# INTRODUCCIÓN A LA CIENCIA DE DATOS

Catedrático: Dr. Irvin Hussein López Nava

**Tarea:** Reporte de lectura

Alumno: Jorge de Jesús Cazares Duran

## REDUCCIÓN DE DIMENSIONALIDAD

Los datos generados cotidianamente suelen tener una alta dimensionalidad, es decir, muchas características o variables, lo que dificulta su análisis y uso. La reducción de dimensionalidad se define por Van Der Maaten, Postma y Van Den Herik (2009) como la reducción de un conjunto de variables aleatorias a un conjunto "principal" de variables. La necesidad de esta reducción surge debido a los efectos negativos de la alta dimensionalidad en el rendimiento de los algoritmos. Aunque los datos se encuentran en un espacio de muchas dimensiones, solo unas pocas dimensiones importantes los describen adecuadamente. La reducción de dimensionalidad busca identificar estas dimensiones clave mientras preserva la estructura original de los datos. Las técnicas de reducción de dimensionalidad se dividen en dos categorías: convexas, que son más sencillas al no tener problemas de óptimos locales y no convexas, que pueden ser más complejas debido a los óptimos locales.

Las técnicas convexas de reducción de dimensionalidad incluyen métodos espectrales y espectrales dispersos. El PCA (Análisis de Componentes Principales) destaca por encontrar los ejes que capturan la mayor varianza en los datos y proyectarlos en un espacio reducido, manteniendo la mayor cantidad de información posible. El PCA con Kernel extiende el PCA para capturar relaciones no lineales mediante una transformación previa. Entre las técnicas espectrales dispersas, los Mapas Espectrales de Laplace construyen un gráfico basado en distancias entre puntos cercanos y utilizan los vectores del Laplaciano para preservar la estructura local en la proyección reducida. En cuanto a las técnicas no convexas, el mapeo de Sammon minimiza la distorsión de distancias entre puntos al proyectarlos en dimensiones reducidas, preservando las relaciones locales. El autocodificador multicapa utiliza redes neuronales para aprender a codificar y decodificar datos de alta dimensión, capturando estructuras no lineales complejas. El LLC (Coordinación Lineal Local) representa cada punto como una combinación lineal de sus vecinos cercanos, manteniendo las relaciones locales y el gráfico de variedades (Manifold Charting) intenta preservar la estructura de los datos en el espacio reducido usando gráficos múltiples que se combinan para ofrecer una representación coherente.

Los autores destacan que las técnicas de reducción de dimensionalidad convexas ofrecen estabilidad y soluciones globales, y las técnicas no lineales, aunque más complejas, suelen tener dificultades para generalizar con datos fuera de la muestra. Cada técnica tiene características que afectan la preservación de la estructura de los datos y su eficiencia, siendo crucial su capacidad de generalización a nuevos datos para aplicaciones prácticas. El artículo concluye que las convexas son consistentes y eficientes, mientras que las no convexas son más flexibles, pero menos estables y fiables fuera del conjunto de datos de entrenamiento. La configuración del experimento y la evaluación de técnicas con bases de datos artificiales y naturales son esenciales para este estudio. La configuración estandarizada permite una comparación consistente, mientras que las bases de datos artificiales proporcionan un entorno controlado para evaluar el rendimiento, además, las naturales reflejan condiciones del mundo real.

El articulo concluye que la selección de técnicas de reducción debe basarse en una evaluación de sus capacidades para preservar la estructura de los datos, su eficiencia computacional y su capacidad de generalización en escenarios reales. Se deben considerar tanto las pruebas en bases de datos artificiales como naturales para una comprensión de fortalezas y limitaciones. Los autores anticipan el desarrollo de nuevas técnicas que se basen en funciones objetivo no convexas y no dependan de gráficos de proximidad. La optimización de las nuevas técnicas debe ser factible tanto computacional como numéricamente para ser útiles en la práctica.

**Reseña**

El artículo presenta una revisión comparativa de técnicas de reducción de dimensionalidad, destacando la importancia en la mejora de la interpretabilidad y eficiencia en el análisis de datos de alta dimensionalidad. Se destaca la necesidad de la configuración experimental, mediante el uso tanto bases de datos artificiales como naturales para evaluar el desempeño de las técnicas. Esta combinación asegura una evaluación completa de cada técnica en diferentes contextos. La comparación entre métodos proporciona una visión detallada de sus fortalezas, como la capacidad para preservar estructuras locales y globales en los datos y sus limitaciones, como problemas de sobreajuste o dificultad para manejar relaciones no lineales. Este enfoque comparativo permite visualizar las aplicaciones actuales y futuras de las técnicas convexas y no convexas, también establece una base sólida para el desarrollo de nuevas metodologías. Las técnicas que se empleen deben superar las limitaciones observadas por los autores, entre ellas la dependencia de gráficos de proximidad o la presencia de soluciones óptimas convencionales y mejorar el rendimiento de los modelos para asegurar una mayor precisión y consistencia en una amplia gama de aplicaciones prácticas.

Referencia:

Van Der Maaten, L.J.P., Postma, E.O. and Van Den Herik, H.J. (2009) Dimensionality Reduction: A Comparative Review. Journal of Machine Learning Research, 10, 1-41.