计算机科学
COMPUTER SCIENCE

表格数据生成技术综述

王永鑫，徐鑫，朱鸿斌

引用本文

王永鑫，徐鑫，朱鸿斌. 表格数据生成技术综述[J]. 计算机科学，2025，52(10)：3-12.

WANG Yongxin, XU Xin, ZHU Hongbin. Survey of Tabular Data Generation Techniques[J]. Computer Science, 2025, 52(10): 3-12.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250800044

表格数据生成技术综述

王永鑫[1,2]　徐　鑫[3]　朱鸿斌[1,2]
1 复旦大学金融科技研究院　上海 200433
2 复旦大学计算与智能创新学院　上海 200433
3 上海立信会计金融学院计算机与人工智能学院　上海 201209
(yongxinwang24@m.fudan.edu.cn)

摘　要　表格数据因在金融、医疗等关键领域广泛应用而具有重要价值。然而,对于表格数据的有效利用,常受到数据稀缺、类别不平衡及隐私法规的严格制约。为应对这些挑战,通过生成模型合成在统计特性上与真实数据高度相似的样本,已成为一种新兴的解决方案,旨在增强数据可用性并保护用户隐私。该领域的技术发展路径从传统的深度学习模型逐步演进至前沿范式。早期的探索以变分自编码器和生成对抗网络为代表,但这些方法常面临训练不稳定和模式坍塌等瓶颈,影响了生成数据的质量。为克服这些难题,扩散模型应运而生,其通过渐进式的去噪过程,在生成高保真度和多样性的样本方面展现出显著优势。尽管如此,这些模型的核心仍是模仿统计分布,缺乏对现实世界常识的理解。为此,最新的研究转向基于大型语言模型的方法,利用其丰富的世界知识,旨在生成不仅统计真实,而且在逻辑与语义上也更合理的合成表格数据。对该领域的系统性回顾,旨在为研究者和从业者提供全面的技术认知,并为不同应用场景下选择最合适的技术路径提供决策参考。

关键词:表格数据生成;大语言模型;生成方法
中图分类号　TP183

Survey of Tabular Data Generation Techniques

WANG Yongxin[1,2],XU Xin[3] and ZHU Hongbin[1,2]
1 Institute of Financial Technology,Fudan University,Shanghai 200433,China
2 College of Computer Science and Artificial Intelligence,Fudan University,Shanghai 200433,China
3 School of Computer Science and Artificial Intelligence,Shanghai Lixin University of Accounting and Finance,Shanghai 201219,China

Abstract　Tabular data holds significant value due to its widespread application in critical domains such as finance and healthcare. However,the effective utilization of tabular data is often constrained by data scarcity,class imbalance,and stringent privacy regulations. To address these challenges,synthesizing samples that are statistically highly similar to real data through generative models has emerged as a novel solution,aiming to enhance data availability and protect user privacy. The technological development path in this field has progressively evolved from traditional deep learning models to cutting-edge paradigms. Early explorations are represented by Variational Autoencoders and Generative Adversarial Networks,but these methods often face bottlenecks such as training instability and mode collapse,affecting the quality of generated data. To overcome these difficulties,diffusion models have emerged,demonstrating significant advantages in generating high-fidelity and diverse samples through a progressive denoising process. Nevertheless,the core of these models remains the imitation of statistical distributions,lacking an understanding of real-world common sense. Consequently,the latest research has shifted towards methods based on Large Language Models (LLMs),leveraging their rich world knowledge to generate synthetic tabular data that is not only statistically authentic but also logically and semantically more reasonable. A systematic review of this field aims to provide researchers and practitioners with a comprehensive understanding of the technology and offer decision-making references for selecting the most appropriate technical path in different application scenarios.

Keywords　Tabular data generation,Large language model,Generative methods

到稿日期:2025-08-12　返修日期:2025-09-20
基金项目:国家自然科学基金青年基金(62306077);国家重点研发计划(2023YFC3305304)
This work was supported by the National Natural Science Foundation of China(62306077) and National Key Research and Development Program of China(2023YFC3305304).
通信作者:朱鸿斌(zhuhb@fudan.edu.cn)

1 引言

表格数据是现实世界应用中的一种基本数据格式，在医疗保健[1]、金融[2]、教育[3]和交通[4]等领域发挥着至关重要的作用。然而，机器学习对表格数据的有效利用，常常受到数据稀缺、类别不平衡和缺失值等固有挑战的制约。同时，随着数据隐私相关法律法规的实施，数据的可用性受到了严格的限制。虽然基于隐私计算的方法，如联邦学习[5]、同态加密[6]、差分隐私[7]，理论上可以在保护数据隐私的同时进行有效的数据分析，但这类方法需要额外的资源来确保数据隐私和安全性，因此往往伴随着庞大的计算和资源消耗。

在此背景下，生成模型提供了一种新的解决方式。图 1 展示了表格数据生成的整体流程，即不同的方法对原始数据集的特征进行学习，之后生成符合原始数据分布的新的合成数据集。通过生成与真实数据在统计特性上相似但不包含用户隐私信息的数据，研究人员能够在不直接使用真实数据的情况下对数据进行分析。通过学习表格数据的分布，这些模型能够生成与真实数据相似的合成样本，从而在缓解隐私顾虑和数据限制问题的同时，增强数据的可用性。早期的探索以传统生成方法为主，涵盖了基于机器学习、变分自编码器(Variational Autoencoder，VAE)[8]和生成对抗网络(Generative Adversarial Network，GAN)[9]的多种技术。这些方法奠定了通过学习数据分布来生成样本的基础，但常面临如 GAN 模型训练不稳定、模式坍塌等挑战，难以保证生成数据的质量和多样性。为解决这些难题，扩散模型方法应运而生，它通过学习一个渐进的去噪过程来生成数据，有效突破了 GAN 的训练瓶颈，在生成高保真度和多样性的样本方面展现出显著优势。然而，无论是传统方法还是扩散模型[10]，其核心仍是模仿数据的统计分布，缺乏对现实世界常识的理解，可能生成不符合逻辑约束的数据。为了弥补这一不足，最新的研究开始转向基于大型语言模型[11]的方法，旨在利用 LLM 预训练中获得的丰富世界知识，生成不仅在统计上真实，而且在逻辑和语义上也更合理的合成表格数据。

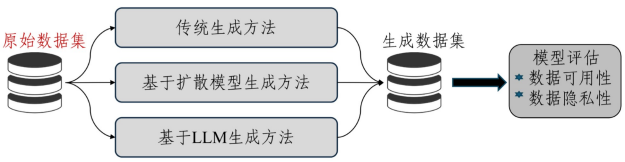

图 1 表格数据生成流程图

Fig. 1 Flow chart of tabular data generation

现有大多数关于表格数据的综述缺乏一个整合了近期技术进展的统一视角。对此，本文进行了结构化回顾，其视角建立在 4 个关键需求之上：统计保真度、下游任务的应用效能、数据保护的隐私程度，以及与领域特定知识的对齐。本文系统性地回顾了从传统的深度学习模型到扩散模型，再到基于大型语言模型的前沿方法的技术发展路径，旨在为研究者和从业者提供全面的技术认知，并为在不同场景下选择合适的技术提供决策参考。

