Universidad Nacional del Altiplano
Facultad de Ingeniería Estadística e Informática
Docente: Fred Torres Cruz
Autor: Jhoan Jeremy Chavez Lima

Trabajo Encargado - N° 003

¿Qué factores consideras al elegir una técnica de escalado en un conjunto de datos con variables muy heterogéneas?

Desarrollo

Respuesta

Al elegir una técnica de escalado en un conjunto de datos con variables muy heterogéneas, es fundamental considerar la naturaleza de las variables, la distribución de los datos y el impacto que tendrá el preprocesamiento en los algoritmos de aprendizaje automático o estadísticos. La heterogeneidad puede darse por diferencias en magnitudes (por ejemplo, una variable en miles y otra en décimas), en unidades de medida o incluso en la forma de las distribuciones. Sin un adecuado ajuste de escala, los algoritmos basados en distancias o gradientes suelen verse sesgados por las variables de mayor rango numérico.

Diversos estudios muestran que la elección del método de normalización o estandarización debe adaptarse al contexto. Por ejemplo, en problemas de **yacimientos de petróleo carbonatado**, se encontró que aplicar transformaciones como Box-Cox mejoraba de manera significativa el desempeño de modelos de predicción de permeabilidad frente a otras técnicas de escalado [1]. Esto muestra que, cuando los datos presentan distribuciones sesgadas o no normales, métodos de transformación estadística son más adecuados que una simple estandarización Z-score.

Asimismo, en contextos donde se trabaja con **datos difusos heterogéneos**, como en la construcción de indicadores multicriterio, se ha propuesto la distancia de reescalado difuso triangular (dTR), que incorpora el reescalado directamente en el cálculo de la métrica, garantizando comparabilidad entre atributos sin necesidad de un paso previo de normalización [2]. Esto refleja que en algunos dominios es posible integrar la normalización dentro del propio método de análisis.

Otro ejemplo proviene del **aprendizaje federado**, donde los datos distribuidos suelen tener diferentes patrones y escalas. Para reducir los efectos de la heterogeneidad en el entrenamiento, Vieira y Campos (2025) desarrollaron el método FedWS, que normaliza los pesos en redes neuronales convolucionales y consigue mayor estabilidad y precisión [3]. Esto evidencia que el escalado no solo es relevante como preprocesamiento, sino que también puede incorporarse dentro de los modelos para reducir la divergencia causada por datos

desbalanceados.

En el campo del **scRNA-seq**, la heterogeneidad de los datos se maneja con modelos que integran normalización y teoría de la información. El enfoque scInfoMaxVAE, por ejemplo, logra preservar mejor la estructura de los datos al considerar la inflación cero y aplicar técnicas de reducción de dimensionalidad más robustas [4].

Finalmente, en aplicaciones industriales basadas en **gráficos de conocimiento**, la heterogeneidad multimodal (texto, imágenes y señales) requiere integrar normalización en la arquitectura del modelo. Zhu et al. (2026) mostraron que la normalización espectral aplicada en un marco adversarial mejora la detección de errores en este tipo de datos complejos [5].

En resumen, la elección de una técnica de escalado no puede hacerse de forma genérica: depende del tipo de variables, de la distribución de los datos y de la naturaleza del problema. En algunos casos, como en los **datos petrofísicos**, transformaciones como Box-Cox resultan más efectivas; en otros, como en **sistemas difusos o de decisión multicriterio**, se diseñan métricas específicas con reescalado integrado; en **aprendizaje federado**, la normalización interna de parámetros permite controlar la heterogeneidad de manera más robusta; y en **aplicaciones biológicas o industriales**, los modelos modernos integran la normalización en su núcleo para garantizar precisión y consistencia.

Referencias

[1] Al-Mudhafar, W. J., Hasan, A. A., Abbas, M. A., & Madera, D. A. (2025). Aprendizaje automático con optimización de hiperparámetros aplicado en el modelado de permeabilidad con soporte de facies en yacimientos de petróleo carbonatado. *Scientific Reports.* Nature Portfolio. https://doi.org/10.1038/s41598-025-95490-0

[2] Soria, E., Valls, A., & Hernández-Lara, A. B. (2026). Distancia de reescalado difuso triangular. *Lecture Notes in Computer Science.* Springer. https://doi.org/10.1007/978-3-032-00891-6_10

[3] Vieira, F., & Campos, C. A. V. (2025). Reducción del impacto de la divergencia de peso mediante la normalización del aprendizaje local en el aprendizaje federado para distribuciones de datos heterogéneas. *Future Generation Computer Systems.* Elsevier. https://doi.org/10.1016/j.futuro.2025.107881

[4] Duy, P. N., Thao, N. P., Le, T., & Van Trinh, L. (2026). Aprovechamiento de la información mutua en los autocodificadores variacionales para mejorar la reducción de la dimensionalidad de los datos de secuenciación de ARN de células individuales: el enfoque scInfoMaxVAE. *Computational Biology and Chemistry.* Elsevier. https://doi.org/10.1016/j.compbiolchem.2025.108637

[5] Zhu, X., Li, Y., Guo, L., Huang, B., & Colmillo, Z. (2026). Detección de errores basados en aprendizaje adversarial para gráficos de conocimiento industrial. *Lecture Notes in Computer Science.* Springer. https://doi.org/10.1007/978-981-96-8892-0_43