Enhancing Adversarial Training Efficiency for Synthetic Tabular Data via Curriculum Learning

Assignee Research

June 11, 2026

Abstract

Machine learning (ML) on tabular data is ubiquitous, yet obtaining abundant high-quality tabular data for model training remains a significant obstacle. Numerous works have focused on tabular data augmentation (TDA) to enhance the original table with additional data, thereby improving downstream ML tasks. Recently, there has been a growing interest in leveraging the capabilities of generative AI for TDA. Therefore, we believe it is time to provide a comprehensive review of the progress and future prospects of TDA, with a particular emphasis on the trending generative AI. Specifically, we prese

1 Introduction

This paper examines: Tabular Data Augmentation for Machine Learning: Progress and Prospects of Embracing Generative AI. Research question: Can the efficiency of adversarial training for synthetic data generation be improved through curriculum learning, as measured by convergence speed and final model accuracy on standard tabular benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

10 papers retrieved. 15 claims extracted; 12 independently verified. Quality review score: 7.3/10.

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
Tabular data is heterogeneous, typically containing both dense numerical and sparse categorical attributes.	✓	0.21
Tabular data exhibits row and column permutation invariance.	×	0.12
Tabular data has a hierarchical organization where cells belong to rows and rows belong to tables.	✓	0.20
Many Tabular Data Augmentation (TDA) tasks involve table pools encompassing millions of tables.	✓	0.23
Table pools often contain tables with inconsistent attribute naming and value formatting.	✓	0.21
Table pools are dynamic and change over time.	×	0.10
TDA methods are broadly categorized into retrieval-based approaches and generation-based approaches.	✓	0.21
Retrieval-based TDA approaches involve retrieving data from table pools.	✓	0.24
Generation-based TDA approaches involve generating new content.	✓	0.24
Generative methods for TDA include statistical approaches such as MICE.	✓	0.16
Generative methods for TDA include deep generative models like diffusion models.	✓	0.17
External data sources used in TDA are referred to as table pools, which can include databases and Web tables.	✓	0.16
In the example scenario, the original training set (TO) yields sub-optimal results due to insufficient data and missing	✓	0.21
Table annotation, such as recovering missing column types, is a preparation step that enhances the TDA process.	×	0.13
Augmentation can be achieved by integrating rows from external data sources (retrieval-based) or by synthesizing new dat	✓	0.15

References

- http://arxiv.org/abs/2407.21523v1
- http://arxiv.org/abs/2402.01204v4
- http://arxiv.org/abs/2504.20900v1