本文第 2 章对表格数据生成问题进行了详细的定义，并从表格数据的异质性、数据不平衡等特征对表格数据进行说明；第 3 章对表格数据生成方法进行了梳理总结，涵盖从传统生成方法到基于扩散模型的生成方法，再到基于 LLM 的生成方法，并对 3 类生成方法进行了对比；第 4 章从模型评估方面展开综述；第 5 章总结了该领域研究面临的挑战，以及对未来研究进行了展望；最后总结全文。

2 表格数据生成问题定义

2.1 表格数据生成任务的问题定义

表格数据生成任务的核心目标是学习一个现有真实表格数据集D_e的内在统计分布，并利用学习到的知识生成一个全新的、在统计特性上高度相似的生成数据集D_s。这一过程可以定义为$D_e=\{(x_e^i,y_e^i)\}_{i=1}^{N_t}$，其中$N_t$是数据集中真实样本的总数，$x_e^i$代表第$i$个样本的所有特征值，$y_e^i$代表该样本对应的标签。整个数据集由一组特征$F=\{f_j\}_{j=1}^{d}$所定义，其中$d$是特征的总数量。

表格数据生成任务的目标是训练一个生成模型p_θ，该模型以真实数据集D_e为输入进行学习。训练完成后，该模型应能生成一个数据集。整个生成过程可被形式化为$D_s \leftarrow p_\theta(D_e)$。

D_s必须遵循与真实数据集D_e相同的结构格式，即拥有相同的特征集合。在实际应用中，完整的流程通常还包含对生成的D_s进行后处理以修正不合逻辑的样本，以及通过一系列评估指标来检验其在数据可用性和隐私保护方面的质量。

2.2 表格数据的特征

与图像、文本等由同质化基本单元构成的同构数据不同，表格数据具有一系列独特的结构和统计特性，这为生成模型的构建带来了巨大挑战。这些特征包括数据类型的混合、复杂的概率分布、内在的列间依赖关系、特殊的结构属性以及常见的数据质量问题，它们共同构成了表格数据生成任务的复杂性。具体来说，表格数据的主要特征如下。

1)异构性。表格数据最核心的特征是其固有的异构性，即数据由多种不同类型的特征列混合而成，每种类型都具有独特的统计属性。常见的特征类型包括数值型(连续或离散)、分类型、二元型及文本型特征。这种混合数据类型的存在，使得对所有特征的联合概率分布进行统一建模变得极为困难。

2)复杂的概率分布。现实世界的表格数据很少遵循简单的概率分布。数值型特征列常常不服从高斯分布，而是表现出偏态、重尾或多模态的特性。同时，分类型特征中严重的类别不平衡现象极为常见。

3)复杂的列间依赖关系。表格数据的各特征列之间并非相互独立，而是存在着基于现实世界逻辑的上下文关联。例如，一个人的教育水平和职业通常是相关的，准确捕捉这些依赖关系对于生成真实可信的数据至关重要，然而这些关系通常是稀疏的，即只有部分特征对之间存在强相关性。

4)结构特性。与文本或时间序列数据不同，表格数据的行与列在物理存储上的顺序通常是任意的，其排列不影响数据本身的含义，即具有顺序不变性。这一特性使得依赖空间局部性的模型难以直接适用。

5)数据质量问题。由于数据采集错误、隐私保护或属性可选等原因，表格数据中普遍存在缺失值，这增加了生成模型

训练的复杂性。缺失机制可分为 3 类：(1)完全随机缺失(MCAR)，缺失与任何变量无关；(2)条件随机缺失(MAR)，缺失与其他观测特征相关；(3)非随机缺失(MNAR)，缺失依赖于自身的潜在值。不同缺失机制对模型训练影响不同：MCAR 主要影响样本有效性；MAR 需要利用其他特征进行合理填充；MNAR 则最具挑战性，可能引入系统性偏差。

3 表格数据生成方法

本章系统地探讨和梳理用于生成合成表格数据的各类主流方法，相关模型和方法如图 2 所示。章节的组织遵循一条从经典到前沿的技术发展路径，对整个领域进行全面介绍。首先，从传统生成方法入手，深入分析经典的基于统计建模的技术(如 Copulas[12]、贝叶斯网络[13])，以及开创了深度学习生成模型先河的 VAE 和 GAN。在此基础上，进一步聚焦于近年来展现出巨大潜力的新兴范式，详细剖析基于扩散模型的生成方法的核心原理与主要技术分支。最后，探索该领域的前沿研究方向，即如何利用大语言模型的强大能力，通过序列化等关键技术来生成高保真的表格数据。

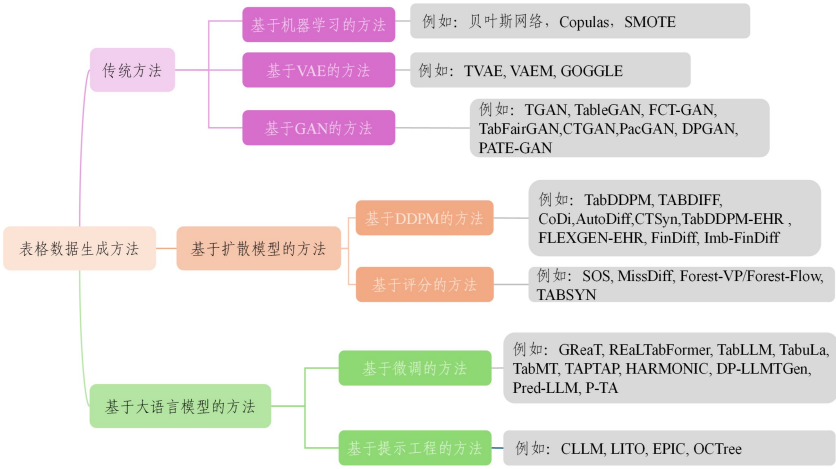

图 2 表格数据生成方法相关模型及方法

Fig. 2 Tabular data generation methods, related models and methods

3.1 传统生成方法

3.1.1 基于机器学习的方法

在特别设计的生成模型(如 VAE，GAN)出现之前，贝叶斯网络、Copulas 等基于统计建模的生成方法已经得到了广泛应用。使用统计模型生成合成表格数据的思想最早可以追溯到 Rubin 的奠基性工作，他提出了为保护数据隐私而创建生成数据集的范式[14]。

贝叶斯网络是一种强大的概率图模型，它通过一个有向无环图来表示变量之间的条件依赖关系。在表格数据生成中，首先从真实数据中学习一个贝叶斯网络的结构和参数，该网络捕捉了各个特征列之间的联合概率分布；然后从这个学习到的网络中进行采样，从而生成新的、与原始数据分布相似的生成数据。Young 等[15]较早探讨了使用贝叶斯网络生成数据的可能性，详细阐述了其在保护数据隐私和支持数据分析方面的潜力。Martins 等[16]提出了一个完全贝叶斯的数据生成与分析框架，该框架基于生成数据统计量的后验预测分布，实现了对不确定性的有效量化。

