计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330742
61(2): 324−337, 2024

基于扩散模型生成数据重构的客户流失预测

杨 斌[1] 王正阳[2] 程梓航[2] 赵慧英[1] 王 鑫[1] 管 宇[2,3] 程新洲[1]
[1](中国联通研究院 北京 100048)
[2](北京邮电大学人工智能学院 北京 100876)
[3](云南省软件工程重点实验室(云南大学) 昆明 650504)
(researcher_yang@outlook.com)

Customer Churn Prediction Based on Generation Data Reconstruction Using Diffusion Model

Yang Bin[1], Wang Zhengyang[2], Cheng Zihang[2], Zhao Huiying[1], Wang Xin[1], Guan Yu[2,3], and Cheng Xinzhou[1]
[1](*China Unicom Research Institute*, *Beijing* 100048)
[2](*School of Artificial Intelligence*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876)
[3](*Yunnan Key Laboratory of Software Engineering* (*Yunnan University*), *Kunming* 650504)

Abstract In the field of data mining, the issue of data imbalance impacting model prediction accuracy is widespread, and also the issue of user privacy protection is neglected. Fake dataset generation has come to light as a crucial remedy for these problems. However, because of the traits of high-dimensional and irrelevant features, it is difficult to generate high-quality data in circumstances where structured data predominate. Considering the successful applications of the diffusion model in image generation task, we aim to apply the diffusion model for the task of customer churn prediction, which is a typical scenario in data mining. we utilize the Gaussian diffusion model and polynomial diffusion model to generate data for numerical and categorical features in customer churn data. Research and analysis have been conducted on the predictive performance and data privacy protection capabilities of our model. We conduct extensive experiments on customer churn data from multiple domains to explore the potential of fusing synthetic dataset and real dataset for data reconstruction. The results demonstrate that the diffusion model can generate high-quality data and improve the performance of various prediction methods, which can help alleviate the issue of data imbalance. Additionally, the data produced by the diffusion model exhibit a distribution that is quite similar to the original dataset, which may be useful for protecting user privacy.

Key words customer churn; diffusion model; user privacy; data generation; categorical features

摘 要 在数据挖掘领域普遍存在数据不平衡影响到模型预测精度的问题,同时还存在未考虑用户隐私保护的问题.生成伪造数据是一种重要的解决方法,但在以结构化数据为主的场景中,由于存在数据特征维度多且不相关等特点,生成高质量的数据存在挑战.考虑到扩散模型在图像生成等任务中被成功应用,以客户流失预测为典型应用场景,尝试将扩散模型应用到客户流失预测任务中.针对该场景数据中的数值型特征和类别型特征,通过高斯扩散模型和多项式扩散模型获得生成数据,并对模型预测效果和数据隐私保护能力进行研究和分析.在多个领域的客户流失数据上进行了大量实验,探索应用生成数据对真实数据融合重构的可能性.实验结果表明基于扩散模型可生成高质量数据,且对多种预测方法均有一定提升,可实现缓解数据不平衡问题.同时,基于扩散模型生成的数据分布更接近真实数据,具有应用于用

收稿日期:2023-09-11;**修回日期**:2023-12-14
基金项目:云南省软件工程重点实验室开放基金项目(2023SE202)
This work was supported by the Open Foundation of Yunnan Key Laboratory of Software Engineering (2023SE202).

户隐私保护的潜在价值.

关键词 客户流失；扩散模型；用户隐私；数据生成；类别特征

中图法分类号 TP391

数据是人工智能技术研究和应用的基础要素. 但在真实的数据挖掘场景中，如客户流失、软件测试等领域[1-5]存在大量正负样本不均衡问题. 除了该类场景客观依然存在正样本天然比负样本少的原因外，更重要的是由于样本的收集、标注成本太高，获取有效的标注样本非常困难. 同时伴随业界对数据安全、隐私保护意识的提升，数据保护被广泛关注，且国内外已出台系列相关的法律法规[6-7]. 为了解决样本不平衡问题，学者们提出了多种方法，其中合成新数据是主流方法之一. 基于合成新数据对原始数据进行融合重构，可获得更多的样本，用于缓解该问题. 同时高质量的合成数据在分布上和真实数据一致，可应用于重构或替代真实样本，具有保护隐私数据的价值.

合成数据的方式常用的有 2 种：数据增强和数据生成. 数据增强在图像领域应用较多[8]. 在以结构化数据为主的数据挖掘领域通常以过采样或插值的方式调整正负样本比[9-11]. 该类方法虽然简单易实现，但需要大量的人工参与，并且合成数据和真实数据分布相去甚远，无法有效应用于数据保护，且对数据挖掘预测效果提升有限. 学者们在图像领域做了很多基于生成对抗网络(generative adversarial network, GAN)的数据生成研究工作[12-13]. 在以结构化数据为主的数据挖掘领域，生成方法研究相对较少，但在近年开始被关注，出现 IRGAN、CTAB-GAN 等方法[14-15]. 由于该领域特征类型较多，且不同维度特征独立，导致生成效果不理想，难以有效应用于对原始数据的融合重构. 近年扩散模型在图像领域大放异彩[16]，但将其应用于结构化数据的研究相对较少. 考虑到客户流失预测是一个典型的正负样本不均衡且涉及用户隐私保护的场景，本文以客户流失预测为实验场景，基于扩散模型生成数据探索在真实数据挖掘场景中的应用价值.

客户流失是指已有客户放弃继续购买企业商品或服务的一种现象. 该现象在企业经营中普遍存在，如电信行业中客户连续多月停止缴费；金融行业中客户中断购买理财产品；互联网场景中客户不再活跃等[1-3]. 客户流失会导致企业失去稳定的客户和收益源，造成巨大损失. 精准的客户流失预测，可发现流失概率大的客户，可通过挽留降低流失. 学者们对客户流失预测进行了大量研究，初期以传统统计学预测方法为主，包括决策树(decision tree, DT)、逻辑回归(logistic regression, LR)、贝叶斯(Bayesian)分类器、支持向量机(support vector machine, SVM)等算法[17-19]. 随着集成学习的引入，随机森林(random forest, RF)、梯度提升决策树(gradient boosting decision tree, GBDT)、Adaboost 和 Stack 等方法被大量引入到对客户流失的预测中. 近年也有工作尝试深度学习，但由于流失客户数据量较少，深度学习模型效果不理想[20-22]. 当前主流的做法还是采用传统的统计方法和集成学习.

客户流失预测在数据方面存在 2 个问题：1)数据不平衡问题. 考虑到用户习惯和适应新产品的成本，客户通常会在较长周期内处于忠诚状态，不会轻易流失. 因此流失客户在数量上相对总客户会非常少，相差几倍甚至几十倍. 2)客户流失数据直接使用真实数据，有用户隐私数据泄露风险. 客户数据通常采集了用户的基础属性和行为数据. 涉及用户隐私的数据需要保护起来，防止泄露对客户造成损失和伤害. 生成数据对不平衡和用户隐私保护都有帮助，但将生成数据应用于客户流失的研究非常少. 本文基于扩散模型生成样本进行数据重构，验证扩散模型在缓解数据不平衡和保护数据隐私的可能性.

本文主要贡献有 3 点：

1)首次提出一种基于扩散模型生成数据重构客户流失数据的方法. 基于扩散模型生成结果针对不同模型设计了多种融合重构策略，充分挖掘扩散模型在缓解数据不平衡问题上的价值.

2)采用扩散模型生成客户流失数据，生成结果在模型效果上和数据分布上与真实数据非常接近. 因此生成数据可作为客户真实数据的替代，在客户流失的用户隐私保护问题上具有很大的潜在价值.

3)在电信、银行、互联网等多个数据集上进行客户流失预测实验. 通过大量的实验和分析，证明基于扩散模型生成的数据有助于提升客户流失模型效果，且能被应用于用户隐私保护.

1 相关工作

本节主要介绍数据生成方法和客户流失预估的

相关研究工作.

1.1 数据生成

数据生成常用方法有简单数据合成和GAN、扩散模型等方法. 在图像领域以Mixup为代表的样本插值方法有着不错的效果[8]. 在数据挖掘领域, Wu等人[23]采用改进的SMOTE(synthetic minority over-sampling technique)解决流失数据的不平衡问题, 应用于电子商务客户流失预测.

