第 62 卷　第 2 期
2026 年 3 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 2
Mar., 2026

DOI:10.13232/j.cnki.jnju.2026.02.010

基于LLM概率提示词的表格数据生成方法

张　爽[1],房　俊[1,2*],欧阳琛[1]

(1.北方工业大学信息学院,北京,100144;
2.大规模流数据集成与分析技术北京市重点实验室,北方工业大学,北京,100144)

摘　要:大语言模型(Large Language Model,LLM)在生成表格数据任务中展现出巨大潜力,但其生成的数据往往难以准确保持数据列间的依赖关系.针对该问题,提出一种基于LLM概率提示词的方法TabProLLM,分别生成表格数据的数值列和分类列.使用高斯混合模型(Gaussian Mixture Model,GMM)切分数值列的概率密度曲线,将其划分为多个正态分布,并基于划分后的正态分布构造概率提示词用于大模型生成数值列数据.对于分类列,以某一数值列为基准进行分区,计算分类列中各类别在不同数值区间的条件概率分布,并根据条件概率分布生成提示词用于生成分类列数据.在提示词生成过程中,还引入相关系数等指标,用于校验生成数据中变量间的依赖关系是否符合原始数据的相关性模式.在10个公开数据集上的实验结果表明,TabProLLM在保证数据隐私性的同时,在SDMetrics工具中的RangeCoverage,CategoryCoverage,KSComplement,TVComplement等多个保真度评估指标上实现了18%左右的性能提升.其相关性指标CorrelationSimilarity与最优模型TabDDPM基本持平,和GPT-4o使用均值方差提示词方法相比,提升约4.1%.同时,在隐私性评估方面,TabProLLM的*DCR*和*NNDR*(取第5百分位数)指标整体表现为最优和次优.

关键词:表格数据生成,大语言模型,提示词,条件概率

中图分类号:TP18　　**文献标志码**:A

A method for generating tabular data based on LLM prompt words

Zhang Shuang[1], Fang Jun[1,2*], Ouyang Chen[1]

(1. School of Information, North China University of Technology, Beijing, 100144, China;
2. Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data,
North China University of Technology, Beijing, 100144, China)

Abstract: Large Language Model (LLM) have demonstrated significant potential in tabular data generation. However, they often struggle to accurately preserve the statistical dependencies between columns. To address this challenge, we propose TabProLLM, a probabilistic prompting framework that separately generates numerical and categorical columns using strategies grounded in probability distributions. For numerical columns, we fit a Gaussian Mixture Model (GMM) to decompose the empirical distribution into multiple Gaussian components. Prompts are then constructed based on these segmented distributions to guide the LLM in generating realistic numerical values. For categorical columns, we condition on a reference numerical column by partitioning its range and computing the conditional probability distribution of each category within each interval. These conditional probabilities are embedded into the prompt design to steer the generation of categorical data consistent with observed inter-variable dependencies. During prompt construction, correlation coefficients and other statistical measures are incorporated to verify that the generated data preserves the correlation structure of the original dataset. Experimental results on 10 public datasets show that TabProLLM, while ensuring strong data privacy, achieves

基金项目:国家重点研发计划(2023YFC3107900),国家自然科学基金(72272140)
收稿日期:2025－11－21
*通信联系人,E-mail:fangjun@ncut.edu.cn

performance gains of 0.5% to 18.3% over existing methods across multiple fidelity metrics in the SDMetrics toolkit, including RangeCoverage, CategoryCoverage, KSComplement, and TVComplement. On the CorrelationSimilarity metric, TabPro-LLM performs comparably to the state-of-the-art TabDDPM model and surpasses GPT-4o (using mean-variance prompts) by approximately 4.1%. Furthermore, in privacy evaluations, TabProLLM achieves top or second-best performance across *DCR* and *NNDR* metrics (evaluated at the 5th percentile), highlighting its robust privacy-preserving capabilities.

