Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

机器合成数据生成与评价方法

硕士研究生 王旭

2026年06月07日

问题回溯

- 问题回溯
  - 前期讲授不流畅，读PPT内容严重
  - 对算法讲解不深入，理解浮于表面
- 相关内容
  - 2024.04.06 徐泽豪《归一化流在表格数据生成中的应用》
  - 2023.08.13 徐泽豪《表格数据生成：GAN模型的演进与未来》

内容提要

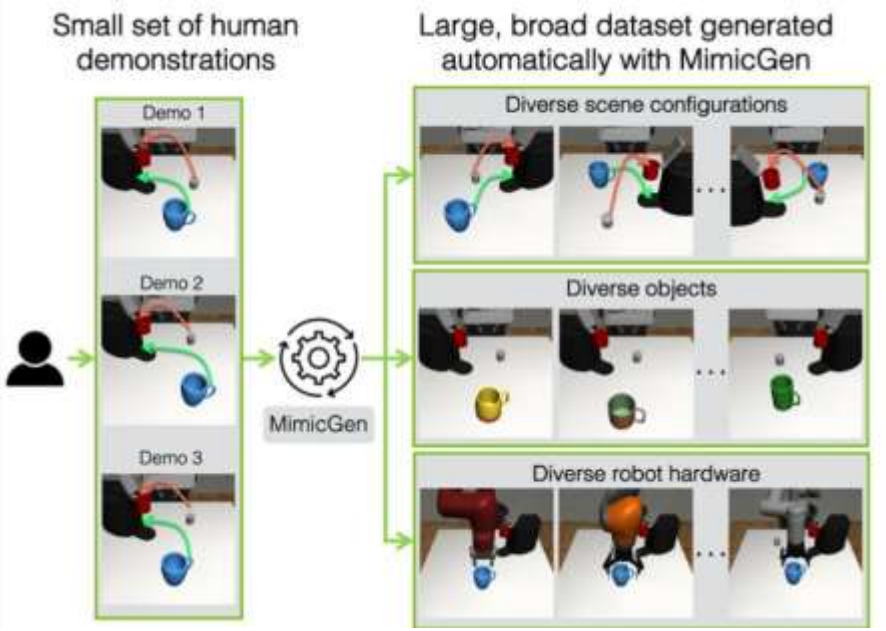
- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - TabDiff
  - How Faithful is your Synthetic Data
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 1. 了解机器合成数据领域的核心概念、主流方法与评估体系
  - 2. 理解主流生成方法的原理与适用场景
  - 3. 了解现有方法的缺陷以及未来发展方向

内涵解析与研究目标

- 内涵解析
  - 合成数据:通过算法生成的、保留原始数据统计特征但不包含真实个人信息的人工数据
  - 三大关键要求:
    - 数据质量: 统计分布、相关性等方面与真实数据一致
    - 隐私保护: 防止合成数据泄露原始数据的敏感信息
    - 泛化能力: 合成数据训练的模型能泛化到真实数据
- 研究目标
  - 系统梳理从统计模型到深度学习的数据合成方法体系
  - 构建分布一致性-多样性-任务相关性三维质量评估框架
  - 分析评估指标与下游模型性能之间的量化关系



研究背景与意义

- 研究背景
  - 数据稀缺与成本高昂
    - 大规模数据采集需高昂成本; 罕见病研究/极端环境监测等场景真实数据极度稀缺
    - 自动驾驶需模拟复杂路况/极端天气, 真实采集危险且不完整
  - 隐私法规限制
    - GDPR(欧盟) / PDPA(中国)法规严格，医疗/金融数据含大量敏感信息，数据泄露可导致个人权益、企业信誉乃至国家安全的损失
  - 数据偏见与不平衡
    - 类别不平衡使分类器难以学习少数类特征模式, 削弱泛化能力
    - 训练数据中的性别/种族偏见导致模型产生歧视性输出
- 研究意义
  - 合成数据技术是平衡"数据开放共享"与"安全合规"的关键技术路径,在医疗、金融、自动驾驶、NLP等领域均具有重大应用价值

研究历史与现状

1993 Rubin等人首次系统提出合成数据概念与多重插补(MI)框架，为数据合成奠定理论基础。该方法通过参数化统计模型估计数据分布并生成模拟数据，在社会科学和医疗研究领域得到广泛应用

2006 Dwork等人引入差分隐私（Differential Privacy, DP）的形式化框架，为隐私保护提供了严格的数学定义。该框架后来与合成数据生成深度融合

2014 Goodfellow等人提出生成对抗网络(GAN)：利用生成器与判别器的对抗博弈训练模型。结合差分隐私等技术为隐私保护数据共享提供了新路径

2019 Xu等人提出CTGAN，基于 WGAN-GP 架构，是表格数据生成领域最具影响力的模型，判别器同时处理多个样本，进一步防止模式坍塌

2019 Xu等人同年发布变分自编码器TVAE，基于与CTGAN相同的VGM预处理架构，具有训练更稳定，对高维数据扩展性好，跨数据集鲁棒（超参数不敏感）等优点

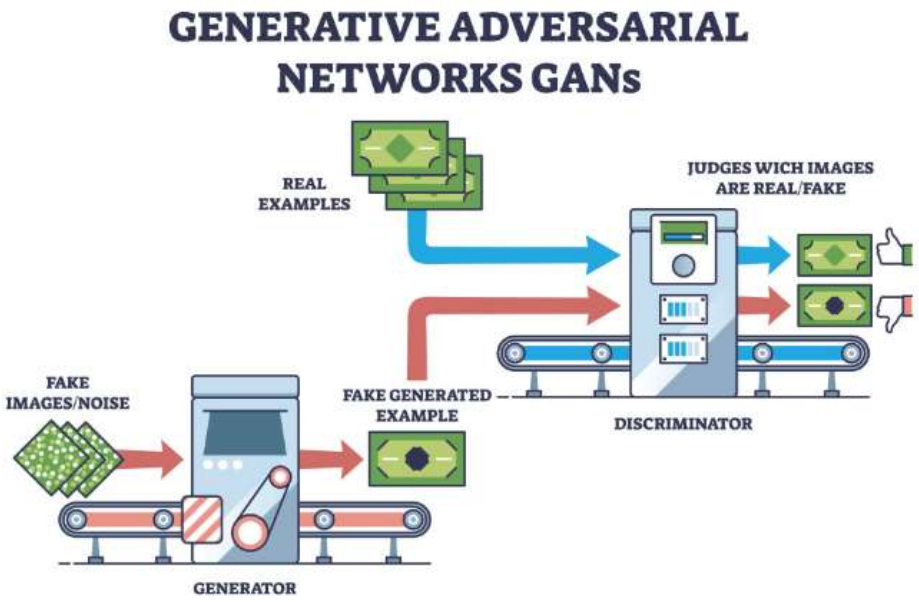
2022 Kotelnikov提出TabDDPM：扩散模型快速成为表格数据生成的新的SOTA范式，在保真度和分布覆盖上持续超越GAN和VAE

2022 Alaa等人提出alpha-Precision(保真度)/beta-Recall(多样性)/Authenticity(泛化性)三维样本级评估框架，基于最小体积集与超球面嵌入理论，首次实现生成模型质量的可解耦评估与样本级审计

2025 Shi等人利用预训练语言模型Transformer架构生成合成数据成为一个快速发展的新方向，即LLM+扩散模型混合方法，自然地处理元数据

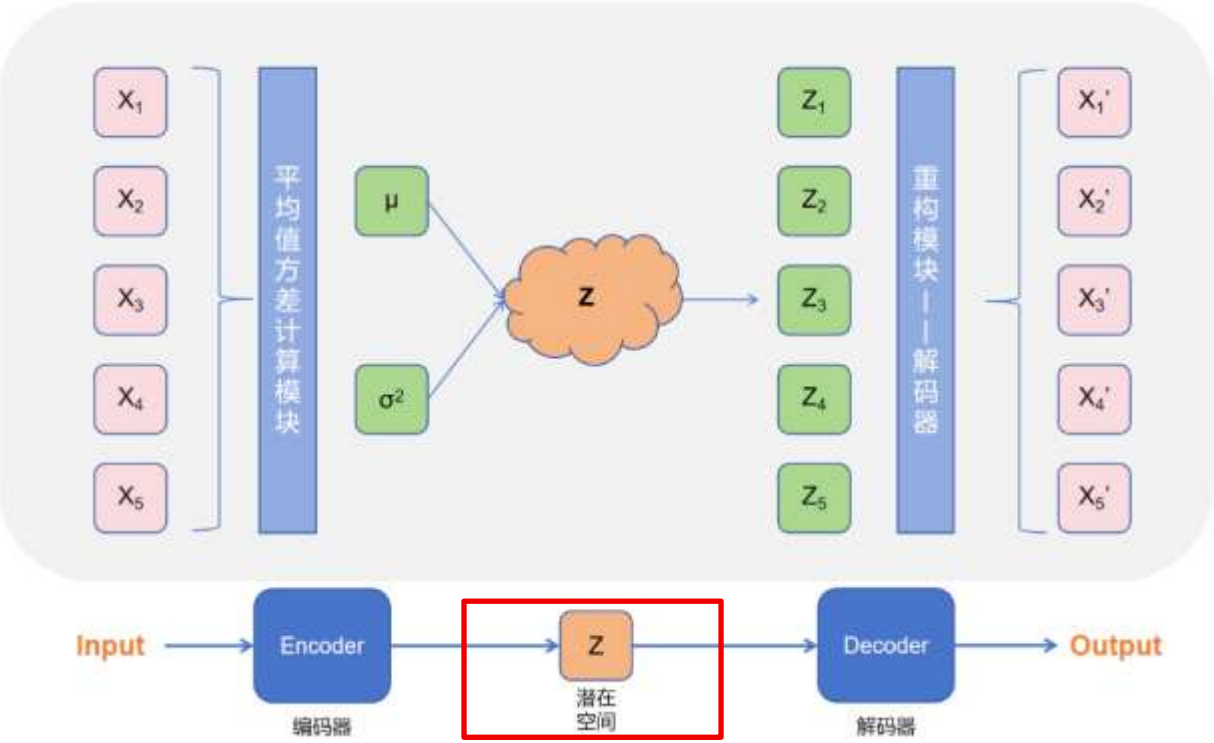
知识基础

- 生成对抗网络 (GAN)
  - 生成网络负责生成模拟数据（大部分情况下是图像），最终目的是“骗过”判别器
  - 判别网络负责判断输入的数据是真实的还是生成的，目的是找出生成器做的“假数据”
  - 训练过程：交替优化生成器和判别器直至纳什均衡
- 主要优势
  - 多模态适用，无需显式密度建模，变体丰富
- 主要缺陷
  - 训练不稳定，模式崩溃，可解释性弱



知识基础

- 变分自编码器 (VAE)
  - VAE是一种生成式模型，通过学习数据的潜在分布实现数据生成
  - 在自编码器AE的基础上引入了概率分布的概念，使潜在表示从固定值变为概率分布，从而增强生成能力和鲁棒性
- 主要优势
  - 训练稳定，计算高效，潜在空间可解释
- 主要缺陷
  - 生成的样本可能较模糊，质量较低

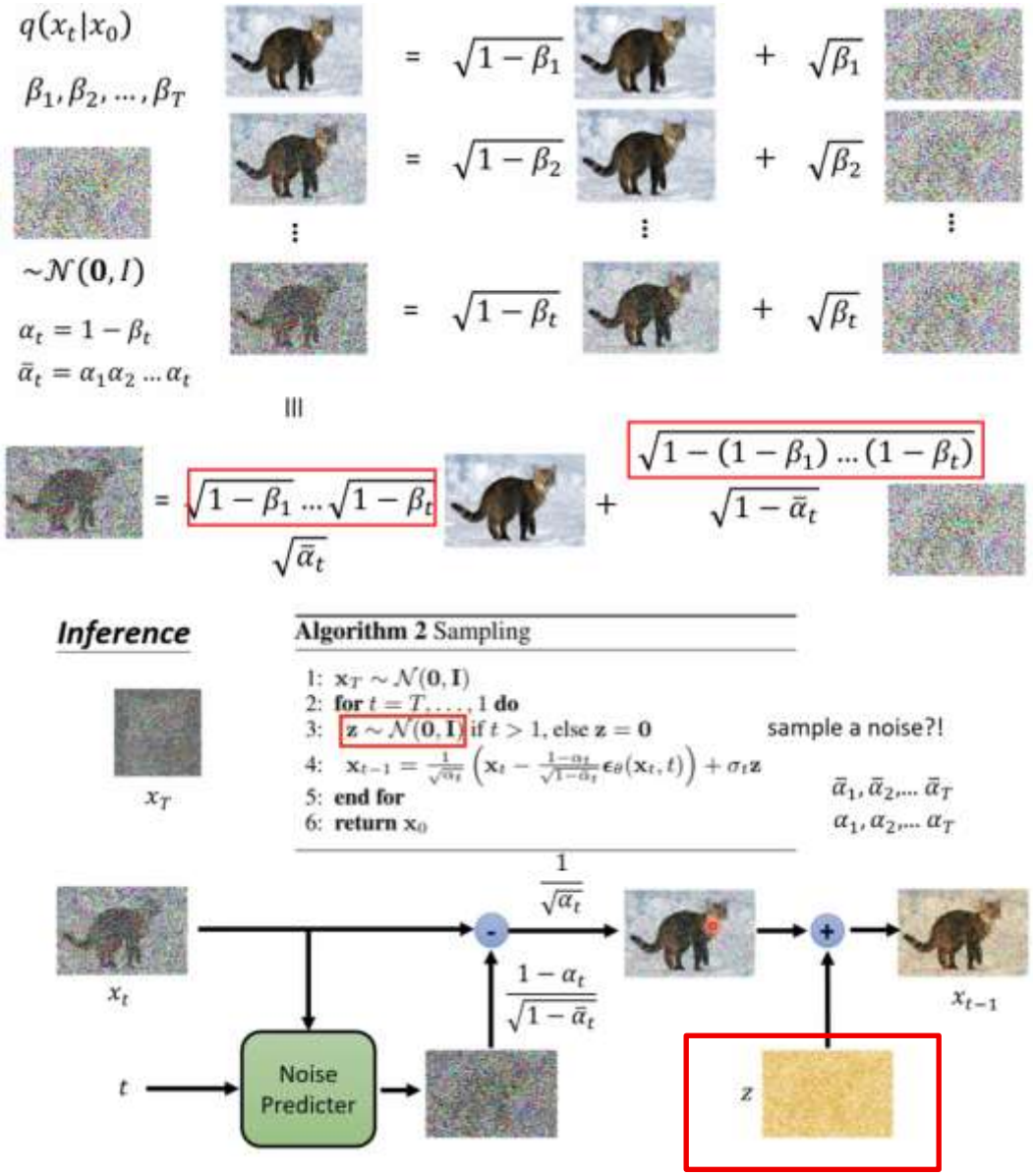


输入数据的编码信息->输入数据的概率分布

知识基础

- 扩散模型(Diffusion Models)
  - 前向过程加噪
    - 将数据逐步加噪至接近高斯分布，干净图与噪声加权生成带噪声图
  - 逆向过程去噪
    - 从纯噪声图逐步去噪得到清晰图的过程
  - 主要优势
    - 质量最高、训练稳定、多模态通用
  - 缺陷
    - 推理慢、计算成本高

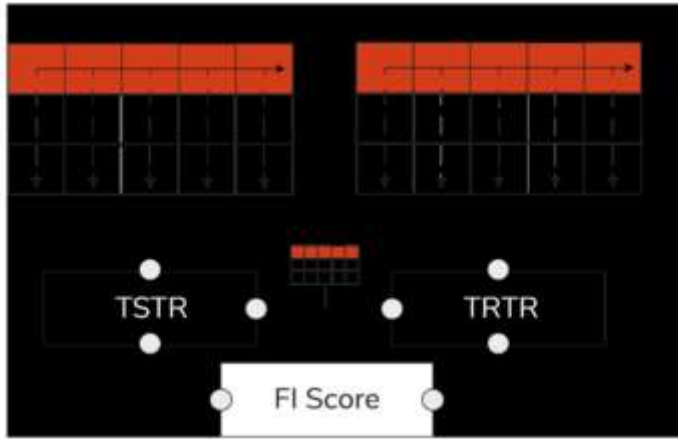
加噪 -> 去噪，无需对抗训练



- 合成数据质量评估 -- 三大核心维度
  - 分布一致性，衡量合成数据与原始数据在统计层面的相似程度
    - 常用度量：KS检验、Wasserstein距离、Hellinger距离
  - 多样性，衡量合成数据是否覆盖了真实数据的所有模式与变化范围
    - 实际意义：高多样性能够帮助模型学习到更完整的决策边界，降低对特定类别的过拟合风险
  - 任务相关性，衡量合成数据在下游机器学习任务中对真实数据的替代能力
    - TSTR范式(Train on Synthetic, Test on Real)：分类F1/AUC | 回归RMSE