Copula 理论的数学基础由 Sklar[17]在其开创性工作中奠定，他证明了任何多维联合分布都可以被唯一分解为边际分布和一个 Copula 函数。然而，将这一理论应用于生成数据的实践则是在数十年后，在金融风险管理等领域首先得到广泛应用。例如 Embrechts 等[18]利用 Copula 来模拟具有复杂依赖关系的变量，这为后来其在通用表格数据生成领域的应用铺平了道路。最近，Restreps[19]提出了一种基于经验 Copula 的非参数化方法来生成合成数据，避免了对数据分布进行参数化假设，提高了方法的普适性。Jutras-Dubé 等[20]探索了如何利用 Copula 框架，结合来自不同数据源的信息，生成更符合现实情况的合成人口数据。Kamthe 等[21]将 Copula 理论与正态化流这一深度学习模型相结合，提出了一个更强大的概率模型，用于学习复杂的数据密度和生成高质量的合成数据。

除了贝叶斯网络、Copulas 两种主流方法，其他的一些统计技术也被用于表格数据生成，尤其是处理特定问题，如数据不平衡问题。SMOTE 虽然主要用于处理非平衡分类问题[22]，但其核心思想是通过在少数类样本间进行线性插值来生成新的合成样本。这是一种基于邻近度的统计生成方法。多元高斯分布是最简单的一种生成模型，它假设数据遵循一个多元正态分布[23]，通过计算真实数据的均值向量和协方差矩阵，可以对这个分布进行参数化，并从中进行采样。然而，这种方法通常过于简化，难以捕捉现实世界数据的复杂性和非高斯特性。

传统统计模型因对先验假设的依赖和在高维空间中学习复杂依赖关系面临困难，应用效果受到限制。为了应对这些由模型结构刚性带来的挑战，深度学习生成模型应运而生，如 VAEs 和 GANs。这类模型无需对数据分布进行显式假设，而是利用神经网络强大的学习能力直接从数据中隐式地捕捉其内在结构，因此在处理复杂、高维的表格数据生成任务时展现出了巨大的优势和潜力。

3.1.2 基于 VAE 的方法

VAE 是深度生成模型领域的另一重要分支。VAE 基于变分推断的原理，通过学习数据的潜在表示来生成新样本。VAE 的原理如图 3 下半部分所示，该框架由两个核心神经

网络组成，即一个编码器和一个解码器。编码器的任务是将真实数据压缩并映射到一个低维的、符合特定先验分布的潜在空间；解码器的任务则是从该潜在空间中采样向量并将其重构为与原始数据分布一致的生成数据。VAE 的训练目标是最大化证据下界(Evidence Lower Bound，ELBO)[24]。

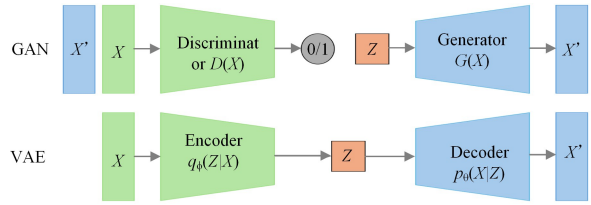

图 3 VAE 和 GAN 原理图

Fig. 3 Schematic diagrams of VAE and GAN

早期适配性模型工作的重点在于，对表格数据进行有效的预处理，使其能被标准的 VAE 架构所接受。TVAE[25]是该方向的代表性工作，它通过独热编码将分类型特征转化为数值型，并设计相应的重构损失函数。尽管 TVAE 在保真度上的表现优于同期的 CTGAN[25]，但由于其解码器在训练时能直接访问真实数据，可能存在隐私泄露的风险。

为了更精确地捕捉表格数据的复杂特性，后续研究在 VAE 的架构上进行了更多创新。VAEM[26]专为处理异构数据而设计，它采用两阶段训练方式：第一阶段学习数据的统一表示，以克服混合数据类型带来的建模难题；第二阶段则在此基础上进行生成模型的训练。GOGGLE[27]则另辟蹊径，它将图神经网络融入 VAE 的编解码器中，通过显式地学习一个表示列间依赖关系的邻接矩阵来指导生成过程。这种方法能够更灵活地捕捉特征间的成对关系，尤其适用于具有复杂关联结构的数据。

尽管基于 VAE 的方法因训练稳定、不易模式崩溃而得到广泛应用，但它们也存在固有的局限性。一个普遍存在的问题是：ELBO 的目标函数中 KL 散度项的强正则化效应，模型倾向于产生平均化的样本，即生成的数值会向均值靠拢，导致生成数据丢失原始数据中的一些锐利细节和极端值，在保真度上可能不及顶尖的 GAN 或扩散模型。

3.1.3 基于 GAN 的方法

GAN 是一种强大的深度生成模型，其核心在于通过对抗过程来优化生成模型。GAN 的基本原理如图 3 上半部分所示，该框架由一个生成器和一个判别器两个相互竞争的神经网络组成。生成器的任务是学习真实数据的内在分布，并产生与真实样本不同的生成数据；而判别器的任务则是尽力分辨出真实数据与生成器产生的生成数据。在这个 min-max 博弈中，生成器不断进化以欺骗判别器，而判别器则不断提升其辨别能力。理想情况下，当训练达到纳什均衡时，生成器便掌握了真实数据的分布规律，其生成的数据分布与真实数据分布基本一致。

研究者对基础 GAN 架构进行了大量研究，这些方法大致可分为以下 3 类。

1)传统 GAN(Traditional GAN)。这类方法的重点在于改进模型架构和数据预处理方式，以适应表格数据的特性。作为该领域的早期代表作，TGAN[28]创新性地采用了长短期记忆网络作为生成器，逐一生成不同列的数值，从而捕捉列间的序列依赖关系。为了处理非高斯和多模态的数值特征，TGAN 引入了模式特异性归一化，即使用高斯混合模型对数值列进行建模和转换。后续的工作进一步拓展了这一方向，例如 TableGAN[29]探索了使用卷积神经网络来处理表格数据，而 FCT-GAN[30]则结合了特征令牌化和傅里叶网络，构建了 Transformer 风格的生成器与判别器。更新的 TabFairGAN[31]则是一个基于 Wasserstein GAN 梯度惩罚的架构，它对数值特征采用分位数变换，对分类型特征采用结合 Gumbel-Softmax 的独热编码，以提升生成质量和公平性。

2)条件 GAN(Conditional GAN，CGAN)。传统 GAN 的一个主要局限在于无法控制生成样本的具体特征。为了解决这一问题，CGAN[32]被广泛采用。CGAN 通过在生成器和判别器的输入中引入一个额外的条件向量，如类别标签，从而实现对生成数据特定特征的控制。这一能力在处理类别不平衡问题时尤为关键，因为它允许模型定向地扩大少数类样本。该领域的标杆性工作 CTGAN[25]正是 CGAN 的深度改良版。CTGAN 有几项关键创新：首先，它采用变分高斯混合模型进行数据预处理，能够自动估计数值列的模态数量；其次，它设计了一种条件向量，并结合按样本训练策略，从而高效地对具有严重不平衡和长尾分布的离散列进行建模；最后，它在判别器中引入了 PacGAN[33]，进一步提升了模型的性能和稳定性。

