TÉCNICA DEL APRENDIZAJE NO SUPERVISADO: CLUSTERING

Nelson Yoel Phuño Cahuana; Daysi Saimira Machaca Condori

INTELIGENCIA COMPUTACIONAL

nelsonyoelpc@gmail.com; zaimira.m1c@gmail.com

El aprendizaje no supervisado es un método de carácter exploratorio que se ocupa de problemas donde los datos no tienen etiquetas, no es necesario supervisar el modelo ya que funciona por si solo para descubrir patrones e información que antes no se había detectado en los datos.

1. CLUSTERING

Clustering o también llamado agrupamiento es una técnica no supervisada, realiza análisis de datos para encontrar una estructura o patrón en un conjunto de datos sin clasificar. La técnica del agrupamiento también es llamada una técnica descriptiva, puesto que su objetivo es describir los datos con la finalidad de descubrir agrupaciones naturales; es decir, los elementos de cada clúster o grupo guardan una similitud o cercanía entre sí, pero son diferentes de los elementos de otros grupos.

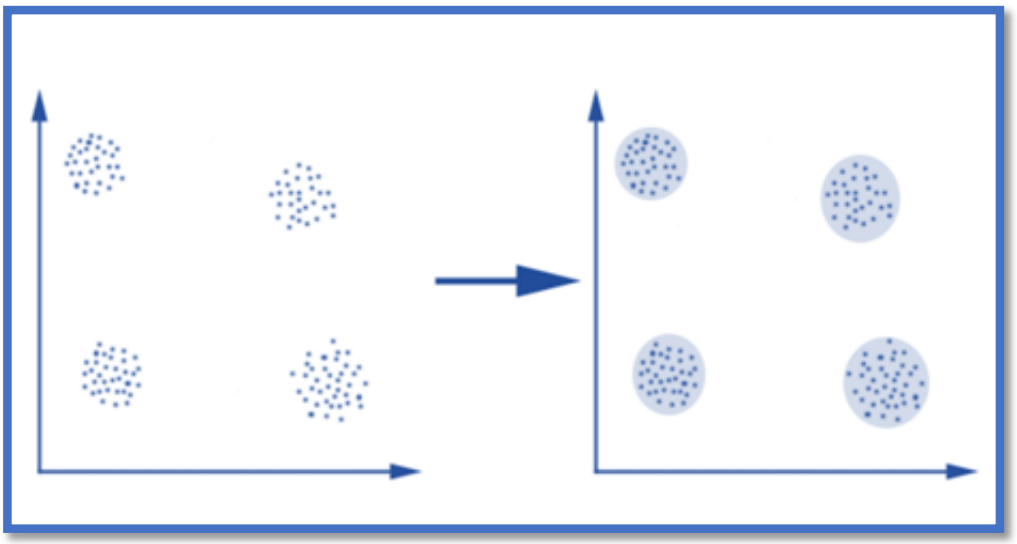

Figura 1: Clustering

Fuente: (Sanatan , 2017)

1.1. Tipos de Clustering

Existen 4 tipos de algoritmos de Clustering:

1.1.1. Clustering Jerárquicos:

Es un algoritmo que crea una jerarquía de agrupaciones para realizar el análisis y existen dos categorías para este tipo de clustering: Aglomerante y divisivo. Para ambos casos, la medida de distancia que se utiliza para generar los clusters es comúnmente la distancia euclidiana.

Métodos aglomerativos o botton-up

- Primero asigna cada elemento a un cluster.
- Después encuentra la matriz de distancias.
- Fusionan los clusters según criterio de similitud.
- Continua este proceso hasta que se forma un solo cluster grande.

Métodos divisorios o top-down

- Se inicia con todos los elementos asignado a un solo cluster.
- Sigue el algoritmo hasta que cada elemento es un cluster individual.

