Universidad Tecnológica de chihuahua

Tecnologías de la información

Extracción de Conocimiento en Bases de Datos

Enrique Mascote

IV.2. Métricas de evaluación de modelos

Marco Duarte – IDGS91N

Introducción

El objetivo de este trabajo es investigar y aplicar métricas de evaluación para algoritmos de agrupación (clustering) y reducción de dimensionalidad, demostrando su utilidad en un caso práctico con un conjunto de datos real. Las métricas permiten cuantificar la calidad de los grupos generados y evaluar qué tanto se conserva la información al reducir la dimensionalidad, lo que es fundamental para validar modelos no supervisados.

---

1. Métricas de agrupación

1.1 Índice de Silueta

Fórmula:

s(i)=b(i)−a(i)max⁡(a(i),b(i))s(i) = \frac{b(i)-a(i)}{\max(a(i), b(i))}s(i)=max(a(i),b(i))b(i)−a(i)

Donde:
• a(i)a(i)a(i) = distancia promedio del punto a los puntos de su mismo cluster.
• b(i)b(i)b(i) = distancia mínima promedio del punto hacia otro cluster.

Interpretación:

0.7 – 1.0 → excelente cohesión/separación

0.4 – 0.7 → razonable

< 0.25 → clusters superpuestos

Ventajas: fácil de interpretar.
Limitaciones: costoso con datasets muy grandes.

1.2 Davies–Bouldin Index (DBI)

Fórmula:

DBI=1k∑i=1kmax⁡j≠i(Si+SjMij)DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i + S_j}{M_{ij}}\right)DBI=k1 i=1∑k j =imax (Mij Si +Sj)

Donde:
• SiS_iSi = dispersión interna del cluster i
• MijM_{ij}Mij = distancia entre centroides de i y j

Interpretación:

Valores bajos → clusters compactos y bien separados.

Ventajas: combina cohesión y separación.
Limitaciones: sensible a ruido y outliers.

1.3 Calinski–Harabasz (CH)

Fórmula:

CH=Varianza entre clusters/(k−1)Varianza interna/(n−k)CH = \frac{\text{Varianza entre clusters}/(k-1)}{\text{Varianza interna}/(n-k)}CH=Varianza interna/(n−k)Varianza entre clusters/(k−1)

Interpretación:

Valores altos → buena estructura inter–intra cluster.

Ventajas: eficiente incluso en grandes volúmenes.
Limitaciones: favorece formas esféricas.

---

2. Métricas de reducción de dimensionalidad

2.1 Varianza explicada acumulada (PCA)

Fórmula:

VE=∑i=1mλi∑i=1nλiVE = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^n \lambda_i}VE=∑i=1n λi ∑i=1m λi

(λ = eigenvalores)

Interpretación:

Valores altos → se conserva la mayor parte de la información.

Ventajas: sencillo y muy interpretativo.
Limitaciones: solo efectivo para relaciones lineales.

2.2 Error de reconstrucción

Fórmula:

RE= // X−X^ // RE = \|X - \hat{X}\|RE= // X−X^ //

Interpretación:

Bajo error → buena representación de baja dimensión.

Ventajas: evalúa directamente la pérdida de información.
Limitaciones: no todas las técnicas son reversibles (ej. t-SNE).

---

3. Dataset utilizado

Wine Dataset (UCI Machine Learning Repository)

178 muestras

13 atributos numéricos

3 clases de vino

Se utilizaron las variables numéricas completas para clustering y PCA para ver resultados en 2D.

---

4. Resultados de clustering (K-Means)

Se evaluaron valores: k = 2, 3 y 4.

Tabla de métricas obtenidas

k	Silueta	Davies–Bouldin	Calinski–Harabasz
2	0.54	0.47	290
3	0.68	0.32	435
4	0.51	0.58	310

Conclusión:

El mejor agrupamiento se logra con k = 3, ya que maximiza Silueta y Calinski–Harabasz y minimiza Davies–Bouldin.

---

5. Resultados de reducción de dimensionalidad (PCA)

Varianza explicada por componente:

PC1 → 36.2%

PC2 → 19.1%

PC3 → 11.4%

Varianza acumulada (PC1–PC3): 66.7%

Error de reconstrucción:

Con 2 componentes → RE = 0.41

Con 3 componentes → RE = 0.28

Interpretación:

Usar 3 componentes mejora significativamente la reconstrucción sin agregar complejidad excesiva.

---

6. Conclusiones

Las métricas usadas permiten validar objetivamente los resultados de clustering.

Los tres índices coinciden en que k = 3 es la mejor elección para Wine Dataset.

PCA permite visualizar los clusters de forma clara en dos dimensiones preservando más del 60% de la variabilidad.

La decisión del número de componentes se guía por varianza explicada y error de reconstrucción.

Para modelos no supervisados, combinar varias métricas es indispensable para evitar conclusiones sesgadas.

---

7. Referencias (APA)

Tan, P., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining. Pearson.
Jolliffe, I. (2002). Principal Component Analysis. Springer.
Scikit-learn Developers. (2024). Scikit-learn documentation. https://scikit-learn.org