Causal Data Augmentation for Cross-Domain Generalization in Tabular Foundation Models

Assignee Research

June 11, 2026

Abstract

Fine-tuning tabular foundation models (TFMs) under data scarcity is challenging, as early stopping on even scarcer validation data often fails to capture true generalization performance. We propose CausalMixFT, a method that enhances fine-tuning robustness and downstream performance by generating structurally consistent synthetic samples using Structural Causal Models (SCMs) fitted on the target dataset. This approach augments limited real data with causally informed synthetic examples, preserving feature dependencies while expanding training diversity. Evaluated across 33 classification datas

1 Introduction

This paper examines: Causal Data Augmentation for Robust Fine-Tuning of Tabular Foundation Models. Research question: Does causal data augmentation enhance the cross-domain generalization of tabular foundation models compared to standard mixture-of-experts fine-tuning strategies?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

11 papers retrieved. 13 claims extracted; 11 independently verified. Quality review score: 8.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
CausalMixFT achieves the highest median improvement of (+0.12$\pm$0.63) over the pre-trained model on 33 classification data	✓	0.27
Default fine-tuning has a variability of $\pm$0.98, while CausalMixFT has a variability of $\pm$0.63.	×	0.15
CausalMixFT ranks first overall in average ranks across datasets, followed by the default fine-tuning baseline.	✓	0.24
Purely synthetic generators, including CTGAN, SCM, TabEBM, TableAugment, and MixedModel, show negative median improvemen	✓	0.22
Early stopping based on limited validation data leads to significant validation set overfitting depending on the fine-tu	✓	0.26
The normalization strategy suggested by Gorishniy et al. [12] is used to compare the performance across different data g	×	0.05
CausalMixFT extends the fine-tuning framework of Bhler et al. [5] by mixing real and causally grounded synthetic sample	✓	0.28
SCM-Based Synthetic Augmentation (CausalMixFT) uses SCMs fitted to the target dataset to generate synthetic data.	✓	0.18
SCMs explicitly encode causal dependencies among features through a directed acyclic graph (DAG) and a set of structural	✓	0.27
The PC and FCI algorithms are used to estimate the structural relations between the features.	✓	0.17
DoWhy's SCM framework with additive noise models is used to sample and fit DAGs.	✓	0.16
Numerical features are modeled with regressors, and categorical features with classifiers in the SCM framework.	✓	0.19
Synthetic samples are generated by sampling exogenous noise and propagating it through the fitted SCM.	✓	0.22

References

- http://arxiv.org/abs/2603.10254v1
- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2512.03307v1