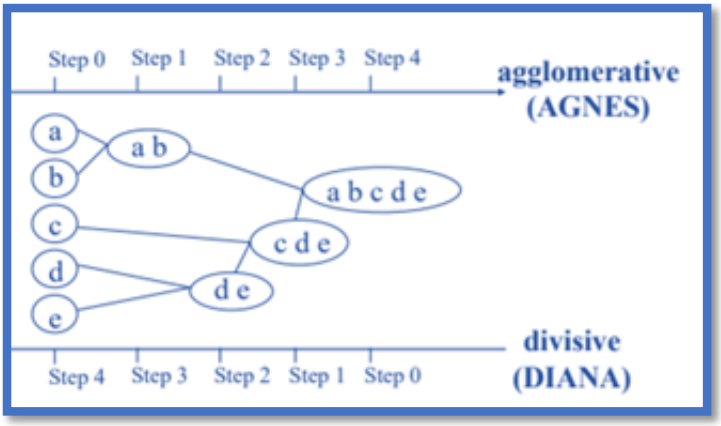

Figura 2: Clustering Jerárquicos

Fuente: (Avila Camacho, 2020)

Los resultados de la jerarquía de grupos se representan en un dendograma.

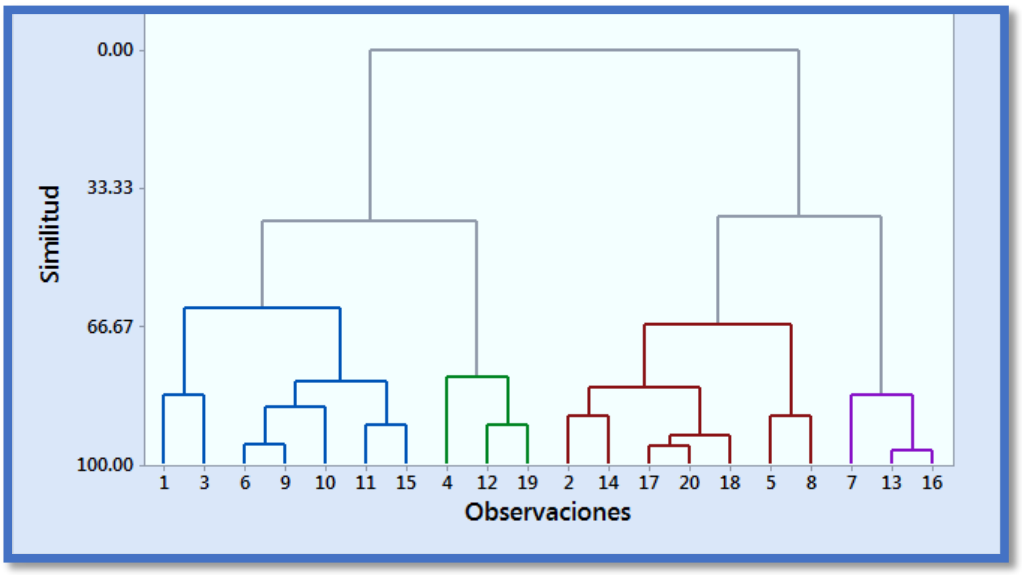

Figura 3: Dendograma

Fuente: (Avila Camacho, 2020)

Algoritmos principales en agrupamiento jerárquico

a. ROCK
b. BIRCH
c. CURE
d. CHAMELEON.

1.1.1.1.¿Cómo funciona el Clustering Jerárquico?

La agrupación jerárquica comienza tratando cada observación como un grupo separado. Luego, ejecuta repetidamente los dos pasos siguientes:

1. Identificar los dos grupos que están más cerca
2. Fusionar los dos grupos más similares.

Este proceso iterativo continúa hasta que todos los clústeres se fusionan. Esto se ilustra en *Figura 2 y 4*.