随着GAN的发展, 在数据挖掘领域有人将GAN应用于数据生成. IRGAN[14]将视觉领域中广泛应用的GAN引入到信息检索, 探索了GAN在离散领域的适用问题. TableGAN[24]使用GAN生成表格数据用于用户隐私保护. CTAB-GAN[15]在GAN的基础上针对含连续变量和类别变量的数据生成进行优化. 近年, 扩散模型被应用于生成领域, 通过对样本的加噪和去噪过程, 有效实现图像生成[25-26].

1.2 客户流失预估

在过去的研究中, 学者们对客户流失预测在多种领域的应用进行了大量的探索, 包括电信、银行、互联网等行业. 预测方法也逐渐成熟, 发展历程主要分为2个阶段:

第1阶段是传统统计学预测方法, 主要包括决策树、逻辑回归、贝叶斯分类器、支持向量机等算法[17-19]. Qiu等人[3]使用逻辑回归来预测电子商务场景下的客户流失. Nie等人[27]分别使用决策树和逻辑回归对银行信用卡用户流失做预测.

第2阶段是集成学习算法的垄断和深度学习的初步探索. 随着集成学习的引入, 随机森林、GBDT、Adaboost和Stack等方法被引入到对客户流失的预测中. 特别是GBDT, 由于其优秀的性能而被广泛应用. Xie等人[28]提出了一种改进的平衡随机森林, 并将其应用于银行客户流失预测. Wu等人[20]在电子商务数据集上提出PCA-AdaBoost模型, 并证明模型的有效性. Liang等人[29]提出基于GBDT和逻辑回归算法的模型预测移动用户流失. 第2阶段的客户流失预测方法主要为集成学习, 应用最多的是GBDT算法.

2 方法介绍

客户流失数据多为表格类数据, 通常包含2类特征, 分别为数值型特征, 包括年龄、收入等; 类别型特征, 包括性别、省份等. 本文的研究路线如图1所示, 为了能够生成客户流失数据, 我们基于高斯扩散模型和多项式扩散模型分别处理数值型和类别型特征数据. 扩散生成模型通常包括前向扩散和逆向生成2个阶段[25-26]. 其中, 高斯扩散模型用于数值型数据的前向扩散阶段, 多项式扩散模型用于类别型的前向扩散阶段; 在逆向阶段, 我们对这2种模型进行结合, 通过神经网络实现逆向生成过程.

2.1 数据预处理

流失客户特征主要分为数值和类别2种类型, 在进入模型之前通常需要进行数据预处理, 主要包括对数值型特征进行归一化处理和对类别型特征进行独热编码处理. 图1中的输入数据$\boldsymbol{x}_0$为预处理后的数据, 包括归一化处理的$\boldsymbol{x}_{\text{num}}$和独热编码处理的$\boldsymbol{x}_{\text{cat}}$.

2.2 扩散模型框架

去噪扩散概率模型(denoising diffusion probabilistic model, DDPM)是一种概率生成模型, 包含前向和逆向2个阶段. 通常假设其前向阶段和逆向阶段均为马尔可夫过程. 前向过程也称前向扩散过程、前向加噪

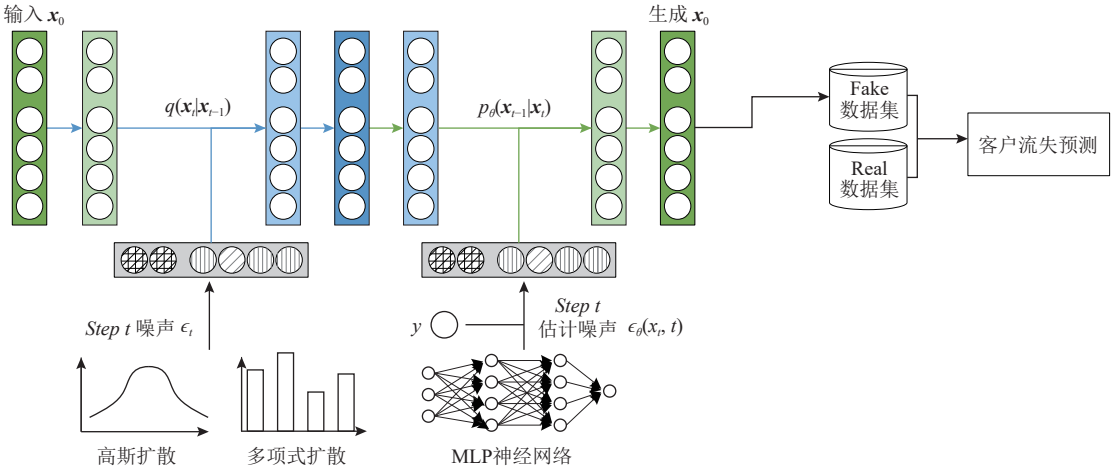


Fig. 1 The schematic diagram of customer churn prediction based on the diffusion model

图1 基于扩散模型的客户流失预测示意图

过程等；逆向过程也称逆向生成过程、逆向去噪过程等[30].

前向扩散过程是一个逐步加噪的过程，向样本$\boldsymbol{x}_0$逐步添加噪声，得到原始数据的每一次扩散后的状态. 当最终加噪步数T足够大时，可以使得原有的样本数据分布$\boldsymbol{x}_T$最终转换为预先定义好的噪声分布，通常选择各向同性的高斯分布，即$q(\boldsymbol{x}_T|\boldsymbol{x}_0)\sim N(0,I)$.

逆向生成过程是一个逐步去噪过程. 整个过程中，若逆向过程中的每一步t能够精确求得条件概率分布$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$，则可以通过逆向迭代不断采样恢复加噪之前的数据$p(\boldsymbol{x}_0)$，完成生成任务. 但是由于分布$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$通常是未知的，所以无法直接求得$p(\boldsymbol{x}_0)$，通常采用神经网络来近似估计其分布.

2.3 基于高斯扩散模型的数值型数据处理

高斯扩散模型是最常用的扩散模型，其针对连续样本空间$(\boldsymbol{x}_t\in\mathbb{R}^n)$如数值型样本进行处理[16,31]. 其前向和逆向阶段均将样本数据分布预定义为高斯分布. 其前向过程用式(1)(2)表示. 其中，$q()$表示前向扩散过程，通过加噪声，直到第T步$q(\boldsymbol{x}_T)$满足高斯分布采样. 其中β用于控制添加噪声的强度.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}):=\mathcal{N}\left(\boldsymbol{x}_t;\sqrt{1-\beta_t}\boldsymbol{x}_{t-1},\beta_t I\right), \tag{1}$$

$$q(\boldsymbol{x}_T):=\mathcal{N}(\boldsymbol{x}_T;0,I). \tag{2}$$

逆向过程$p()$通常采用构建由θ参数化的神经网络来近似估计其分布. 假设$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$是逆向过程的概率分布，且服从于高斯分布，考虑到从$\boldsymbol{x}_t$去第$t$步噪声估计$\boldsymbol{x}_{t-1}$，其均值$\mu_\theta$和方差$\sum_\theta$均可看作$\boldsymbol{x}_t$和$t$的函数，即式(3). 为便于后续计算和减少神经网络训练难度，将方差$\sum_\theta$设定为不需要参与神经网络训练且与时间相关的常数，因此在训练时，仅使用神经网络训练均值μ_θ即可. 如式(4)所示，$\mu_\theta(\boldsymbol{x}_t,t)$可以看成$\boldsymbol{x}_t$和$\epsilon_\theta(\boldsymbol{x}_t,t)$的函数. 其中$\epsilon_\theta(\boldsymbol{x}_t,t)$为$\boldsymbol{x}_t$的估计噪声，$\alpha$和$\beta$都是常数.

$$p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t):=\mathcal{N}(\boldsymbol{x}_{t-1};\mu_\theta(\boldsymbol{x}_t,t),\Sigma_\theta(\boldsymbol{x}_t,t)), \tag{3}$$

$$\mu_\theta(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t,t)\right). \tag{4}$$

在实际计算中，误差公式多被简化为噪声的估计误差，如式(5)所示，其中ϵ是前期加的噪声，假定其符合高斯分布.

$$L_t^{\text{simple}}=\mathbb{E}_{\boldsymbol{x}_o,\epsilon,t}\|\epsilon-\epsilon_\theta(\boldsymbol{x}_t,t)\|_2^2. \tag{5}$$

2.4 基于多项式扩散模型的类别型数据处理

多项式扩散模型是近年被提出用于处理类别型数据的扩散模型[32-33]. 多项式扩散模型需要类别特征用独热方式进行编码. 假定类别数据$\boldsymbol{x}_t\in\{0,1\}^K$用独热方式表示，$K$表示该类别特征有$K$个值.

