## 一种相关属性生成和纵向联邦填补的参与方样本生成方法

摘要：为了解决纵向联邦学习场景中由某些参与方相对样本缺少而导致的多方样本对齐后联合样本不足的问题，本文提出了一种基于相关属性生成和纵向联邦填补的参与方样本生成方法，简称FedPSG-CAG。该方法采取一种“先生成一些高相关属性值，再用纵向联邦填补方法进行填补”的策略进行参与方缺少样本生成。由于高相关的属性列在生成模型学习的时候能更好的捕获到它们之间的数据分布，本文方法首先计算和获取样本缺少参与方的对齐样本集合中具有高相关性的属性列，训练本地生成模型，用于生成缺少样本中的相关属性值。然后，为了利用多个参与方的数据关联性，在保护数据隐私的同时提高数据生成效果，本文构建了基于GANs的纵向联邦填补模型框架，用于生成参与方缺少样本中的剩余属性值。在该填补框架中，对联邦填补过程、损失函数、训练过程进行了重新设计。在多个公开数据集上，实验充分地验证和分析了our proposed method优于目前可用于参与方样本生成的表现SOTA的基线模型。

关键字：参与方样本生成，纵向联邦学习，相关属性生成，数据填补

# 1 引言

在许多实际应用中，多个组织协同地利用其数据来构建和训练更强大的机器学习模型。例如，银行和电子商务公司可以通过整合各自不同的特征集，改进对共同客户群的信用风险预测。纵向联邦学习（Vertical Federated Learning, VFL）[1]是一种隐私保护的分布式机器学习方法，适用于多个参与方拥有相同样本群体但不同特征的场景。其核心思想是在不直接共享原始数据的情况下，通过加密技术和安全计算协议，实现跨组织的协同建模。在纵向联邦学习过程中，具有相同样本ID但特征不同的多方加密对齐样本称为联合样本。联合样本集利用来自多方的更多特征帮助机器学习模型训练。在机器学习中，除了样本特征外，样本量很重要。样本量直接影响模型的泛化能力、训练稳定性。然而，在纵向联邦学习很多的应用场景中，某些参与方的一些样本相对于其他参与方是缺少的。此时，相对缺少的这些样本无法进行多方加密样本对齐，从而使联邦机器学习模型训练的联合样本较少。这将导致纵向联邦机器学习模型，尤其是复杂模型（如联邦深度学习模型），在训练过程中可能会产生过拟合等问题。我们需要更多的高质量的多方联合训练样本来避免这些问题。

为了解决由某些参与方样本缺少而导致的多方样本对齐后联合样本不足的问题，可采用生成联合样本的方法。目前，有一些基于纵向联邦学习（VFL）的样本生成方法可以实现基于多个参与方的联合样本生成。例如：基于GANs的纵向联邦数据生成方法[2,3 , 4, 5]，和基于（MRFs）的合成表格数据生成框架[6]。但机器学习模型的训练效果不仅与样本的数量有关，更与样本数据的质量等多种因素有关。采用纵向联邦生成方法得到的是一个全新的联合样本，其各方数据都是合成的非真实数据。并且，当某些参与方样本缺少比很高时，其他参与方多出来的未对齐样本也无法加入到联合样本集中，使得用于训练纵向联邦生成模型的联合样本很少，无法训练出表现非常优秀的纵向联邦生成模型。因此，生成联合样本的方法无法确保获得高质量的联合样本。

除了生成联合样本的方法，还可以将其他参与方多出来的未对齐样本保留，为样本缺少的参与方生成样本。这种方式可以使生成的缺少样本与多出来的未对齐样本进行样本对齐，以此获得更多联合样本。这也是本论文探讨的关键问题和技术。目前，有一些生成方法可用于为样本较少的参与者生成样本，例如，基于生成对抗网络的生成方法（GANs）[7]、基于自动编码器的生成方法（AEs）[8]和基于去噪扩散概率模型的生成方法（DDPMs）[9]。这些方法能够通过学习潜在的数据分布创建合成样本来生成高质量数据。 在为样本缺少的参与方生成样本时，这些方法仅从自己一方的数据中学习，不考虑其他参与方数据的关系和影响。但在纵向联邦应用场景中，联合样本中每一方的数据与其他方的数据之间总是存在一定的相关性。尽管如上述例举的一些基于深度学习的生成方法性能良好，但仅在本方内部应用的模型忽略了多方数据相关性对生成结果的作用和影响。

针对上述问题，本文提出了一种基于相关属性生成和纵向联邦填补的参与方缺少样本生成方法（FedPSG-CAG），旨在为多方样本对齐时样本缺少的参与方生成样本，从而提供更多高质量的联合训练样本。我们的主要贡献如下：

1. 本文设计了一种“先在样本缺少方通过属性相关性生成部分样本属性值，再通过纵向联邦填补生成剩余部分样本属性值”的策略，既保证了样本缺少参与方中高相关性属性列在数据生成时的准确性，又可以通过联合其他方数据对样本缺少参与方中低相关性属性列进行有效的填补生成，以获得更高质量的参与方样本。
2. 本文提出了基于GANS的纵向联邦填补框架，并设计了纵向联邦下的填补方法、损失函数优化、训练过程，可用于纵向联邦场景下的多方缺失数据填补。而在本文中，该框架用于对样本缺少参与方中低相关性属性列进行数据填补，以此达到生成完整参与方样本的目的。
3. 我们进行了充分的实验验证，以表明我们的方法在解决参与方缺少样本生成问题以及对纵向联邦分类模型训练的有效性、鲁棒性等方面达到了最先进的性能。

# 2 相关工作

2.1 数据生成方法

除了基于早期机器学习模型的一些数据生成方法，目前表现比较优秀的数据生成方法主要有基于自编码的生成模型、生成对抗网络（Generative Adversarial Networks, GAN）、扩散模型（Diffusion Probabilistic Models，DDPM）等。基于自编码的生成模型，如：自编码器AE[10]、变分自编码器VAE[11]等。L. Xu [12]等人提出的是对变分自编码器（VAE）的改进方法，通过对隐变量与表格数据特征的联合分布进行专门建模，以实现对连续和离散特征的有效生成与重构。GAN[7]是 I. J. Goodfellow 等人在2014年提出的模型，GAN受博弈论的启发，内部有生成器（G）和判别器（D）两个网络，两个网络相互对抗博弈达到有效生成数据的目的。M. Mirza 等人[13]提出的CGANs（Conditional Generative Adversarial Nets）在生成模型和判别模型的建模中均引入条件变量，即数据的标签，将GANs从无监督学习变成有监督学习。2019年，L. Xu 等人[12]提出了一种基于Conditional GAN的CTGAN来生成表格数据，用于对表格数据分布和样本行进行建模。2021年，Z. Zhao 等人[14]在CTGAN的基础上，提出了CTAB-GAN用于对不同数据类型的建模，包括连续变量、分类变量、混合特征变量等类型。DDPM[9]是 J. Ho 等人提出的扩散概率模型，通过正向的逐步加噪过程将数据分布转化为标准正态分布，并学习逆向去噪过程，从高斯噪声中生成目标数据，生成质量高效率相对较低。在此基础上，A. Kotelnikov 等人[15]提出了一种基于扩散模型的表格数据生成方法TabDDPM，通过特定的噪声添加和去噪过程，有效捕捉了数值型与类别型特征间的复杂关系。

