UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

EXTRACCION DE CONOCIMIENTO EN BASES DE DATOS

IV.1. ALGORITMOS DE AGRUPACIÓN

IDGS91N

PRESENTA:

SEBASTIÁN ACOSTA ORTIZ

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 30 de noviembre de 2025

Introducción

En el campo de la minería de datos y el aprendizaje automático, el análisis no supervisado tiene un papel fundamental para descubrir patrones ocultos en conjuntos de datos. Dos de las herramientas más relevantes dentro de este enfoque son el clustering (agrupación) y la reducción de dimensionalidad.
El clustering se utiliza para agrupar objetos similares sin necesidad de etiquetas, revelando estructuras naturales en los datos. La reducción de dimensionalidad, por otro lado, permite simplificar datasets con muchas variables, eliminando redundancia y ruido, facilitando la visualización y mejorando la eficiencia de muchos algoritmos.

Este reporte describe los principales algoritmos de cada categoría, su funcionamiento, sus ventajas y limitaciones, además de ejemplos prácticos de aplicación.

Algoritmos de agrupación (Clustering)

A continuación, se presentan tres algoritmos representativos: K-means, clustering jerárquico aglomerativo y DBSCAN.

K-means

Principio de funcionamiento

K-means divide los datos en k grupos minimizando la distancia entre los puntos y el centroide del grupo.
El proceso sigue estos pasos:

1. Seleccionar *k* centroides iniciales.
2. Asignar cada punto al centroide más cercano.
3. Recalcular los centroides promediando los puntos asignados.
4. Repetir hasta que no cambien las asignaciones.

Parámetros clave

- k: número de clusters.
- Métrica de distancia: típicamente euclidiana.
- Iteraciones máximas.

Ventajas

- Rápido y eficiente en datasets grandes.
- Fácil de implementar.
- Resultados interpretables.

Limitaciones

- Requiere conocer *k* de antemano.
- Sensible a valores atípicos.
- Solo detecta grupos esféricos.

Ejemplo de aplicación (pseudocódigo)

```
Inicializar k centroides al azar
Mientras no converja:
    Para cada punto:
        asignar al centroide más cercano
    Recalcular centroides
Retornar grupos
```

Clustering jerárquico aglomerativo

Principio de funcionamiento

Este método construye una jerarquía de grupos mediante un enfoque bottom-up.

Proceso:

1. Cada punto inicia como un cluster individual.
2. Se fusionan iterativamente los dos clusters más similares.
3. El proceso continúa hasta llegar a un solo cluster o hasta un nivel deseado.

Parámetros clave

- Métrica de distancia (euclidiana, Manhattan).
- Método de enlace:
  - Single-linkage
  - Complete-linkage
  - Average-linkage
  - Ward

Ventajas

- No requiere elegir k inicialmente.
- Representación visual mediante dendrogramas.
- Detecta estructuras complejas.

Limitaciones

- Computacionalmente más costoso ($O(n^2)$).
- No funciona bien con datasets muy grandes.
- Difícil corregir fusiones tempranas incorrectas.

Ejemplo (flujo simplificado)

```
Crear un cluster por cada punto

Calcular matriz de distancias

Mientras queden más de un cluster:

    fusionar clusters más cercanos

Dibujar dendrograma
```

2.3 DBSCAN (Density-Based Spatial Clustering)

Principio de funcionamiento

DBSCAN forma grupos con base en densidades, agrupando puntos cercanos y detectando ruido.

Elementos clave:

- Puntos núcleo: tienen suficientes vecinos.
- Puntos borde: están cerca de un núcleo.
- Ruido: no cumplen con densidad mínima.

Parámetros clave

- eps: radio del vecindario.
- min_samples: puntos mínimos para ser núcleo.

Ventajas

- Detecta clusters de formas arbitrarias.
- Maneja ruido de forma natural.
- No requiere especificar k.

Limitaciones

- Sensible a la escala de los datos.
- Difícil ajustar eps y min_samples.
- Problemas en datasets con densidad variable.