3)差分隐私 GAN(Differentially Private GAN，DP-GAN)。在处理金融、医疗等领域的敏感数据时，隐私保护是首要考量。差分隐私为数据生成过程提供了可严格量化的隐私保障。将 DP 与 GAN 结合的核心思想是在训练过程中对判别器的梯度进行扰动。具体而言，差分隐私随机梯度下降(DP-SGD)[34]算法被广泛采用，其关键步骤是在每次迭代中，对每个样本的梯度进行裁剪以限制其影响，然后向聚合后的梯度中添加经过精确校准的高斯噪声。DPGAN[34]正是这一思想的直接应用。作为一种替代方案，PATE-GAN[35]采用了私有知识聚合框架。该框架训练一个由多个教师判别器组成的集成模型，然后向教师的聚合标签中注入噪声，用带噪标签来训练一个最终的学生判别器，从而实现隐私保护。

3.2 基于扩散模型的生成方法

尽管基于 GAN 的方法极大地推动了表格数据生成技术的发展，但其训练过程不稳定、模式崩溃以及对多模态分布表示不佳等问题依然是研究的难点，这也激发了学术界对扩散模型等的更新及更稳定生成范式的探索。

为应对 GAN 等模型在训练稳定性上的挑战，扩散模型[10]作为一种新兴的生成范式，近年来在生成数据领域获得了广泛关注。这类模型通过一个独特的去噪机制，能够有效捕捉复杂的数据分布，并在多个方面展现出超越传统生成模型的潜力。目前，应用于表格数据生成的扩散模型主要可分为两大技术路线：基于去噪扩散概率模型(Denoising Diffusion Probabilistic Model，DDPM)的方法和基于评分(Score-based)的方法。

3.2.1 扩散模型原理

扩散模型原理示意图如图 4 所示，将数据生成过程构建为一个包含两个核心阶段的马尔可夫过程：一个前向加噪过

程和一个反向去噪过程。前向过程是一个固定的、无需学习的扩散阶段。在此过程中,原始数据在预设的多个时间步长内被逐步、迭代地注入噪声,直至其完全转化为一个纯粹的先验噪声分布。反向过程是模型学习的关键所在。模型学习逆转上述加噪过程,即从一个纯噪声样本出发,通过一系列逐步的去噪操作,最终重构出符合原始数据分布的样本。整个学习过程通过优化变分下界来完成。

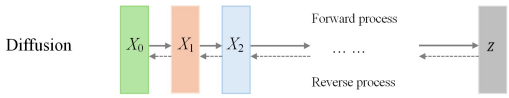

图 4 扩散模型原理图

Fig. 4 Schematic diagrams of diffusion model

3.2.2 基于 DDPM 的方法

研究者对 DDPM[36]的基础框架进行了多种改造,使其适应表格数据独特的结构与挑战。在通用表格数据生成方面,TabDDPM[37]是将 DDPM 成功应用于此领域的开创性工作之一,其性能在多个方面超越了当时的 GAN 和 VAE 模型。为处理表格的异构性,后续工作提出了更精细的方案。例如,TABDIFF[38]设计了一种联合连续时间的扩散过程,以同时处理数值和分类特征,CoDi[39]则为这两类特征分别训练了两个相互约束的扩散模型。另一种有效的策略是在一个更规整的潜在空间中执行扩散,如 AutoDiff[40]和 CTSyn[41]先利用自编码器将表格行映射到低维潜在空间,再进行扩散和去噪处理。此外,DDPM 在特定领域的应用也取得了显著成果。在医疗领域,TabDDPM-EHR[42]和 FLEXGEN-EHR[43]被用于生成高质量的电子健康记录。在金融领域,FinDiff[44]专注于生成高保真的金融表格数据,其改进版 Imb-FinDiff[45]则进一步解决了金融场景中常见的类别不平衡问题。

3.2.3 基于评分的方法

基于评分的生成模型提供了另一种视角。这类模型不直接学习去噪的转换函数,而是学习一个评分函数,即数据在任意噪声水平下对数密度的梯度。该过程通常通过随机微分方程(Stochastic Differential Equation,SDE)来描述。数据生成是通过求解一个逆向时间 SDE 来实现的,该方程利用训练好的评分网络来引导样本从纯噪声逐渐向真实数据分布演化。在表格数据应用中,SOS[46]是首个专为处理类别不平衡问题而设计的基于评分的采样方法。MissDiff[47]通过引入特定的损失函数,使模型能够直接在含有缺失值的数据集上进行训练。更进一步地,一些工作探索了非神经网络的评分函数。例如,Forest-VP/Forest-Flow[48]创新性地采用 XGBoost 梯度提升树来建模评分函数,充分利用了其天然适合处理表格数据的优势。混合模型也是一个重要的研究方向,例如 TABSYN[49]将评分扩散过程置于一个由 VAE 构建的潜在空间内执行,结合了两种模型的优点。

基于 DDPM 的方法与基于评分的方法各有其侧重。DDPM 框架通过前向加噪与反向去噪过程实现数据生成,适合处理高维和复杂依赖的数据结构。其优势在于训练稳定,生成样本多样性高,可以通过联合连续时间扩散或潜在空间映射处理异构特征。然而,DDPM 对计算资源需求较高,且对于缺失值和类别不平衡问题需要额外改造。相比之下,基于评分的方法通过学习数据在不同噪声水平下的对数密度梯度进行生成,能够灵活利用非神经网络模型(如梯度提升树)处理表格数据,且在缺失值处理和类别不平衡问题上具有天然优势,但其训练过程对梯度估计精度敏感,可能影响信息保真度。综上,DDPM 更适合追求高保真度和复杂分布建模的场景,而评分方法在处理缺失、类别不平衡或计算资源受限的情况下表现更优。针对具体应用,可通过联合潜在空间映射、异构特征分组训练或引入额外损失函数等策略,对两类方法进行改进,以进一步提升生成数据的质量和适用范围。

3.3 基于大语言模型的生成方法

随着大语言模型的崛起,数据生成领域迎来了一次重要的范式转移。与以往为表格数据设计专门生成架构的传统思路不同,新范式旨在利用 LLM 强大的预训练能力,将结构化数据的生成问题巧妙地转化为 LLM 所擅长的文本序列生成任务。基于 LLM 的生成方法如图 5 所示。

