Universidad Tecnológica de Chihuahua

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

U3E1. Análisis Supervisado

Enrique Mascote
RICARDO ALONSO RIOS MONRREAL

Introducción

En la era del Big Data, dos de los desafíos más grandes en la extracción de conocimiento (Data Mining) son la inmensa cantidad de registros no etiquetados y el exceso de variables o características por registro. Aquí es donde entra el aprendizaje no supervisado.

Los algoritmos de agrupación (Clustering) nos permiten descubrir estructuras ocultas y segmentar datos en grupos homogéneos sin conocimiento previo de las categorías. Por otro lado, la reducción de dimensionalidad ataca la "maldición de la dimensionalidad", simplificando conjuntos de datos complejos al reducir el número de variables aleatorias bajo consideración, preservando la información esencial. Este reporte analiza tres técnicas clave de agrupamiento y dos de reducción, explorando sus mecanismos, ventajas y aplicaciones prácticas.

Algoritmos de Agrupación (Clustering)

Para este análisis se han seleccionado: K-Means, DBSCAN y Clustering Jerárquico Aglomerativo.

K-Means (K-Medias)

Este es el algoritmo de particionamiento más utilizado debido a su simplicidad y velocidad.

- Principio de funcionamiento:

El algoritmo busca dividir el conjunto de datos en K grupos predefinidos, donde cada punto pertenece al grupo cuyo centroide (media) es el más cercano. Funciona iterativamente:

1. Se inicializan K centroides aleatoriamente.
2. Se asigna cada punto al centroide más cercano (usualmente distancia Euclidiana).
3. Se recalcula la posición del centroide promediando los puntos asignados a él.
4. Se repiten los pasos 2 y 3 hasta que los centroides no cambien (convergencia).

- Parámetros clave:
    - n_clusters (k): El número de grupos deseados.

    - init: Método de inicialización (ej. k-means++ para acelerar la convergencia).
    - max_iter: Número máximo de iteraciones.
- **Ventajas y limitaciones:**
    - *Ventajas:* Es muy rápido y escala bien con grandes datasets. Fácil de implementar.
    - *Limitaciones:* Requiere especificar K de antemano (no lo descubre solo). Es sensible a valores atípicos (outliers) y asume que los clústeres son esféricos; falla con formas complejas.
- Ejemplo de aplicación (Lógica/Pseudocódigo):

Caso: Segmentación de clientes de un supermercado según Gasto Anual e Ingresos.

INICIO K-Means

Definir k = 3 (Bajo, Medio, Alto perfil)

Seleccionar 3 clientes al azar como centroides iniciales

REPETIR

Para cada cliente en la Base de Datos:

Calcular distancia a los 3 centroides

Asignar cliente al grupo del centroide más cercano

Para cada grupo (1, 2, 3):

Calcular el promedio de Gasto e Ingresos

Mover el centroide a ese promedio

HASTA QUE los centroides ya no se muevan

FIN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

A diferencia de K-Means, DBSCAN agrupa puntos basándose en la densidad, lo que le permite encontrar formas arbitrarias y descartar ruido.

- Principio de funcionamiento:

El algoritmo clasifica los puntos en tres tipos:

1. *Núcleo:* Tiene al menos un número mínimo de vecinos (min_samples) dentro de un radio eps.
2. *Borde:* Está dentro del radio de un punto núcleo pero no tiene suficientes vecinos propios.
3. Ruido: No cumple ninguna de las anteriores.