Keywords: tabular data generation, large language model, prompt words, conditional probability

表格数据作为结构化数据的一种核心形式，在金融风控、医疗健康、电子商务等多个实际场景中具有重要的应用价值[1]. 在实际应用中，获取真实且可用的表格数据面临诸多挑战，包括数据隐私保护、样本数量有限、类别分布不均衡等问题，限制了数据驱动方法的应用[2]. 为了缓解上述问题，近年来学术界广泛探索表格数据生成方法，旨在不泄露真实信息的前提下生成结构合理且具备统计特性的替代数据.

随着大语言模型(Large Language Model, LLM)的快速发展，其强大的语言理解与生成能力为表格数据生成开辟了新的技术路径[3]. LLM能够通过自然语言提示灵活生成结构化信息，显著降低了模型设计与训练成本[3]，但基于LLM的表格生成方法在保持多列之间的联合分布结构方面仍存在不足，尤其在数据列与列间的依赖关系建模方面表现不佳[4]，导致生成数据列间的依赖关系与原始数据不一致，影响其在后续任务中的可用性.

针对当前LLM表格数据生成方法在变量间依赖关系建模方面存在的准确性不足问题，本文提出了一种基于大语言模型概率提示词的表格数据生成方法(Tabular Data Generation via Probabilistic Prompting with LLM, TabProLLM)，构建了面向数值型与分类变量的两种提示词生成框架.

现有的LLM表格生成方法多依赖直接示例提示或统计描述提示，而本文提出的方法将概率分布结构显式编码为提示词，通过高斯混合模型(Gaussian Mixture Model, GMM)概率切分与条件概率提示，使LLM能够更精确地捕捉变量间的复杂依赖关系. 这一概率提示框架为解决LLM生成表格数据时联合分布保持不佳的问题提供了新路径.

1 相关工作

表格数据生成技术的发展经历了从传统统计模型到深度生成模型的持续演进. 早期方法多依赖于统计建模技术，如Copula函数[5]和贝叶斯网络[6]，在建模变量间依赖关系方面具有一定优势，但在面对多模态、长尾或高度非线性分布的数据时，其生成质量往往较差，难以满足实际应用需求[7].

深度学习技术的发展，使GAN (Generative Adversarial Networks)，VAE (Variational Auto-encoder)以及扩散模型等方法广泛用于表格数据生成任务. 例如，CTGAN[8]通过引入模式特定归一化技术，增强了模型对非高斯分布数据的建模能力，CTAB-GAN+[9]通过改进的条件生成机制有效缓解了类别不平衡问题. 这类模型通常依赖大量的数据预处理步骤，容易引发信息损失和维度灾难[10]. TabDDPM[11]将扩散过程引入结构化数据建模，通过噪声扰动与逐步去噪过程有效逼近复杂数据分布，其采用对连续和分类变量分别建模的策略，易破坏变量间的联合依赖结构，导致生成数据在保持多变量相关性方面仍存在不足. 此外，深度学习模型通常需要进行大量训练，过程耗时且计算资源消耗较高.

近年来，大语言模型的生成能力为表格数据生成提供了全新范式，相关研究主要分为两类路径：微调式方法与提示驱动式方法. 微调方法，如GReaT[12]基于预训练语言模型进行有监督的任务适配，通过“特征名 is 特征值”的方式将表格数据序列化为自然语言，避免了特征预定义和预处理. 尽管通过微调的方法取得了一定成效，但其普遍存在训练开销大、开发周期长等问题[13]. 相比之下，提示词方法因无需训练和对模型参数进行修改，具备零样本或少样本适应性. 典型方法

如Barr et al[14]提出的基于GPT-4o的大模型零样本表格数据生成框架，无需预训练和微调，仅通过自然语言提示（包含数据的统计特性，如均值、标准差、相关性等）即可实现高保真的数据生成．但GPT-4o对于偏态和非正态数据，模型生成的样本难以精确捕捉原始数据的分布细节，而且，有时会出现数值违反实际定义边界（如负值）的情况．TabLLM[15]等将表格列结构编码为自然语言提示来引导模型生成整行或单列数据，展现出良好的灵活性．多数提示词方法仍偏重“直接示例提示”或“固定模板提示”，未能深入融合数据的概率结构信息，导致模型在复杂变量依赖结构保持上的表现有明显不足．综上，现有的基于LLM提示词的表格数据生成方法虽然避免了训练成本，但其提示词设计多停留在统计描述层面，如均值、标准差，未能深入编码变量间的概率依赖结构．因此，将概率语义显式融入提示词以引导LLM精确生成符合联合分布的数据，成为当前关键挑战．