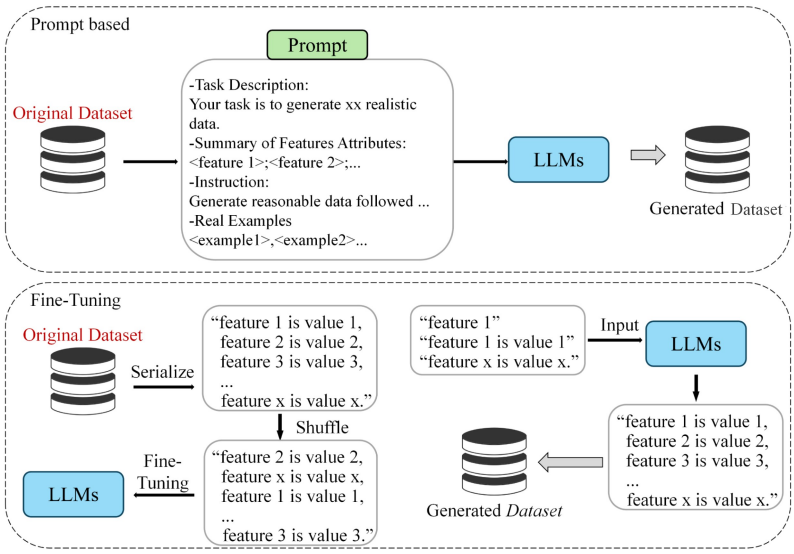

图 5 基于大语言模型的生成方法

Fig. 5 Generative methods based on LLMs

3.3.1 核心机制

基于 LLM 的表格数据生成方法的核心机制在于实现了一次根本性的范式转换：将一个结构化的数据建模问题巧妙地重构为一个 LLM 所擅长的序列到序列的文本生成任务。这一转换的实现依赖于序列化技术，即将表格中由行列构成的二维数据，连同其表头，一同转化为一维的文本字符串。一旦表格数据被表示为文本形式，生成过程便可以充分利用 LLM 强大的、在海量语料上预训练所获得的语言理解和生成能力。与仅学习数值间统计关系的传统生成模型不同，LLM 能够理解并利用特征名称的自然语言语义。因此，该机制的本质就是将序列化作桥梁，将问题域从数值空间迁移到语言空间，从而释放 LLM 在捕捉复杂依赖、理解语义上下文以及进行常识推理方面的巨大潜力。

3.3.2 关键技术

将大型语言模型应用于结构化表格数据的处理与生成，其首要且最关键的步骤是将表格的二维结构转换为模型能够理解的一维文本序列，这一过程被称为序列化。由于 LLM 的内在架构是为处理和生成顺序文本而设计的[50]，因此序列化远非简单的格式转换，它是一个直接影响信息保真度的核心环节。一个设计得当的序列化策略能够有效保留表格的结构、单元格语义以及列间复杂依赖关系，反之则可能导致关键信息丢失或引入干扰性的因素，从而严重制约模型在下游任务中的性能表现。因此，对现有序列化技术进行梳理与分析，是释放 LLM 在结构化数据领域潜力的基础。

目前，学术界与工业界主要探索了 4 种核心的序列化技术。1）分隔符分隔格式，以 CSV 和 TSV 为代表。这种方法将每一行数据直接转换为由逗号或制表符连接的字符串。其优点是序列效率高、文本简洁，但代价是丢失了列的明确语义信息，需要模型从上下文或额外的提示中进行推断，这在处理少样本表格分类任务时是一大挑战[51]。2）键值对格式，以 JSON 为典型代表。它将每一行转换为一个结构化对象。这种方式通过将值与列名显式绑定，极大地增强了语义保真度，并能很好地支持嵌套等复杂结构。3）标记语言格式，如 XML 或 HTML。它利用结构化标签来精确定义表格的二维布局。这种方法在保留表格物理结构方面效果较好，对需要理解单元格位置和跨度的任务尤为重要，早期的表格预训练模型如 TaBERT[52] 就验证了其有效性。4）自然语言描述，即将表格行转化为流畅的句子。这种文本化方法最符合 LLM 的预训练范式，能够利用强大的文本到文本模型来生成高度连贯的描述[53]，但其挑战在于冗余度高，且从生成的文本中准确解析出结构化数据非常困难。

在实际应用中，这些技术之间的选择是一个复杂的权衡过程。分隔符格式简洁高效但语义缺失；键值对格式在语义和结构间取得了良好平衡；标记语言格式能最精确地保留布局；自然语言描述最符合 LLM 的本性，但冗余度最高。因此没有任何一种策略能在所有场景下都表现最佳，最优选择高度依赖于具体任务、数据复杂度以及所用模型的特性[54]。此外，安全性也是一个不容忽视的考量，因为不安全的序列化实现可能为提示注入攻击提供可乘之机，恶意构造的单元格内容可能会劫持模型的行为，对下游应用构成威胁[55]。

综上所述，对序列化方法的深刻理解与审慎选择，是成功将 LLM 应用于表格数据分析与生成领域的基石，并为后续的提示工程与模型微调等高级技术奠定了关键基础。

3.3.3 基于微调的方法

微调（Fine-tuning）是将一个通用的预训练 LLM 适配到特定表格数据分布的直接方式。该方法在序列化后的表格数据上对 LLM 的全部或部分参数进行进一步训练，使其能够精确地学习目标数据的统计特征。GReaT[54] 是该领域的开创性工作之一，它通过在序列化后的表格数据上微调 GPT-2 等自回归模型来进行数据生成。REaLTabFormer[56] 在 GReaT 的基础上进行了扩展，使其能够处理关系型数据库中的多表生成任务。TabLLM[51] 同样采用微调范式，并系统性地评估了不同序列化方法对少样本分类性能的影响，验证了通过参数高效微调可以取得优于传统方法的性能。TabuLa[57] 在微调前增加了一个对比学习的预训练阶段，以帮助模型学习更有效的表格数据表示。TabMT[58] 采用基于 BERT 的掩码变换器设计，能够有效处理异构数据字段并处理缺失值问题。TAPTAP[59] 利用表格数据预训练来增强模型的表格预测能力，之后可以通过微调生成高质量数据以支持隐私保护、小样本、数据插补和不平衡分类等多种应用。HARMONIC[60] 构建了一个基于 K 近邻思想的指令微调数据集，旨在让 LLM 学习行与行之间的关系，从而在保护隐私的同时提升生成效果。DP-LLMTGen[61] 专注于隐私保护，它采用一个为表格数据专门设计的损失函数，并通过两阶段的微调过程，将差分隐私机制融入 LLM 训练中。Pred-LLM[62] 通过预训练 LLM 生成表格特征，并进一步提示 LLM 生成其标签。P-TA[63] 创新地引入了强化学习，利用类似 GAN 的机制来指导 LLM 微调，以优化生成数据的概率分布，使其更接近真实数据。AIGT[64] 是一种基于提示增强的方法，它利用元数据信息作为提示来生成合成表格数据集，并提出了长令牌分区算法，使其能够对任意规模的表格数据进行建模。

3.3.4 基于提示工程的方法

基于提示工程的上下文学习（In-context Learning，ICL）是另一种更轻量级的范式，它无需修改 LLM 的任何参数。这种方法通过构建一个包含任务描述和少量样本示例的提示来引导一个“冻结”的 LLM 直接生成新的数据。CLLM[65] 主要利用 GPT-4 等前沿 LLM 的先验知识，在小样本场景下进行数据增强。其核心是一个基于学习动态的策划机制，通过计算置信度和不确定性指标来过滤生成的样本，以确保最终获得高质量的数据集。LITO[66] 提出了一个用于表格数据过采样的框架。该框架通过逐步掩盖多数类样本中的重要特征并提示 Distill-GPT2 等模型进行插补，从而将多数类样本渐进式地转化为少数类样本。EPIC[67] 专注于解决类别不平衡问题，它通过在提示中构建平衡且分组的样本并采用独有的特征映射，来引导 GPT-3.5 等模型准确地为所有类别生成数据。OCTree[68] 利用 LLM 进行树状推理，通过优化特征生成来支持下游任务。

3.4 对比分析

传统方法的核心优势在于其数学基础扎实和结果可解释性强。例如，贝叶斯网络能够显式地刻画特征之间的条件

依赖,Copulas 能够灵活建模边际分布与联合分布之间的关系。这使得传统方法在对隐私敏感、需要透明推理的应用场景中具有一定价值。然而,这类方法通常依赖强先验假设,难以应对高维非线性依赖关系,且生成样本的多样性有限,在复杂表格任务中表现受限。