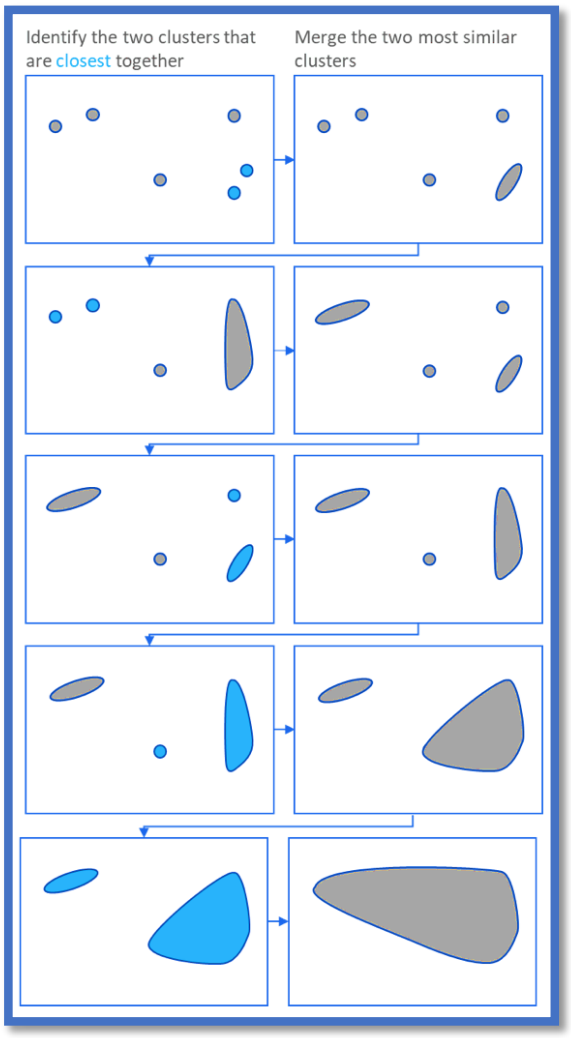

Figura 4: Clustering Jerárquico

Fuente: (Bock, 2016)

1.1.2. Agrupamiento por particiones

Son métodos de agrupación que se utilizan para clasificar las observaciones, dentro de un conjunto de datos, en varios grupos según su similitud. Los algoritmos requieren que el analista especifique el número de conglomerados que se generarán.

Algoritmos principales en agrupamiento por particiones

a. K-means
b. K-medianas
c. K-medoids: Algoritmo PAM, Algoritmo Clara
d. K-modas

1.1.2.1.¿Cómo funciona el agrupamiento por particiones?

A partir de un conjunto de datos se realiza una división en cierta cantidad de clusters (k)

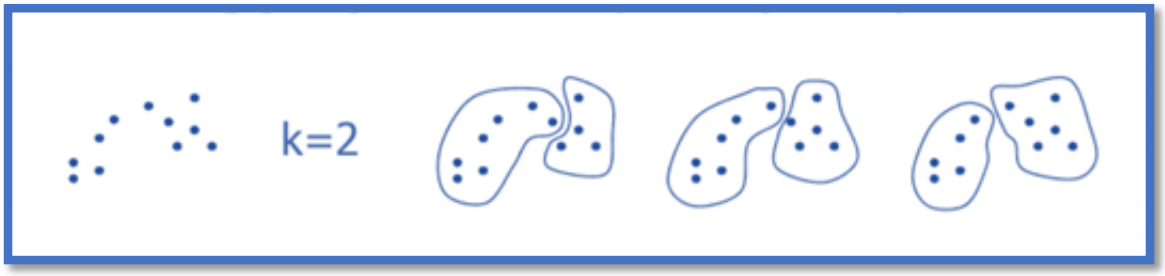

Figura 5: Agrupamiento por particiones

Fuente: (Avila Camacho, 2020)

1.1.3. Métodos basados en densidad

Los algoritmos basados en densidad localizan zonas de alta densidad separadas por regiones de baja densidad.

Algoritmos Basados en Densidad

a. DBSCAN
b. Optics
c. DenClue

1.1.3.1.Cómo funciona el método basados en densidad

Comienza seleccionando un punto t arbitrario, si t es un punto central, se empieza a construir un cluster alrededor de él, tratando de descubrir componentes denso-conectadas; si no, se visita otro objeto del conjunto de datos.

