UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

Extracción de Conocimiento de Base de Datos

Algoritmos de Agrupación

IDGS91N

Alumno:

Erick Fabian Terrazas Hernandez

Docente:

Enrique Mascote

Chihuahua, Chih., 30 de noviembre de 2025

Introducción

La evaluación de modelos de agrupación y reducción de dimensionalidad es fundamental para medir la cohesión, separación y calidad de las representaciones reducidas. Las métricas permiten validar objetivamente si un modelo encuentra patrones útiles o si pierde información importante al reducir dimensiones.

1. Métricas de Agrupación

1.1 Índice de Silueta

Definición:

Mide la calidad de los clústeres comparando la cohesión interna y la separación respecto a otros clústeres.

Fórmula:

$S(i) = (b(i) - a(i)) / \max(a(i), b(i))$

Interpretación:

- Cercano a 1 → excelente separación
- Cercano a 0 → frontera entre clústeres
- Negativo → mala asignación

Ventajas / Limitaciones:

- Fácil de interpretar
- Costoso en datasets grandes

1.2 Davies–Bouldin

Definición:

Evalúa la similitud entre clústeres: busca minimizar la dispersión interna y maximizar la distancia entre ellos.

Interpretación:

- Valor bajo = mejor desempeño

Ventajas: automático

Limitaciones: favorece formas esféricas

1.3 Calinski–Harabasz

Definición:

Relación entre dispersión inter-clúster e intra-clúster.

Interpretación:

- Valor alto = mejor estructura de clústeres

2. Métricas de Reducción de Dimensionalidad

2.1 Varianza explicada acumulada (PCA)

Definición:

Cantidad de variabilidad conservada en las componentes principales.

Interpretación:

- Valores altos → mejor preservación

2.2 Error de reconstrucción

Definición:

Mide cuánto cambian los datos después de reducir y reconstruir.

Interpretación:

- Valor bajo → buena reconstrucción

3. Caso de Estudio y Aplicación

Dataset utilizado: Iris

- 150 instancias
- 4 atributos numéricos
- Se usan solo datos numéricos (no etiquetas)

3.1 Clustering aplicado: K-means (k=3)

Código utilizado

```
6
7   # Cargar dataset Iris
8   data = load_iris()
9   X = data.data
10
11  # K-means
12  kmeans = KMeans(n_clusters=3, random_state=42)
13  labels = kmeans.fit_predict(X)
14
15  # Métricas
16  sil = silhouette_score(X, labels)
17  db = davies_bouldin_score(X, labels)
18  ch = calinski_harabasz_score(X, labels)
19
20  print("Índice de Silueta:", sil)
21  print("Davies-Bouldin:", db)
22  print("Calinski-Harabasz:", ch)
23
24  # Gráfica en PCA para visualizar clusters
25  from sklearn.decomposition import PCA
26  pca = PCA(n_components=2)
27  X_pca = pca.fit_transform(X)
28
29  plt.scatter(X_pca[:,0], X_pca[:,1], c=labels)
30  plt.title("Clusters usando K-means (visualizados con PCA)")
31  plt.xlabel("Componente 1")
32  plt.ylabel("Componente 2")
33  plt.show()
```

Tabla de resultados obtenidos

Métrica	Valor
Índice de Silueta	0.55
Davies–Bouldin	0.63

Calinski–Harabasz	560.2

3.2 Reducción aplicada: PCA

Código utilizado

```python
# PCA a 2 componentes
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)
X_reconstructed = pca.inverse_transform(X_pca)

# Varianza explicada
explained = pca.explained_variance_ratio_
print("Varianza explicada por PC1 y PC2:", explained)
print("Varianza explicada acumulada:", explained.sum())

# Error de reconstrucción
error = np.mean((X - X_reconstructed)**2)
print("Error de reconstrucción:", error)

# Gráfica de varianza explicada
plt.bar(["PC1", "PC2"], explained)
plt.title("Varianza explicada por componente (PCA)")
plt.ylabel("Proporción de varianza")
plt.show()
```

Resultados de PCA

- PC1: 72.7%
- PC2: 23.0%
- Varianza acumulada: **95.7%**
- Error de reconstrucción: **0.042**

4. Comparativa y Análisis

Las métricas de agrupación evaluaron correctamente que los clústeres formados por K-means presentan buena cohesión y separación.

Las métricas de reducción mostraron que PCA conserva casi toda la información, permitiendo visualizar los clústeres de manera efectiva.

5. Conclusiones

La combinación de técnicas de agrupación y reducción de dimensionalidad permite descubrir patrones y visualizar mejor los datos.

Las métricas empleadas proporcionan una cuantificación clara del desempeño de cada técnica, facilitando su comparación y selección.

6. Referencias (APA)

Jain, A. K. (2010). *Data clustering: 50 years beyond K-means*. Pattern Recognition.

Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer.