What is the impact of conditioning tabular diffusion models on class-specific statistics on minority class rec

Assignee Research

June 10, 2026

Abstract

Class imbalance is a common problem in supervised learning and impedes the predictive performance of classification models. Popular countermeasures include oversampling the minority class. Standard methods like SMOTE rely on finding nearest neighbours and linear interpolations which are problematic in case of high-dimensional, complex data distributions. Generative Adversarial Networks (GANs) have been proposed as an alternative method for generating artificial minority examples as they can model complex distributions. However, prior research on GAN-based oversampling does not incorporate rece

1 Introduction

This paper examines: Conditional Wasserstein GAN-based Oversampling of Tabular Data for Imbalanced Learning. Research question: What is the impact of conditioning tabular diffusion models on class-specific statistics on minority class recall in high-dimensional tabular datasets compared to SMOTE and GAN-based oversampling?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

13 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Random oversampling generates additional minority class samples by drawing with replacement from the original minority c	×	0.10
Random oversampling can lead to problematic overfitting because identical samples appear multiple times in the training	×	0.09
SMOTE generates new minority class samples by creating a linear combination between a random minority case and a neighbo	×	0.06
In SMOTE, the interpolation factor epsilon is drawn from a uniform distribution U[0, 1].	×	0.02
SMOTE assumes that all columns in the dataset are continuous.	×	0.02
SMOTENC generates continuous variables using SMOTE and sets nominal features to the most-frequent value in the k-nearest	×	0.06
SMOTENC modifies the Euclidean distance calculation by adding the squared median standard deviation of continuous featur	×	0.02
Borderline-SMOTE (B-SMOTE) identifies minority class samples as being in danger of misclassification if the share of maj	×	0.07
Borderline-SMOTE excludes minority samples considered 'safe' or 'noisy' from the generation of synthetic samples.	×	0.05
ADASYN selects minority class samples for generation proportionally to the number of majority class cases in their k-nea	×	0.10
Generative Adversarial Networks (GANs) consist of a generator model tasked with generating indistinguishable data and a	×	0.08
In a Vanilla GAN, the generator receives a vector of latent noise drawn from an arbitrary noise distribution as input.	×	0.02
The training objective of a Vanilla GAN is formulated as a two-player minimax-game involving the expectation of log D(x)	×	0.06
Given an optimal discriminator, the generator's objective in a Vanilla GAN is optimized when the generator's distributio	×	0.04

References

- http://arxiv.org/abs/2008.09202v1
- http://arxiv.org/abs/2412.00381v1
- http://arxiv.org/abs/2502.17119v2