EXTRACCIÓN DE CONOCIMIENTOS EN BASES DE DATOS

ING. LUIS ENRIQUE MASCOTE CANO

Universidad Tecnológica de Chihuahua

ALGORITMOS DE AGRUPACIÓN

Lic. Ricardo Hernández Martínez

Fecha de Entrega: 30/11/25

Introducción

El análisis de datos contemporáneo requiere técnicas capaces de identificar patrones, simplificar información y revelar estructuras ocultas dentro de grandes volúmenes de información. Entre estas técnicas destacan **el clustering** y **la reducción de dimensionalidad**, pilares fundamentales en la extracción de conocimiento.

El **clustering** permite agrupar datos según similitud sin necesidad de etiquetas previas, facilitando la segmentación, detección de anomalías y comprensión de estructuras ocultas. Por otro lado, la **reducción de dimensionalidad** busca representar datos de alta dimensionalidad mediante menos variables preservando la mayor cantidad de información posible, permitiendo visualización, mejora del rendimiento de algoritmos y eliminación de ruido.

Ambos enfoques resultan esenciales en aplicaciones como análisis de clientes, bioinformática, visión computacional y sistemas de recomendación.

Algoritmos de Agrupación (Clustering)

A continuación se describen tres algoritmos representativos: **K-means**, **Clustering Jerárquico Aglomerativo** y **DBSCAN**.

1. K-means

Principio de funcionamiento

K-means busca dividir los datos en *k* grupos minimizando la distancia entre los puntos y los centroides de cada clúster. El proceso iterativo es:

1. Elegir *k* centroides iniciales.
2. Asignar cada dato al centroide más cercano.
3. Actualizar los centroides como la media de los puntos asignados.
4. Repetir 2 y 3 hasta convergencia.

Parámetros clave

- *k*: número de clústeres.
- Inicialización (p. ej., *k-means++*).
- Número máximo de iteraciones.

Ventajas

- Rápido y eficiente en grandes conjuntos.
- Fácil de interpretar.

Limitaciones

- Requiere definir *k* previamente.
- Sensible a valores atípicos.
- Solo captura clústeres esféricos.

Ejemplo de aplicación (pseudocódigo)

```
inicializar k centroides
mientras no converja:
   asignar cada punto al centroide más cercano
```

    recalcular centroides
return clusters

2. Clustering Jerárquico Aglomerativo (HAC)

Principio de funcionamiento

Construye una jerarquía de clústeres mediante un enfoque "bottom-up":

1. Cada punto comienza como un clúster individual.
2. Se fusionan los dos clústeres más similares.
3. Continúa hasta formar un solo clúster o la cantidad deseada.

Usa medidas como *single linkage*, *complete linkage* o *average linkage*.

Parámetros clave

- Métrica de distancia (euclidiana, Manhattan...).
- Criterio de enlace.
- Umbral o número de clústeres deseado.

Ventajas

- No requiere número de clústeres a priori.
- Genera un dendrograma fácil de interpretar.

Limitaciones

- Alto costo computacional.
- Difícil de ajustar en grandes conjuntos.

Ejemplo de aplicación (diagrama conceptual)

iniciar cada punto como clúster
mientras existan clústeres por unir:
    encontrar clústeres más cercanos
    fusionarlos
return dendrograma

3. DBSCAN

Principio de funcionamiento

Agrupa puntos según densidad. Identifica regiones densas como clústeres y puntos aislados como ruido.

Categoriza puntos en:

- Núcleo
- Borde
  Ruido

Parámetros clave

- *eps*: radio de vecindad.
- *min_samples*: mínimo de puntos para formar un clúster.

Ventajas

- Detecta clústeres de forma arbitraria.
- Maneja ruido y outliers.
- No requiere *k*.

Limitaciones

- Sensible a la elección de *eps*.
- Difícil de usar en alta dimensionalidad.

Pseudocódigo simple

```
para cada punto no visitado:
   obtener vecinos en eps
   si suficientes vecinos:
      crear clúster(expansión por densidad)
   sino:
      marcar como ruido
```

Algoritmos de Reducción de Dimensionalidad

Se describen dos métodos: **PCA** y **t-SNE**.

1. Análisis de Componentes Principales (PCA)

Fundamento matemático

PCA encuentra nuevas variables (componentes principales) que son combinaciones lineales de las originales y que maximizan la varianza.

Pasos:

1. Estandarización.
2. Cálculo de la matriz de covarianza.
3. Obtención de autovalores y autovectores.
4. Selección de componentes con mayor varianza.

Parámetros clave

- Número de componentes.
- Escalado previo.

Ventajas

- Reduce ruido.
- Facilita visualización en 2D/3D.
- Acelera modelos.

Limitaciones

- Solo capta relaciones lineales.
- Difícil de interpretar componentes.

Ejemplo simple

X estandarizado -> matriz covarianza -> autovectores -> proyección en componentes

2. t-SNE

Fundamento conceptual

t-SNE busca preservar la estructura local proyectando datos en 2D o 3D. Convierte distancias en probabilidades y minimiza divergencia KL entre espacios.

Parámetros clave

- *perplexity*
- *learning rate*
- Número de iteraciones

Ventajas

- Produce visualizaciones claras.
- Excelente para datos no lineales.

Limitaciones

- Alto costo computacional.
- No es útil para modelado predictivo.
- Resultados no deterministas.

Ejemplo conceptual

calcular probabilidades locales -> proyectar en espacio reducido -> optimizar

Comparativa y Conclusiones

Cuándo usar clustering vs. reducción de dimensionalidad

Objetivo	**Clustering**	**Reducción de Dimensionalidad**
Agrupar datos	✔️	✖️

Visualización	✔️ (limitada)	✔️
Eliminar ruido	✖️	✔️
Encontrar patrones	✔️	✔️
Preparación para otros algoritmos	Parcial	✔️

Prioridad según situación práctica

- Si se desea **segmentar clientes**, **detectar anomalías** o **agrupar patrones**, se usa clustering.
- Si se necesita **reducir complejidad**, **mejorar rendimiento** o **visualizar datos**, se usa reducción de dimensionalidad.
- En muchos flujos, PCA o t-SNE se usan **antes** de aplicar clustering.

Conclusiones Generales

Las técnicas de clustering permiten descubrir estructuras naturales en los datos, mientras que los métodos de reducción de dimensionalidad permiten simplificarlos sin perder información esencial. Usadas conjuntamente, potencian la capacidad analítica en entornos donde la exploración de datos es clave. Su elección depende del objetivo analítico: **descubrir grupos** o **reducir complejidad**. Ambas resultan indispensables en la minería de datos moderna.

Referencias

Libros y artículos fundamentales

- Aggarwal, C. C., & Reddy, C. K. (2014). *Data clustering: Algorithms and applications*. Chapman and Hall/CRC.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666.