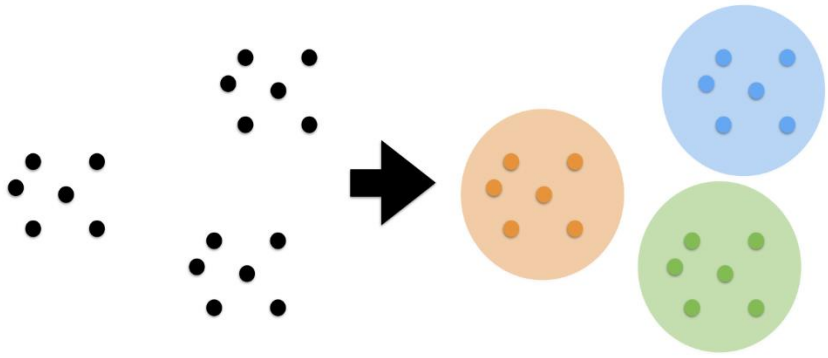
Centro Universitario de Ciencias Exactas e Ingenierías

Universidad de Guadalajara

Actividad 9: Algoritmos de agrupamiento avanzado

Aprendizaje Máquina



Integrantes:

- Paulina Amezcua García
- Josué Gael Magaña Corona
- Samuel David Pérez Brambila

Profesora: Karla Ávila Cárdenas

Sección: D01

Fecha de Entrega: 03 de Noviembre de 2024

Introducción

El clustering o agrupamiento es “un algoritmo de machine learning no supervisado que organiza y clasifica diferentes objetos, puntos de datos u observaciones en grupos o clusters basados en similitudes o patrones.” (IBM, s.f.)

Antes de profundizar en el clustering, es importante definir qué es el aprendizaje no supervisado. Según INESDI (2022), es “un tipo de aprendizaje automático o machine learning en el que los modelos aprenden a partir de conjuntos de datos sin etiquetar sobre el que se les permite actuar sin supervisión”.

Con este contexto sobre el aprendizaje no supervisado, podemos seguir abordando el tema del clustering. Existen múltiples algoritmos de clustering, cada uno con distintas formas de definir un cluster. Los distintos enfoques funcionarán bien para diferentes tipos de modelos en función del tamaño de los datos de entrada, la dimensionalidad de los datos, la rigidez de las categorías y el número de conglomerados dentro del conjunto de datos. Vale la pena señalar que un algoritmo puede funcionar muy bien para un conjunto de datos y muy mal en otro. Según IBM (s.f.), cinco de los enfoques más utilizados son:

- Clustering basado en centroides: El clustering basado en centroides es un tipo de método de clustering que divide o divide un conjunto de datos en grupos similares en función de la distancia entre sus centroides. El centroide de cada clúster es la media o la mediana de todos los puntos del clúster, en función de los datos.
- Clustering jerárquico: El clustering jerárquico, a veces denominado clustering basado en la conectividad, agrupa los puntos de datos en función de la proximidad y la conectividad de sus atributos. Este método determina los clústeres en función de la proximidad de los puntos de datos entre sí en todas las dimensiones. La idea es que los objetos que están más cerca están más estrechamente relacionados que los que están lejos unos de otros.
- Clustering basado en distribución: El clustering basado en la distribución, a veces denominado clustering probabilístico, agrupa los puntos de datos en función de su distribución de probabilidad. Este enfoque supone que hay un proceso que genera distribuciones normales para cada dimensión de los datos que crea los centros de clústeres. Se diferencia del clustering basado en centroides en que no utiliza una métrica de distancia como una distancia euclidiana o de Manhattan. En su lugar, los enfoques basados en la distribución buscan una distribución bien definida que aparezca en cada dimensión.
- Clustering basado en densidad: El clustering basado en la densidad funciona mediante la detección de áreas donde se concentran puntos y donde están separados por áreas que están vacías o escasas. A diferencia de los

enfoques basados en centroides, como las medias K, o los enfoques basados en la distribución, como la maximización de expectativas el clustering basado en la densidad puede detectar clústeres de una forma arbitraria. Esto puede ser extremadamente útil cuando los clústeres no están definidos en torno a una ubicación o distribución específica.

- Clustering basado en cuadrícula: Los algoritmos de clustering basados en cuadrículas no se utilizan con tanta frecuencia como en los cuatro enfoques anteriores, pero pueden resultar útiles en clústeres de alta dimensión, ya que otros algoritmos de agrupamiento pueden no tener el mismo rendimiento. En este enfoque, el algoritmo divide un conjunto de datos de alta dimensión en celdas. A cada celda se le asigna un identificador único llamado ID de celda y todos los puntos de datos que se encuentran dentro de una celda se consideran parte del mismo clúster.

Hay muchas áreas de aplicación donde el clustering es una herramienta valiosa para la minería de datos o el análisis exploratorio de datos, acorde con IBM (s.f.) encontramos:

- Detección de anomalías: El clustering puede ayudar a descubrir anomalías mediante la medición de los puntos de datos que no se incluyen en la estructura de clustering definido por el análisis de clústeres. Los puntos de datos que pertenecen a clústeres pequeños o muy dispersos o que están lejos de su clúster asignado se pueden considerar anomalías.
- Investigación de mercado: Al tratar de comprender qué perfiles de clientes o subconjuntos de mercados podrían existir, el clustering puede ser una herramienta poderosa para ayudar a realizar la segmentación de clientes. Es posible que pueda combinar datos demográficos con datos de comportamiento del cliente para encontrar qué tipos de características y patrones de compra se correlacionan con mayor frecuencia.
- Segmentación de imágenes: Las imágenes pueden tener sus píxeles agrupados en una variedad de formas que pueden ayudar a cortar la imagen en diferentes secciones para separar un primer plano de un fondo, detectar objetos usando similitudes en color y brillo, o dividir imágenes en regiones de interés para su posterior procesamiento. Con las imágenes, los métodos de clustering procesan los píxeles de la imagen y definen áreas dentro de la imagen que representan el clúster.

