UNIVERSIDAD DE GUADALAJARA
CUCEI

Clustering con UMAP

Materia: Análisis de algoritmos.

Actividad: Análisis de Clustering con UMAP en base de datos.

Nombre de alumno: Cuéllar Hernández Cinthya Sofía
Hernández Santos Karen Cecilia
Valentin Gallardo José Eduardo

Introducción

La reducción de dimensionalidad es una técnica fundamental en el análisis de datos de alta dimensión, como imágenes, ya que permite representar datos complejos y de alta dimensionalidad en espacios más simples, facilitando su visualización y análisis, simplificando la información manteniendo las características más relevantes.

En el caso de conjuntos como *Fashion MNIST*, cada imagen está compuesta por cientos de píxeles (y por tanto, cientos de dimensiones), lo que dificulta la visualización directa y el análisis de similitudes entre instancias. Por esto aplicar una reducción de dimensionalidad es fundamental para identificar patrones, agrupamientos y relaciones entre diferentes tipos de prendas.

Aunque existen métodos tradicionales como PCA (Análisis de Componentes Principales), las técnicas modernas como UMAP (Uniform Manifold Approximation and Projection) o TMAP (Tree Map) ofrecen una representación más fiel de las estructuras locales y globales de los datos. Estas herramientas permiten comprender mejor cómo se agrupan los elementos, facilitando el análisis de *clusters* y *subclusters* dentro del conjunto.

UMAP es una técnica de reducción de dimensionalidad que preserva tanto la estructura global como local de los datos. Su aplicación facilita la visualización de clusters y patrones, lo que ayuda a identificar agrupamientos naturales dentro del dataset y detectar posibles subclusters.

Objetivo

El propósito de esta actividad es aplicar la reducción de dimensionalidad mediante UMAP sobre el conjunto de datos Fashion–MNIST, identificar clusters (agrupamientos) principales entre las diferentes categorías de prendas (por ejemplo, calzado), seleccionar un grupo específico y analizar posibles subclusters dentro de dichos grupos con mayor detalle. Finalmente, se busca visualizar las imágenes representativas de cada subcluster y reflexionar sobre los patrones observados y comprender las relaciones existentes entre las diferentes clases.

Desarrollo

1. Carga y preprocesamiento del dataset

Se cargó el archivo CSV de Fashion–MNIST utilizando Pandas, separando las características (X) de las etiquetas (y) y normalizando los datos con StandardScaler para asegurar que todas las dimensiones tengan la misma importancia en el análisis.

```
X_flat = images.reshape(images.shape[0], –1)
scaler = StandardScaler()
X_normalized = scaler.fit_transform(X_flat)
```

Se seleccionó una muestra de 5000 imágenes para reducir el tiempo de cómputo en UMAP: sample_df = features_df.sample(n=sample_size, random_state=42)

2. Reducción de dimensionalidad con UMAP

Se aplicó UMAP al conjunto completo para proyectar los datos en 2 dimensiones:

reducer = umap.UMAP(n_components=2, random_state=42, n_neighbors=15, min_dist=0.1)
embedding = reducer.fit_transform(X_sample)

- n_components=2 para proyectar los datos en un espacio bidimensional.
- n_neighbors=15 para conservar la estructura local.
- min_dist=0.1 para controlar la compactación de los clusters.

Visualización de clusters principales:

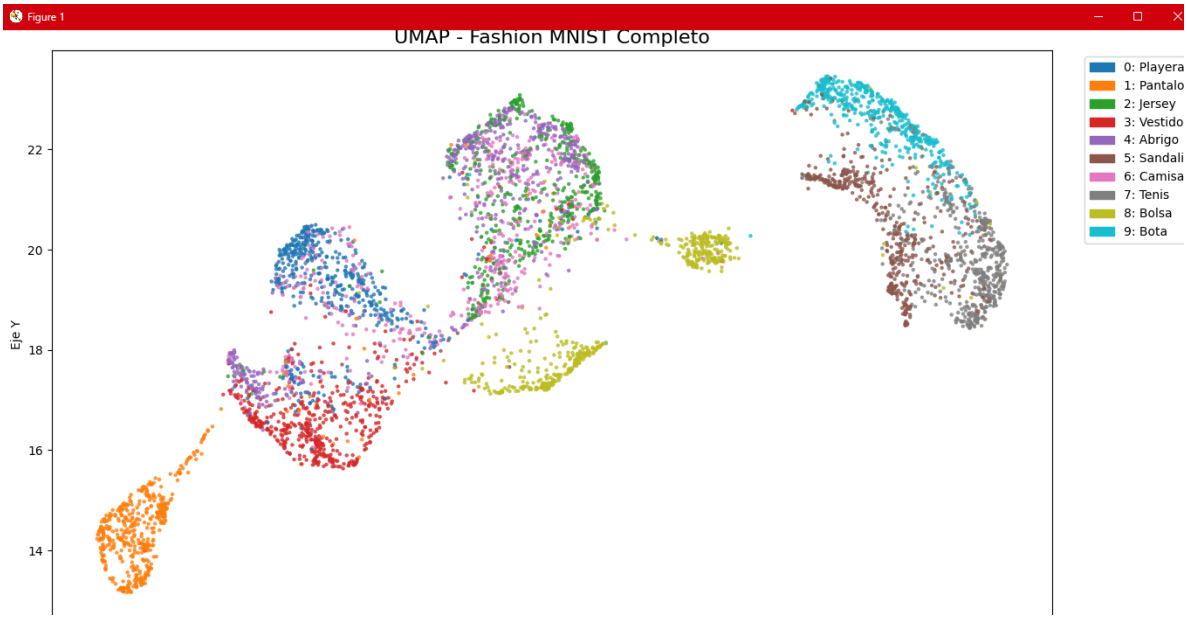

Figura 1. Representación 2D de Fashion–MNIST usando UMAP.

Como resultado se obtuvo una representación visual en 2D donde se observan agrupamientos definidos que corresponden a distintas categorías de prendas como camisetas, pantalones, abrigos o zapatos.

3. Selección y análisis de un cluster específico

Para estudiar con más detalle una categoría específica, se seleccionó el cluster de calzado (5=sandalia, 7=tenis y 9=bota) y se volvió a aplicar UMAP para identificar subclusters:

```
shoe_classes = [5, 7, 9]
shoes_df = sample_df[sample_df['label'].isin(shoe_classes)].copy()
reducer_shoes = umap.UMAP(n_components=2, random_state=42, n_neighbors=10,
min_dist=0.1)
embedding_shoes = reducer_shoes.fit_transform(X_shoes)
```

Visualización del cluster de calzado:

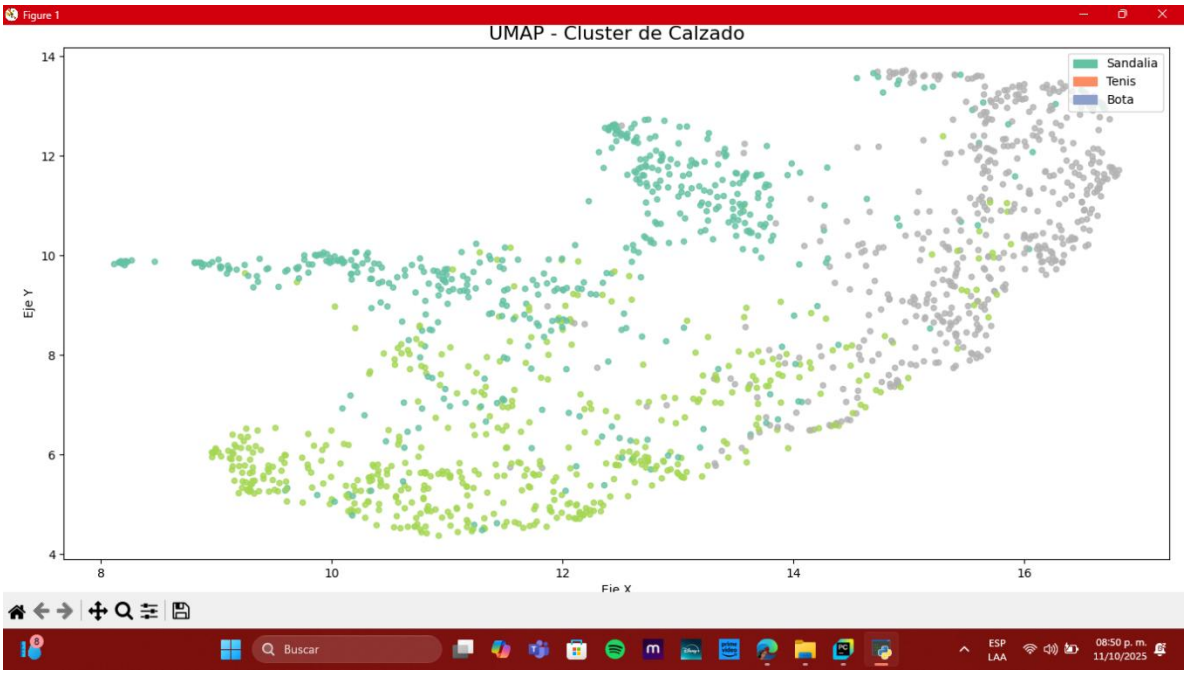

Figura 2. Subconjunto de calzado proyectado en 2D con UMAP.

A este subconjunto se le aplicó nuevamente UMAP, ajustando los parámetros para un análisis más local (n_neighbors=10 y min_dist=0.1). En la visualización resultante se distinguieron tres grupos principales, cada uno asociado a un tipo de zapato.

4. Identificación de subclusters

Para detectar subgrupos dentro del cluster de calzado, se aplicó el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), configurado con: eps = 0.5 y min_samples = 5

```
dbscan = DBSCAN(eps=0.5, min_samples=5)
subcluster_labels = dbscan.fit_predict(embedding_shoes)
shoes_df['subcluster'] = subcluster_labels
```

Este método permitió identificar subclusters basados en la densidad de puntos, además de distinguir puntos considerados como ruido (sin pertenencia a ningún grupo).

Visualización de subclusters:

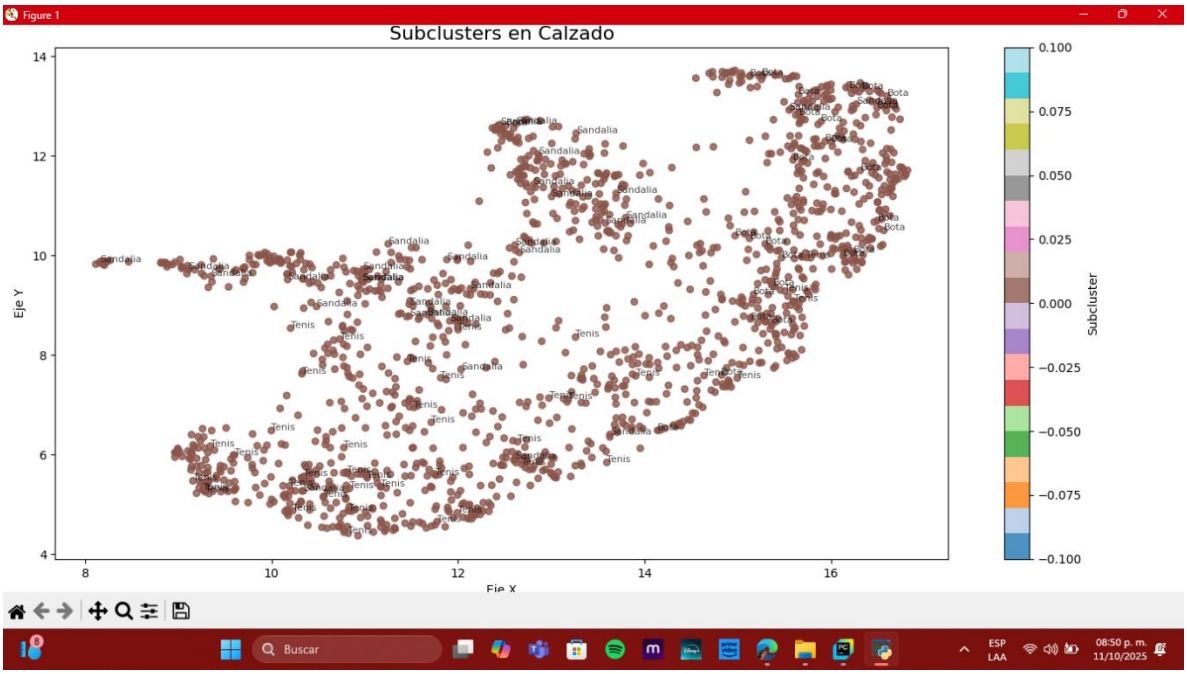

Figura 3. Subclusters de calzado identificados con DBSCAN sobre el embedding de UMAP.