扩散模型作为近年来兴起的生成范式,在捕捉复杂分布和多模态特征方面展现出卓越能力。其通过逐步加噪和去噪的过程来学习数据分布,避免了模式崩溃等困扰 GAN 的方法学难题,生成样本在统计保真度和分布一致性上往往优于传统方法。然而,这类模型训练和采样过程的计算成本较高,生成效率相对不足,并且模型的黑箱特性使其可解释性有限。因此,扩散模型更适合在对数据质量要求极高的任务中使用,但其在实时性或资源受限的应用中存在瓶颈。

与上述两类方法不同,基于 LLM 的生成方法代表了一种范式转换,即通过序列化技术将二维表格数据转化为一维文本序列,使生成问题得以在语言建模框架下求解。这种方法不仅能够利用特征名称中的语义信息,还能通过提示工程实现灵活的条件控制,从而在小样本、跨领域迁移和语义约束较强的任务中展现独特优势。此外,LLM 能够结合外部知识和上下文进行推理,使得生成的结果在逻辑合理性和语义一致性方面优于纯粹的统计建模方法。然而,LLM 生成也面临"幻觉"问题,可能产生与事实不符或内部逻辑不一致的数据;同时,其对序列化策略和提示设计的依赖较强,缺乏严格的统计一致性保障。因此,将其语义优势与稳健的统计建模方法结合,是未来值得探索的方向。

结合以上不同方法的特点,在表格数据生成方法的选择过程中,应综合考量研究目标、数据特征以及资源约束等因素。若研究侧重于模型可解释性与计算效率,且数据维度较低或变量依赖关系相对简单,则基于统计建模的方法更具适用性;若任务要求生成结果在分布保真度和多样性方面达到较高水准,尤其是在高维复杂数据场景下,则基于扩散模型的生成方法是更优选择;若应用强调跨领域迁移、小样本学习,或在语义合理性与逻辑一致性上具有较高要求,则基于大语言模型的方法更具优势。此外,GAN 与 VAE 等深度生成模型可视为统计建模与扩散模型之间的折中方案,能够在一定程度上兼顾样本多样性与建模能力。

4 模型评估维度与核心指标

4.1 数据可用性

数据可用性旨在评估生成数据在支持下游任务、还原统计特性以及遵守领域知识等方面的有效性,主要包含以下 3 个子维度。

1)机器学习效率。这是衡量生成数据可用性最核心的指标。其评估遵循一个被广泛采用的协议,即在生成数据上训练,在真实数据上测试(Train on Synthetic, Test on Real, TSTR)[69]。具体流程是:首先,将真实数据集划分为训练集和测试集;然后,在训练集上训练生成模型以产生生成数据集;最后,使用生成数据集来训练一个下游预测模型,并在真实的测试数据集上评估其性能。如果基于生成数据训练的模型性能与基于真实数据训练的模型性能相当,则认为生成数据具有较高的机器学习效率。常用的评估指标包括分类任务中的 AUC(Area Under the Curve)和 F1 分数,以及回归任务中的 RMSE(Root Mean Square Error)。

2) 统计保真度。该维度衡量生成数据在多大程度上保留了真实数据的统计特性。评估通常从 3 个层次进行开展。(1)列级保真度:逐列比较真实数据与生成数据的边缘分布。对于数值型特征,常用柯尔莫哥洛夫-斯米尔诺夫检验(Kolmogorov-Smirnov Test, KST)[70]和 Wasserstein 距离;对于分类型特征,则常用全变差距离(Total Variation Distance, TVD)[71]和杰森-香农散度。(2)列间保真度:评估特征对之间的依赖关系是否被保留。常用方法是计算并比较真实数据与生成数据的相关性矩阵,例如使用皮尔森相关系数来衡量数值特征对之间的关系。(3)联合分布保真度:评估整体的联合分布相似性,这是最具挑战性的部分。常用的指标包括在模拟数据上计算的似然拟合分数,以及更通用的 α-Precision(衡量保真度)和 β-Recall(衡量多样性)。

3)知识对齐性。该维度评估生成数据是否遵守已知的领域知识或逻辑约束。例如,在医疗数据中,患者年龄不应为负数,或最低血红蛋白水平不应高于最高水平。常用的评估指标是约束违反率(Constraint Violation Rate, CVR),它计算了合成样本中违反至少一条预定义规则的样本所占的百分比。此外,还有约束违反覆盖率(Constraint Violation Coverage, CVC)等更细粒度的指标。

4.2 数据隐私安全

合成表格数据的另一重要目标是,在金融、医疗等隐私敏感领域生成数据时,避免从原始数据集中泄露可能识别个人身份的敏感信息。因此,必须通过隐私保护指标对生成数据进行评估,以确保实现隐私保护。这些指标旨在量化真实数据中的记录在多大程度上能从生成数据中被重新识别或推断。其设计目的是评估生成数据是否因直接模仿真实记录或通过间接统计推断而无意间泄露敏感信息。主要的隐私保护评估技术包括以下 4 类。

1)成员推断攻击:最经典的隐私攻击模型之一。攻击者试图判断某一个体的真实记录是否存在于用于训练生成模型的原始数据集中。一个高成功率的成员推断攻击表明,生成模型可能对训练数据产生了记忆或过拟合,从而存在隐私泄露风险。

2)属性推断攻击:在此场景下,攻击者在已知目标个体部分信息的情况下,试图利用生成数据集来推断其缺失的敏感属性。较高的推断准确率意味着较低的隐私保护水平,表明生成数据泄露了个体属性间的关联信息。

3)最近记录距离:计算每个合成样本与其在真实数据集中最近邻的距离。较小的 DCR(Distance to Closest Record)值可能意味着合成样本与真实记录过于相似,重识别风险较高。

4)最近邻距离比:该指标是 DCR 的扩展,它计算每个合成样本的最近邻与次近邻的距离之比。一个较低的比值表明合成样本周围的真实数据点分布较为密集,降低了其被唯一识别的风险。

上述指标均属于静态评估方法,主要用于分析生成数据

与训练数据之间的直接关联。然而,在实际应用中,攻击者可能结合外部公开数据或跨数据源信息进行动态攻击,从而增强重识别或属性推断能力。虽然本文未对这些动态场景进行实验评估,但相关研究表明,可通过差分隐私、私有知识聚合、潜在空间扰动或噪声注入等技术,对生成数据实施隐私增强,从而降低潜在泄露风险。

本节主要介绍了静态隐私评估指标及其概念应用,同时概述了可能的动态攻击场景及可选的隐私增强方法,为后续研究提供了方向和参考。

5 挑战与展望

尽管基于 LLM 的表格数据生成方法取得了显著进展,开辟了利用先验知识和语义理解能力的新范式,但该领域仍处于快速发展的早期阶段,面临着诸多独特的挑战。这些挑战不仅涉及模型本身的技术局限性,也关乎其在实际应用中的稳健性、可靠性和安全性。与此同时,这些挑战也为未来的研究指明了极具潜力的发展方向。