多项式前向扩散过程，定义扩散采样分布$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$为一个类别分布$CAT$，在原始数据基础上通过均匀增加噪声的方式进行扩散. 如式(6)所示，可以看出每一步扩散对应的K个类分别引入很小的均匀噪声(β)，在上一步状态$\boldsymbol{x}_{t-1}$上以很大概率$(1-\beta)$进行采样，保留上一步的信息. 待t步扩散之后，可以得到$\boldsymbol{x}_t$如式(7)所示. 其中$\alpha_t=1-\beta_t$ ，$\overline{\alpha}_t=\prod_{\tau=1}^{t}\alpha_\tau$.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}):=CAT\left(\boldsymbol{x}_t;(1-\beta_t)\boldsymbol{x}_{t-1}+\frac{\beta_t}{K}\right), \tag{6}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0)=CAT\left(\boldsymbol{x}_t;\overline{\alpha}_t\boldsymbol{x}_0+\frac{1-\overline{\alpha}_t}{K}\right), \tag{7}$$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)=CAT\left(\boldsymbol{x}_{t-1}|\theta_{\text{post}}(\boldsymbol{x}_t,\boldsymbol{x}_0)\right), \tag{8}$$

$$\theta_{\text{post}}(\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{\tilde{\theta}}{\sum_{k=1}^{K}\tilde{\theta}_k},$$

$$\tilde{\theta}=\left[\alpha_t\boldsymbol{x}_t+\frac{1-\alpha_t}{K}\right]\odot\left[\overline{\alpha}_{t-1}\boldsymbol{x}_0+\frac{1-\overline{\alpha}_{t-1}}{K}\right].$$

和高斯扩散模型类似，$\boldsymbol{x}_{t-1}$的逆向估计分布$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$无法直接计算，但其可用后验概率$q(\boldsymbol{x}_{t-1};\boldsymbol{x}_t,\boldsymbol{x}_0)$表示，如式(8)所示. 因此，逆向分布$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$可以被参数化为$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t,t))$，其中$\hat{\boldsymbol{x}}_0$表示估计的$\boldsymbol{x}_0$，可以通过神经网络进行估计.

2.5 扩散模型噪声估计

本文结合高斯扩散模型和多项式扩散模型建模数值型和类别特征，在前向加噪声过程中分别通过高斯和多项式扩散进行采样. 在逆向生成过程中，考虑到输入信息除了输入数据$\boldsymbol{x}_0$，还应包括扩散步数t和类别信息y，二者对噪声分布都有影响，如式(9)~(11)所示. 其中扩散步数信息采用传统的正弦时间嵌入方法[11,16,34].

$$t_{emb}=Linear(SiLU(Linear(SinTimeEmb(t)))), \tag{9}$$

$$y_{emb}=Embedding(y), \tag{10}$$

$$\boldsymbol{x}=Linear(\boldsymbol{x}_0)+t_{emb}+y_{emb}. \tag{11}$$

由于深度学习中多层感知机(multilayer perceptron, MLP)多被应用于拟合输出，本文逆向去噪生成过程采用MLP进行噪声分布估计，如式(12)(13)所示，输出和输入相同维度的特征作为估计噪声.

$$MLP(\boldsymbol{x})=Linear(MLPBlock(\ldots(MLPBlock(\boldsymbol{x})))), \tag{12}$$

$$MLPBlock(\boldsymbol{x})=Dropout(ReLu(Linear(\boldsymbol{x}))). \tag{13}$$

在训练阶段，为了能够同时对高斯扩散模型和多项式扩散模型进行优化，我们基于2种模型的损

失函数设计了联合损失函数如式(14)所示,其中式(14)右侧L_t^{simple}为高斯扩散模型损失和类别特征的平均KL散度(Kullback–Leibler divergence, KLD),L_t^i为每个多项式扩散项的KL散度,C为类别特征数目.

$$L_t^{\text{DDPM}} = L_t^{\text{simple}} + \frac{1}{C}\sum_{i\leqslant C} L_t^i. \tag{14}$$

2.6 流失预测方法

为了能够验证基于扩散生成模型在客户流失场景下的预测效果和隐私保护能力,本文在客户流失预测中选择了2种模式,分别是弱分类器模式和强分类器模式,包括经典的分类算法和GBDT[2,21].其中,对于弱分类器本文选择了客户流失预测场景中的经典算法,包括决策树、贝叶斯、支持向量机和逻辑回归等.我们选择了客户流失场景中被广泛应用的GBDT作为强分类器,其中,Catboost是一种主流的GBDT实现方法,能够在大部分数据上达到最好效果[21].

3 实验设置

3.1 数据集

为了能够对不同场景的客户流失进行建模,我们选择了3种不同场景的客户流失数据,分别为:

1)电信客户流失数据集①.该数据集来自Kaggle的电信流失用户预测数据.其共有7 043条电信客户数据,其中客户流失数据有1869条.每条数据包括21个特征,其中3个数值特征和18个类别特征.在实验时我们过滤了用户ID特征"customer ID",保留了其他17个类别特征.

2)银行流失数据集②.该数据来自Kaggle的银行信用卡流失数据集.该数据集共有10 127条银行客户数据,其中客户流失数据有1 627条.每条数据包括23个特征,其中16个数值特征和7个类别特征.在实验时我们过滤了用户ID特征"CLIENTNUM",并按照数据说明删去了最后2个无关特征,只保留14个数值特征和6个类别特征.

3)互联网产品流失数据集③.该数据是来自阿里云天池比赛的直播电商用户流失行为预测数据集.该数据集共有5 630条直播电商相关客户数据,其中客户流失数据有948条.每条数据包括19个特征,其中9个数值特征和10个类别特征,在实验时我们过滤了用户ID特征"Customer ID",保留了其他9个类别特征.

在训练时,每个数据集都按照8∶2的原则分为训练集和测试集.

3.2 实验环境及超参设置

实验以Pytorch为基础框架,在NVIDIA GPU 3050Ti上进行训练.根据经验设置,实验中扩散模型的总步数设置为1 000, Batchsize设置为4 096, Epoch设置为30 000,优化器采用Adam算法,学习率设置为0.001.实验中,Catboost算法的学习率设置为0.03,计算叶子节点值时候的迭代次数设置为8. Logistic Regression算法的正则化项的类型设置为l2,其正则化强度的倒数设置为10,模型的最大迭代次数设置为1 000.

3.3 评估指标

实验中,为了评估最终的模型分类效果,本文采用了*AUC*(area under curve)、*ACC*(accuracy)、Logloss以及*F*1值等指标进行评估,其相关计算方式可参考文献[35].

4 实验结果及分析

4.1 生成数据展示

我们通过扩散模型DDPM在电信、银行、互联网数据集上生成了新的客户流失数据.以电信数据集为例,真实数据(REAL)和DDPM、SMOTE以及CTAB-GAN的生成结果如表1所示.从表1可以看出,DDPM、SMOTE, CTAB-GAN这3种方法均能生成与REAL相似的结果.但比较具体生成的特征细节,可以发现DDPM生成的数据和REAL更接近,特别是数值型特征;SMOTE和CTAB-GAN则存在大量的整型数据生成后成为浮点型的问题;从随机抽取的样本上看,CTAB-GAN的生成数据相对比较发散,和REAL差异较大,这也比较符合GAN的算法特点.

4.2 生成结果对比分析

为了研究在客户流失场景下扩散模型和其他生成方法的生成数据效果对比,我们对REAL以及不同方法生成的数据进行采样,并在单个独立特征和多个联合特征分布上分别进行展示.