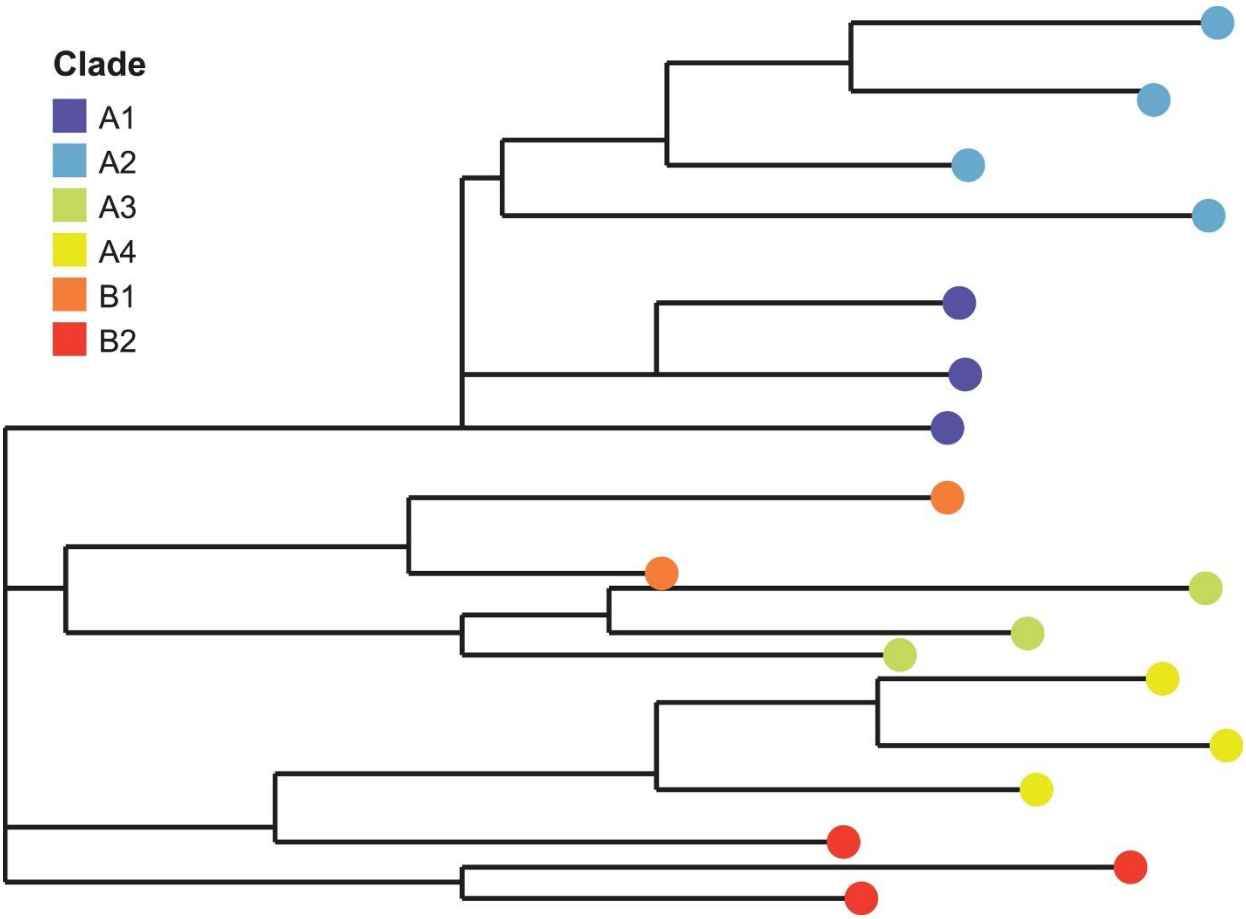
El clúster se expande conectando puntos núcleo vecinos.

- **Parámetros clave:**
    - eps: La distancia máxima para considerar dos puntos como vecinos.
    - min_samples: El número mínimo de puntos para formar una región densa.
- **Ventajas y limitaciones:**
    - *Ventajas:* No requiere especificar el número de clústeres. Es robusto a outliers (los marca como ruido). Encuentra formas no convexas (ej. lunas o anillos).
    - *Limitaciones:* Tiene problemas si los clústeres tienen densidades muy diferentes. Es difícil determinar el valor óptimo de eps en espacios de alta dimensión.
- Ejemplo de aplicación:

Caso: Detección de zonas de alta criminalidad en una ciudad (puntos GPS).

- Si muchos reportes de crimen ocurren en un radio de 200m (eps), se forma una "zona roja".
- Los crímenes aislados en zonas rurales se marcan como ruido (incidentes aislados) y no distorsionan el análisis.

Clustering Jerárquico (Aglomerativo)



Construye una jerarquía de clústeres, usualmente visualizada mediante un dendrograma.

- Principio de funcionamiento:

Sigue un enfoque "bottom-up" (de abajo hacia arriba):

1. Cada punto comienza siendo su propio clúster.
2. Se encuentran los dos clústeres más cercanos y se fusionan en uno solo.
3. Se repite el paso 2 hasta que todos los puntos están en un único gran clúster.

El usuario puede "cortar" el árbol (dendrograma) a la altura deseada para obtener el número de grupos.

- **Parámetros clave:**
    - linkage: Criterio de unión (Ward minimiza varianza, Complete usa la distancia máxima, Single usa la mínima).

    - metric: Métrica de distancia (Euclidiana, Manhattan).

- **Ventajas y limitaciones:**
    - *Ventajas:* El dendrograma ofrece una visualización rica de la estructura de datos. No asume un número fijo de clústeres.
    - *Limitaciones:* Es computacionalmente costoso, por lo que no sirve para Big Data.
- Ejemplo de aplicación:

Caso: Taxonomía biológica.

    - Agrupar especies animales basándose en características genéticas. El dendrograma mostraría cómo las especies se unen en géneros, luego en familias, órdenes, etc.

Algoritmos de Reducción de Dimensionalidad

Se han seleccionado PCA y t-SNE por ser los estándares en reducción lineal y visualización no lineal, respectivamente.

Análisis de Componentes Principales (PCA)

Es una técnica lineal que transforma los datos a un nuevo sistema de coordenadas.

- Fundamento matemático/conceptual:

PCA busca encontrar las "direcciones" (componentes principales) en las que los datos varían más.

1. Calcula la matriz de covarianza de los datos.
2. Obtiene los vectores propios (eigenvectors) y valores propios (eigenvalues).
3. Proyecta los datos originales sobre los vectores propios que tienen los valores propios más altos (mayor varianza).

