MBA05

Clustered Interpretability: Bridging Global and Local Explanations in AI

Matt Baucum[1], Meysam Rabiee[2]

[1]Colorado State University, Fort Collins, Colorado, USA. [2]University of Colorado Denver, Denver, Colorado, USA

Abstract

This paper introduces the novel concept of clustered interpretability to enhance the transparency and fairness of AI models, especially in high-stakes domains like healthcare, finance, and justice. Traditional interpretability techniques often miss critical subgroup differences by focusing on global or local explanations. To address this, we propose a framework that partitions datasets into meaningful clusters, introducing the notions of cluster fit (alignment with SHAP values) and parsimony (simplicity of feature-outcome relationships). We also present a multi-objective framework for optimizing both cluster fit and parsimony, along with a novel solution filtering algorithm that ensures fairness across clusters. The framework's effectiveness is demonstrated using a healthcare dataset on substance use disorder (SUD) treatment, showing that prioritizing parsimony yields simpler interpretability patterns and uncovers significant clinical insights and social implications not highlighted by traditional methods. This approach bridges the gap between global and local interpretability, offering a balanced and scalable solution for identifying nuanced patterns within and across subpopulations.

Conference Track

MIS and Business Analytics