1)独立特征对比.我们对单个独立特征进行采

① https://www.kaggle.com/datasets/blastchar/telco-customer-churn

② https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

③ https://tianchi.aliyun.com/dataset/124814

Table 1 Generation Results Presentation for Telco Customer Churn Datasets

表 1 电信客户流失数据集生成结果展示

生成方法	特征集										
REAL	Male	0	…	2	Yes	No	DSL	…	Mailed check	53.85	108.15
	Female	0	…	8	Yes	Yes	Fiber optic	…	Electronic check	99.65	820.5
DDPM	Male	0	…	1	Yes	Yes	Fiber optic	…	Electronic check	95.810	135.310
	Female	1	…	1	Yes	Yes	Fiber optic	…	Electronic check	75.708	75.535
	Female	0	…	1	Yes	No	DSL	…	Mailed check	43.954	45.141
	Male	0	…	15	Yes	Yes	Fiber optic	…	Credit card (automatic)	89.232	1 370.882
SMOTE	Female	0	…	6.058	Yes	No	Fiber optic	…	Electronic check	89.304	581.600
	Female	0	…	10.409	Yes	Yes	DSL	…	Electronic check	72.897	749.778
	Male	0	…	57.535	No	No phone service	DSL	…	Electronic check	52.085	2 966.530
	Male	0	…	68.014	Yes	Yes	Fiber optic	…	Credit card (automatic)	111.845	7 754.865
CTAB-GAN	Female	1	…	25.814	Yes	No	Fiber optic	…	Credit card (automatic)	72.452	107.121
	Male	0	…	7.578	Yes	No	DSL	…	Mailed check	45.844	105.699
	Male	0	…	28.817	Yes	Yes	DSL	…	Credit card (automatic)	77.303	756.299
	Female	1	…	26.574	Yes	No phone service	DSL	…	Electronic check	57.699	781.075

样，针对 3 个不同场景数据集分别随机挑选 2 个特征，包括 1 个数值型和 1 个类别型. 生成数据和 REAL 的分布统计结果如图 2 所示. 我们可以看出 DDPM 相对其他方法和 REAL 分布更相似. 对于数值型特征，在 3 个场景上 CTAB-GAN 的效果最不理想，和真实分布差异较明显. 在类别型特征上，SMOTE 效果最不理想，在某些类别上生成数据与 REAL 分布差异较明显，且在生成数量比较少的类别值时可能会出现生成数据量远小于 REAL 量的问题，这与其算法进行插值有关.

2)联合特征对比. 我们对特征相关性也进行了研究. 对数值型特征我们统计了生成数据和 REAL 的皮尔逊相关系数；对类别型特征计算 Theil's U 统计量；对数值与类别间特征关系计算 Correlation Ratio[15]，合成数据与 REAL 的特征相关性矩阵差异结果如图 3 所示. 图中颜色深浅代表对应生成数据和 REAL 在特征相关性上的差异，颜色越深代表差距越大. 我们可以看出和其他生成方法相比，DDPM 的生成结果更加接近 REAL，生成效果更显著，CTAB-GAN 的生成效果相对最差.

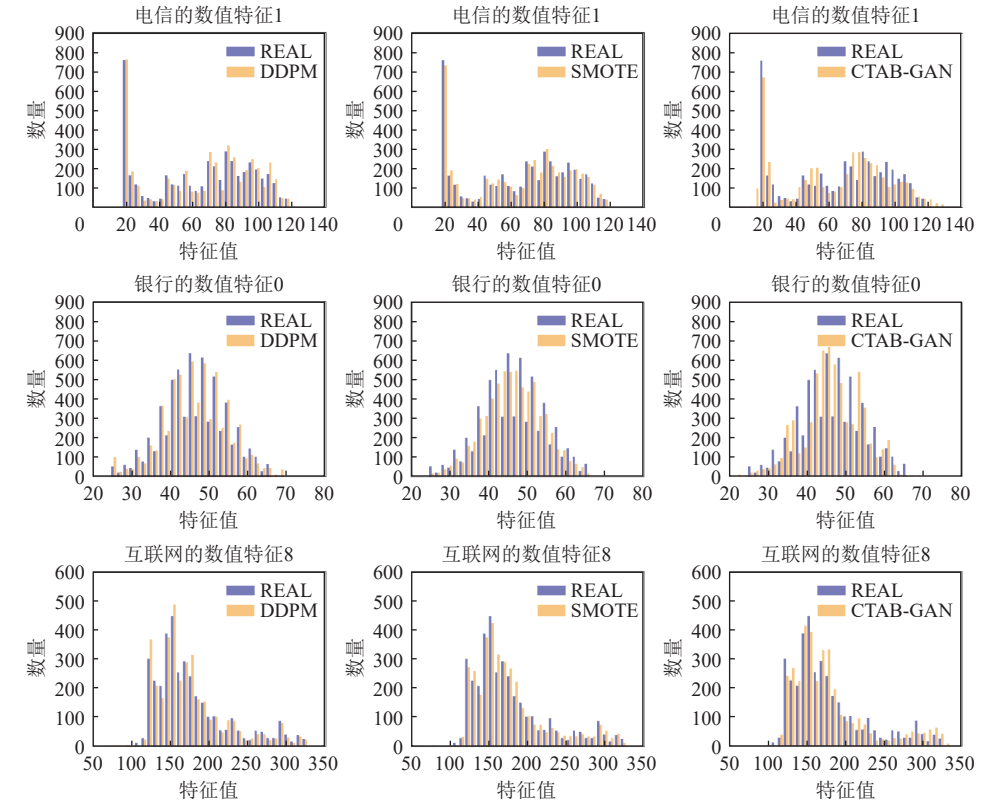
4.3 数据重构预测效果对比

为了验证生成数据对客户流失预测的重构效果，本文分别在电信、银行、互联网 3 个场景进行对比. 本文设计了 4 种训练集数据重构方式：1)用生成数据等量代替真实训练集集，记为 Fake±；2)只用生成数据等量代替正样本，负样本仍为真实数据，记为 Fake+；3)对第 2 种融合方式的负样本数据进行欠采样，使得正负样本量比为 1:1，记为 Fake+_1:1；4)只用生成的正样本数据和所有真实数据融合，使得正样本数据量增加至和真实负样本数据量相当，正负样本量比为 1:1，记为 Merged+_1:1.

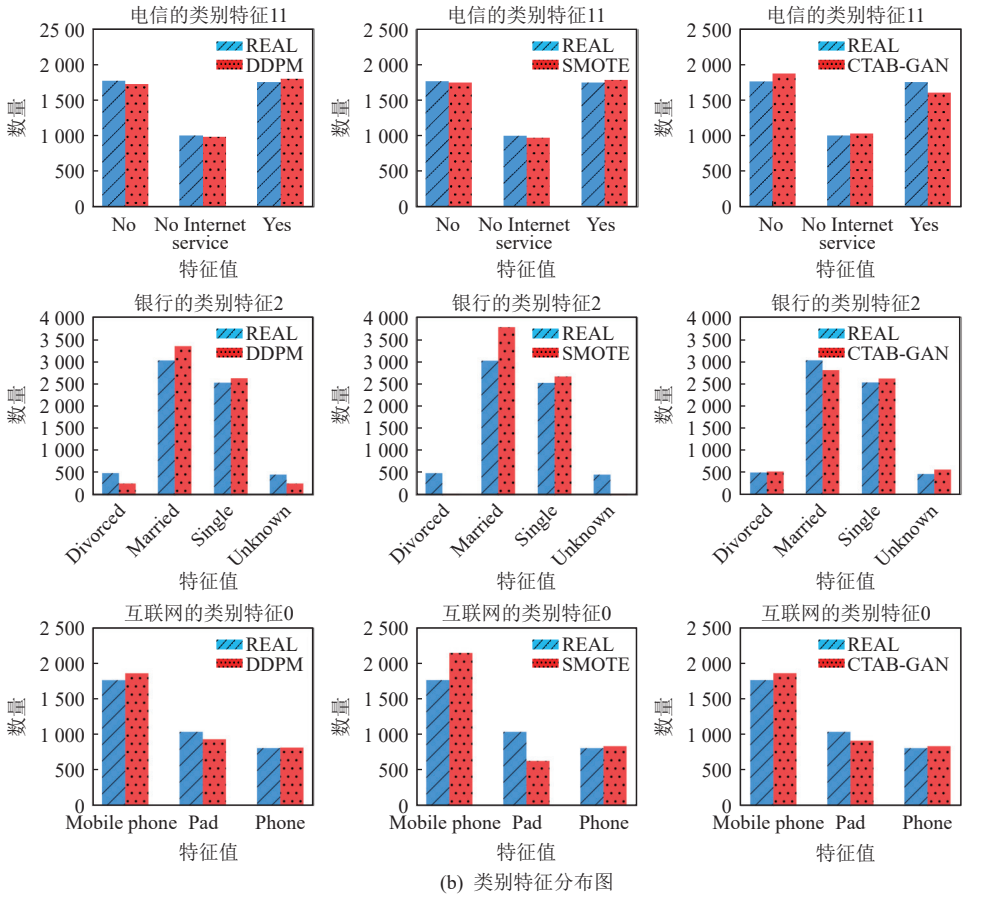
考虑到弱分类器和强分类器对客户流失模型的预测效果不一样，本文选择了决策树、贝叶斯分类器、支持向量机和逻辑回归 4 种弱分类器，对其结果取平均值；同时选择 GBDT 作为强分类器，采用广泛应用的 Catboost 算法实现. 我们采用这种模式验证扩散模型在弱分类器和强分类器上的预测效果，实验结果如表 2 和表 3 所示.