$$- AUC = \sum_{i=2}^{m} \frac{(x_i - x_{i-1}) * (y_i - y_{i-1})}{2}$$

$$- RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



TabDiff: A Mixed-type Diffusion Model for Tabular Data Generation

算法原理 TIPO

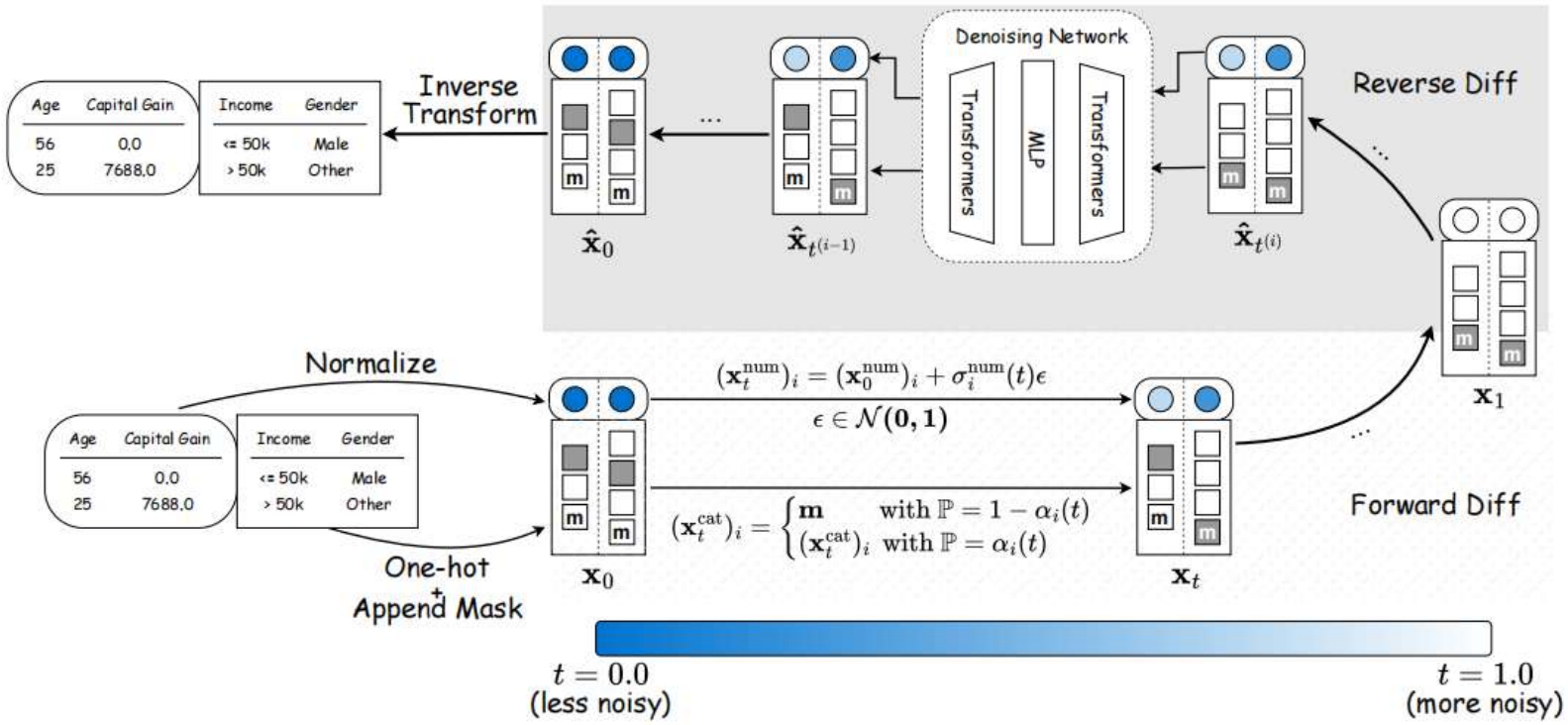
T	目标	高质量混合类型表格生成
I	输入	特征包含数值和类型的表格数据集
P	处理	1. 联合前向扩散(高斯SDE+掩码扩散) 2. 端到端训练(θ,ρ,k联合优化) 3. 随机采样(前向扰动+反向去噪)
O	输出	高质量合成表格数据集

P	问题	表格特征高度异质，每列是独立的数据模态，现有方法受限于编码开销或离散化精度
C	条件	在连续时间扩散框架下统一处理异构类型 + 自适应分配模型容量
D	难点	1.噪声调度灵活性与稳定性平衡 2.类别特征在掩码扩散中"一旦解码不可修正"——早期错误导致列间关联信息永久丢失
L	水平	ICLR 2025（CCF-A）

算法原理 TabDiff 联合扩散框架

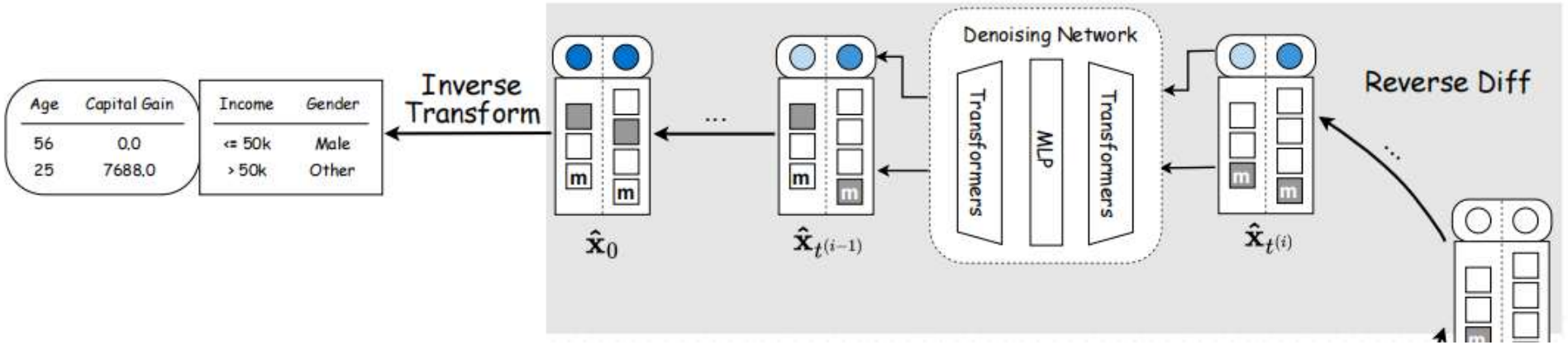
- 表格数据集T
  - $T = \{x\} = \{[x^{num}, x^{cat}]\}$
    - $x^{num} \in R^{M^{num}}$
    - $x^{cat} \in \{0,1\}^{(C_j+1)}$

- TabDiff创新
  - 联合连续时间扩散，数值和类别在同一时间进行
  - 特征级可学习噪声调度
  - 随机采样器
  - 无分类器引导



算法原理

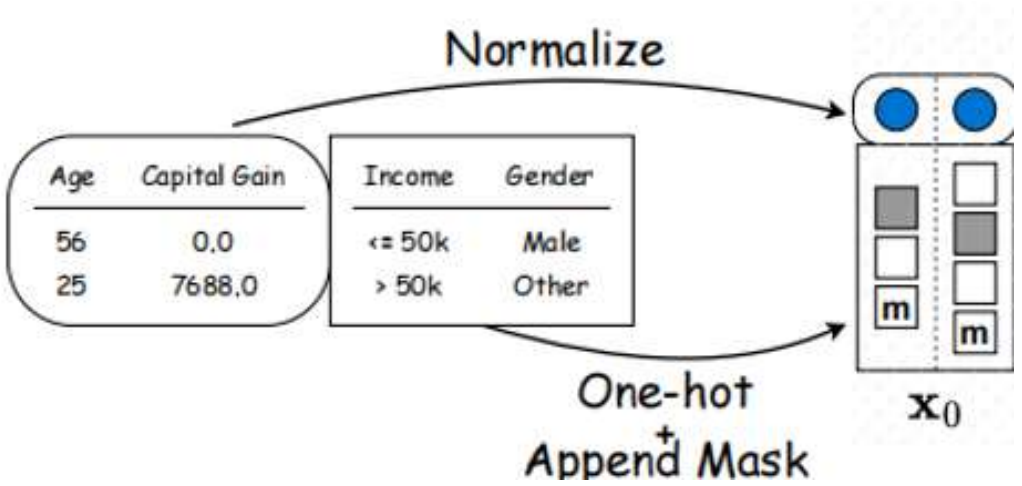
- 去噪网络架构
  - 列级投影
    - 每列独立通过 Linear 层投影到 d=4 维向量，保证所有列初始时被同等对待
  - Transformer 编码器（2层），时间条件 MLP（5层），Transformer 解码器（2层）
  - 输出投影
    - 恢复各列原始维度：数值→1 维标量（预测噪声ε），类别→softmax（预测 X_0 概率）