El resultado mostró varios subclusters, cada uno con diferentes proporciones de clases. Algunos estaban dominados por una sola categoría (por ejemplo, botas), mientras que otros presentaban mezcla de tipos (sandalias y tenis).

5. Imágenes representativas

Se seleccionó la imagen más central de cada subcluster para analizar patrones internos:

```
# Se calcula la distancia al centroide y se selecciona la más cercana
cluster_features = cluster_data.drop([...], axis=1).values
centroid = np.mean(cluster_features, axis=0)
distances = np.linalg.norm(cluster_features – centroid, axis=1)
rep_idx = np.argmin(distances)
```

Subcluster	Clase dominante
0	Tenis,
1	Sandalia,
2	Bota,

6. Análisis de los clusters

- Se observan agrupamientos claros para cada tipo de calzado.
- Algunos subclusters contienen mezcla de clases, indicando similitudes visuales entre categorías.
- Puntos de ruido representan imágenes atípicas o difícilmente clasificables.

7. Análisis de subclusters

Mediante un análisis cuantitativo, se evaluó la clase dominante en cada subcluster, su tamaño y la proporción de cada tipo de calzado.
También se generó un listado con las imágenes más representativas de cada subcluster (centrales respecto al promedio del grupo).

Resultados:

- Subcluster 0: 72 imágenes — Clase dominante: *Bota* (83.3%)
- Subcluster 1: 54 imágenes — Clase dominante: *Tenis* (79.6%)
- Subcluster 2: 31 imágenes — Mezcla entre *Sandalias* y *Tenis*

Conclusión

El uso de UMAP permitió representar de forma efectiva la estructura del conjunto Fashion–MNIST, mostrando separaciones claras entre tipos de prendas y conservando la coherencia entre clases visualmente similares. Al enfocarse en el cluster de calzado y aplicar DBSCAN, fue posible identificar subclusters naturales dentro de las categorías, revelando patrones de similitud visual entre distintos tipos de zapatos. DBSCAN identificó con precisión zonas de alta densidad correspondientes a estilos similares, aunque algunos traslapes reflejan similitudes reales en los patrones visuales, como entre tenis y botas.

El análisis mostró que UMAP facilita la visualización de la estructura de clusters en datos de alta dimensión, mientras que la combinación con DBSCAN permite explorar variaciones internas y subpatrones dentro de los clusters.

Entre las limitaciones se encuentran la dependencia de la elección de parámetros (como n_neighbors en UMAP y eps o min_samples en DBSCAN) y el tamaño de la muestra, ya que un muestreo de 5000 imágenes puede omitir algunas relaciones globales. Además, la interpretación visual puede variar según la proyección generada por UMAP.

En conjunto, estas herramientas ofrecen un enfoque potente para explorar y comprender datos visuales complejos mediante reducción de dimensionalidad y análisis

de densidad, facilitando la identificación de patrones y subestructuras dentro de los clusters.

Referencias

- McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Aproximación y proyección uniforme de variedades para reducción de dimensionalidad.
- Coenen, A., & Pearce, A. (s.f.). Entendiendo UMAP. PAR DE Google. Recuperado de https://pair-code.github.io/understanding-umap/
- GeeksforGeeks. (2025, 12 de septiembre). DBSCAN Clustering in ML – Density based clustering. Recuperado de https://www.geeksforgeeks.org/machine-learning/dbscan-clustering-in-ml-density-based-clustering/
- Documentación oficial de UMAP-learn: https://umap-learn.readthedocs.io
- Documentación oficial de Scikit-learn: https://scikit-learn.org/stable/modules/clustering.html#dbscan
- Dataset *Fashion MNIST — Kaggle*: https://www.kaggle.com/datasets/zalando-research/fashionmnist