从表 2 可以看出，弱分类器的实验在电信、银行、互联网 3 个场景中，生成数据重构的数据集的模型效果均有所提升. 针对不同训练集融合重构方式，我们发现 DDPM 在 Fake±方式下(用生成数据全量等比例替换真实数据)和 REAL 效果一致，其他方法则是提升. 这说明针对弱分类器，DDPM 生成的数据更接近 REAL 分布.

强分类器 Catboost 的实验结果如表 3 所示. DDPM 生成的数据重构后的效果基本和 REAL 的效果一样，这个结论和弱分类器实验保持一致. 说明 DDPM 生成的数据和 REAL 分布非常接近，并且与分类器无关. 其他方法如 CTAB-GAN 和 SMOTE 均略逊于 REAL 的效果. 由于银行数据集在强分类器上的分类实验效果非常理想，接近上限，因此 DDPM、SMOTE, CTAB-



(a) 数值特征分布图



(b) 类别特征分布图

Fig. 2 Distribution of independent features in generated data and real data

图 2 关于生成数据和真实数据的独立特征分布

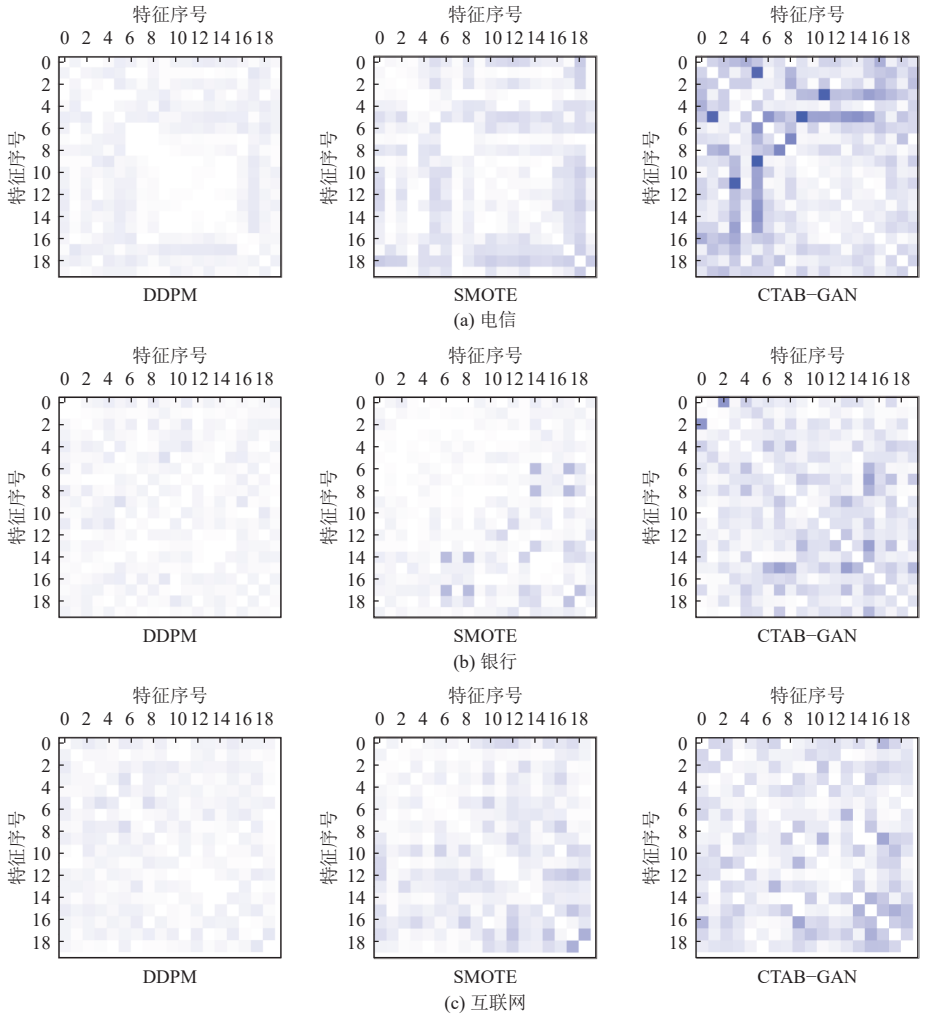


Fig. 3 Difference of correlation matrix between generated data and real data

图 3 生成数据与真实数据的相关性矩阵差异

Table 2 Effect of Data Reconstruction Based on Week Classifiers Algorithm

表 2 基于弱分类器算法的数据重构效果

模型	融合方法	电信		银行		互联网	
		*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*
	REAL	0.599 4	0.724 6	0.938 0	0.739 3	0.588 1	0.728 9
DDPM	Merged_1:1	0.761 4	0.767 1	0.848 1	0.830 3	0.809 3	0.823 8
	Fake+_1:1	0.738 7	0.772 7	0.878 3	0.809 9	0.754 8	0.809 5
	Fake+	0.674 5	0.756 2	0.939 1	0.744 4	0.569 1	0.720 6
	Fake±	0.600 9	0.727 3	0.937 8	0.731 9	0.607 7	0.744 2
SMOTE	Merged_1:1	0.789 4	0.791 7	0.869 1	0.853 2	0.844 1	0.854 0
	Fake+_1:1	0.784 8	0.812 0	0.900 5	0.850 4	0.795 7	0.842 7
	Fake+	0.789 2	0.792 1	0.948 0	0.796 4	0.685 1	0.792 9
	Fake±	0.627 4	0.744 4	0.945 8	0.782 3	0.749 6	0.831 9
CTAB-GAN	Merged_1:1	0.780 2	0.762 4	0.826 3	0.781 0	0.822 4	0.808 6
	Fake+_1:1	0.733 3	0.763 4	0.853 2	0.746 4	0.764 1	0.807 5
	Fake+	0.638 5	0.741 2	0.935 6	0.660 4	0.673 5	0.774 0
	Fake±	0.677 8	0.766 3	0.932 8	0.630 5	0.660 5	0.766 6

Table 3 Effect of Data Reconstruction Based on Catboost Algorithm

表 3 基于 Catboost 算法的数据重构效果

模型	融合方法	电信		银行		互联网	
		*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*
	REAL	0.738 8	0.854 6	0.945 6	0.994 8	0.873 7	0.952 8
DDPM	Merged_1:1	0.723 2	0.853 4	0.907 4	0.993 1	0.870 3	0.948 4
	Fake+_1:1	0.700 8	0.828 5	0.881 2	0.985 8	0.744 9	0.860 9
	Fake+	0.752 8	0.853 4	0.932 0	0.987 0	0.778 6	0.888 1
	Fake±	0.736 7	0.854 2	0.924 1	0.983 6	0.807 6	0.889 3
SMOTE	Merged_1:1	0.760 3	0.845 6	0.946 6	0.994 3	0.868 7	0.947 5
	Fake+_1:1	0.727 4	0.830 9	0.923 2	0.986 9	0.763 8	0.852 5
	Fake+	0.685 1	0.836 8	0.905 4	0.986 9	0.650 5	0.861 6
	Fake±	0.694 6	0.846 9	0.939 6	0.992 5	0.808 0	0.895 5
CTAB-GAN	Merged_1:1	0.723 7	0.837 0	0.952 6	0.995 7	0.851 4	0.943 1
	Fake+_1:1	0.678 0	0.777 0	0.622 1	0.819 0	0.667 9	0.755 5
	Fake+	0.505 9	0.769 9	0.454 4	0.861 1	0.629 6	0.772 9
	Fake±	0.707 9	0.817 5	0.702 5	0.944 8	0.723 6	0.840 5

GAN 这 3 种方法在该数据集上的表现也非常接近.