2　方法

TabProLLM方法主要分为三个阶段：数值列建模与提示词生成、分类列建模与提示词生成以及数据生成与校验．整体流程如图1所示．

2.1　数值列建模与提示词生成　实验发现LLM在生成表格数据时，更擅长还原单峰且对称的正态分布，而具有偏态或多峰特征的复杂分布往往难以准确复现[16]．为此，引入GMM将每个数值列整体分布拆解为多个近似正态子分布．如图1a所示，使用GMM对每个数值列的概率密度函数（PDF）进行建模，将复杂的概率密度曲线拆解为若干局部的正态分布，每个分布具有独立的均值和方差参数，对偏态、多峰等复杂分布形态进行精细建模．给定数值列x，其GMM形式为：

$$P(x)=\sum_{k=1}^{N}\pi_k N\left(x\mid\mu_k,\sigma_k^2\right) \tag{1}$$

其中，π_k为划分后第k个正态子分布的权重，μ_k和σ_k^2分别为该分布的均值与方差．

为了避免各列独立生成引发的结构性失真，提升生成数据的多维结构一致性，进一步引入原始数值列间的相关系数矩阵作为先验知识，共同构建自然语言提示词．如图2所示，数值列提示词主要由统计属性、相关性结构信息和生成约束三部分组成．统计属性部分描述各变量的分布形式，如每列由若干个正态子分布（均值、方差及权重）构成；相关性结构部分通过计算皮尔逊相关系