2.2 数据填补方法

当样本中某些数据元素出现缺失时，无法给机器学习任务（如分类、预测等）的模型训练提供更多优质训练样本，需要对样本中的缺失数据元素进行生成填补。目前主要的数据填补方法包括基于统计的填补方法、基于传统机器学习的填补方法、基于深度学习的填补方法等。统计填补方法主要利用统计特征对缺失值进行填充，例如：均值填补（Mean Imputation）[16]，K近邻填补（K-Nearest Neighbors Imputation, KNNI）[17]。传统机器学习预测填补方法则尝试通过学习复杂的特征关系来预测缺失值。XGBoost填补（XGBI）[18]用XGBoost构建预测模型，对缺失值进行回归或分类预测，其适用性与模型性能密切相关。MissForest填补（MissFI）[19]基于随机森林迭代预测缺失值，可在非线性关系下保持较高填补精度。链式方程多重插补（MICE）[20]通过线性回归或广义线性模型迭代地为每个变量创建填补模型，适合多重插补场景。基于深度学习的填补算法包括：基于自编码器的填补方法，基于GANs的填补方法和基于扩散模型的填补方法。自编码器（Autoencoders, AEs）[21]通过编码-解码结构将数据嵌入潜在空间，并在解码时填补缺失值，可在低维表示中保持主要数据特征。J. Yoon 等人[22]将GAN用于缺失数据生成填补，提出了GAIN模型。2021年，S. E. Awan 等人[23]受CGAN的启发，提出了条件生成对抗填补网络（Conditional Generative Adversarial Imputation Network, CGAIN），将样本中分类标签的类别信息作为生成器和判别器的输入，在解决样本不平衡问题的同时，对样本的缺失部分进行了缺失数据填补。Y. Wang 等人[24]提出了伪标签条件生成对抗网络（Pseudo-label Conditional Generative Adversarial Imputation Networks, PC-GAIN），通过引入伪标签作为条件信息，增强生成模型对数据分布的捕捉能力。X. Miao 等人[25]提出的VGAIN是一个在GAN基础上融合变分自编码器（VAE）思想的填补模型。TabCSDI[26]通过扩散过程的前向添加噪声与反向去噪过程，逐步重构缺失值，实现对数据分布更灵活和更平滑的逼近。

2.3 纵向联邦数据生成和数据填补方法

联邦学习（Federated Learning, FL）最初由 B. McMahan 等人[27]提出，是一种分布式机器学习框架。在数据隐私和安全保护下，联邦学习允许多个参与方在不共享原始数据的情况下，协同训练一个全局模型。针对不同的应用场景，联邦学习模型将分为横向联邦学习、纵向联邦学习等不同类别。其中，纵向联邦学习[1，28]（Vertical Federated Learning, VFL）适用于不同参与方拥有相同样本但具有不同属性特征的数据分布场景。

在纵向联邦学习的应用场景中，当训练数据集中联合样本量不足时，可以通过样本生成的方式来进行数据增强。J. Zhang 等人[2]提出一种基于生成对抗网络(GAN)的纵向联邦学习数据增强方法FedDA。X. Jiang 等人 [3]提出了VertiGAN框架,包括一个多输出全局生成器和多个本地判别器,生成一个高效用的合成集成数据集。GTV[4]是专门针对表格数据GAN的VFL框架，用于生成高保真合成表格数据。X. Yuan 等人 [5] 提出了另一个基于纵向联邦学习（VFL）的生成对抗网络——VFLGAN。W. Lin 等人[6]提出了VertiMRF，用于在纵向联邦学习环境中生成合成数据，并为所有共享信息提供差分隐私保护。

在实际应用中，当多个本地数据集的样本中某些数据元素出现缺失时，无法给纵向联邦机器学习任务（如联邦的分类、预测等）的模型训练提供更多优质训练样本。并且，在纵向联邦下的各方数据之间存在相关性。因此，也需要纵向联邦的数据填补模型对这些缺失的数据元素进行填补。W. Du 等人[29]在其论文中指出缺失值填补的研究在纵向联邦应用场景中的探索仍然较为缺乏。并提出了一种隐私保护下的纵向联邦KNN特征填补方法，但通过KNN计算的方式无法满足高质量样本数据填补的需求。

综上所述，目前有比较多优秀的生成方法、填补方法，但在纵向联邦场景下的数据生成方法和数据填补方法却不多见。尽管已有研究展示了在VFL中训练如生成对抗网络（GAN）等其他模型以生成数据的可行性，但在不同实际应用场景中仍面临诸多挑战。

# 3 问题描述

假设在一个给定的纵向联邦学习场景中，有N个数据持有者（N个参与方）和一个中央服务器。为了清楚地说明本文解决的问题和所采用的方法，我们以A、B两个参与方为例，它们的可信合作方C作为中央服务器。A方和B方持有敏感数据，在纵向联邦学习时，需要在保留数据隐私的前提下协作地进行模型训练。

我们将预处理技术应用于 A、B方持有的数据，包括数据清理、规范化和特征编码。类别型特征使用 one-hot 编码处理，数值特征使用标准化缩放进行归一化。经过数据预处理后，A方和B方根据其样本ID空间进行对齐。为了在样本对齐过程中保护数据隐私，我们使用基于盲RSA的私密集合交集（PSI）协议[ 30]进行加密样本对齐。该协议使得所有参与方可以安全地计算他们数据集的交集，而不会泄露任何关于他们持有的样本的额外信息。

设A方数据集为 ，，其中 ， 是A方的第i个样本， 是第i个样本的第m个属性，，，a表示A方样本的属性个数， 表示A方的样本数量。设B方数据集为， ，其中 ， 是B方的第i个样本， 是第i个样本的第n个属性，， ，b表示B方样本的属性个数， 表示B方的样本数量。A方和B方采用加密样本对齐技术对其数据集 和 进行对齐，获得 个对齐样本，是A方样本的样本ID集合，是B方样本的样本ID集合。A方的 - 个样本，无法与B方进行样本对齐。B方的 - 个样本，无法与A方进行样本对齐。A方对齐的样本集表示为，未对齐的样本集为。B方对齐的样本集表示为，未对齐的样本集为。如图1所示，的样本数量和是相同的，且它们的样本ID是一致的。

因此，本文解决的问题是：在纵向联邦学习场景，联合利用多方数据的关联性，生成A方中相对于的缺少样本，和B方中相对于的缺少样本，使其与A或B方中的未对齐样本进行对齐，获得更多基于真实数据的联合样本，以提高纵向联邦学习模型训练的有效性。