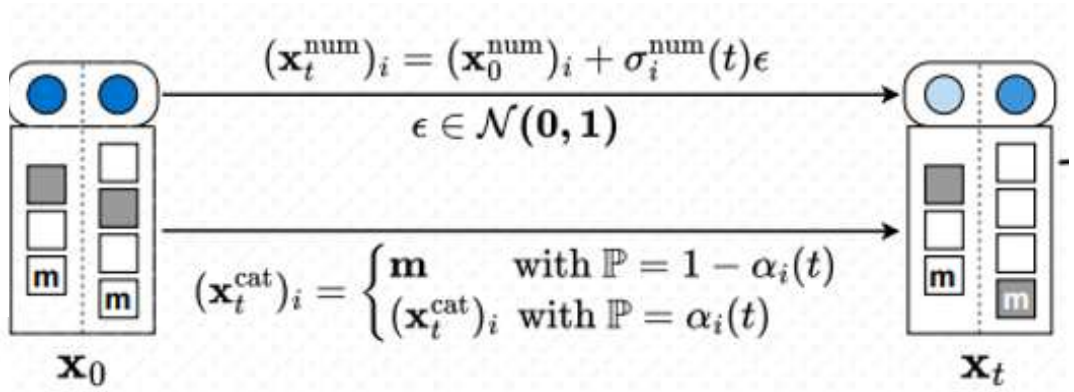


d=4 的选择足够表达单列信息，又不至于引入过多参数

算法原理 混合类型联合扩散

- 联合前向扩散
  - $q(X_t|X_0) = q(X_t^{num}|X_0^{num}, \sigma_0^{num}(t)) \cdot q(X_t^{cat}|X_0^{cat}, \sigma^{cat}(t))$
  - 两个模态使用不同的噪声调度 σ^{num} 和 σ^{cat}，但在同一连续时间 t下联合扰动
- 数值特征 — Gaussian SDE 扩散前向公式
  - $X_t^{num} = X_0^{num} + \sigma_0^{num}(t)\epsilon,\ \epsilon \sim N(0, I_{M^{num}})$
  - 为数值特征逐步添加高斯噪声
- 类别特征 — 掩码（吸收态）扩散
  - $q(X_t|X_0) = cat(X_t; \alpha_t X_0 + (1 - \alpha_t m))$
  - 为类别特征添加掩码





算法原理 特征级可学习噪声调度

- 设计动机
  - 表格数据的不同列具有完全不同的边际分布（"年龄"vs"收入"），不同于图像或文本，需要每个特征有独立的噪声衰减曲线，让模型自适应分配容量
- 数值特征ρ —Power-Mean 调度
  - $\sigma_{\rho_i}^{\text{num}}(t) = \left(\sigma_{\min}^{\frac{1}{\rho_i}} + t(\sigma_{\max}^{\frac{1}{\rho_i}} - \sigma_{\min}^{\frac{1}{\rho_i}})\right)^{\rho_i}$
  - 用ρ参数控制，让模型自己学
- 类别特征k — Log-Linear 调度
  - $\alpha_{k_j}^{\text{cat}}(t) = 1 - t^{kj}$
  - $\sigma_{k_j}^{\text{cat}}(t) = -\log\left(1 - \left((1-\delta)\cdot t^{k_j} + \delta\right)\right)$
- 参数优化
  - 参数集 {θ, ρ, k} 通过反向传播同时优化

Algorithm 1 Training

repeat
  Sample $\mathbf{x}_0 \sim p_0(\mathbf{x})$
  Sample $t \sim U(0,1)$
  Sample $\boldsymbol{\epsilon}_{\text{num}} \sim \mathcal{N}(0, \mathbf{I}_{M\text{num}})$
  $\mathbf{x}_t^{\text{num}} = \mathbf{x}_0^{\text{num}} + \boldsymbol{\sigma}^{\text{num}}(t)\boldsymbol{\epsilon}_{\text{num}}$
  $\boldsymbol{\alpha}_t = \exp(-\boldsymbol{\sigma}^{\text{cat}}(t))$
  Sample $\mathbf{x}_t^{\text{cat}} \sim \mathrm{q}(\mathbf{x}_t|\mathbf{x}_0, \boldsymbol{\alpha}_t)$ Eq. (6)
  $\mathbf{x}_t = [\mathbf{x}_t^{\text{num}}, \mathbf{x}_t^{\text{cat}}]$
  Take gradient descent step on $\nabla_{\theta,\rho,k}\mathcal{L}_{\text{TABDIFF}}$
until converged

网络取自参数θ取自transformer去噪网络

算法原理 混合类型随机采样器

- 核心问题
  - 类别特征在掩码扩散中一旦被解码（unmask），后续步骤不再更新，导致早期解码错误永久保留
- 解决方案 — 随机采样器
  - 传统采样：噪声 → 去噪 → 去噪 → 去噪 … → 干净
  - 随机采样：噪声 → 加噪→去噪 → 加噪→去噪 … → 干净
  - 每步先往回扰动，错误可以被重新翻成[MASK]，重新解码

实验设计　实验设置

- 评估指标

指标	衡量目标	类别
Shape	单列边际分布（KST + TVD）	保真度
Trend	列间相关性（Pearson + Contingency）	
α-Precision	合成样本的忠实度	
β-Recall	合成样本的覆盖率	
MLE	机器学习效率 (AUC/RMSE)	下游
Imputation	缺失值填补准确率	
DCR	最近记录距离（50%最佳）	隐私

- 数据集

Dataset	# Rows	# Num	# Cat	# Max Cat	# Train	# Validation	# Test	Task
Adult	48,842	6	9	42	28,943	3,618	16,281	Classification
Default	30,000	14	11	11	24,000	3,000	3,000	Classification
Shoppers	12,330	10	8	20	9,864	1,233	1,233	Classification
Magic	19,019	10	1	2	15,215	1,902	1,902	Classification
Beijing	43,824	7	5	31	35,058	4,383	4,383	Regression
News	39,644	46	2	7	31,714	3,965	3,965	Regression
Diabetes	101,766	9	27	716	61,059	2,0353	20,354	Classification

- 基线方法
  - GAN：CTGAN（2019）
  - VAE：TVAE（2019）、GOGGLE（2021）
  - AR-LM：GReaT（2024）
  - 扩散模型
    - STaSy（2021）、CoDi（2021）、TabDDPM（2022）、TABSYN（2024）

实验结果 保真度对比

- Shape列密度误差率证明TABDIFF在各列的边际分布恢复上显著优于所有基线

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Average
CTGAN	16.84± 0.03	16.83±0.04	21.15±0.10	9.81±0.08	21.39±0.05	16.09±0.02	9.82±0.08	15.99
TVAE	14.22±0.08	10.17±0.05	24.51±0.06	8.25±0.06	19.16±0.06	16.62±0.03	18.86±0.13	15.97
GOGGLE	16.97	17.02	22.33	1.90	16.93	25.32	24.92	17.91
GReaT	12.12±0.04	19.94±0.06	14.51±0.12	16.16±0.09	8.25±0.12	OOM	OOM	14.20
STaSy	11.29±0.06	5.77±0.06	9.37±0.09	6.29±0.13	6.71±0.03	6.89±0.03	OOM	7.72
CoDi	21.38±0.06	15.77± 0.07	31.84±0.05	11.56±0.26	16.94±0.02	32.27±0.04	21.13±0.25	21.55
TabDDPM	1.75±0.03	1.57± 0.08	2.72±0.13	1.01±0.09	1.30±0.03	78.75±0.01	31.44±0.05	16.93
TabSyn [1]	0.81±0.05	**1.01±0.08**	1.44±0.07	1.03±0.14	1.26±0.05	**2.06±0.04**	1.85±0.02	1.35
TabDiff	**0.63±0.05**	1.24±0.07	**1.28±0.09**	**0.78±0.08**	**1.03±0.05**	2.35±0.03	**0.89±0.23**	**1.17**
Improv.	22.2% ↓	0.0% ↓	11.11% ↓	14.29% ↓	18.25% ↓	0% ↓	46.39% ↓	13.3% ↓