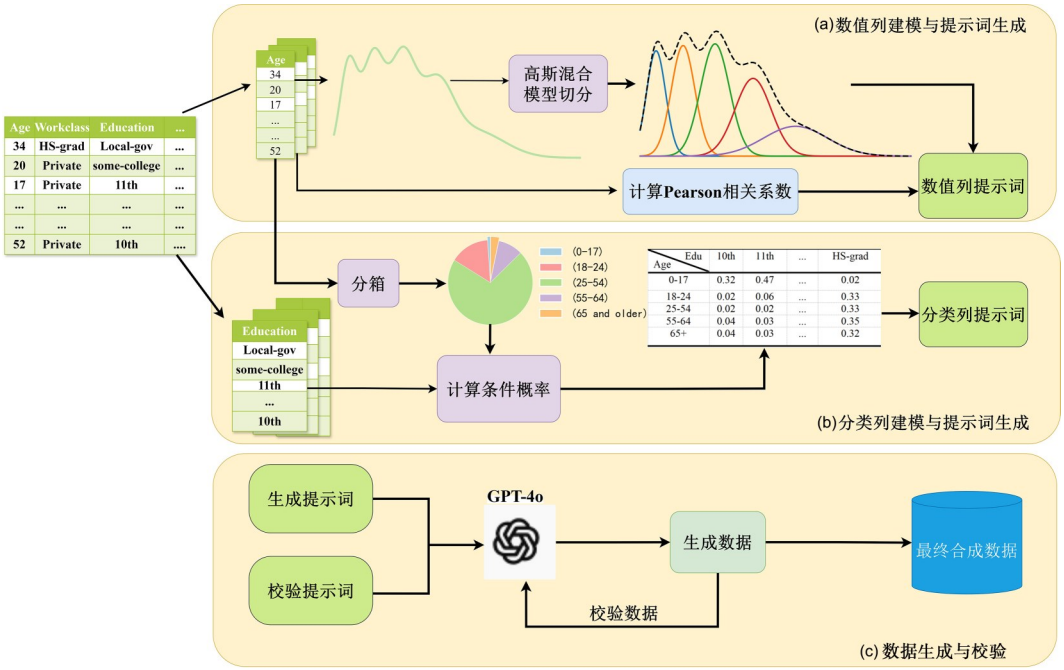

图1　TabProLLM方法的整体框架图

Fig. 1　Overall framework of the TabProLLM method

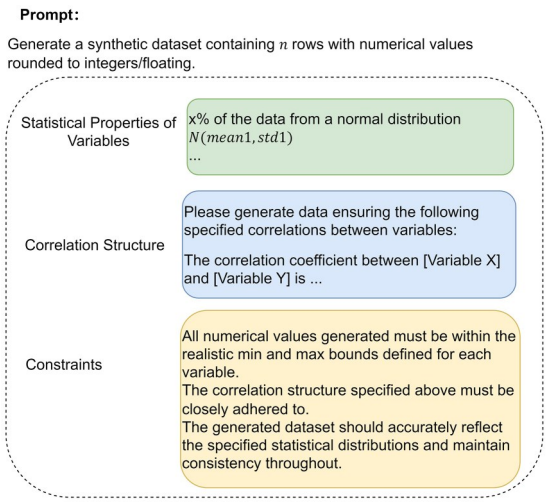

图 2　数值列的提示词模板

Fig. 2　Prompt template for numerical columns

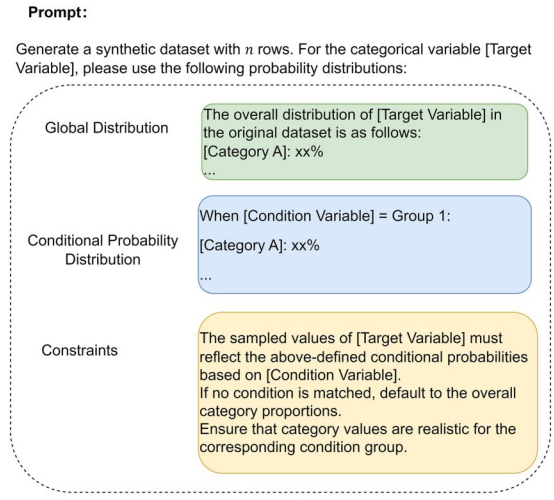

图 3　分类列的提示词模板

Fig. 3　Prompt template for categorical columns

数明确各列间的依赖关系;生成约束部分则确保所有生成的数值在各自定义的上下限区间内,并严格符合前述的相关结构和统计特性.

2.2　分类列建模与提示词生成　分类列的生成基于数值列作为条件变量进行建模,以保持分类变量与数值变量之间的潜在依赖结构(图 1b). 具体步骤如下.

(1)基准列选择与分箱:选择生成的一个数值列作为条件变量,根据其数值范围进行分箱(等宽分箱或者根据数据列的实际意义),划分为多个区间 $B_1, B_2, \cdots, B_k$.

(2)条件概率估计:对于每个分类列 C 和每个数值区间 B_i,计算其条件概率:

$$P\left(c_j \middle| x \in B_i\right) = \frac{\sum_{n=1}^{N} I\left(x_n \in B_i\right) \cdot I\left(c_n = c_j\right)}{\sum_{n=1}^{N} I\left(x_n \in B_i\right)} \tag{2}$$

其中,N 为数据样本总数,$I(\cdot)$ 为指示函数,条件满足返回 1,不满足返回 0,$c_j \in C$ 为某一具体分类的取值.

(3)提示词构造与生成:融合当前分类列的整体类别分布及其基于条件变量的条件概率分布. 分类变量的提示词结构同样由三部分组成:全局分布、条件概率分布与生成约束(图 3).

2.3　数据生成与校验　在确保生成数据的统计分布特征与取值范围严格符合预定义要求的基础上,进一步对列与列之间的相关系数进行一致性校验,引入相关系数差异阈值作为判定标准,当生成数据中的相关系数与目标值之间的差异超过该阈值时,视为不符合要求. 该阈值作为一个可调节的超参数,其设定对生成结果的结构一致性具有重要影响,具体的调优策略见 3.2. 校验的提示词模板如图 4 所示.