1)幻觉问题与事实一致性:LLM 固有的幻觉问题在表格数据生成中尤为突出。模型可能会生成在统计上看似合理,但与事实或常识相悖的数据。与仅学习统计分布的传统模型不同,LLM 被期望利用其世界知识生成逻辑上正确的数据,然而幻觉问题的存在使得这一优势难以完全保证。如何确保生成数据的事实一致性和知识对齐性,是当前 LLM 方法面临的核心挑战和未来需要研究的方向。

2)数值处理能力的局限性:LLM 本质上是为处理离散的文本符号而设计的,其对连续数值的理解和处理能力相对较弱,模型可能难以精确地复现复杂数值特征的非高斯分布、重尾或多模态特性。现有的方法通常将数值特征进行分位数离散化,转换为"高""中""低"等文本描述,但这可能导致精度损失,特别是在需要进行精细数值比较和趋势分析的金融、科学计算等领域。

3)数据隐私与模型记忆的权衡:虽然生成合成表格数据旨在保护用户隐私,但经过微调的大型模型存在记忆训练数据方面的风险。攻击者可能通过成员推断攻击等手段,从模型生成的内容中推断出原始训练集中的敏感信息。因此,如何在提升生成数据质量和可用性的同时,提供严格、可量化的隐私保障,是一个亟待解决的难题。尽管已有 DP-LLMTGen[61] 探索将差分隐私与 LLM 微调相结合,但如何在整个深度生成领域实现最先进的隐私保护,仍是活跃的研究前沿。

4)增强的知识对齐与逻辑约束:未来的研究需要开发更有效的机制,将领域知识和逻辑约束深度整合到 LLM 的生成过程中。虽然在历史上知识对齐受到的关注有限,但它正逐渐成为生成真实数据的关键因素,并有望成为该领域的基本要求。

5)模型可解释性:现有生成模型多为黑箱,其不透明的决策过程易继承并放大数据偏见,威胁公平性与可信度。未来研究的关键在于提升透明度,例如探索应用思维链等方法来揭示模型的推理过程,以确保其在关键领域应用的可靠性。

结束语 本文旨在回顾应对数据稀缺与隐私挑战的表格数据生成技术。其技术路径从早期的变分自编码器(VAEs)和生成对抗网络(GANs)演进至为解决其训练不稳定问题而兴起的扩散模型,后者在生成高保真度样本上优势显著。然而,这些模型仅模仿统计分布,因此最新的研究范式转向基于大型语言模型(LLMs)的方法,旨在利用其世界知识生成逻辑与语义更合理的合成数据。尽管 LLM 方法潜力巨大,但仍面临幻觉、数值处理能力有限及隐私安全等挑战。未来的研究需在知识对齐、模型可解释性等方面取得突破,以推动该领域的进一步发展。

参 考 文 献

[1] SHAILAJA K, SEETHARAMULU B, JABBAR M A. Machine learning in healthcare: A review [C]// 2018 2nd International Conference on Electronics, Communication and Aerospace Technology(ICECA). IEEE, 2018: 910-914.

[2] CAO L. AI in finance: challenges, techniques, and opportunities [J]. ACM Computing Surveys, 2022, 55(3): 1-38.

[3] COMBRINK H M E, MARIVATE V, ROSMAN B. Comparing synthetic tabular data generation between a probabilistic model and a deep learning model for education use cases [J]. arXiv: 2210.08528, 2022.

[4] SUN C, LI S, CAO D, et al. Tabular learning-based traffic event prediction for intelligent social transportation system [J]. IEEE Transactions on Computational Social Systems, 2022, 10(3): 1199-1210.

[5] LI L, FAN Y, TSE M, et al. A review of applications in federated learning [J]. Computers & Industrial Engineering, 2020, 149: 106854.

[6] ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes: Theory and implementation [J]. ACM Computing Surveys, 2018, 51(4): 1-35.

[7] FRIEDMAN A, SCHUSTER A. Data mining with differential privacy [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010: 493-502.

[8] KINGMA D P, WELLING M. Auto-encoding variational Bayes [J]. arXiv: 1312.6114, 2013.

[9] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. 2014: 2672-2680.

[10] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics [C]// International Conference on Machine Learning. PMLR, 2015: 2256-2265.

[11] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [J]. arXiv: 2303.18223, 2023.

[12] SKLAR A. Random variables, joint distribution functions, and copulas [J]. Kybernetika, 1973, 9(6): 449-460.

[13] RABAEY P, DELEU J, HEYTENS S, et al. Clinical reasoning over tabular data and text with bayesian networks [C]// International Conference on Artificial Intelligence in Medicine. Cham: Springer, 2024: 229-250.

[14] RUBIN D B. Statistical disclosure limitation [J]. Journal of Offi-

cial Statistics,1993,9(2):461-468.

[15] YOUNG J,GRAHAM P,PENNY R. Using Bayesian networks to create synthetic data [J]. Journal of Official Statistics,2009,25(4):549-567.

[16] MARTINS L N A,GONÇALVES F B,GALLETTI T P. Generation and analysis of synthetic data via Bayesian networks:a robust approach for uncertainty quantification via Bayesian paradigm [J]. arXiv:2402.17915,2024.

[17] SKLAR M. Fonctions de répartitionàn dimensions et leurs marges [J]. Annales de l'ISUP,1959,8(3):229-231.

[18] EMBRECHTS P,MCNEIL A,STRAUMANN D. Correlation and dependence in risk management:properties and pitfalls [M]// Risk Management:Value at Risk and Beyond. 2002:176-223.

[19] RESTREPO J P. Nonparametric generation of synthetic data using copulas [J]. Electronics,2023,12(7):1601.

[20] JUTRAS-DUBÉ P,AL-KHASAWNEH M B,YANG Z C,et al. Copula-based synthetic population generation [J]. arXiv:2302.09193,2023.

[21] KAMTHE S,ASSEFA S,DEISENROTH M. Copula flows for synthetic data generation [J]. arXiv:2101.00598,2021.

[22] CHAWLA N V,BOWYER K W,HALL L O,et al. SMOTE:synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research,2002,16:321-357.

[23] GOODMAN N R. Statistical analysis based on a certain multivariate complex Gaussian distribution(an introduction) [J]. The Annals of Mathematical Statistics,1963,34(1):152-177.

[24] ALEMI A,POOLE B,FISCHER I,et al. Fixing a broken ELBO [C]// International Conference on Machine Learning. PMLR,2018.

[25] XU L,SKOULARIDOU M,CUESTA-INFANTE A,et al. Modeling tabular data using conditional GAN [C]// Proceedings of the 33rd Internationa Conference on Neural Information Processing Systems. 2019:7335-7345.

[26] MA C,TSCHIATSCHEK S,TURNER R,et al. VAEM:a deep generative model for heterogeneous mixed type data [C]// Advances in Neural Information Processing Systems. 2020:11237-11247.

[27] LIU T,QIAN Z,BERREVOETS J,et al. Goggle:Generative modelling for tabular data by learning relational structure [C]// The 11th International Conference on Learning Representations. 2023.

[28] XU L,VEERAMACHANENI K. Synthesizing tabular data using generative adversarial networks [J]. arXiv:1811.11264,2018.

[29] PARK N,MOHAMMADI M,GORDE K,et al. Data synthesis based on generative adversarial networks [J]. arXiv:1806.03384,2018.