Es así que, en el presente trabajo se elaborará un cuadro comparativo que sintetiza los puntos clave de diferentes enfoques del clustering, como el clustering jerárquico, el clustering basado en centroides (como k-means), el clustering basado en densidad (como DBSCAN), entre otros. De esta manera, se podrán diferenciar de manera asertiva, facilitando la comprensión de las características particulares de cada enfoque y su aplicabilidad según el tipo de datos y los objetivos del análisis.

Contenido de la Actividad

	Basados en prototipos	Jerárquicos	Basados en densidad
Descripción	Este enfoque se basa en el uso de una serie de representantes o prototipos que determinan diferentes comportamientos en los datos. Los grupos se definen en función de la distancia entre los patrones (datos) y los prototipos. Un nuevo patrón se asigna al prototipo más cercano. El algoritmo más conocido en esta familia es K-means, que itera entre la asignación de puntos a los prototipos y la recalculación de los prototipos como la media de los puntos asignados. Una variante es el K-medoids, donde los prototipos (medoids) son puntos de datos reales, lo que mejora la interpretabilidad.	Los algoritmos jerárquicos construyen agrupamientos de manera secuencial. Pueden ser de dos tipos: aglomerativos, donde se comienza con cada punto como un clúster separado y se van fusionando grupos, o divisivos, donde se comienza con todos los datos en un solo clúster y se va dividiendo en clústeres más pequeños. Esta jerarquía de agrupamientos se representa en un dendrograma, que permite visualizar las relaciones entre los clústeres.	Estos algoritmos definen los clústeres como áreas de alta densidad de patrones separadas por áreas de baja densidad. Uno de los algoritmos más utilizados en esta familia es DBSCAN (Density-Based Spatial Clustering of Applications with Noise). En este método, un clúster se define en torno a un punto núcleo si existen al menos un número mínimo de patrones en su vecindad (definido por minPuntos y eps). Se permite que los clústeres tengan formas arbitrarias, y los puntos que no cumplen con estos criterios se clasifican como ruido o puntos frontera.
Ventajas	• Bajo coste computacional, al poder establecer un límite de iteraciones. • Crea con eficacia el número de grupos que se le especifique. • Simplicidad, es fácil de implementar y comprender.	• Más usado cuando se sabe que hay cierta jerarquía entre los datos. • Es bueno para grupos que no tienen formas elípticas. • Permite observar clústeres a diferentes niveles, lo que es útil para obtener diferentes perspectivas.	• Los grupos pueden tener cualquier forma. • Puede separar zonas de alta densidad de otras de baja. • Es robusto frente a casos atípicos.

			• No es necesario especificar el número de grupos.
Desventajas	• El resultado de la media de variables y no se corresponde con ningún dato, perdiendo interpretabilidad. • Los grupos se pueden solapar. • Sensibilidad a outliers, los datos atípicos pueden afectar significativamente los prototipos.	• Alto coste computacional para grandes conjuntos de datos. • Baja resistencia a datos atípicos. • Una vez construido un grupo, no se puede deshacer.	• Dependencia con los parámetros clave (minPuntos y eps). • El orden en que se presentan los patrones al algoritmo impacta directamente en la estructura de los grupos obtenida. • Tiene dificultades en conjuntos de datos donde la densidad varía mucho entre regiones.
Ejemplo de algoritmo	K-means, K-medoids	Enlace Único, Enlace Completo, Método de Ward	DBSCAN

Conclusiones

En conclusión, comprender el funcionamiento del clustering es fundamental para aprovechar al máximo su capacidad de análisis y agrupamiento de datos en machine learning o aprendizaje máquina. Desde enfoques basados en centroides hasta enfoques más complejos como el clustering jerárquico o el clustering basado en densidad, cada técnica ofrece ventajas específicas para abordar distintos tipos de problemas. La correcta aplicación de estos algoritmos no solo facilita el descubrimiento de patrones ocultos en los datos, sino que también juega un papel crucial en distintas áreas como la detección de anomalías, la segmentación de mercados y el procesamiento de imágenes.

Elaborar un cuadro comparativo que sintetice los puntos clave de estas técnicas permite estructurar y clarificar las diferencias entre los enfoques de clustering, ayudando a destacar aspectos como las fortalezas de cada método y las situaciones más adecuadas para su aplicación. Al presentar estas diferencias de manera visual, el cuadro facilita una comprensión más profunda y concreta, permitiendo que los desarrolladores o expertos seleccionen la técnica más apropiada según el problema a resolver.

Por lo tanto, este trabajo no solo resalta la relevancia del clustering en el análisis de datos, sino también la importancia de diferenciar claramente los distintos enfoques y comprender sus características específicas. Identificar las fortalezas, limitaciones y aplicaciones de cada técnica permite seleccionar la más adecuada para cada problema, maximizando la efectividad del análisis. Esta diferenciación es clave para abordar con precisión problemas complejos, logrando así una aplicación más efectiva y adecuada de los algoritmos de clustering en diversos contextos o problemáticas, optimizando su potencial en escenarios prácticos.

Referencias

- IBM (s.f.). *¿Qué es el clustering?*. IBM. Recuperado el 24 de Octubre de 2024 de: https://www.ibm.com/es-es/topics/clustering
- INESDI (2022, 2 de noviembre). *¿Qué es el aprendizaje no supervisado y cuándo usarlo?*. INESDI Business School. Recuperado el 24 de Octubre de 2024 de: https://www.inesdi.com/blog/que-es-aprendizaje-no-supervisado/