El resultado es que el primer componente (PC1) retiene la mayor parte de la información, el PC2 la siguiente mayor, etc., y todas son ortogonales (no correlacionadas).

- **Parámetros clave:**
    - n_components: Número de dimensiones a mantener (o porcentaje de varianza explicada deseada, ej. 0.95).
- **Ventajas y limitaciones:**

    - *Ventajas:* Elimina la multicolinealidad. Reduce el ruido. Es rápido y determinista.
    - *Limitaciones:* Solo captura relaciones lineales. Se pierde la interpretabilidad directa de las variables originales (las nuevas variables son combinaciones matemáticas de las anteriores).
- Ejemplo ilustrativo:

Caso: Compresión de imágenes.

    - Una imagen de 100x100 píxeles tiene 10,000 dimensiones.
    - Aplicando PCA, podemos quedarnos con los primeros 50 componentes que explican el 90% de la varianza (formas básicas, sombras).
    - Se reduce el tamaño del archivo drásticamente perdiendo muy poca calidad visual.

t-SNE (t-Distributed Stochastic Neighbor Embedding)

Es una técnica no lineal utilizada principalmente para visualización de datos de alta dimensión.

- Fundamento matemático/conceptual:

Convierte las similitudes entre puntos de datos en probabilidades conjuntas. Intenta minimizar la divergencia Kullback-Leibler entre la probabilidad conjunta del espacio de alta dimensión (original) y la del espacio de baja dimensión (mapa).

Básicamente: si dos puntos están cerca en el espacio original de 100 dimensiones, t-SNE hace todo lo posible para que estén cerca en el gráfico 2D.

- **Parámetros clave:**
    - perplexity: Relacionado con el número de vecinos cercanos que se consideran (balance entre estructura local y global).
    - learning_rate: Velocidad de ajuste del algoritmo.
- **Ventajas y limitaciones:**
    - *Ventajas:* Excelente para visualizar clústeres complejos que PCA no puede separar (ej. espirales). Preserva muy bien la estructura local.
    - *Limitaciones:* Es estocástico (el resultado cambia cada vez si no se fija la semilla). Muy lento computacionalmente. No preserva bien las distancias globales ni la densidad.
- Ejemplo ilustrativo:

Caso: Visualización de dígitos manuscritos (Dataset MNIST).

- Entrada: Imágenes de 784 píxeles (dimensiones).
- Salida: Un gráfico de dispersión (scatter plot) 2D.
- Resultado: Se verán 10 "islas" claramente separadas, correspondientes a los dígitos del 0 al 9, algo que PCA mezclaría.

Comparativa y Conclusiones

Tabla Comparativa: Clustering vs. Reducción de Dimensionalidad

Característica	Agrupación (Clustering)	Reducción de Dimensionalidad
Objetivo	Encontrar grupos de *registros* (filas) similares.	Reducir el número de *variables* (columnas) o características.
Salida	Una etiqueta de grupo para cada dato (ej. Grupo A, Grupo B).	Un nuevo set de datos con menos columnas (ej. PC1, PC2).
¿Cuándo usar?	Segmentación de mercado, detección de anomalías, organización documental.	Preprocesamiento antes de entrenar modelos, visualización 2D/3D, compresión.
Dependencia	A menudo se beneficia de una reducción de dimensionalidad previa.	Puede usarse independientemente o como paso previo al clustering.

Situaciones prácticas

En un flujo de trabajo típico de Ciencia de Datos, la reducción de dimensionalidad suele tener prioridad temporal. Si tenemos un dataset con 1,000 variables (columnas) y aplicamos K-Means directamente, sufriremos la "maldición de la dimensionalidad": en espacios de muchas dimensiones, todas las distancias tienden a parecerse, haciendo que el clustering sea ineficaz. Por tanto, primero aplicamos PCA para bajar a 50 componentes, y luego aplicamos K-Means sobre esos componentes para obtener segmentos de calidad.

Conclusiones Generales

El dominio de estos algoritmos es fundamental para el ingeniero en tecnologías de la información. Mientras que los algoritmos supervisados dependen de datos etiquetados (que son caros y escasos), las técnicas no supervisadas como el Clustering y la Reducción de Dimensionalidad nos permiten explotar la vasta cantidad de datos crudos disponibles. La elección correcta (ej. usar DBSCAN para datos geoespaciales o PCA

para pre-procesar imágenes) define el éxito de un proyecto de extracción de conocimiento.

Referencias

Awan, A. A. (2025, January 21). *Comprender la reducción de la dimensionalidad*. https://www.datacamp.com/es/tutorial/understanding-dimensionality-reduction

Lab, B. B. (2022, June 30). *Los algoritmos de agrupación del machine learning: ¿qué son y cuándo se utilizan?* The Black Box Lab. https://theblackboxlab.com/algoritmos-de-agrupacion-machine-learning/