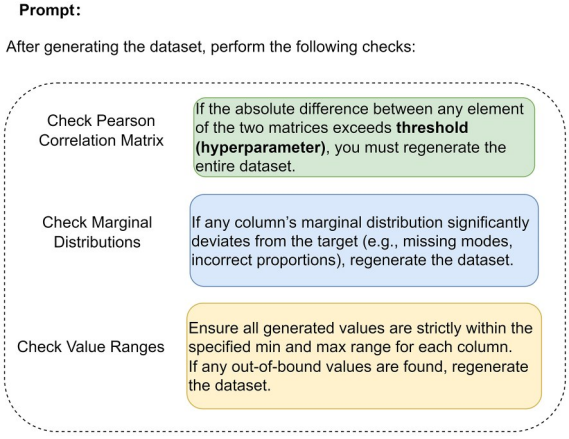

图 4　校验的提示词模板

Fig. 4　Validation prompt template

3　实验

3.1　数据集选取　选取表格数据生成领域的 10 个具有代表性且公开可获取的表格数据集,涵盖分类与回归两类主流任务类型,涉及医疗、金融、房地产、交通等多个领域,具有良好的多样性和通用性. 数据集的详细信息如表 1 所示.

3.2　超参数调优　为了确定校验步骤中相关系数差异阈值的最优参数,选择在 0.01～0.20 进行

表1　数据集的详细信息

Table 1　Detailed information of the datasets

数据集	样本量	类型	数值列数	分类列数
Adult	48842	分类	6	8
Iris	150	分类	4	1
Diabetes	768	分类	8	1
California House	20640	回归	8	0
Abalone	4177	回归	7	1
Fish Measurements	85	回归	6	2
German Credit	1000	分类	7	13
Heart Disease	303	分类	8	5
Real Estate Valution	414	回归	6	0
Winequality Red	1599	回归	11	0

探索，以0.05为间隔设置了一系列中间阈值，以深入分析和详细评估不同容忍度对于生成次数、时间成本和数据质量的具体影响．最终选取0.1作为阈值来进行校验，调优结果如表2所示．

3.3　对比方法　选取四种代表性生成模型作为对比基线，涵盖GAN，VAE，Diffusion Model和LLM驱动方法，具体包括CTGAN，TVAE，TabDDPM以及Barr et al[14]的基于GPT-4o的LLM提示词生成方法，代表当前表格数据生成领域中不同建模范式的最新进展．

3.4　评估指标　借助Neha Patki的表格数据生成开源评估工具包SDMetrics[17]，其详细信息如表3所示．从保真度（Fidelity）、相关性（Correlation）和隐私性（Privacy）三个维度对生成结果进行量化分析，具体如下．

（1）保真度．除了计算SDMetrics中多个常用指标外，还采用核密度估计图（KDE图）[18]展示数值变量的边缘分布拟合情况，通过对比原始数据

表2　阈值的调优结果

Table 2　Experimental results of threshold tuning

阈值	0.01	0.05	0.10	0.15	0.20
生成次数	20	19	1	1	2
所需时间	0.73	1.07	0.07	0.25	0.10
实验结果（平均相关系数差异）	0.02	0.02	0.02	0.21	0.08

与生成数据的KDE曲线，观察TabProLLM方法在处理高斯混合分布、多峰分布等复杂分布形态时的拟合效果．

（2）相关性．使用SDMetrics工具中的CorrelationSimilarity指标来评估各方法在保持生成数据中变量间相关性结构方面的能力．

（3）隐私性．使用NewRowSynthesis、距离最近纪录（Distance to Closest Record，*DCR*）和最近邻距离比（Nearest Neighbour Distance Ratio，*NNDR*）三项指标．参考Zhao et al[19]的做法，对于*DCR*和*NNDR*指标，分别取其第5百分位数作为评估值，以增强隐私性分析的鲁棒性．

3.5　实验结果　为了避免偶然性影响实验结果，所有方法均在五个公开数据集上分别重复进行五次独立实验，并对各项评估指标取平均值，确保评估结果的稳健性和代表性．各方法的综合实验结果如表4所示，表中向上的箭头表示该项指标数值越高越好，黑体字表示性能最优．

由表可见，TabProLLM在保真度与隐私性两方面均展现出明显优势．

（1）保真度．TabProLLM在SDMetrics中各指标上均接近于1，而且，其大多数指标都优于其他模型，展现了更好的全局分布拟合能力．在KDE图中（以Diabetes数据集为例，如图5所示），

