UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERÍAS

INGENIERO EN COMPUTACIÓN

ANÁLISIS DE ALGORITMOS

MACIEL VARGAS OSWALDO DANIEL

GARCÍA SALDIVAR HUGO GABRIEL

Actividad en equipos #03

Análisis de Clustering con TMAP en base de datos

I. Introducción

Explicación sobre la importancia de la reducción de dimensionalidad

En la época actual de la tecnología los profesionales del area se enfrentan a con conjuntos de datos que son complejos, enormes y en ocasiones difíciles de interpretar con sistemas convencionales. Los datos en la actualidad llevan consigo cientos o miles de variables que pueden describir una enorme cantidad de funciones u objetivos como lo son una imagen, un perfil de usuario, un algoritmo, entre otros. A pesar de su potencial valor al contener múltiples variables en un solo grupo dato, esto plantea un reto crucial llamado "la maldición de la dimensionalidad". Este problema se basa en que la cantidad de datos necesarios para generalizar correctamente aumenta exponencialmente al añadir dimensiones (variables) a un modelo, haciendo que los datos se vuelvan dispersos y dificultando el análisis, la búsqueda de patrones y la validación de modelos de aprendizaje automático.

La dimensionalidad de un conjunto de datos se refiere simplemente al número de características o variables que se utilizan para describir cada observación. Una imagen de 28x28 píxeles, como las del dataset Fashion MNIST, tiene una dimensionalidad de 784 (una por cada píxel). Si bien más dimensiones pueden significar más información, también traen consigo problemas significativos:

- Dispersión de Datos: A medida que aumenta el número de dimensiones, el espacio se vuelve vasto y los puntos de datos se alejan cada vez más entre sí, volviéndose "solitarios". Esto hace que los algoritmos que dependen de la densidad, como el clustering, pierdan efectividad.
- Pérdida de Significado en las Distancias: En espacios de muy alta dimensión, las distancias entre los puntos tienden a volverse uniformes. Es decir, la distancia al vecino más cercano y al más lejano se vuelve casi la misma, haciendo que conceptos como "proximidad" o "similitud" pierdan su significado.
- Costo Computacional: El tiempo de procesamiento y la memoria requerida por los algoritmos de aprendizaje automático a menudo crecen exponencialmente con el número de dimensiones, haciendo que el análisis sea lento o inviable.

Por ello y para poder tener un mejor control sobre los datos con los que se estan trabajando es plantea utilizar métodos de reducción de dimensionalidad. Esta es técnica de procesamiento de datos que disminuye el número de características (dimensiones) en un conjunto de datos,

transformándolos a un espacio de menor dimensión sin perder información significativa ayudando a optimizar y facilitar el uso y evaluación de miles o millones de datos.

Para esta práctica estaremos utilizando TMAP que es una de las herramientas para visualizar y entender el análisis de clústeres de una forma sencilla e intuitiva.

El uso de TMAP en análisis de clústers

TMAP en comparación con técnicas clásicas como PCA (Análisis de Componentes Principales) permite un diseño especializado en la visualizaciones, entendimiento y exploraciones de datos de alta dimensión. TMAP se encarga de posicionar un grupo de puntos similares entre si entre un espacio cercano, pero además con esta lógica construye un grafo que revela todas las relaciones entre los elementos existentes, esta función permite analizar de una mejor manera los clústeres abordar y actuar conforme a la estructura de los datos.

El proceso de TMAP se puede resumir en los siguientes pasos:

1. Indexación Rápida (LSH): TMAP usa una técnica llamada Locality Sensitive Hashing (LSH) para encontrar rápidamente "vecinos aproximados" de cada punto.
2. Construcción del Grafo: Con la información de los vecinos, TMAP construye un grafo, una red donde los nodos son nuestros datos (las imágenes) y las aristas conectan los nodos que son similares entre sí.
3. Cálculo del Árbol de Expansión Mínima (MST): El algoritmo encuentra el "esqueleto" de ese grafo, un subgrafo que conecta todos los puntos sin formar ciclos y con el menor "costo" posible. Este árbol es el que genera las conexiones, ramas y hojas que son tan informativas visualmente.
4. Visualización 2D: Finalmente, TMAP utiliza un algoritmo de layout de grafos para dibujar esta estructura de árbol/grafo en un espacio 2D, creando el mapa final.

De esta forma TMAP nos permite visualizar los clústeres lejanos de un grupo de datos de una mejor manera y con un mejor entendimiento, además tener la información ya separada de esta manera es muy útil para aplicar sub-clusters en ella con el objetivo de encontrar relaciones específicas de cada grupo específico de datos.

Finalmente, TMAP nos permite visualizar de una manera muy agradable a la vista como es la transición entre clúster, mostrándonos como dos clústeres que en principio deberían de ser completamente diferentes tambien tienen ciertas similitudes entre ellos, pero no por ello se trata de

un mismo clúster.

II. Objetivos

En proceso…

III. Desarrollo

En proceso…

IV. Conclusión

Oswaldo Maciel:

Hugo Garcia:

V. Referencias

- Fashion MNIST. (2017, 7 diciembre). Kaggle. https://www.kaggle.com/datasets/zalando-research/fashionmnist
- Probst, D. (s. f.). tmap - Visualize big high-dimensional data. https://tmap.gdb.tools/