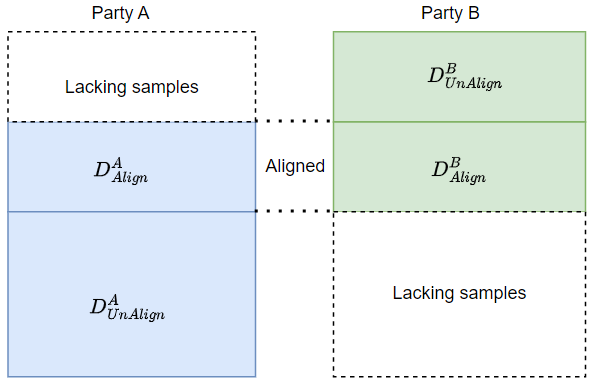
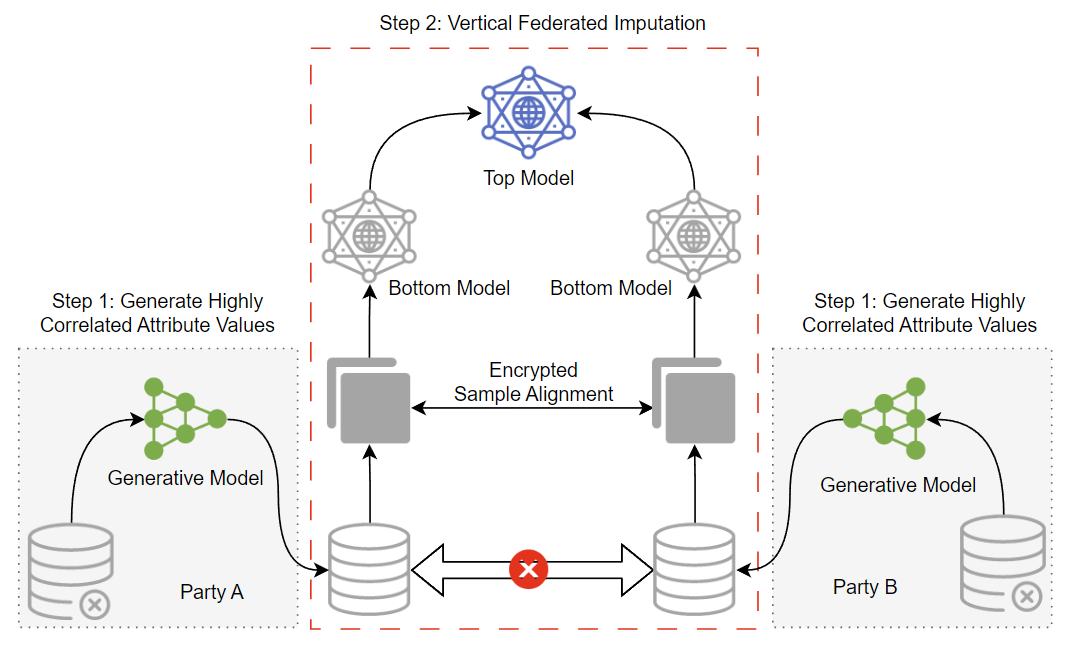


图1 参与方缺少样本示意图

Figure 1 Diagram of participants lacking samples

# 4 Methods

该模型的整体结构可以概括为两个阶段：第一阶段侧重于本地生成高相关属性值。各参与方（如A方与B方）基于其本地数据中高度相关的属性列，训练并使用各自的生成模型先行补全这部分关键属性，以确保后续处理时拥有更丰富的初始信息。第二阶段引入纵向联邦填补框架，通过加密样本对齐、安全多方计算等手段，将各方本地补全后的数据中仍有缺失的属性提交到底部模型和顶部模型中进行联合生成，利用GANs等生成对抗结构来补全剩余的缺失属性。底部模型在各参与方本地进行特征表示和深度学习，顶部模型则在安全聚合端完成全局联邦填补，从而确保隐私不泄露的前提下，最大程度发挥多方数据的协同效益。



4.1 相关属性生成

由于高相关的属性列在生成模型学习的时候能更好的关注和捕获到各属性之间的数据分布，因此，本文将计算和获取样本缺少参与方的对齐样本集合中具有高相关性的属性列，用于相关属性生成模型的学习和训练。以B方的相关属性生成为例，详细过程如下：

计算B方的高相关属性采用Spearman相关性系数[31]：对于B方第m个属性列，第n个属性列，，分别循环遍历、中的每个属性值、，并对其赋排名。排名从1开始，排名相同的值使用平均排名（即将相同值的排名求平均），得到每个属性值、的排名、。计算与之间的差值：

根据差值计算Spearman相关性系数，得到相关性系数：

其中，为B方对齐样本集中的样本数，即。

根据公式（2）求出B方所有属性之间的相关性系数矩阵MB：

设需生成的B方相关属性列集合为空，循环遍历相关性矩阵中的每一个值，找出当前矩阵中最大的相关性系数Max()，并据此确定具有强相关性的属性对)。如果)属性对中的或都不存在于中，则将或加入中；如果或已存在于中，则将该属性对中的另一个属性列加入中，并将的值重置为 0。重复以上过程，直到中的B方相关属性生成列数达到设定的要求为止。表1和表2展示了本文实验中Bank和Credit Dataset在B方不同相关属性生成列数下的，以及每加入一个属性列所涉及的属性对和相关性系数。

当获得B方相关属性列集合后，将其在B方进行本地生成模型的训练，并将训练好的生成模型用于生成B方缺少样本中的相关属性值。可采用的生成模型例如：CTGAN、TableGAN、CTABGAN、VAE和TabDDPM等。

同理，计算A方所有属性之间的相关性系数矩阵，以及相关属性列集合。将在A方进行本地生成模型的训练，用于生成A方缺少样本中的相关属性值。

在多个参与方的应用场景中，如果某个或某些参与方也存在相对于其他方的样本缺少，也采用上述过程生成这个或这些参与方缺少样本中相关属性。生成这些参与方缺少样本的相关属性值后，基于纵向联邦学习的填补模型可以生成样本中的剩余属性值（如第4.2节所述），最终完成参与方缺少样本的生成。