- Trend列间相关性误差率（TABDIFF 最突出的优势）

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Average
CTGAN	20.23±1.20	26.95±0.93	13.08±0.16	7.00±0.19	22.95±0.08	5.37±0.05	18.95±0.34	16.36
TVAE	14.15±0.88	19.50±0.95	18.67±0.38	5.82±0.49	18.01±0.08	6.17±0.09	32.74±0.26	16.44
GOGGLE	45.29	21.94	23.90	9.47	45.94	23.19	27.56	28.18
GReaT	17.59±0.22	70.02±0.12	45.16±0.18	10.23±0.40	59.60±0.55	OOM	OOM	44.24
STaSy	14.51±0.25	5.96±0.26	8.49±0.15	6.61±0.53	8.00±0.10	3.07±0.04	OOM	7.77
CoDi	22.49±0.08	68.41±0.05	17.78±0.11	6.53±0.25	7.07±0.15	11.10±0.01	29.21±0.12	23.21
TabDDPM	3.01±0.25	4.89±0.10	6.61±0.16	1.70±0.22	2.71±0.09	13.16±0.11	51.54±0.05	11.95
TabSyn	1.93±0.07	2.81±0.48	2.13±0.10	0.88±0.18	3.13±0.34	1.52±0.03	3.90±0.04	2.33
TabDiff	**1.49±0.16**	**2.55±0.75**	**1.74±0.08**	**0.76±0.12**	**2.59±0.15**	**1.28±0.04**	**2.20±0.16**	**1.80**
Improve.	22.8% ↓	9.3% ↓	18.3% ↓	13.6% ↓	4.4% ↓	15.8% ↓	37.3% ↓	22.6% ↓

实验结果　下游任务

- 下游任务性能
  - MLE — 机器学习效率（用合成数据训练 XGBoost → 在真实测试集评估）

Methods	Adult	Default	Shoppers	Magic	Beijing	News[1]	Diabetes	Average Gap
	AUC ↑	AUC ↑	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓	AUC ↑	%
Real	.927±.000	.770±.005	.926±.001	.946±.001	.423±.003	.842±.002	.704±.002	0.0
CTGAN	.886±.002	.696±.005	.875±.009	.855±.006	.902±.019	.880±.016	.569±.004	23.7
TVAE	.878±.004	.724±.005	.871±.006	.887±.003	.770±.011	1.01±.016	.594±.009	20.2
GOGGLE	.778±.012	.584±.005	.658±.052	.654±.024	1.09±.025	.877±.002	.475±.008	42.1
GReaT	.913±.003	.755±.006	.902±.005	.888±.008	.653±.013	OOM	OOM	13.3
STaSy	.906±.001	.752±.006	.914±.005	.934±.003	.656±.014	.871±.002	OOM	10.9
CoDi	.871±.006	.525±.006	.865±.006	.932±.003	.818±.021	1.21±.005	.505±.004	30.2
TabDDPM	.907±.001	.758±.004	.918±.005	.935±.003	.592±.011	4.86±3.04	.521±.008	11.95
TabSyn	.909±.001	**.763±.002**	.914±.004	**.937±.002**	.580±.009	**.862±.024**	.684±.002	6.78
TabDiff	**.912±.002**	**.763±.005**	**.921±.004**	.936±.003	**.555±.013**	.866±.021	**.689±.016**	**5.76**

实验结果 无分类器引导（CFG）

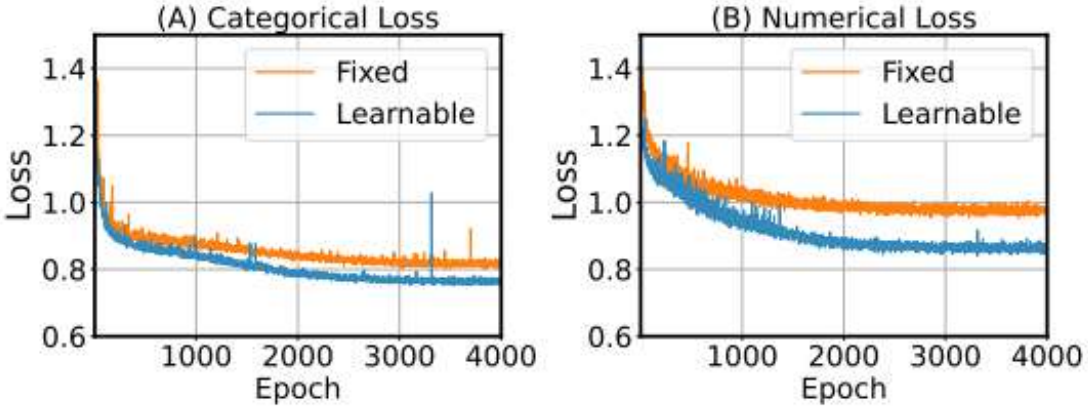
- 应用场景：给定部分已知特征 y，预测缺失特征 x
  - 传统填补：均值/中位数/XGBoost → 无法建模列间关系
- CFG 原理
  - 在“条件预测”和“无条件预测”之间做加权组合，引导强度 ω 控制偏向
  - TabDiff 填补：仅对 x 进行去噪采样，保持 y 固定 → 利用全表联合分布做填补
- 实验结果
  - 原始 TABDIFF（在所有列上训练）

Methods	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Avg. Improv.
	AUC ↑	AUC ↑	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓	AUC ↑	%
Predicted by XGBoost	92.7	77.0	92.6	94.6	0.423	0.842	70.4	0.0
Impute with TabSyn	93.1	86.7	**96.5**	91.3	**0.386**	0.818	66.6	4.99
Impute with TabDiff + CFG ($\omega = 0.0$)	92.5	91.6	95.7	92.5	0.424	0.828	66.0	3.76
Impute with TabDiff + CFG ($\omega = 0.6$)	**93.2**	**91.7**	96.4	**93.0**	0.414	**0.815**	**66.9**	**5.60**

实验结果　消融实验

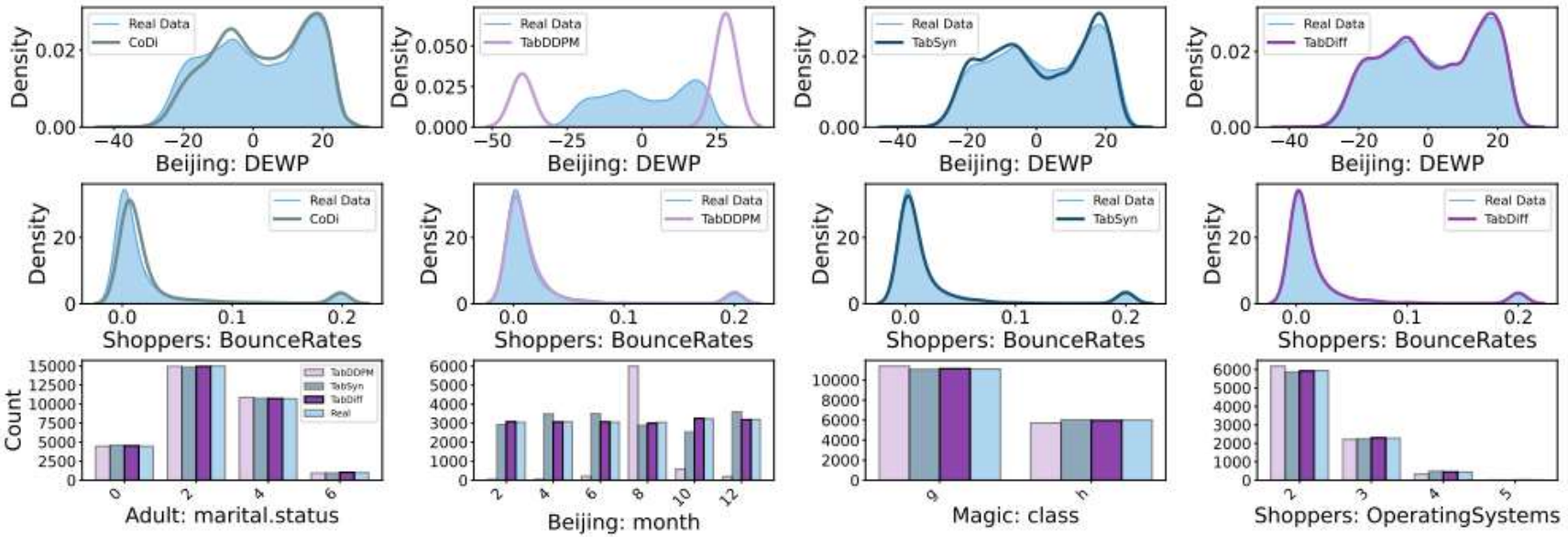
- 消融设计
  - 随机采样器的贡献: Fix.+Det. → Fix.+Sto.
    - 前向扰动有效减少累积解码误差
  - 可学习噪声调度的贡献: Fix. → Learn.
    - 无论用 Det. 还是 Sto.，Learn. 都优于 Fix.
  - 组合叠加效果: Learn.+Sto. 达最优
    - 两大设计互补——调度决定"关注什么"，采样器决定"能否修正错误"
- 训练损失曲线
  - 验证了"自适应分配模型容量"的设计动机