表3　评估指标

Table 3　Evaluation metrics

	指标	适用数据类型	指标说明
Fidelity	RangeCoverage	数值型	评估生成数据是否覆盖真实数据的最小值与最大值范围
	CategoryCoverage	分类型	检查生成数据是否包含了真实数据中的所有类别
	KSComplement	数值型	使用Kolmogorov-Smirnov检验对比真实与生成数据的分布差异，值越高表示越相似
	TVComplement	分类型	使用全变差距离评估生成数据与真实数据在类别分布上的差异，值越大越好
Correlation	CorrelationSimilarity	数值型、分类型	比较生成数据与真实数据中变量间的相关性强度与方向
Privacy	NewRowSynthesis	所有类型	测量生成数据中是否出现与真实数据完全重复的行，用于隐私评估

表 4 实验结果

Table 4 Experimental results

	指标	CTGAN	TVAE	TabDDPM	GPT-4o	TabProLLM
Fidelity	RangeCoverage(↑)	0.931±0.003	**0.966±0.021**	0.469±0.004	0.947±0.002	0.953±0.003
	CategoryCoverage(↑)	0.945±0.032	0.940±0.202	0.957±0.071	0.891±0.105	**0.998±0.012**
	KSComplement(↑)	0.817±0.022	0.956±0.011	0.951±0.050	0.787±0.018	**0.977±0.006**
	TVComplement(↑)	0.958±0.031	0.874±0.021	0.951±0.014	0.815±0.014	**0.963±0.201**
Correlation	CorrelationSimilarity(↑)	0.911±0.001	0.886±0.012	**0.986±0.029**	0.947±0.011	0.982±0.03
Privacy	NewRowSynthesis(↑)	1±0	1±0	1±0	1±0	1±0
	DCR(↑)	0.338±0.001	0.243±0.103	0.375±0.079	0.167±0.051	**0.421±0.208**
	NNDR(↑)	0.211±0.005	0.407±0.012	**0.566±0.022**	0.184±0.001	0.545±0.013

TabProLLM 生成的数值列分布曲线与真实数据高度重叠,能复现原数据中的多模态特征、主峰位置及分布形态,表明其边缘分布拟合能力较强.

(2)相关性. TabProLLM 的 CorrelationSimilarity 达到 0.982,与 TabDDPM (0.986)接近,优于 GPT-4o (0.947)约 4.1%. 各数据集在不同模型下的 CorrelationSimilarity 如表 5 所示,表中黑体字表示性能最优. 由表可见,TabProLLM 在绝大多数数据集上的相关性都优于其他生成模型.

(3)隐私性. 模型的 NewRowSynthesis 都稳定为 1,TabProLLM 的平均 *DCR* (0.421)指标最优,其 *NNDR* 表现次优,进一步体现出较高的数据新颖性与隐私保护能力. 各数据集详细的 *DCR* 和 *NNDR* 如表 6 所示,表中黑体字表示性能最优.