表1 Bank Dataset中B方不同相关属性生成列数的

|  |  |  |  |
| --- | --- | --- | --- |
| 的列数 | 加入 的属性 | 新加入属性所在的属性对 | 相关系数 |
| 1 | ['pdays'] | ['pdays', 'previous'] | 0.98564 |
| 2 | ['pdays', 'previous'] | ['pdays', 'previous'] | 0.98564 |
| 3 | ['pdays', 'previous', 'month'] | ['month', 'contact'] | 0.82003 |
| 4 | ['pdays', 'previous', 'month', 'contact'] | ['month', 'contact'] | 0.82003 |
| 5 | ['pdays', 'previous', 'month', 'contact', 'poutcome'] | ['month', 'poutcome'] | 0.5566 |
| 6 | ['pdays', 'previous', 'month', 'contact', 'poutcome', 'campaign'] | ['month', 'campaign'] | 0.1474 |
| 7 | ['pdays', 'previous', 'month', 'contact', 'poutcome', 'campaign', 'day'] | ['campaign', 'day'] | 0.13958 |
| 8 | ['pdays', 'previous', 'month', 'contact', 'poutcome', 'campaign', 'day','duration'] | ['day','duration'] | 0.108 |

表2 Credit Dataset中生成B方不同相关属性列数的

|  |  |  |  |
| --- | --- | --- | --- |
| 的列数 | 加入 的属性 | 新加入属性所在的属性对 | 相关系数 |
| 1 | ['credit\_limit'] | ['credit\_limit','credits\_num'] | 0.829 |
| 2 | ['credit\_limit', 'credits\_num'] | ['credit\_limit','credits\_num'] | 0.829 |
| 3 | ['credit\_limit', 'credits\_num', 'property'] | ['credit\_limit', 'property'] | 0.79865 |
| 4 | ['credit\_limit','credits\_num','property','history\_credit'] | ['credits\_num','history\_credit'] | 0.50602 |
| 5 | ['credit\_limit','credits\_num','property','history\_credit', 'saving\_account'] | ['property','saving\_account'] | 0.30679 |
| 6 | ['credit\_limit','credits\_num','property','history\_credit','saving\_account','income\_installment\_rate'] | ['credit\_limit','income\_installment\_rate'] | 0.303 |
| 7 | ['credit\_limit','credits\_num','property','history\_credit','saving\_account','income\_installment\_rate', 'account\_status'] | ['saving\_account','account\_status'] | 0.2427 |
| 8 | ['credit\_limit','credits\_num','property','history\_credit','saving\_account','income\_installment\_rate', 'account\_status', 'credit\_purpose'] | ['income\_installment\_rate', 'credit\_purpose'] | 0.10287 |
| 9 | ['credit\_limit','credits\_num','property','history\_credit','saving\_account','income\_installment\_rate', 'account\_status', 'credit\_purpose','period'] | ['account\_status','period'] | 0.0627 |
| 10 | ['credit\_limit','credits\_num','property','history\_credit','saving\_account','income\_installment\_rate','account\_status','credit\_purpose','period','other\_debtor'] | ['history\_credit','other\_debtor'] | 0.0467 |
| 11 | ['credit\_limit','credits\_num','property','history\_credit','saving\_account','income\_installment\_rate','account\_status','credit\_purpose','period','other\_debtor','installment\_plans'] | ['account\_status','installment\_plans'] | 0.0429 |

## 4.2 基于GANs的纵向联邦数据填补方法

当我们为参与方缺少样本生成相关属性值后，还需要对样本的剩余属性值进行生成以获得完整的样本信息。为此，本文提出了一种基于生成对抗网络的纵向联邦填补框架，在各方数据安全隐私保护下，使用生成对抗网络（GAN）协同多方数据对样本缺少的参与方的剩余属性值的进行联合填补。该纵向联邦填补框架包括各方底部模型，以及顶层模型，整个架构设计如图2所示。在整个纵向联邦填补过程中，我们使用全同态加密[32]（Cheon-Kim-Kim-Song, CKKS）进行数据加密运算。

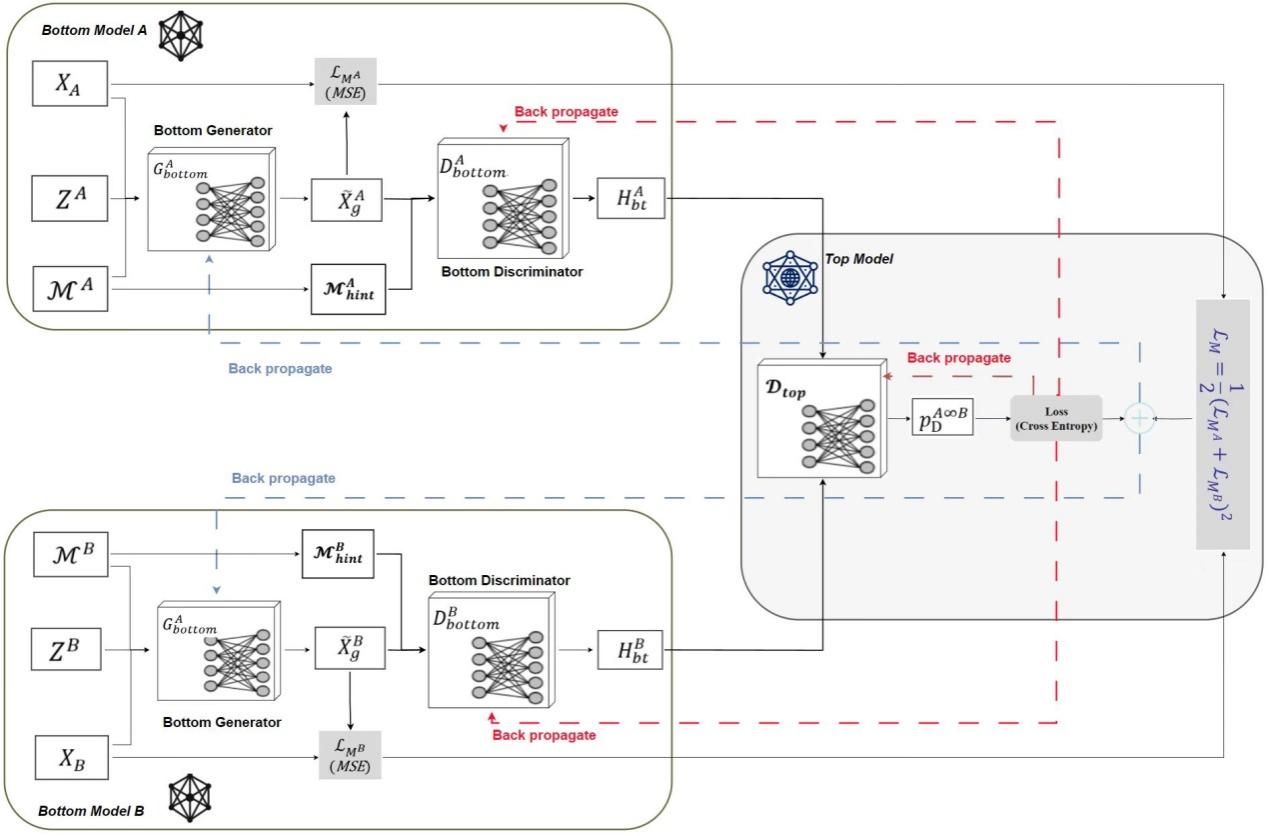


Fig.2：The framework of vertical federated imputation Based on GANs

（1）掩码向量（Mask Hint）

在数据填补过程中，掩码向量扮演着至关重要的角色。掩码向量指示了填补运算过程样本中剩余需填补的属性的位置，各方生成器利用这一信息在联合样本中各方已知数据基础上生成剩余属性值。