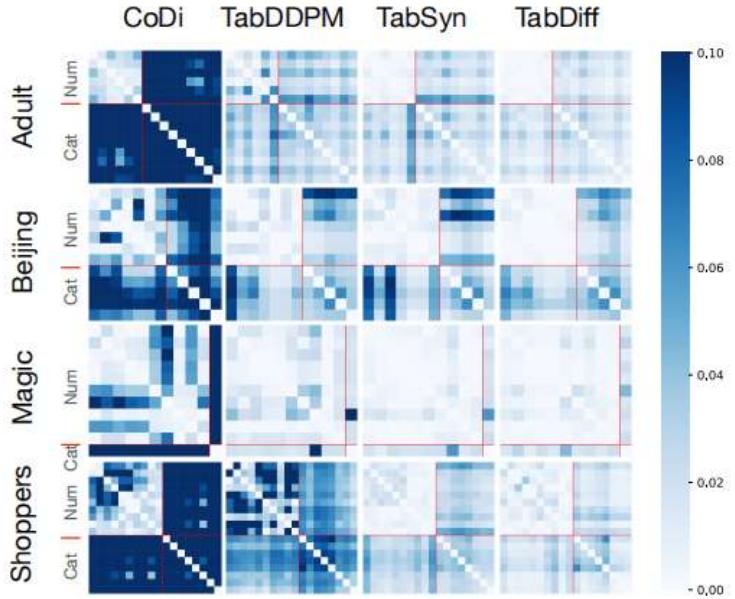
Method	Shape	Trend
TabSyn	1.35	2.33
TabDiff-Fix.+Det.	1.39	2.29
TabDiff-Fix.+Sto.	1.20	1.93
TabDiff-Learn.+Det.	1.24	1.92
TabDiff-Learn.+Sto.	1.17	1.80



实验结果 可视化与鲁棒性分析

- 单列密度可视化和列间相关性热力图





- 采样步数鲁棒性
  - TabDiff方法仅需5步就达到基本可用质量

	TabSyn		TabDiff	
Steps	Shape	Trend	Shape	Trend
5	34.09	49.30	12.51	22.15
10	1.99	3.92	1.55	3.36
25	0.84	1.96	0.62	1.50
50	0.81	1.95	0.63	1.49
100	0.82	1.94	0.64	1.53

算法总结

- 算法贡献
  - 特征级可学习噪声调度
  - 混合类型随机采样器
- 算法不足
  - 训练速度可进一步优化，网络架构仍有精简空间
  - 当前主要面向无条件生成 + 单列填补，多列联合填补未深入探索

How Faithful is your Synthetic Data?
Sample-level Metrics for Evaluating and Auditing Generative Models

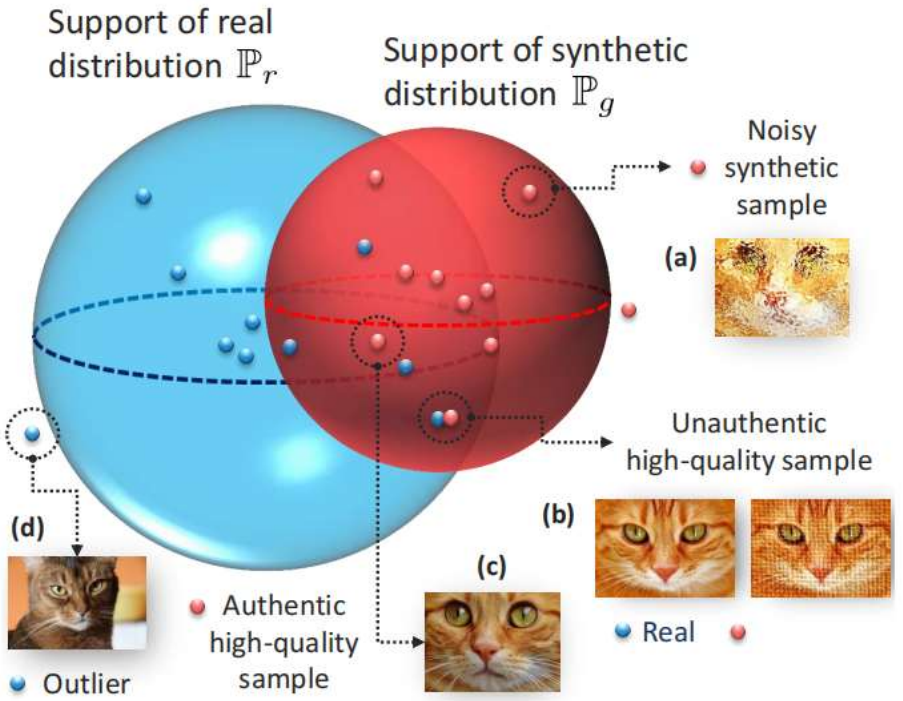
算法原理 TIPO

T	目标	构建域无关、模型无关、三维解耦的评估框架；
I	输入	一组真实数据集与相应的合成数据集
P	处理	1.评估嵌入：训练单类神经网络将真实数据压缩为超球面分布，使 α-支撑集自然化为同心球 2.三维评分：α-Precision 合成样本是否落入真实 α-支撑集；β-Recall 真实样本是否被合成 β-支撑集覆盖；Authenticity 似然比假设检验判断是否为训练数据拷贝 3.模型审计：丢弃低精度/非真实样本，后处理提升合成数据整体质量
O	输出	审计后的合成数据集

P	问题	现有生成模型评估指标（FID、IS等）仅为图像应用设计，跨领域诊断能力有限
C	条件	构建一个既能量化分布级差异（整体质量）、又能进行样本级评判（单样本好坏）的统一评估框架
D	难点	模型可在不创新任何样本的情况下取得完美保真度和多样性
L	水平	ICML 2022（CCF A）