4.4 基于误分样本的数据重构

在分类问题中, 相较于易分类的样本, 误分样本对提升模型性能更有价值. 因此鉴于 Catboost 在生成数据集上的效果和真实数据效果持平, 没有明显提升, 我们提出基于误分样本提升模型效果的重构策略. 该方法基于生成数据的误分类样本重构真实数据, 增加误分类样本的数量和多样性.

考虑到基于扩散模型生成的数据与真实数据分布一致, 我们应用真实数据训练的 Catboost 对基于扩散模型生成的数据进行预测, 选出误分类的生成样本, 然后将误分的生成样本与真实数据进行融合得到重构数据集. 重构数据集中误分样本的比例和多样性都得到了提升. 基于重构数据集进行训练并在真实的测试集上进行评估, 结果如表 4 所示. 其中 Fake_Best 是表 3 中未加入误分样本策略的基于生成数据在测试集上表现最好的重构效果; REAL+是真实数据增加了误分样本重构后的数据集. 可以看出基于误分样本进行数据重构的策略在 3 个数据集上均对强分类器 Catboost 效果有提升, 说明基于扩散模型的误分样本的数据重构对强分类器 Catboost 是可行的.

5 讨 论

5.1 生成数据规模对模型效果的影响

为了探索基于扩散模型的生成数据规模对模型分类效果的影响, 本文基于扩散模型生成真实训练集数据量 1, 10, 30, 50, 70 和 100 倍的合成数据(fakeall)用于训练 Catboost 模型, 结果如表 5 所示. 在 3 个数据集中, 增大样本规模都可以使分类器的性能得以提升. 在电信和互联网数据集中, 当数据样本增加至原训练样本的 70 倍时模型效果最优, 且均高于 REAL 的训练结果. 银行数据集的生成数据量从 1 倍增加到 100 倍的过程中, 模型效果持续提升. 由于扩散模型生成的数据在整体分布上更符合真实

Table 4 Effect of the Fusion of Real Samples and Misclassified Generated Samples for Catboost

表 4 基于融合真实样本和误分生成样本的 Catboost 效果

数据集	电信			银行			互联网		
	ACC	*F*1	*AUC*	*ACC*	*F*1	*AUC*	*ACC*	*F*1	*AUC*
REAL	0.807 4	0.738 8	0.854 6	0.970 1	0.945 6	0.994 8	0.940 1	0.873 7	0.952 8
Fake_Best	0.805 1	0.736 7	0.854 2	0.942 7	0.907 4	0.993 1	0.929 6	0.870 3	0.948 4
REAL+	**0.810 0**	**0.743 1**	**0.859 0**	**0.971 1**	**0.947 0**	**0.995 3**	**0.958 6**	**0.916 5**	**0.972 5**

注: 加粗数字表示最佳结果.

Table 5 Effect of Generated Amount of Data on Effectiveness

表 5 生成数据量对效果的影响

融合方法	电信		银行		互联网	
	*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*
REAL	0.738 8	0.854 6	**0.945 6**	**0.994 8**	0.873 7	0.952 8
1x_fakeall	0.736 7	0.854 2	0.924 1	0.983 6	0.807 6	0.889 3
10x_fakeall	0.707 2	0.838 4	0.923 9	0.985 5	0.864 6	0.951 1
30x_fakeall	0.756 6	0.859 5	0.927 7	0.986 6	0.865 2	0.955 6
50x_fakeall	0.702 0	0.844	0.927 9	0.987 2	0.870 8	0.954 3
70x_fakeall	**0.772 2**	**0.873 1**	0.924 9	0.986 8	**0.875 7**	**0.957 4**
100x_fakeall	0.707 6	0.844 2	0.931 2	0.988 2	0.868 7	0.956 2

注: 加粗数字表示最佳结果.

样本的分布规律. 当样本数量增加时, 数据的多样性也会随之提升, 因此可以获得更好的预测效果. 实验表明通过增大生成数据规模提升模型效果是可行的.

5.2 用户隐私保护能力分析

客户流失数据涉及到大量的用户数据, 建模时被直接暴露, 未对用户隐私数据进行保护, 存在对客户造成伤害和损失的风险. 本文探索了将生成数据应用于用户隐私保护的潜力. 我们设计了 2 种指标用于定量研究生成数据和真实数据之间的差异, 评估生成数据用于隐私保护的能力[15].

1)最近样本距离(distance to closest record, DCR). 对于每条生成数据, 计算其与真实数据间的欧氏距离, 将其中最近的距离作为 DCR, 其值越小代表生成数据和真实数据越接近, 复制真实数据的信息越多, 数据隐私信息保护的能力也就越弱. 反之, DCR 值越大代表生成的数据和真实数据的差异也就越大, 表明生成数据越接近新数据, 而不是简单地复制原始数据的信息, 因此想要通过这种生成数据的方式复原原始数据的难度也越大, 隐私信息保护能力越强.

2)最邻近距离比例(nearest neighbour distance ratio, NNDR)是度量最近的距离和第二近的距离的比例, 其取值范围为 [0,1]. 和 DCR 一样, NNDR 值越大表示有更好的隐私保护能力; 越小的 NNDR 值意味着和真实数据越接近, 保护性越差.

我们计算了生成数据的隐私指标, 结果如表 6 所示. 可以看出, 在 3 个数据集中, 和 SMOTE 相比, DDPM 生成的数据在 DCR 和 NNDR 上在 3 个客户流失场景下均要优于 SMOTE, 说明 DDPM 的数据更适合应用到隐私保护中. 但和 CTAB-GAN 相比, DDPM 在这 2 个隐私指标上要略逊一筹. 经过分析, CTAB-GAN 在生成结果上和真实数据分布差异较大, 导致了其客户流失模型预测效果较差. 但由于生成数据与真实数据不一致, 会造成很难通过生成数据复原原始数据, 保护能力更好. 由此可以看出, DDPM 生成的数据在生成质量和隐私保护 2 个方面具有更好的平衡.

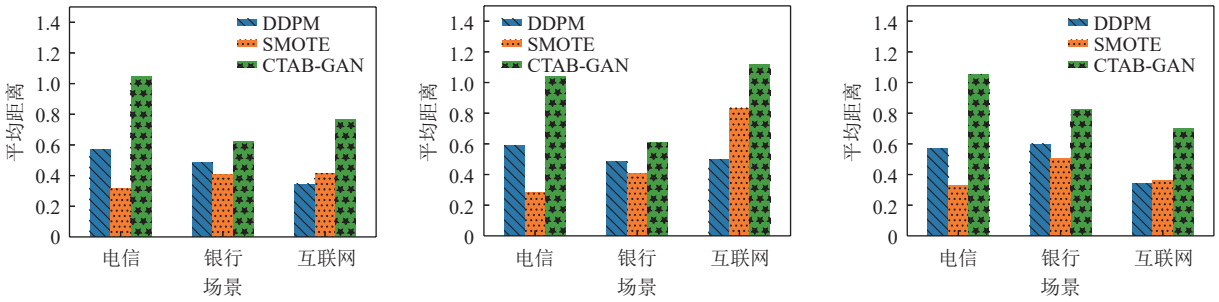
Table 6 Analysis of Privacy Protection Capability

表 6 隐私保护能力分析

数据集	模型	DCR	NNDR
电信	DDPM	0.004 1	0.029 4
	SMOTE	0.001 6	0.012 4
	CTAB-GAN	0.053 1	0.194 4
银行	DDPM	0.269 9	0.513 0
	SMOTE	0.059 3	0.158 4
	CTAB-GAN	0.336 8	0.656 6
互联网	DDPM	0.050 6	0.078 8
	SMOTE	0.043 4	0.113 8
	CTAB-GAN	0.221 2	0.344 1

为了能够展示生成数据和隐私保护的关系, 如图 4 所示我们统计了生成结果与原始数据的平均欧式距离作为生成数据与真实数据的相似度指标. 分别就所有样本、正样本和负样本统计, 平均距离越大, 相似性越小; 平均距离越小, 相似性越大.