设A方样本的特征属性编码向量为，B方样本的特征属性编码向量表示为。在本文中，基于GANs的联邦填补模型对数值型变量归一化使用Min-Max归一化，使用One-Hot对类别变量进行编码。设A方掩码向量为，，表示A方样本向量维度数。B方掩码向量为，，表示B方样本向量的维度数。



将A方、B方剩余需填补的样本数据分别表示为A方、B方的缺失数据，用向量、表示：

， （4）

即

，（5）

其中，表示向量间的元素级乘法；



（2）底部模型

在底部模型中，我们为每个参与方（如图2中的A方和B方）分别设计了生成器（Generator）和判别器（Discriminator）。各底部生成器的主要任务是根据各参与方本地数据的联邦加密协作学习为样本缺少的参与方的缺失数据生成合理的填补数据，而判别器则负责判断生成的数据是否接近真实数据，从而对生成器进行反馈优化。需要注意的是，在第 4 部分的实验中，我们提及了多种基于GAN的纵向联邦填补模型，包括 VF-GAIN、VF-VGAIN 和 VF-CGAIN。这些纵向联邦填补模型在整体架构上相似，主要区别在于各自的底层模型的网络结构。它们的各本地网络结构与原本的填补模型（GAIN、VGAIN、CGAIN）的结构一致。GAIN、VGAIN、CGAIN的具体模型结构和实现细节可以参考这些论文 [22]、[25]、[23]。

在联合学习过程中，A方底部生成器输入包括：缺失样本数据、掩码向量和随机噪声向量。B方底部生成器输入包括：缺失样本数据、掩码向量和随机噪声向量。A、B方底部生成器的生成结果为：

（6）

由A方生成的样本数据和B方生成的样本数据构成的联合样本用表示:

（7）

其中，因为两方数据有数据安全隐私要求，表示纵向联邦学习下的连接，它不是基于A、B方数据的直接连接，只是为了方便联合样本的表示。

由A方、B方生成的向量、和原缺失样本向量、的观测数据，得到填补后样本、：

， （8）

此时的A、B两方可以看成是一个基于联合样本执行的填补。生成填补后的联合样本用表示，由A方数据和B方数据构成，如公式(9)所示:

（9）

和B方生成的中的部份，将用于联邦学习过程中生成器总损失的运算。

为了使判别器专注于评估特定部分的填充质量，确保更准确地区分观测数据和填充数据，本文引入了GAIN的提示掩码来帮助训练模型。提示掩码依赖于，即定义一个超参数作为随机抽样的概率，使掩码向量中值为1的概率随机置为0。本文中，超参数的取值参照论文模型中的最优值0.9[33]。

判别器底部网络采用全连接神经网络结构，主要包括包含两个隐藏层，每个隐藏层使用LeakyReLU激活函数，并在其中采用dropout以一定概率忽略部分神经元以减少模型的过拟合现象。本文中，LeakyReLU的负斜率设置为0.2，Dropout比例设置为0.5。A方底部判别器的输出是由和A方提示掩码得到的向量，用表示，B方底部判别器的输出是由和B方提示掩码得到的向量，用表示：

（10）

（3）顶部模型

顶部模型在纵向联邦填补框架中主要用于协作 A 方和 B 方的数据填补过程，在各方底部判别器的基础上，通过一个顶层判别器对生成填补数据的质量进行进一步评估和优化。服务器端判别器顶部网络采用全连接网络，拼接输入A、B两方底部判别器网络输出的隐藏向量，包含两个隐藏层，每个隐藏层使用LeakyReLU激活函数。本文中，LeakyReLU的负斜率设置为0.2，Dropout比例设置为0.5。将A、B两方底部判别器输出的向量和输入服务器端判别器顶部网络，得到输出的概率向量：

(11)

其中，

表示由A、B两方底部判别器得到的向量、在纵向联邦学习下的连接。表示由A、B两方提示掩码向量、在纵向联邦学习下的连接。 是鉴别器区分联合样本中元素的估算数据和真实数据的概率向量。

（4）损失函数和训练过程

整个纵向联邦填补的损失函数包括生成器损失函数和判别器损失函数。整个优化过程参照生成对抗网络（GAN）的对抗训练机制。具体而言，生成器通过最小化生成数据与真实数据之间的差异，逐步提高生成数据的质量和真实性；而判别器则通过最大化区分真实数据和生成数据的准确性，增强其判别性能。这种训练目标可以形式化地表示为：

（12）

首先，我们固定生成器，训练判别器。每轮从联合样本集中选择最小批次样本量，最小批次中的每个样本，包括A方、，B方、。

判别器的总损失函数为：

（13）

为第条样本的提示掩码向量中第个元素，为第条样本中第个元素。需特别说明的是，对于VF-CGAIN，由于其底部模型与CGAIN一致，需要加入类别标签，即，，其中，表示第条样本的标签向量。

紧接着，我们固定判别器，训练生成器。对于纵向联邦学习下的联合样本，我们每轮从中选择最小批次样本量，对于最小批次中的每个样本，其中A方包括、、，B方包括、、。

生成器的总损失函数为：

（14）

其中，

A方的损失函数为：

B方的损失函数为：

其中，为该批次联合样本集中的第条样本。在A方，为A方第条样本中的第个元素，为A方观测数据，为A方的生成数据，和分别为第条样本和的第个元素。在B方，为B方第条样本中的第个元素，为B方观测数据，为B方的生成数据，和分别为第条样本和的第个元素。A、B两方损失函数中的为超参数，在本文实验中超参数取值参照GAIN方法中的超参数设定。

基于GANs的纵向联邦填补方法的训练过程如表3所示。设作为条件向量C的标签列在B方，顶部模型在B方。表3中，、分别表示A方的底部生成器权重参数和底部判别器权重参数，、分别表示B方的底部生成器权重参数和底部判别器权重参数，表示顶部判别器权重参数。、分别表示A方和B方的随机掩码。顶层判别器网络第一层的权重梯度，分别表示A方和B方底部判别器网络最后一层的权重梯度。表示全同态加密（CKKS）。

表3 基于GANs的纵向联邦填补模型的训练步骤

|  |  |  |  |
| --- | --- | --- | --- |
| 步骤 | A方 | B方 | C方 |
| 步骤1 | 初始化、 | 初始化、、 | 创建加密密钥对，并将公共密钥发送给A方和B方 |
| 步骤2 | 计算、，并将、和发送B方 | 计算、并进行加密得到、。将和联合为，和联合成，计算、、和并将、发送给C方 |  |
| 步骤3 | 初始化，利用计算+，并将其发送给C方。  计算+，并将其发送C方。 | 初始化，计算+，  将拆分成为和，利用计算+，并将其发送给C方，将发送给A方.  计算+，并将其发送C方。 | 解密和、+、+、+、+、+，将+，+发送给A方，将+、+、+发送给B方 |
| 步骤4 | 更新、 | 更新、、 |  |