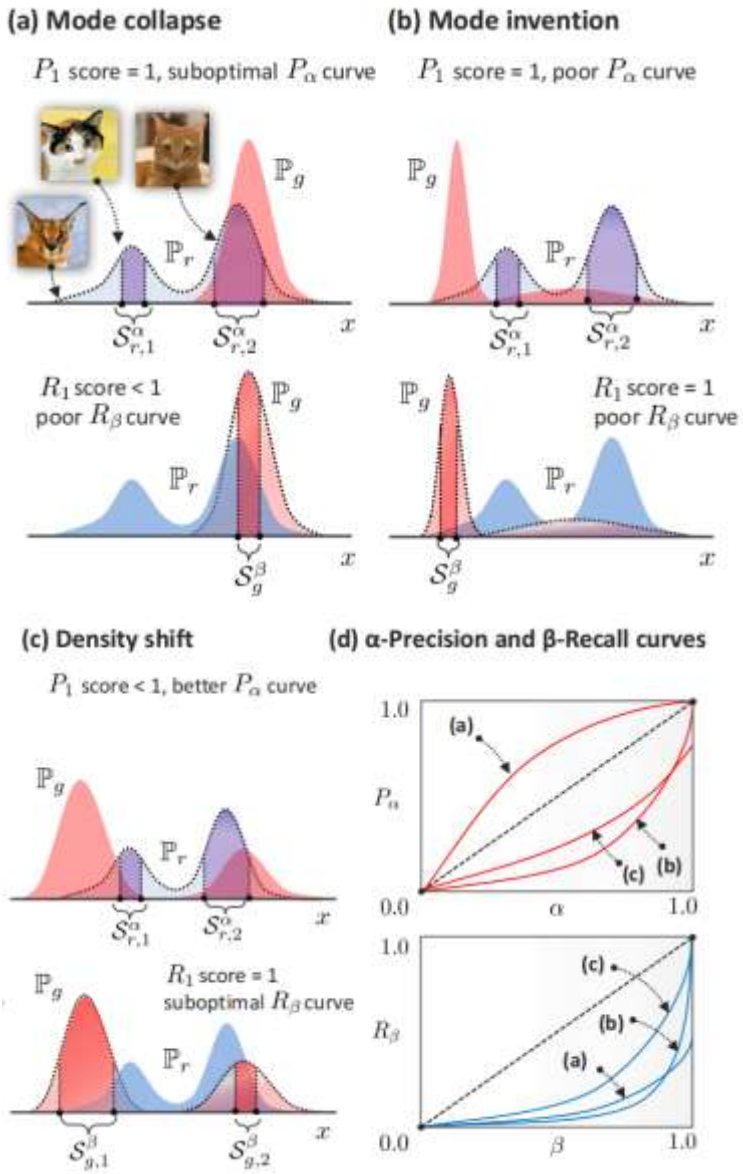
算法原理 三维评估框架总览

- 三维评估指标
  - E ≜ (α-Precision, β-Recall, Authenticity)
  - 保真度（α-Precision）
    - 合成样本是否"真实且典型"，是噪声/离群
  - 多样性（β-Recall）
    - 合成样本是否覆盖了真实数据的所有变异，有无遗漏的模式
  - 泛化性（Authenticity）
    - 合成样本是模型"发明"的新样本，还是训练数据的含噪拷贝



算法原理

- α-支撑集 (Minimum Volume Set)
  - $\mathcal{S}^{\alpha} \triangleq \min_{s \subseteq \mathcal{S}} V(s),\ s.t.\ \mathbb{P}(s) = \alpha$
- α-Precision（保真度）
  - $P_{\alpha} \triangleq \mathbb{P}(\widetilde{X}_g \in \mathcal{S}_r^{\alpha}),\ \text{for}\ \alpha \in [0,1]$
  - 合成样本中落入真实分布 α-支撑集的比例
- β-Recall（多样性/覆盖率）
  - $R_{\beta} \triangleq \mathbb{P}(\widetilde{X}_r \in \mathcal{S}_g^{\beta}),\ \text{for}\ \beta \in [0,1]$
  - 真实样本中落入生成分布 β-支撑集的比例
- P_{α} 和 R_{β} 曲线 vs 传统 Precision/Recall
  - 传统指标只检查支撑集是否重叠 → 密度完全不匹配也能拿满分
  - P_{α}/R_{β} 曲线对密度校准敏感



算法原理

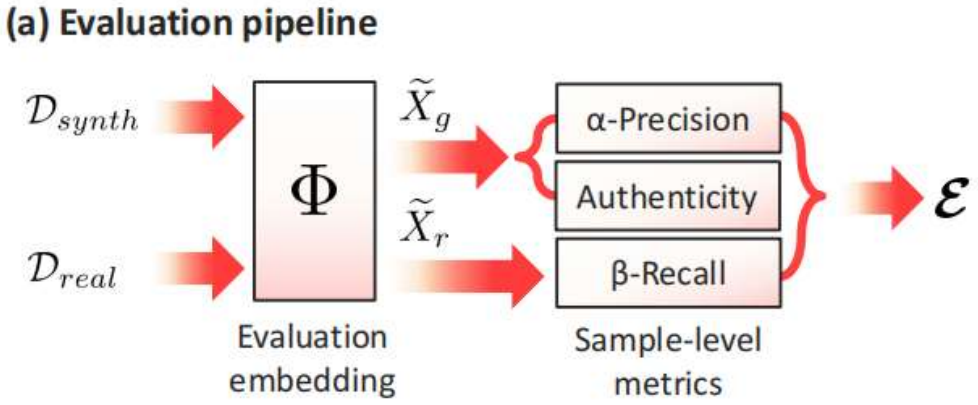
- **Authenticity**
  - **Authenticity 独立于 Precision/Recall，是第三个必不可少的评估维度**
  - **模型可以通过直接重采样训练数据 → Precision = 1, Recall = 1, 但 A = 0**
- **数学模型**
  - $\mathbb{P}_g = A \cdot \mathbb{P}'_g + (1 - A) \cdot \delta_{g,\epsilon}$
- **假设检验**
  - H_1: $A_j = 1$ **样本 j是创新的（authentic）**
  - H_0: $A_j = 0$ **样本 j是训练数据的拷贝**
  - $$\Lambda(\widetilde{X}_{g,j}) = \frac{\mathbb{P}(\widetilde{X}_{g,j} \mid A_j = 1)}{\mathbb{P}(\widetilde{X}_{g,j} \mid A_j = 0)} = \frac{\mathbb{P}'_g(\widetilde{X}_{g,j})}{\delta_{g,\epsilon}(\widetilde{X}_{g,j})}$$
  - **A=1 → 所有样本都是创新的；A≈0 → 几乎全是拷贝**

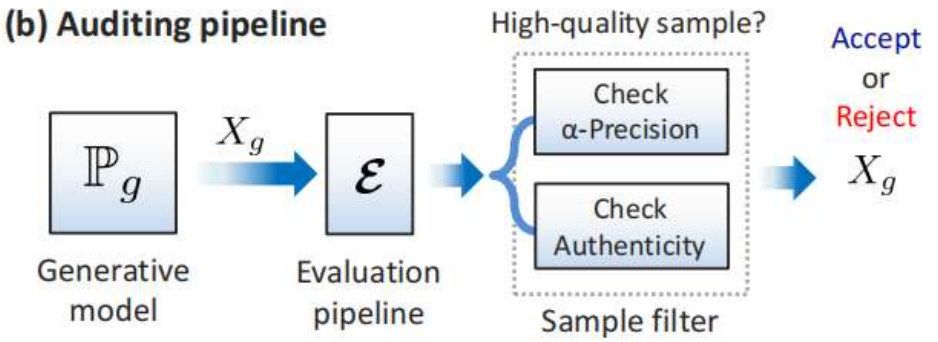
算法原理 理论完备性与评估嵌入

- **Theorem 1（完备性定理）**
  - $P_\alpha/\alpha = R_\beta/\beta = 1,\ \forall \alpha,\beta \in [0,1] \ \Leftrightarrow\ \mathbb{P}_g = \mathbb{P}_r$
  - 即α-Precision 和 β-Recall 在所有 α, β 上都达到最优（恒等于 α 和 β），当且仅当生成分布与真实分布完全相同
- 评估嵌入Φ
  - 原始空间中计算 $\mathcal{S}^\alpha \triangleq \min_{s \subseteq \mathcal{S}} V(s),\ s.t.\ \mathbb{P}(s) = \alpha$ 是困难的优化问题
  - 解决思路：设计嵌入 Φ 使支撑集变成简单的几何形状
- 单类神经网络超球面嵌入- $\ell_i = r^2 + \frac{1}{\nu}\max\{0, \| \Phi(X_{r,i}) - c_r \|^2 - r^2\}$
  - 将真实数据"挤入"以 c_r 为中心、r 为半径的最小超球面
- 合成支撑集的 k-NN 估计
  - $f_R(\widetilde{X}_{r,i}) = 1$ 最近的典型合成样本在 k-NN 距离内，对每个真实样本，检查其 k 近邻范围内是否有 β-支撑集内的合成样本

算法原理　评估与审计流水线

- 评估流水线
  - 训练评估嵌入 Φ（单类 NN）将数据映射到超球面特征空间
  - 在嵌入空间中计算三个二元分类器
    - 合成样本是否在真实 α-支撑内,聚合得 P_α
    - 真实样本是否被生成 β-支撑覆盖,聚合得 R_β
    - 合成样本是否不紧贴训练数据,聚合得 A
  - 扫描 $\alpha, \beta \in [0,1]$ 获得完整曲线
- 审计流水线
- 审计的四个关键特性
  - 不需修改或重训生成模型
  - 黑盒友好，不需访问模型内部，只需能采样
  - 充当拒绝采样器，持续采样直到凑够高质量样本
  - 可与任何生成模型配合使用