从图 4 可以看出, 基于 SMOTE 生成的结果与真实数据的平均距离最小, 即相似度最大; CTAB-GAN 则反之, 其相似度最小. DDPM 的平均距离在数据整体上和不同子类别上均处于中游水平, 其相似度水平居中. 我们分析认为: SMOTE 有直接插值的思想, 会保留较多的真实样本信息, CTAB-GAN 则是生成了大量不太相关的信息. 相对而言, DDPM 生成的数据更符合真实分布, 且具有一定的隐私保护能力. 这个结论也符合基于 DCR 和 NNDR 的评估结果. 综上



(a) 全量样本的平均距离 (b) 正样本的平均距离 (c) 负样本的平均距离

Fig. 4 Average distance between generated data and real data

图 4 生成数据与真实数据的平均距离

所述, DDPM 相对其他方法在模型效果和隐私保护 2 个方面做到了比较理想的平衡.

5.3 生成数据解释性分析

本文选择电信数据集中的真实样本和生成样本, 通过模型解释性的方式对模型决策机制进行剖析, 基于 SHAP(SHapley Additive exPlanations)实现[36]. 通过 Shapley 值判断特征对模型决策的贡献进而实现决策解释. 本文设计 2 种模式: 1)通过全局解释性方法, 对比真实数据和生成数据在重要特征选择上的相似度. 2)通过局部解释性方法, 对比单个生成流失样本和其最接近的真实样本在模型决策上的相似性.

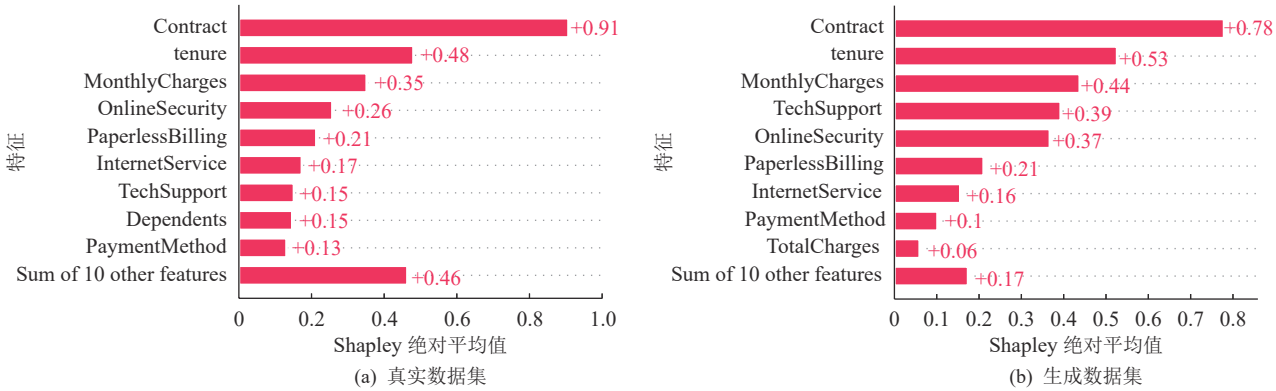
我们对生成的所有样本和真实样本集进行 Global 级别的全局解释, 计算了全部样本的 Shapley 值的绝对平均值. 图 5 直观地展示了在真实和生成数据集上特征对模型决策的重要程度. 在一共 19 个特征的数据中 Top 9 的特征重叠度非常高, 只有 1 个特征不同, 且特征排序也几乎一致. 这说明基于 DDPM 生成的数据在全局上和真实数据非常接近, 在模型决策时的表现和真实数据几乎一致.

对于 Local 模式下的解释性, 本文首先使用计算最近样本距离方法随机挑出原始样本, 并找到生成数据集对应的最近生成样本, 分别计算这 2 个样本的 Shapley 值, 生成瀑布图如图 6 所示. 瀑布图中, 图标长度说明特征的影响大小, 可以看出对应单个样本级别, 生成数据和真实数据与全局解释性上的表现非常统一, 在重要的特征上几乎完全一致, 仅在特征的顺序上有一些区别, 这也能反映出生成数据和其最相似的真实数据的区别.

总之, 由解释性分析结果可知, Shapley 值度量的特征重要性符合人工判断, 说明模型解释是合理的. 同时, 生成数据与原始数据的特征分布具有高度一致性, 且各特征对预测模型的影响也非常相似, 说明基于扩散模型生成数据能够作为补充样本应用于流失预测模型中.

6 结论与展望

本文以客户流失预测为典型场景, 研究了基于扩散模型探索数据不平衡改进和用户隐私数据保护等问题. 我们结合高斯和多项式扩散模型对客户流失数据进行生成, 获得和真实流失客户数据分布接近的生成数据, 并对生成数据的质量进行展示和定



(a) 真实数据集 (b) 生成数据集

Fig. 5 Global mode: Comparison of the importance of features for the different datasets

图 5 全局模式: 不同数据集中特征重要性对比

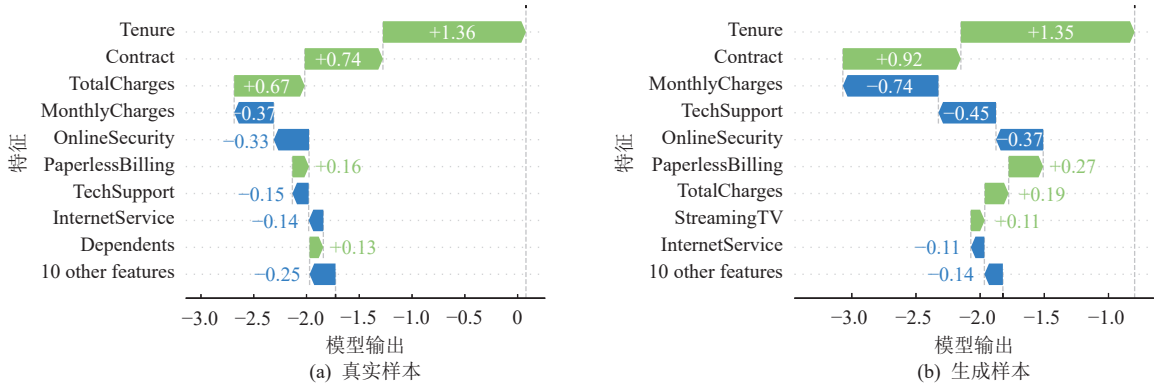


Fig. 6 Local mode: Comparison of the importance of features for a single Sample

图 6 局部模式：单个样本特征重要性对比

量评估. 本文还对生成数据和真实数据的融合方式进行研究，分析了生成数据在不同性能的分类器下的表现和客户隐私保护能力. 本文在多个领域的客户流失数据上进行实验，结果表明扩散模型可生成高质量数据，与真实数据分布一致，且生成数据对客户流失预测模型效果均有一定的提升. 考虑到基于扩散模型的生成数据与真实数据的分布一致，我们认为基于扩散模型生成的数据可以达到完全替换真实数据的目的，可应用于用户隐私数据保护.

基于本文研究，未来我们会对不同的数据融合重构方式进行探索，寻找更加通用的融合策略. 同时会对生成数据的隐私保护潜力进行深入挖掘.

作者贡献声明：杨斌提出了算法思路和实验方案，并撰写论文；王正阳和程梓航负责完成实验并撰写论文；赵慧英、王鑫和管宇参与文献调研和讨论；程新洲提出指导意见并修改论文.