# 4 实验

## 4.1 数据集和数据准备

为了验证我们所提方法在样本缺少方样本生成的有效性，本文采用了四个数据集from the UCI repository[34 ]进行了充分的实验：

① 银行营销数据集[ ]涉及葡萄牙一家银行机构的直接营销活动。数据集包含 45,211 个实例和 16 个属性，以及一个 ID 列和一个标签列。该数据集的目的是对客户是否会认购定期存款进行分类。

② 德国信贷数据集 [ ]源自信用评分系统，目标是将申请人划分为信用 “良好 ”或 “不良”。数据集包含 1000个样本，每个样本包括 21 个属性，涵盖财务、人口和社会特征： 13 个分类属性特征、7 个数字属性特征和一个标签列。

**③ Letter Recognition 数据集**[ ]用于字符识别任务，包含20,000个实例和16个属性，每个实例包括手写字母图像的统计特征，如尺寸、形状和轮廓等。该数据集旨在将每个样本分类为26个英文字母之一。

**④ Online News Popularity 数据集**[ ]用于预测在线新闻文章的流行度，包含39,644个实例和60个属性。该数据集涉及文章内容、社交媒体互动（如分享次数、点赞数）、发布时间等信息。

实验中，将这四个数据集分别称为“银行”、“信贷”、“字母”和 “新闻”。在“银行”和“信贷”数据集的实验中，根据它们特征所有权，在A方和B方之间都进行了垂直划分，A方包含客户信息，B方包含银行信息。具体来说，在银行数据集中，A方除 ID 列的属性有8 列，而B方除 ID 列的属性有 8 列。在信贷数据集中，A方除 ID 列的属性有9列，而B方除 ID 列的属性有 11 列。在“字母”和 “新闻”数据集的实验中，我们仅模拟了按原有属性列顺序均分给A、B两方的纵向联邦场景，以进一步展示除“银行”和“信贷”数据集之外的其他验证性实验结果。

同时，为了验证所提方法的有效性，我们重点针对图1中当=∅的情况进行了充分的实验论证。≠∅的情况可进行同理推断。当=∅时，本文实验中为A方和B方设置了不同的相对样本比例，即B方相对于A方的样本缺失比例。样本缺失比例为 0.2 表示B方有 80% 的样本可以与A方对齐，而 20% 的样本在B方相对于A方是缺失的。缺失比例为 0.5 表示B方有 50% 的样本可以与A方对齐，而 50% 的样本是缺失的。缺失比例为 0.8 表示B方只有 20% 的样本能够与A方对齐，而 80% 的样本在B方是缺失的。我们的方法旨在生成B方相对于A方缺失的样本，确保由两方构建的联合样本集保留A方的全部样本。本文方法通过基于更多真实数据增加联合数据集的样本量，在保证样本量相同的情况下，提高联合样本集的数据质量。在本文方法中，LackR-B表示B方相对于A方的样本缺少比，Cnum表示B方相关属性的生成列数。

## 4.2 实验一：FedPSG-CAG中各种不同设定的实验比较和分析

（1）实验一的实验设计

实验设定一：固定方法中所采用的生成模型和填补模型，设置不同的B方相关属性生成列数，以验证本文方法在参与方生成样本任务中的最终效果。其中，FedPSG-CAG中的GM采用TabDDPM生成B方缺少样本的相关属性值，TabDDPM[15]是目前文献中生成效果表现SOTA的生成模型（本文实验二也进行了验证）；生成B方样本剩余属性值的填补模型采用VF-GAIN。

实验设定二：固定方法中B方相关属性的生成列数，采用不同填补模型生成B方样本剩余属性值，以验证在不同的填补模型下，本文方法在参与方生成样本任务中的最终效果。其中，B方相关属性的生成列数是由实验设定一实验结果中的最佳列数确定。FedPSG-CAG中的GM仍采用TabDDPM。用于对比的填补模型为目前效果较好的几种模型：Mean[16 ]、MissFI（其中随机森林中树的棵数设为100）[19 ]、MICE[20 ]、GAIN[22 ]、VGAIN[ 25]、CGAIN[23 ]。实验中，将它们在纵向联邦框架下进行部署和实现，被表示为VF-MissFI，VF- MICE，VF- GAIN， VF- VGAIN, VF- CGAIN。Mean的填补方式是对需填补的当前列计算均值进行填补，其填补过程不涉及参考和结合其他方的数据，所以没有纵向联邦计算。

实验设定三：固定方法中B方相关属性的生成列数，验证多方纵向联邦填补和仅本地非联邦填补对B方缺少剩余属性值的生成效果。一种方式为非联邦填补：基于B方原有数据和使用GM生成获得的B方相关属性列，对B方缺少样本剩余属性列进行数据填补；另一种方式为联邦填补：基于A方、B方原有数据和使用GM生成获得的B方相关属性列，纵向联邦地对B方剩余属性列进行数据填补。FedPSG-CAG中的GM仍采用TabDDPM。填补模型采用了实验设定二的实验结果中较优的GAIN、CGAIN、VGAIN的联邦和非联邦实现。

本实验用于评估样本生成效果的评估指标均为RMSE[35 ]，用于表示生成样本中数据与真实数据之间的平均偏差程度。实验采用了4.1小节提及的‘Bank’ and ‘Credit’两个数据集，这两个数据集代表了样本规模相对较大和较小的两种情况。本实验中，基于GANs的纵向联邦/非联邦填补模型，以及本文方法中涉及的生成模型，它们的训练迭代总数都设为 10,000 次，epoch 数量为 10，学习率设置为0.001，优化器使用Adam。

（2）实验一的结果分析

表4、表5和图3展示了实验设定一的实验结果。当Cnum=0时，表示没有在B方先用GM生成属性列，B方缺少样本的所有属性均采用联邦填补模型生成；当Cnum=1时，表示在B方先用GM生成B方缺少样本的1个属性列，样本的剩余属性采用联邦填补模型生成；以此类推...。如表4、表5和图3所示，在B 方不同的样本缺少比下，当Cnum为4、5、6时，FedPSG-CAG得到的RMSE值表现更优，如表1和表2中的红色和蓝色字体表示的结果。而Cnum=0时，即B方缺少样本的所有属性均采用联邦填补模型的方式进行生成，其RMSE值的结果较差。这说明为样本缺少的参与方生成样本时，采取“先生成一些高相关属性值，再用纵向联邦填补方法进行填补”的方式是可行的。并且，根据B方样本的属性相关性，能更多的为B方缺少样本生成一些相关性高的属性，再进行联邦填补，生成获得的样本的RMSE更低，更利于得到接近真实的样本集。但是，值得注意是的，随着B方相关属性生成列数的增加，如表4、表5和图3所示，生成样本的RMSE也会有所增加，这说明并非生成的B方相关属性列数越多越好，一些非高相关属性列更适合用联邦填补的方式进行生成。因为，生成过多的属性列，特别是关联性较低的属性，可能会导致生成模型在学习过程产生误差，降低生成数据的质量。而采用联邦填补方法生成非高相关属性列，可以充分利用多方参与者的数据协同作用，以联合填补的形式提升生成数据的准确性和一致性。因此，通过生成部份高相关性属性列，结合联邦填补方法生成填补剩余属性列，能够在一定程度上保证生成样本的数据质量。