[30] ZHAO Z,BIRKE R,CHEN L Y. FCT-GAN:Enhancing table synthesis via fourier transform [J]. arXiv:2210.06239,2022.

[31] RAJABI A,GARIBAY O O. Tabfairgan:Fair tabular data generation with generative adversarial networks [J]. Machine Learning and Knowledge Extraction,2022,4(2):488-501.

[32] MIYATO T,KOYAMA M. cGANs with projection discriminator [J]. arXiv:1802.05637,2018.

[33] LIN Z,KHETAN A,FANTI G,et al. PacGAN:The power of two samples in generative adversarial networks [J]. IEEE Journal on Selected Areas in Information Theory,2020,1(1):324-335.

[34] XIE L,LIN K,WANG S,et al. Differentially private generative adversarial network [J]. arXiv:1802.06739,2018.

[35] JORDON J,YOON J,VAN DER SCHAAR M. PATE-GAN:Generating synthetic data with differential privacy guarantees [C]// International Conference on Learning Representations. 2018.

[36] HO J,JAIN A,ABBEEL P. Denoising diffusion probabilistic models [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020:6840-6851.

[37] KOTELNIKOV A,BARANCHUK D,RUBACHEV I,et al. Tabddpm:Modelling tabular data with diffusion models [C]// International Conference on Machine Learning. PMLR,2023:17564-17579.

[38] SHI J,XU M,HUA H,et al. Tabdiff:a multi-modal diffusion model for tabular data generation [J]. arXiv:2410.20626,2024.

[39] LEE C,KIM J,PARK N. CODI:Co-evolving contrastive diffusion models for mixed-type tabular synthesis [C]// International Conference on Machine Learning. PMLR,2023:18940-18956.

[40] SUH N,LIN X,HSIEH D Y,et al. Autodiff:combining auto-encoder and diffusion model for tabular data synthesizing [J]. arXiv:2310.15479,2023.

[41] LIN X,XU C,YANG M,et al. Ctsyn:A foundational model for cross tabular data generation [J]. arXiv:2406.04619,2024.

[42] CERITLI T,GHOSHEH G O,CHAUHAN V K,et al. Synthesizing mixed-type electronic health records using diffusion models [J]. arXiv:2302.14679,2023.

[43] HE H,HAO W,XI Y,et al. A Flexible Generative Model for Heterogeneous Tabular {EHR} with Missing Modality [C]// The 12th International Conference on Learning Representations. 2024.

[44] SATTAROV T,SCHREYER M,BORTH D. Findiff:Diffusion models for financial tabular data generation [C]// Proceedings of the 4th ACM International Conference on AI in Finance. 2023:64-72.

[45] SCHREYER M,SATTAROV T,SIM A,et al. Imb-FinDiff:Conditional Diffusion Models for Class Imbalance Synthesis of Financial Tabular Data [C]// Proceedings of the 5th ACM International Conference on AI in Finance. 2024:617-625.

[46] KIM J,LEE C,SHIN Y,et al. Sos:Score-based oversampling for tabular data [C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022:762-772.

[47] OUYANG Y,XIE L,LI C,et al. Missdiff:Training diffusion models on tabular data with missing values [J]. arXiv:2307.00467,2023.

[48] JOLICOEUR-MARTINEAU A,FATRAS K,KACHMAN T. Generating and imputing tabular data via diffusion and flow-

based gradient-boosted trees [C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2024: 1288-1296.
[49] ZHANG H, ZHANG J, SRINIVASAN B, et al. Mixed-type tabular data synthesis with score-based diffusion in latent space [J]. arXiv: 2310.09656, 2023.
[50] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 1877-1901.
[51] HEGSELMANN S, BUENDIA A, LANG H, et al. Tabllm: Few-shot classification of tabular data with large language models [C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2023: 5549-5581.
[52] YIN P, NEUBIG G, YIH W, et al. TaBERT: Pretraining for joint understanding of textual and tabular data [J]. arXiv: 2005.08314, 2020.
[53] KALE M, RASTOGI A. Text-to-text pre-training for data-to-text tasks [J]. arXiv: 2005.10433, 2020.
[54] BORISOV V, SEßLER K, LEEMANN T, et al. Language models are realistic tabular data generators [J]. arXiv: 2210.06280, 2022.
[55] GRESHAKE K, ABDELNABI S, MISHRA S, et al. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection [C]// Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. 2023: 79-90.
[56] SOLATORIO A V, DUPRIEZ O. Realtabformer: Generating realistic relational and tabular data using transformers [J]. arXiv: 2302.02041, 2023.
[57] ZHAO Z, BIRKE R, CHEN L Y. Tabula: Harnessing language models for tabular data synthesis [C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore: Springer, 2025: 247-259.
[58] GULATI M, ROYSDON P. TabMT: Generating tabular data with masked transformers [C]// Advances in Neural Information Processing Systems. 2023: 46245-46254.
[59] ZHANG T, WANG S, YAN S, et al. Generative table pre-training empowers models for tabular prediction [J]. arXiv: 2305.09696, 2023.
[60] WANG Y, FENG D, DAI Y, et al. HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection [C]// Advances in Neural Information Processing Systems. 2024: 100196-100212.
[61] TRAN T V, XIONG L. Differentially private tabular data synthesis using large language models [J]. arXiv: 2406.01457, 2024.
[62] NGUYEN D, GUPTA S, DO K, et al. Generating realistic tabular data with large language models [C]// 2024 IEEE International Conference on Data Mining (ICDM). IEEE, 2024: 330-339.
[63] YANG S, YUAN C, RONG Y, et al. P-ta: Using proximal policy optimization to enhance tabular data augmentation via large language models [J]. arXiv: 2406.11391, 2024.
[64] ZHANG M, XIAO Z, LU G, et al. Aigt: AI generative table based on prompt [J]. arXiv: 2412.18111, 2024.
[65] SEEDAT N, HUYNH N, VAN BREUGEL B, et al. Curated LLM: Synergy of LLMs and data curation for tabular augmentation in low-data regimes [J]. arXiv: 2312.12112, 2023.
[66] YANG J Y, PARK G, KIM J, et al. Language-interfaced tabular oversampling via progressive imputation and self-authentication [C]// The Twelfth International Conference on Learning Representations. 2024.
[67] KIM J, KIM T, CHOO J. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models [C]// Advances in Neural Information Processing Systems. 2024: 31504-31542.
[68] NAM J, KIM K, OH S, et al. Optimized feature generation for tabular data via llms with decision tree reasoning [C]// Advances in Neural Information Processing Systems. 2024: 92352-92380.
[69] FEKRI M N, GHOSH A M, GROLINGER K. Generating energy data for machine learning with recurrent generative adversarial networks [J]. Energies, 2019, 13(1): 130.
[70] BERGER V W, ZHOU Y Y. Kolmogorov-smirnov test: Overview [EB/OL]. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06558.
[71] TAO L, XU S, WANG C H, et al. Discriminative estimation of total variation distance: A fidelity auditor for generative data [J]. arXiv: 2405.15337, 2024.

WANG Yongxin, born in 2001, postgraduate. His main research interest is tabular data generation.

ZHU Hongbin, born in 1991, assistant professor, is a member of CCF (No. Q5992M). His main research interests include generative AI, graph learning and financial technology.

(责任编辑:李亚辉)