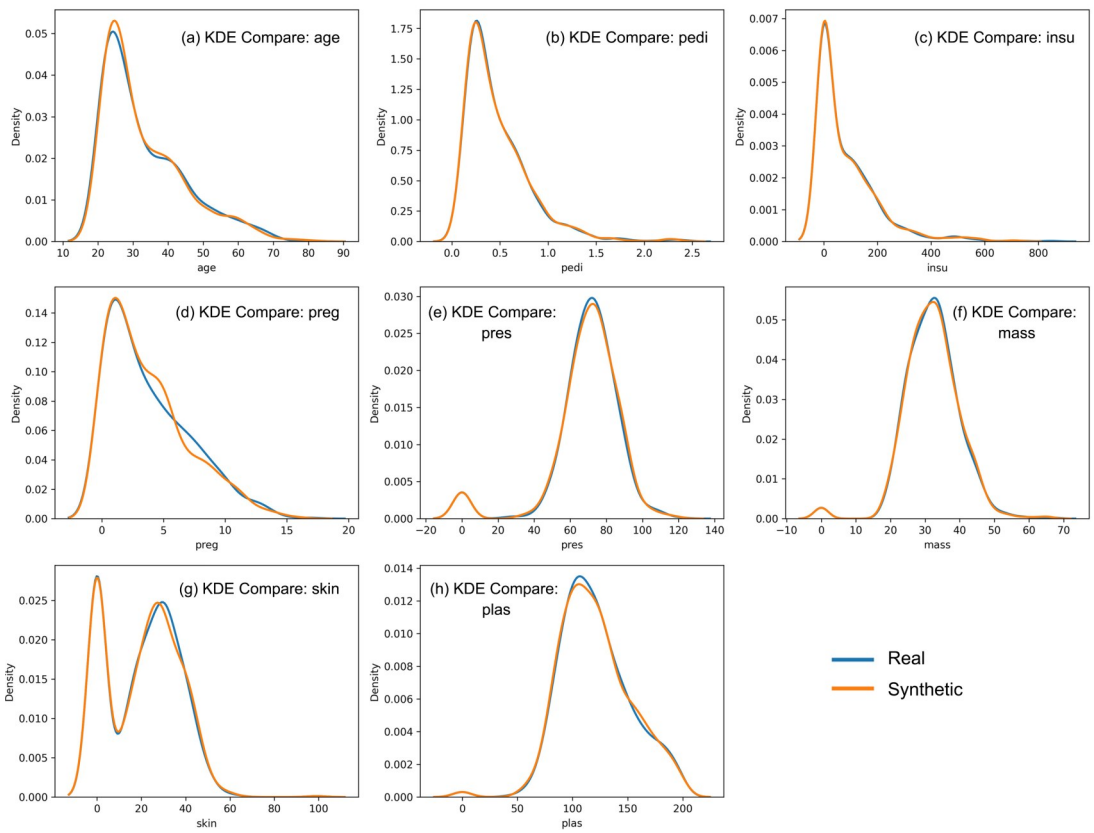

图 5 Diabetes 数据集真实数据与生成数据 KDE 图

Fig. 5 KDE plot of real vs. generated data on the Diabetes dataset

表5　CorrelationSimilarity的值

Table 5　CorrelationSimilarity values

	CTGAN	TVAE	TabDDPM	GPT-4o	TabProLLM
Adult	0.910	0.895	**0.999**	0.915	0.991
Iris	0.925	0.902	0.965	0.928	**0.973**
Diabetes	0.932	0.910	0.968	0.934	**0.978**
California House	0.899	0.899	**0.998**	0.918	0.970
Abalone	0.915	0.914	0.998	0.929	**0.999**
Fish Measurements	0.918	0.908	0.976	0.922	**0.980**
German Credit	0.920	0.893	0.981	0.931	**0.986**
Heart Disease	0.935	0.920	**0.992**	0.917	0.969
Real Estate Valution	0.905	0.905	0.979	0.935	**0.981**
Winequality Red	0.921	0.914	0.990	0.931	**0.993**

表6　*DCR*与*NNDR*的值(取第5百分位数)

Table 6　*DCR* and *NNDR* values (at the 5th percentile)

	DCR					*NNDR*				
	CTGAN	TVAE	TabDDPM	GPT-4o	TabPro-LLM	CTGAN	TVAE	TabDDPM	GPT-4o	TabPro-LLM
Adult	0.070	**0.210**	0.207	0.012	0.003	0.132	**0.435**	0.132	0.102	0.143
Iris	**0.576**	0.085	0.110	0.045	0.121	0.155	0.371	0.402	0.112	**0.440**
Diabetes	0.420	0.034	0.031	0.025	**0.606**	0.146	0.318	0.626	0.095	**0.651**
California House	0.220	0.160	0.202	0.068	**0.295**	0.129	0.412	0.547	0.110	**0.705**
Abalone	0.370	**0.531**	0.122	0.059	0.216	0.140	0.390	0.529	0.116	**0.664**
Fish Measurements	0.136	0.079	**0.234**	0.112	0.141	0.269	0.459	**0.784**	0.211	0.503
German Credit	0.501	0.138	**0.715**	0.079	0.697	0.352	0.327	**0.612**	0.352	0.331
Heart Disease	0.288	0.268	0.896	0.757	**0.965**	0.507	0.552	0.535	0.147	**0.603**
Real Estate Valution	0.420	0.382	**0.435**	0.302	0.359	0.106	0.426	0.605	0.116	**0.652**
Winequality Red	0.379	0.543	0.798	0.211	**0.807**	0.173	0.378	**0.892**	0.480	0.756