**表4** 不同的B方相关属性生成列数in FedPSG-CAG的RMSE结果（**Bank Dataset**）

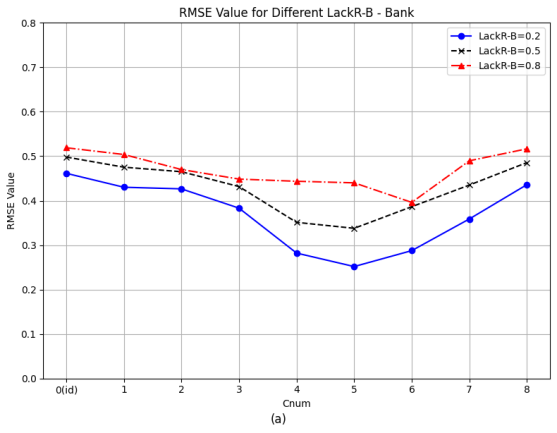
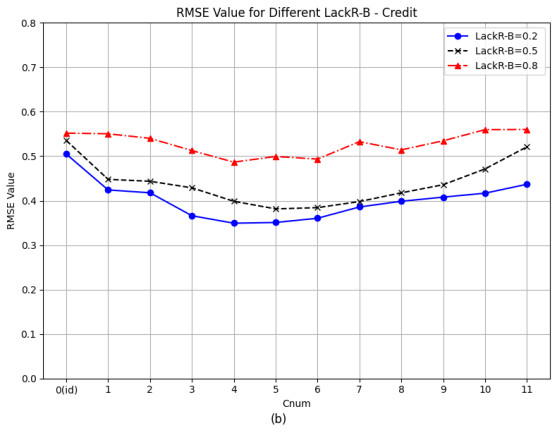
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cnum  LackR-B | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.2 | 0.4615 | 0.4302 | 0.4265 | 0.3832 | 0.2821 | 0.2518 | 0.2876 | 0.3588 | 0.4360 |
| 0.5 | 0.4981 | 0.4753 | 0.4651 | 0.4317 | 0.3512 | 0.3378 | 0.3864 | 0.4351 | 0.4856 |
| 0.8 | 0.519 | 0.5038 | 0.4700 | 0.4485 | 0.4436 | 0.4401 | 0.3619 | 0.4897 | 0.5166 |

**（注：红色字体表示最优，蓝色字体表示次优）**

**表5** 不同的B方相关属性生成列数in FedPSG-CAG的RMSE结果（**Credit Dataset**）

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cnum  LackR-B | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 0.2 | 0.5048 | 0.4241 | 0.4177 | 0.3661 | 0.3493 | 0.3509 | 0.3604 | 0.3858 | 0.3986 | 0.4079 | 0.4168 | 0.4368 |
| 0.5 | 0.5354 | 0.4479 | 0.4436 | 0.4289 | 0.3986 | 0.3815 | 0.3843 | 0.3980 | 0.4176 | 0.4358 | 0.4714 | 0.521 |
| 0.8 | 0.5518 | 0.5504 | 0.5402 | 0.5127 | 0.4866 | 0.4994 | 0.4935 | 0.5327 | 0.5142 | 0.5142 | 0.5597 | 0.5602 |

**（注：红色字体表示最优，蓝色字体表示次优）**

** **

**图3** 不同的B方相关属性生成列数的RMSE折线图by VF-PSG: (a) Bank Dataset; (b) Credit Dataset

表6展示了实验设定二的实验结果。实验中，B方相关属性生成列数为实验设定一实验结果中相对较优的列数---5列。在B 方不同的样本缺少比下，在实验设定二中，FedPSG-CAG的最优和次优RSME结果都出现在当填补模型为VF-GAIN、VF-VGAIN、VF-CGAIN时。其中，Bank数据集的最优结果出现在填补模型为VF-VGAIN、VF-CGAIN时；而Credit数据集的最优结果出现在填补模型为VF-VGAIN时。这些结果表明，基于GAN改进的纵向联邦填补模型的有效性。两个数据集最差RSME结果出现在当填补模型为Mean时，因为Mean的填补方式是一种对当前属性列所有值的统计计算，其填补过程不参考和结合其他方的数据，且均值计算得到的填补值本身误差较大。

**表6** FedPSG-CAG中不同填补模型的RMSE结果（**Bank and Credit Dataset**）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Imputation  Dataset&LackR-B | | Mean | VF-MissFI | VF-MICE | VF-GAIN | VF-VGAIN | VF-CGAIN |
| Bank | 0.2 | 0.4077 | 0.4036 | 0.4214 | 0.2518 | 0.2372 | 0.2691 |
| 0.5 | 0.4744 | 0.4563 | 0.4047 | 0.3378 | 0.2997 | 0.2722 |
| 0.8 | 0.5287 | 0.4776 | 0.4520 | 0.3468 | 0.3537 | 0.3832 |
| Credit | 0.2 | 0.4815 | 0.4481 | 0.4512 | 0.3509 | 0.3495 | 0.4376 |
| 0.5 | 0.5006 | 0.4516 | 0.4911 | 0.3815 | 0.3610 | 0.4522 |
| 0.8 | 0.5437 | 0.4928 | 0.5210 | 0.4994 | 0.4621 | 0.4657 |

**（注：红色字体表示最优，蓝色字体表示次优）**

表7展示了实验设定三的实验结果。首先，我们将Cnum分别设置为1列、5列、8列，使用GM为B方生成缺少样本的相关属性列后，再以联邦和非联邦填补方式对B方缺少样本的剩余属性列进行生成。表7中的联邦填补方式是指利用A、B两方原有和B方生成属性列进行基于VF-GAIN模型的填补生成，非联邦填补方式是指仅利用B方原有和B方生成属性列进行基于本地GAIN模型的填补生成。如表7的实验结果所示，采用GM生成B方缺少样本的相关属性列后，采用基于VF-GAIN的联邦填补方式为B方生成剩余属性，全面优于基于本地GAIN模型的非联邦填补方式。因为，基于联邦的填补方式在填补生成过程中更多地引入多方数据进行协作学习，更多数据特征的加入将会提高联合样本中B方缺少数据的生成质量。同时，当Cnum=5时，在B 方不同的样本缺少比下，FedPSG-CAG中联邦和非联邦填补方式的实验结果都优于生成重要属性的列数为1列和8列时，这与实验设定一的结论是一致的。为了进一步验证在实验设定三下不同填补模型的实验效果，设置Cnum=5，采用GAIN、VGAIN、CGAIN进行了联邦和非联邦填补的对比。如表8所示，在FedPSG-CAG中基于VF-GAIN、VF-VGAIN、VF-CGAIN的联邦填补方式均全面优于基于本地GAIN、VGAIN、CGAIN的非联邦填补方式，这于表7的实验结论一致。特别地，基于VF-VGAIN的FedPSG-CAG是三种基于GANs的联邦填补方法中表现较优的。