Ejemplo de aplicación

Para cada punto:

    obtener vecinos dentro de eps

    si vecinos >= min_samples:

        etiquetar como núcleo

Expandir cluster desde puntos núcleo

Ignorar puntos sin densidad (ruido)

Algoritmos de reducción de dimensionalidad

Se presentan **PCA** y **t-SNE**, dos técnicas ampliamente usadas.

PCA (Análisis de Componentes Principales)

Fundamento matemático

PCA transforma un conjunto de variables correlacionadas en nuevas variables llamadas **componentes principales**, que:

- Son combinaciones lineales de las originales.
- Capturan la mayor varianza posible.
- Se obtienen mediante descomposición en valores propios de la matriz de covarianza.

Parámetros clave

- **Número de componentes (n_components).**
- **Método de normalización.**

Ventajas

- Reduce dimensionalidad preservando varianza.
- Elimina redundancia.
- Rápido y matemáticamente elegante.

Limitaciones

- Solo captura relaciones lineales.
- Difícil interpretar componentes.
- Sensible a escalas.

Ejemplo conceptual

Estandarizar datos

Calcular matriz de covarianza

Obtener eigenvectores y valores propios

Proyectar datos en componentes principales

t-SNE (t-Distributed Stochastic Neighbor Embedding)

Fundamento conceptual

t-SNE busca conservar las **relaciones locales** entre puntos al proyectarlos en 2D o 3D. Convierte distancias reales en **probabilidades de similitud** y minimiza la divergencia Kullback-Leibler entre distribuciones.

Parámetros clave

- **Perplexity:** tamaño efectivo del vecindario.
- **Learning rate.**
- **Número de iteraciones.**

Ventajas

- Visualizaciones 2D/3D altamente interpretables.
- Excelente preservación de estructura local.
- Útil para explorar datasets complejos como imágenes o embeddings.

Limitaciones

- No mantiene estructura global.
- Alto costo computacional.
- No sirve para modelos predictivos (solo visualización).

Ejemplo simplificado

```
Calcular probabilidades de similitud en alta dimensión

Inicializar puntos en 2D

Ajustar posiciones reduciendo KL-divergence

Mostrar scatterplot 2D
```

Comparativa y conclusiones

Clustering vs. Reducción de dimensionalidad

ASPECTO	CLUSTERING	REDUCCIÓN DE DIMENSIONALIDAD
OBJETIVO	Agrupar datos según similitud	Simplificar datos conservando estructura
TIPO DE TÉCNICA	Descubrimiento de patrones	Preprocesamiento / visualización
ENTRADA	Datos en alta dimensión	Datos originales
SALIDA	Etiquetas de grupos	Espacio reducido (2-50 dims)
USO TÍPICO	Segmentación, detección de patrones	Visualización, mejora de modelos

Cuando usar cada uno

- Clustering:
    - Segmentación de clientes
    - Detección de grupos naturales
    - Análisis exploratorio profundo
- Reducción de dimensionalidad:
    - Visualizar datos complejos
    - Acelerar modelos supervisados
    - Eliminar ruido y multicolinealidad

Conclusiones

Los algoritmos de agrupación permiten identificar estructuras latentes en los datos, mientras que los métodos de reducción de dimensionalidad ayudan a simplificar datasets complejos sin perder información relevante. Ambos enfoques son complementarios y frecuentemente se utilizan juntos: por ejemplo, aplicando PCA para reducir dimensiones y después K-means para agrupar.

Referencias

Scikit-learn. (s/f). Recuperado el 30 de noviembre de 2025, de Scikit-learn.org website: https://scikit-learn.org/stable/

Virtual UTCH - Tecnologías de la Información. (s/f). Recuperado el 30 de noviembre de 2025, de Edu.mx website: https://ti.utch.edu.mx/mod/page/view.php?id=11427

What is K-Means algorithm and how it works. (s/f). Recuperado el 30 de noviembre de 2025, de Towardsmachinelearning.org website: https://towardsmachinelearning.org/k-means/