4　结论

融合概率提示与混合分布建模的LLM表格数据生成方法TabProLLM能够有效生成边缘分布准确、变量间结构关系一致的高保真表格数据.当前工作主要关注对结构性表格数据本身的高质量生成,没有对标签信息进行显式建模,生成过程中没有包含标签列,也没有充分考虑生成数据对下游监督学习任务(如分类、回归)的有效性支持.此外,TabProLLM的提示词设计仍依赖于手工构建统计特征.未来将探索提示自动生成或学习优化策略,以提升方法的通用性与自动化水平.

参考文献

[1] Marcinkevičs R, Vogt J E. Interpretable and explainable machine learning: A methods‐centric overview with concrete examples. WIREs Data Mining and Knowledge Discovery, 2023, 13(3): e1493.

[2] Salmi M, Atif D, Oliva D, et al. Handling imbalanced medical datasets: Review of a decade of research. Artificial Intelligence Review, 2024, 57(10): 273.

[3] 魏博伦,张贤坤.面向扩散模型的电子健康档案数据生成研究综述.计算机应用研究,2024,41(12):3521－3532.

[4] Lu W Z, Zhang J, Fan J, et al. Large language model

for table processing: A survey. Frontiers of Computer Science, 2025, 19(2): 192350.

[5] Sidorenko A. A note on statistically accurate tabular data generation using large language models. https://arxiv.org/abs/2505.02659, 2025－05－06.

[6] Bourouis S. Recent advances in statistical mixture models: Challenges and applications// Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods. Lisbon, Spain: SciTePress, 2023: 312－319.

[7] Wang A X, Chukova S S, Simpson C R, et al. Challenges and opportunities of generative models on tabular data. Applied Soft Computing, 2024, 166: 112223.

[8] Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional GAN// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2019: 7335－7345.

[9] Zhao Z L, Kunar A, Birke R, et al. CTAB-GAN+: Enhancing tabular data synthesis. Frontiers in Big Data, 2024, 6: 1296508.

[10] Tazwar S M, Knobbout M, Quesada E H, et al. Tab-VAE: A novel VAE for generating synthetic tabular data// Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods. Lisbon, Spain: Science and Technology Publications, 2024: 17－26.

[11] Kotelnikov A, Baranchuk D, Rubachev I, et al. Tabddpm: Modelling tabular data with diffusion models// Proceedings of International Conference on Machine Learning. Cambridge, United Kingdom: PMLR, 2023: 17564－17579.

[12] Borisov V, Seßler K, Leemann T, et al. Language models are realistic tabular data generators. https://arxiv.org/abs/2210.06280, 2023－04－22.

[13] Han Z Y, Gao C, Liu J Y, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. https://arxiv. org/abs/2403.14608, 2024－09－16.

[14] Barr A A, Rozman R, Guo E. Generative adversarial networks vs large language models: A comparative study on synthetic tabular data generation. https://doi.org/10.48550/arXiv.2502.14523, 2025－02－20.

[15] Hegselmann S, Buendia A, Lang H, et al. Tabllm: Few-shot classification of tabular data with large language models// Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. Cambridge, United Kingdom: PMLR, 2023: 5549－5581.

[16] Recasens P G, Gutierrez A, Torres J, et al. In-context bias propagation in LLM-based tabular data generation. https://arxiv.org/abs/2506.09630, 2025－06－11.

[17] Patki N, Wedge R, Veeramachaneni K. The synthetic data vault// 2016 IEEE International Conference on Data Science and Advanced Analytics. Montreal, Canada: IEEE, 2016: 399－410.

[18] Sun C, Dumontier M. Generating unseen diseases patient data using ontology enhanced generative adversarial networks. NPJ Digital Medicine, 2025, 8(1): 4.

[19] Zhao Z L, Kunar A, Birke R, et al. CTAB-GAN: Effective table data synthesizing// Proceedings of the 13th Asian Conference on Machine Learning. Cambridge, United Kingdom: PMLR, 2021: 97－112.

（责任编辑 杨可盛）