参 考 文 献

[1] Yao Bo. Research and application of customers churn prediction model[D]. Xi'an: Northwest University, 2017 (in Chinese)
(姚博. 客户流失预测模型研究及其应用[D]. 西安：西北大学, 2017)

[2] Jain H, Khunteta A, Srivastava S. Churn prediction in telecommunication using logistic regression and logit boost[J]. Procedia Computer Science, 2020, 167: 101−112

[3] Qiu Yanfang, Li Chen. Research on e-commerce user churn prediction based on logistic regression[C]// Proc of the 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conf (ITNEC). Piscataway, NJ: IEEE, 2017: 87−91

[4] Xing Ying, Lin Wanting, Lin Xueyan, et al. Cross-project defect prediction based on two-phase feature importance amplification[J]. Computational Intelligence and Neuroscience, 2022. https://www.hindawi.com/journals/cin/2022/232044

[5] Zeng Fuping, Lin Wanting, Xing Ying, et al. A cross-project defect prediction model using feature transfer and ensemble learning[J]. Tehnički Vjesnik, 2022, 29(4): 1089−1099

[6] Qian Wenjun, Shen Qingni, Wu Pengfei, et al. Research progress on privacy-preserving techniques in big data computing environment[J]. Chinese Journal of Computers, 2022, 45(4): 669−701(in Chinese)
(钱文君, 沈晴霓, 吴鹏飞, 等. 大数据计算环境下的隐私保护技术研究进展[J]. 计算机学报, 2022, 45(4): 669−701)

[7] Zhao Jingxin, Yue Xinghui, Feng Chongpeng, et al. Survey of data privacy security based on general data protection regulation[J]. Journal of Computer Research and Development, 2022, 59(10): 2130−2163 (in Chinese)
(赵景欣, 岳星辉, 冯崇朋, 等. 基于通用数据保护条例的数据隐私安全综述[J]. 计算机研究与发展, 2022, 59(10): 2130−2163)

[8] Zhang Hongyi, Cisse M, Dauphin Y, et al. mixup: Beyond empirical risk minimization[C]//Proc of the 6th Int Conf Learn Represent (ICLR). Vancouver, BC, Canada: OpenReview.net, 2018: 1−13

[9] 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

[10] Ai-jun L, Peng Z. Research on unbalanced data processing algorithm base tomeklinks-smote[C]//Proc of the 3rd Int Conf on Artificial Intelligence and Pattern Recognition. New York: ACM, 2020: 13−17

[11] Bogaert M, Delaere L. Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art[J]. Mathematics, 2023, 11(5): 1137

[12] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139−144

[13] He Huang, Yu P S, Wang Changhu. An introduction to image synthesis with generative adversarial nets[J]. arXiv preprint, arXiv: 1803. 04469, 2018

[14] Wang Jun, Yu Lantao, Zhang Weinan, et al. IRGAN: A minimax game for unifying generative and discriminative information retrieval models[C]//Proc of the 40th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2017:

515−524

[15] Zhao Zilong, Kunar A, Birke R, et al. CTAB-GAN: Effective table data synthesizing[C]// Proc of Asian Conf on Machine Learning. https://www.acml-conf.org/2021/

[16] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840−6851

[17] Yang Bin, Li Huilai, Xing Ying et al. Directed search based on improved whale optimization algorithm for test case prioritization[J]. International Journal of Computers Communications & Control. https://www.univagora.ro/jour/index.php/ijccc/article/view/5049

[18] Ying Weiyun, Qin Zheng, Zhao Yu, et al. Support vector machine and its application in customer churn prediction[J]. Systems Engineering-Theory & Practice, 2007, 27(7): 105−110 (in Chinese)
(应维云, 覃正, 赵宇, 等. SVM 方法及其在客户流失预测中的应用研究 [J]. 系统工程理论与实践, 2007, 27(7): 105−110)

[19] Ahn J, Hwang J, Kim D, et al. A survey on churn analysis in various business domains[J]. IEEE Access, 2020, 8: 220816−220839

[20] Wu Zengyuan, Jing Lizheng, Wu Bei, et al. A PCA-AdaBoost model for e-commerce customer churn prediction[J]. Annals of Operations Research, 2022: 1−18

[21] Prokhorenkova L, Gusev G, Vorobev A, et al. Catboost: Unbiased boosting with categorical features[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2018, 31: 6638−6648

[22] Pekel Ozmen E, Ozcan T. A novel deep learning model based on convolutional neural networks for employee churn prediction[J]. Journal of Forecasting, 2022, 41(3): 539−550

[23] Wu Xiaojun, Meng Sufang. E-commerce customer churn prediction based on improved SMOTE and AdaBoost[C]// Proc of the 13th Int Conf on Service Systems and Service Management (ICSSSM). Piscataway, NJ: IEEE, 2016: 1−5

[24] Park N, Mohammadi M, Gorde K, et al. Data synthesis based on generative adversarial networks[J]. arXiv preprint, arXiv: 1806.03384, 2018

[25] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models[C]// Proc of the Int Conf on Machine Learning. Virtual: PMLR, 2021: 8162−8171

[26] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2021, 34: 8780−8794

[27] Nie Guangli, Rowe W, Zhang Lingling, et al. Credit card churn forecasting by logistic regression and decision tree[J]. Expert Systems with Applications, 2011, 38(12): 15273−15285

[28] Xie Yaya, Li Xiu, Ngai E W T, et al. Customer churn prediction using improved balanced random forests[J]. Expert Systems with Applications, 2009, 36(3): 5445−5449

[29] Liang Jiafu, Qiu Xinyong. Research on user churn monitoring technology based on GBDT and LR algorithm[J]. Journal of Hebei Software Institute, 2021, 23(3): 1−4 (in Chinese)
(梁家富, 邱新泳. 基于 GBDT 和 LR 算法的用户流失监控技术研究 [J]. 河北软件职业技术学院学报, 2021, 23(3): 1−4)

[30] Yang Guangkai. Fingerprint image generation method based on diffusion model[J]. Journal of the Hebei Academy of Sciences, 2023, 40(1): 13−18+66 (in Chinese)
(杨光锴. 基于扩散模型的指纹图像生成方法 [J]. 河北省科学院学报, 2023, 40(1): 13−18+66)

[31] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics [C]// Proc of the Int Conf on Machine Learning. Lille, France: PMLR, 2015: 2256-2265

[32] Kotelnikov A, Baranchuk D, Rubachev I, et al. TabDDPM: Modelling tabular data with diffusion models[J]. arXiv preprint, arXiv: 2209.15421,2022

[33] Hoogeboom E, Nielsen D, Jaini P, et al. Argmax flows and multinomial diffusion: Learning categorical distributions [C]// Advances in Neural Information Processing Systems. New York: Curran Associates, 2021, 34: 12454−12465

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2017 [2023-08-01]. https://proceedings.neuri ps. cc/paper_ files/paper/2017/hash/3f5ee243547dee91fbd053-clc4a845aa-Abstract. html

[35] Yang Bin, Li Haoling, Teng Sikai, et al. Attentional interactive encoder network focused on aspect for sentiment classification[J]. Electronics, 2023, 12(6): 1329

[36] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C] //Advances in Neural Information Processing Systems. New York: Curran Associates. http://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract,html



Yang Bin, born in 1986. PhD. Member of CCF. His main research interests include data mining and natural language processing.

杨　斌, 1986 年生. 博士. CCF 会员. 主要研究方向为数据挖掘、自然语言处理.



Wang Zhengyang, born in 2003. Undergraduate. His main research interests include data mining and artificial intelligence.

王正阳, 2003 年生. 本科生. 主要研究方向为数据挖掘、人工智能.



Cheng Zihang, born in 2003. Undergraduate. His main research interests include data mining and artificial intelligence.

程梓航, 2003 年生. 本科生. 主要研究方向为数据挖掘、人工智能.



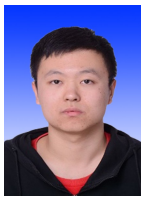
Zhao Huiying, born in 1991. PhD, postdoctoral fellow of China United Network Communications Group Co., Ltd. Her main research interests include graph neural network and autonomous networks.

赵慧英, 1991 年生. 博士, 中国联通集团企业博士后. 主要研究方向为图神经网络、自智网络.



Wang Xin, born in 1988. Master. Her main research interests include artificial intelligence and graph neural network.

王　鑫, 1988 年生. 硕士. 主要研究方向为人工智能、图神经网络.



Guan Yu, born in 1998. Master. His main research interests include software test, data mining, and artificial intelligence.

管　宇, 1998 年生. 硕士生. 主要研究方向为软件测试、数据挖掘、人工智能.



Cheng Xinzhou, born in 1978. Professor level senior engineer. His main research interests include network intelligent operations and artificial intelligence.

程新洲, 1978 年生. 教授级高级工程师. 主要研究方向为网络智能运营、人工智能.