- 实验指标

指标	衡量内容	类别
FID	Inception特征空间 P_r 与 P_g的距离	统计散度
Wasserstein Distance	最优传输距离	
Parzen Window Likelihood	核密度估计下的对数似然	密度估计
Precision/Recall (P_1/R_1)	支撑集重叠（传统方法）	P-R分析
Density/Coverage (D/C)	含密度加权的支撑集分析	
IP_α / IR_β/ A	密度校准 + 泛化性	P-R+泛化

- 数据集

Name	Type	n	d	Embedding	d_{emb}
SIVEP-GRIPE	Tabular	6882	25	-	-
AmsterdamUMCdb	Time-series	7695	70	Seq-2-Seq	280
MNIST	Image	10000	784	InceptionV3	2048

实验结果　对比试验

- 实验设计
  - 4 个模型各自生成合成数据集 → 对每个合成数据集训练 Logistic 回归预测死亡率 → 在真实测试集上评估 AUC
- Ground-Truth 排序，各合成数据下游 AUC 的排序
- 密度 vs 支撑集
  - 只比较支撑集的指标（$P_1/R_1/D/C$）排名与下游效用几乎无关
    - 即使支撑集重叠，密度偏移 → 合成数据中的"协变量偏移" → 预测模型性能差
  - P_α与下游 AUC 的相关系数 = 0.71，而 FID 仅为 0.38
  - 密度校准（而不仅是支撑集重叠）才是合成数据下游效用的决定因素

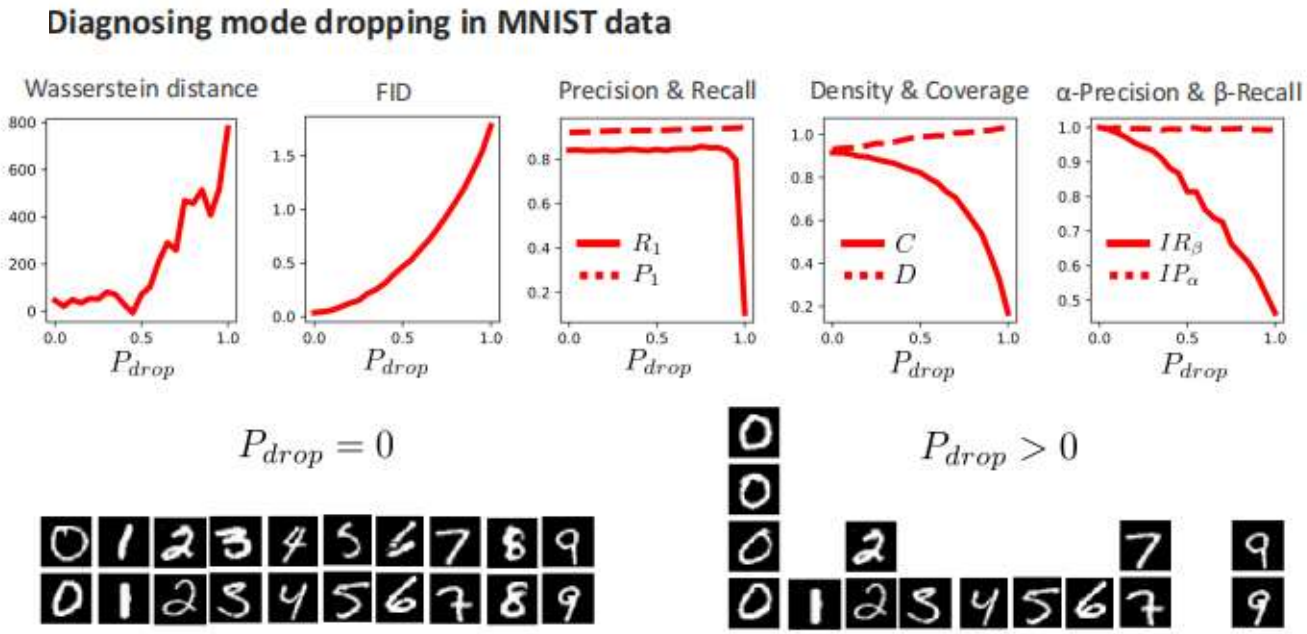
(a) Ranking generative and predictive models

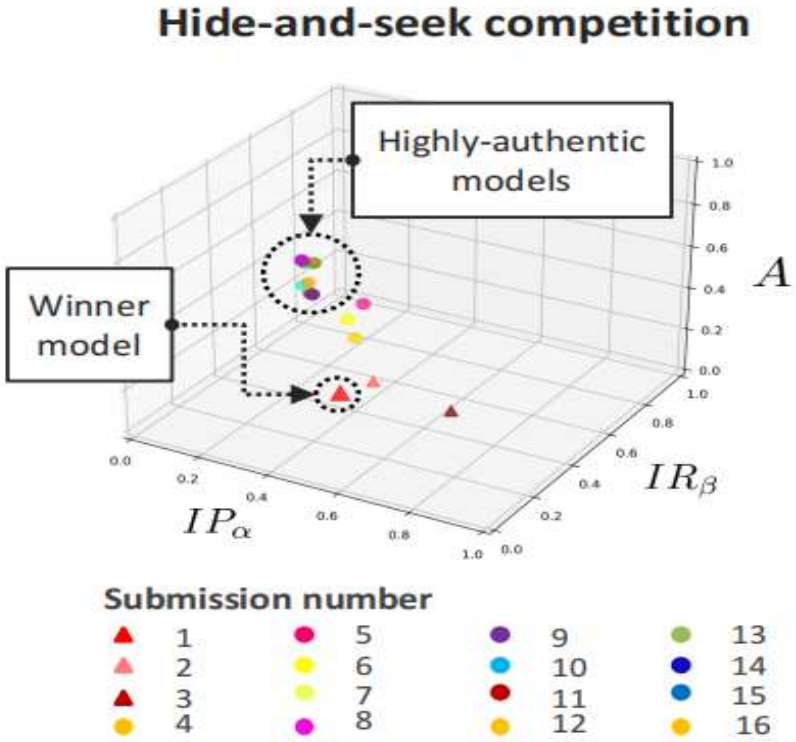
Ground-truth ranking: ✖ ADS-GAN ● WGAN-GP ▲ VAE ■ GAN

Metric	Generative models ranking	AUC-ROC
FID	✖ ● ■ ▲	0.79 ± 0.02
PW	✖ ● ■ ▲	0.79 ± 0.02
W	● ✖ ■ ▲	0.76 ± 0.02
P_1	▲ ● ✖ ■	0.55 ± 0.03
R_1	✖ ● ■ ▲	0.79 ± 0.02
D	▲ ● ✖ ■	0.55 ± 0.03
C	● ✖ ■ ▲	0.76 ± 0.02
IP_α	✖ ● ▲ ■	0.79 ± 0.02
IR_β	✖ ● ■ ▲	0.79 ± 0.02

实验结果　审计、隐私权衡与跨领域诊断

- 超参数优化
- 模型审计 — 后处理提升
- MNIST 模式坍塌诊断
- Hide-and-Seek 竞赛重评估





算法总结

- 算法贡献
  - 三维完备评估框架
    - 可诊断传统指标盲区的失败模式：mode collapse, mode invention, density shift
    - α-Precision（保真度）+ β-Recall（多样性）+ Authenticity（泛化性）
    - Theorem 1 保证三维最优
  - 样本级评分 → 模型审计
    - 对黑盒模型友好，无需修改模型内部
  - 域无关 & 模型无关
- 算法不足
  - 部分数据模态需要预训练嵌入（图像）或自训练嵌入（时序/表格）
  - k-NN 的 k 参数无理论闭式解（实践中 k=5 通常有效）
  - 部分数据模态需要预训练嵌入（图像）或自训练嵌入（时序/表格）

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - TabDiff & Faithfulness
    - 生成能力突破
    - 评估体系革新
    - 端到端实用闭环
    - 降低应用风险（Authenticity 检测拷贝 + DCR 隐私保护）
    - 域无关 & 模型无关
- 未来展望
  - 生成与评估协同优化
  - LLM 时代的高维表格合成（扩散框架 + 大模型语义理解）

参考文献

[1] Shi J, Xu M, Hua H, et al. TabDiff: A Mixed-Type Diffusion Model for Tabular Data Generation[C]. The Twelfth International Conference on Learning Representations (ICLR), 2025.

[2] Alaa A M, van Breugel B, Saveliev E S, et al. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models[C]. Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162, 2022: 290-306.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！