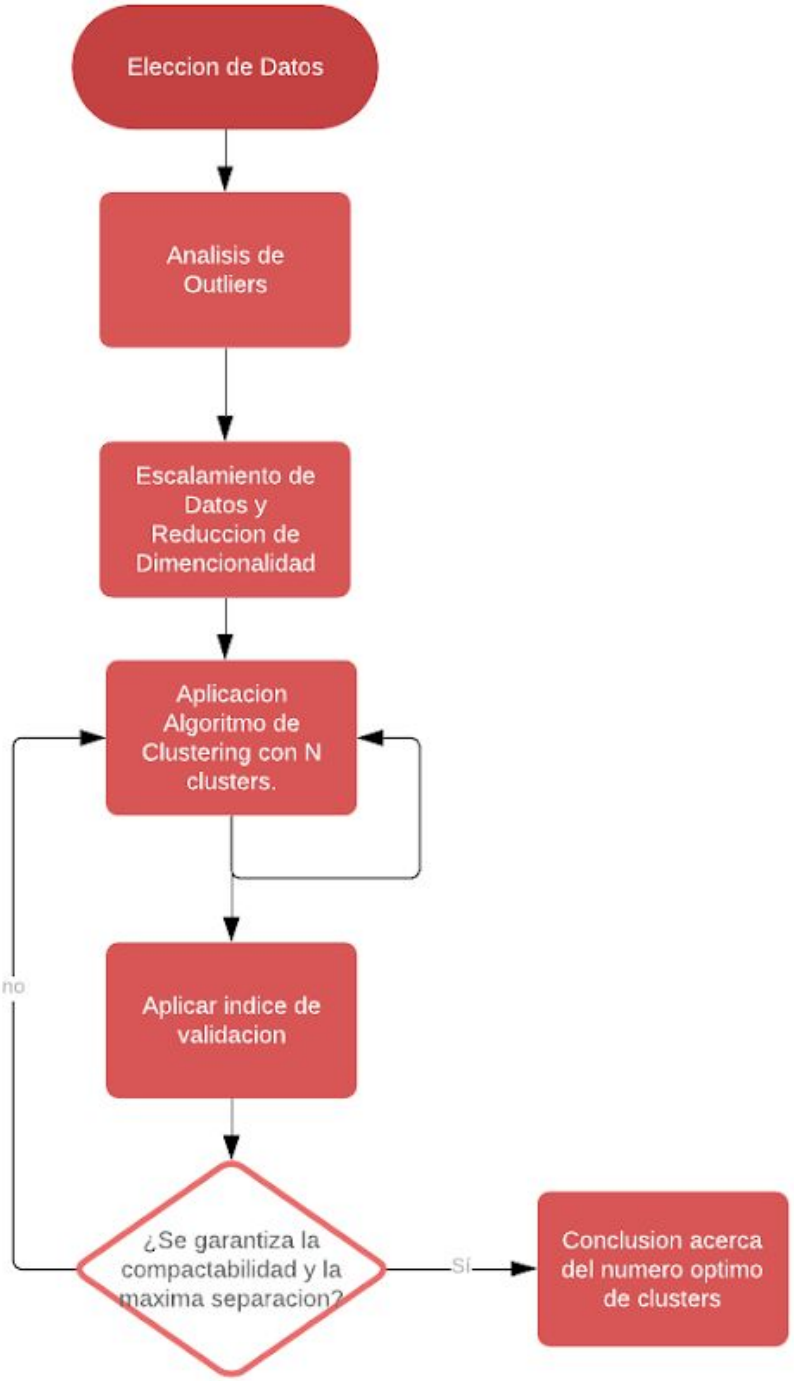
Metodología

Para la realización de los objetivos específicos en el tiempo estimado del proyecto se plantea una serie de etapas que nos permitirán culminar el proyecto de manera eficiente:

- Se buscará y seleccionará bases de datos ofrecidas por MEData.
- Una vez seleccionado los datos de nuestro proyecto procederemos a darle un tratamiento que nos permitirá trabajar y agruparlos de manera correcta, este tratamiento conlleva realizar una ingeniería de características, la cual consiste en realizar eliminación de datos faltantes y análisis de outliers[2], es decir eliminar datos que no nos aportan información. Esto se realiza en dos etapas principales; primero se realizará un escalamiento de datos y después una reducción de dimensionalidad.
- Trabajando con los datos ya tratados, se procederá con la fase de agrupamiento o clustering. En esta fase procederemos buscando, analizando y aplicando diferentes algoritmos de clustering tales como K-MEANS, DBSCAN, Gustafson–Kessel means (GK) y de manera opcional Fuzzy C-Means, agglomerative clustering, spectral clustering, entre otros.
- Por último utilizaremos los índices de validación interna anteriormente planteados que permitan evaluar la calidad del agrupamiento de datos para realizar una comparación que ayude a seleccionar el mejor modelo.
- Realizadas ya todas las actividades anteriores se eligen los mejores modelos para cada grupo de datos permitiendo a los expertos en las bases de datos de MEData en el mejoramiento del análisis de datos y encontrar nueva información que amplíe las soluciones a un tipo de problemática de la ciudad.

Metodologia

Proyector Integrador I



Cronograma

	Semana															
Actividad	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Planteamiento de propuesta																
Búsqueda de material bibliográfico																
Elección de base de datos																
Análisis de outliers y datos faltantes																
Escalamiento de datos y reducción de dimensionalidad																
Aplicación de algoritmos de clustering																
Validación																

Presupuesto

RUBROS	**FUENTES**		**TOTAL**
	UdeA	***Otras fuentes (indique cuáles)***	
PERSONAL	0	0	0
EQUIPOS	0	4.000.000	4.000.000
SOFTWARE	0	0	0
MATERIALES	0	100.000	100.000
SALIDAS DE CAMPO	0	0	0
MATERIAL BIBLIOGRÁFICO	0	0	0
PUBLICACIONES,	0	0	0

PATENTES O REGISTRO DE SOFTWARE			
SERVICIOS TÉCNICOS	0	0	0
VIAJES	0	0	0
TOTAL	**0**	**4.100.000**	**4.100.000**