Son aquellos tales que en su vecindad de radio Eps, hay una cantidad de puntos mayor o igual que un umbral MinPts especificado. Un punto borde o frontera tiene menos puntos que MinPts en su vecindad, pero pertenece a la vecindad de un punto central. Un punto ruido (noise) es aquel que no es ni central ni borde. La Figura 3 ilustra cada uno de esos conceptos: si MinPts es mayor o igual a 4 y menor o igual a 6, A es un punto central, B es un punto borde y C es un punto ruido.

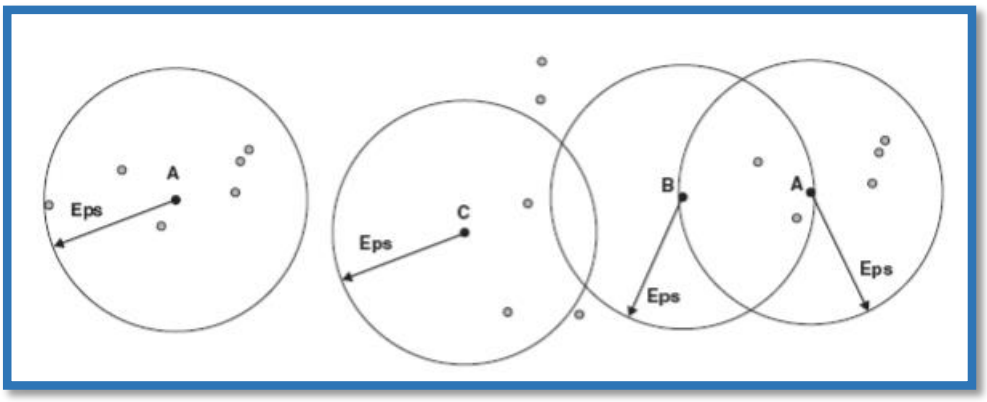

Figura 6: Definiciones de punto central, borde y ruido.

Fuente: (Pascual González , 2010)

2. ¿CÓMO SE MIDE SU EFICACIA?

Para el análisis de los resultados de una clusterización se deben tener en cuenta varios aspectos para la validación de los resultados del algoritmo:

- La determinación de la tendencia de agrupación de los datos (es decir, si existe realmente una estructura no aleatoria).
- Determinar el número correcto de clústers.
- Evaluar la calidad de los resultados de la agrupación sin información externa.
- Comparar los resultados obtenidos con la información externa.
- Comparar dos agrupaciones distintas para determinar cuál es mejor.

3. ¿A QUÉ TIPOS PROBLEMAS MEJOR RESPONDE?

- Análisis de datos estadísticos
- Reconocimiento de patrones
- Recuperación de información
- Compresión de datos
- Aprendizaje automático.

4. EJEMPLO

- Segmentación de clientes
- Análisis en redes sociales
- Agrupación de documentos

5. CONCLUSIONES

Una de las técnicas mas importantes y de los mas usados del aprendizaje no supervisado es el clustering ya que es muy útil para encontrar características de ciertos elementos para su agrupación, también es muy útil para el acceso secuencial de grandes cantidades de datos.

BIBLIOGRÁFIA

Avila Camacho, J. (2020). *Clustering Jerárquico con Python*. Obtenido de https://www.jacobsoft.com.mx/es_mx/clustering-jerarquico-con-python/#:~:text=Para%20representar%20los%20resultados%20de,en%20una%20matriz%20de%20distancias.

Bock, T. (2016). *What is Hierarchical Clustering?* DISPLAYR.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* Obtenido de http://oreilly.com

Pascual González , D. (2010). *Algoritmos de Agrupamiento basados en densidad y Validación de clusters*. Departament de Llenguatges I Sistemas Informátics Universitat Jaume I, Castellón. Obtenido de http://www.cerpamid.co.cu/sitio/files/Damaris

Sanatan , M. (2017). *Unsupervised Learning and Data Clustering*. Towards Data Science. Obtenido de https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a

Sancho Caparrini, F. (2021). *Medir la eficacia de un aprendizaje*. Obtenido de http://www.cs.us.es/~fsancho/?e=231

Shai Shalev, S., & Shai Ben, D. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. Obtenido de https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/