REPORTE INVESTIGACIÓN

Act. 03 Análisis de Clustering con TMAP en base de datos

Análisis de Algoritmos Sección: D06

Integrantes:
López Galván Melanie Montserrat
De la Mora Villaseñor Diego Gabriel
Lopez Esparza Angel Emanuel

Maestro:
Jorge Ernesto Lopez Arce Delgado

Funcionamiento de TMAP

1. Introducción

El presente documento tiene como objetivo explicar qué es TMAP, cómo funciona esta técnica de reducción de dimensionalidad y qué herramientas son necesarias para su implementación en proyectos de análisis de datos. TMAP es una herramienta poderosa y visualmente atractiva, utilizada principalmente para visualizar relaciones entre grandes volúmenes de datos de alta dimensión.

2. ¿Qué es TMAP?

TMAP (Tree-based Manifold Approximation and Projection) es una técnica de reducción de dimensionalidad cuyo propósito principal es representar datos complejos y con numerosas características (por ejemplo, los píxeles de una imagen) en un espacio bidimensional o tridimensional de manera coherente y comprensible.

Por ejemplo, en un conjunto de datos como Fashion-MNIST, cada imagen de prenda de ropa tiene 784 píxeles, lo que equivale a un punto en un espacio de 784 dimensiones. TMAP proyecta esos puntos en un mapa 2D, agrupando los elementos similares entre sí.

La principal ventaja de TMAP es su capacidad para preservar tanto la estructura local como la global de los datos. Esto permite no solo agrupar correctamente los elementos parecidos (por ejemplo, todas las sandalias juntas), sino también mostrar cómo se relacionan los diferentes grupos entre sí (por ejemplo, los zapatos se ubican cerca de los botines, pero lejos de las camisetas).

3. Funcionamiento de TMAP

El proceso de TMAP puede explicarse en cuatro etapas principales, que permiten construir el mapa final a partir de los datos originales:

3.1. Indexación (LSH Forest)

TMAP utiliza un método de indexación denominado Locality Sensitive Hashing (LSH) para encontrar los “vecinos más cercanos” de cada punto sin necesidad de comparar

todos los datos entre sí. Este proceso actúa como un índice de libro, que permite acceder rápidamente a los elementos relacionados, mejorando significativamente la eficiencia.

3.2. Construcción del Grafo

Con la información obtenida en el paso anterior, TMAP construye un grafo o red de conexiones entre cada punto y sus vecinos más cercanos. Cada punto puede imaginarse como una ciudad conectada mediante carreteras con las ciudades más próximas.

3.3. Creación del Árbol de Expansión Mínima (MST)

A partir del grafo, TMAP selecciona únicamente las conexiones más importantes mediante la creación de un Árbol de Expansión Mínima (Minimum Spanning Tree - MST). Este árbol conserva la conectividad de todos los puntos, pero utilizando el menor número posible de enlaces. El resultado es una estructura que representa la 'columna vertebral' o la estructura fundamental de los datos.

3.4. Visualización del Mapa

Finalmente, TMAP proyecta el árbol obtenido en un espacio bidimensional, utilizando un algoritmo de diseño (layout algorithm) que mantiene cercanos los puntos conectados en el árbol. El resultado es una visualización clara y ramificada, característica de las representaciones producidas por TMAP.

4. Librerías Necesarias y Configuración

Para la implementación de TMAP en un entorno de Python, es necesario instalar una serie de librerías esenciales para la carga, manipulación y visualización de los datos.

Librería	Descripción	Instalación
tmap	Implementa el algoritmo TMAP y genera el mapa de proyección.	pip install tmap
pandas	Permite la carga y manipulación de archivos CSV (como el de Fashion-MNIST).	pip install pandas

numpy	Facilita las operaciones numéricas y es una dependencia común en ciencia de datos.	pip install numpy
scikit-learn	Proporciona herramientas para preprocesamiento, escalado y comparación de modelos.	pip install scikit-learn
matplotlib	Se utiliza para mostrar las imágenes o resultados de manera visual.	pip install matplotlib
faerun	Librería complementaria para crear visualizaciones interactivas exportables en formato HTML.	pip install faerun

5. Conclusión

TMAP constituye una herramienta moderna y eficiente para explorar y visualizar grandes volúmenes de datos de alta dimensionalidad. Su capacidad para mantener la estructura global y local de los datos lo convierte en una alternativa valiosa frente a otros métodos de proyección, especialmente en el análisis de conjuntos de datos complejos como Fashion-MNIST.