表7 相关属性生成后，GAIN联邦和非联邦的RMSE结果，when Cnum=1,5,8 （Bank and Credit Dataset）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VF-Imputed(Y/N)  Dataset(LackR-B) | | 1 | | 5 | | 8 | |
| VF-GAIN | GAIN | VF-GAIN | GAIN | VF-GAIN | GAIN |
| Bank | 0.2 | **0.4302** | 0.4477 | **0.2518** | 0.3305 | **0.4360** | 0.4372 |
| 0.5 | **0.4753** | 0.491 | **0.3378** | 0.3524 | **0.4856** | 0.5072 |
| 0.8 | **0.5038** | 0.5149 | **0.3468** | 0.4401 | **0.5166** | 0.5418 |
| Credit | 0.2 | **0.4241** | 0.5542 | **0.3509** | 0.5095 | **0.3986** | 0.5513 |
| 0.5 | **0.4479** | 0.5923 | **0.3815** | 0.5108 | **0.4176** | 0.5467 |
| 0.8 | **0.5504** | 0.5791 | **0.4994** | 0.5314 | **0.5142** | 0.5551 |

表8 相关属性生成后，不同填补模型联邦和非联邦的RMSE结果，when num=5（Bank and Credit Dataset）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VF-Imputed(Y/N)  Dataset(LackR-B) | | 5列 | | 5列 | | 5列 | |
| VF-GAIN | GAIN | VF-VGAIN | VGAIN | VF-CGAIN | CGAIN |
| Bank | 0.2 | **0.2518** | 0.3305 | **0.2372** | 0.2764 | **0.2691** | 0.3531 |
| 0.5 | **0.3378** | 0.3524 | **0.2997** | 0.3236 | **0.2722** | 0.3336 |
| 0.8 | **0.3468** | 0.4401 | **0.3537** | 0.386 | **0.3832** | 0.4018 |
| Credit | 0.2 | **0.3509** | 0.5095 | **0.3495** | 0.4289 | **0.4376** | 0.4618 |
| 0.5 | **0.3815** | 0.5108 | **0.3610** | 0.4573 | **0.4522** | 0.5003 |
| 0.8 | **0.4994** | 0.5314 | **0.4621** | 0.4846 | **0.4657** | 0.5059 |

**（注：红色字体表示最优，蓝色字体表示次优）**

## 4.3 实验二：FedPSG-CAG与SOTA基线模型的比较和分析

为了验证本文方法的有效性，在实验一的基础上，我们将FedPSG-CAG与直接用生成模型为B方生成缺少样本的方法进行实验对比。对比的基线模型为目前数据生成领域中表现为SOTA的模型： CTGAN[ 12]、TableGAN[36 ]、CTAB-GAN[ 14]、VAE[12 ] 和 TabDDPM[15 ]。实验也设置了B方相对于A方的不同缺少比。本实验样本生成效果的评估指标均也为RMSE，用于表示生成样本中数据与真实数据之间的平均偏差程度。为了进行更充分的实验验证，本实验也设置了两组实验。

（1）实验二的实验设计

第一组实验是在4.1小节提及的‘Bank’ and ‘Credit’两个数据集上进行的。这两个数据集代表了样本规模相对较大和较小的两种情况。在本组实验中，对于FedPSG-CAG的实验设置有以下说明：根据实验一中实验设定一的实验结果表 1和表2得知，在B 方不同的缺少比下，当Cnum=5时，RMSE值表现较优。又根据实验一中实验设定二的实验结果表3和实验设定三的实验结果表5 得知，在B 方不同的缺少比下，基于GANs的联邦填补模型，RMSE值表现较优。因此，在实验的FedPSG-CAG方法中，Cnum设置为5，联邦填补模型包括实验一中相对较优的VF-GAIN和VF-VGAIN。

第二组实验是在4.1小节提及的‘Letter’ and ‘News’两个数据集上进行的。这两个数据集代表了除金融之外的不同应用场景和样本特征数。根据实验二第一组实验结果，在本组实验的FedPSG-CAG中，Cnum设置为5，联邦填补模型包括实验一中相对较优的VF-GAIN和VF-VGAIN。

在本实验中，FedPSG-CAG 设置的生成模型GM为仍为TabDDPM（本实验也同时验证了TabDDPM作为本地生成模型的最优效果）。当Cum= 5时，用FedPSG-CAG-5表示。本实验中，各生成方法的训练迭代总数都设为 10,000 次，epoch 数量为 10，学习率设置为0.001，优化器使用Adam。

（2）实验二的结果分析

实验二第一组实验的实验结果如表9所示。在B 方不同的缺少率下，直接用生成模型为B方生成样本时，TabDDPM生成数据的RMSE值是较优的。在表9中，FedPSG-CAG-0（VF-GAIN）表示没有在B方先用GM生成属性列，B方缺少样本的所有属性均采用VF-GAIN联邦填补模型生成。在Bank和Credit数据集中，当缺少比较高时，缺少样本所有属性全部用联邦填补生成的方式会优于B方本地模型生成的方式。但是在缺少比较低时，B方本地模型生成可以更好的学到本地数据的分布用于缺少样本的生成。而如表9结果所示，无论B方缺少比高低，FedPSG-CAG-5的RMSE都优于直接用生成模型为B方生成样本，以及所有属性全部用联邦填补生成。这再次表明采取“先生成一些高相关属性值，再用纵向联邦填补方法进行填补”的方式进行参与方缺少样本生成是有效的。

实验二第二组实验的实验结果如表7所示。在本组实验中，我们为了模拟A、B两方，对‘Letter’ and ‘News’数据集进行了数据样本及其特征的纵向划分。虽然它们不像‘Bank’ and ‘Credit’数据集，A、B两方的特征能代表真实的特征所有权，但我们仍能从实验结果中观察到本文方法在不同应用场景和不同数据集中的有效性。如表10所示，在B 方不同的缺少率下，直接用生成模型为B方生成样本时，仍是TabDDPM生成数据的RMSE值是较优的。特别地，这两个数据集在进行纵向联邦划分时是模拟原有属性列顺序均分给A、B两方，数据的相关性和属性列数，与第一组实验有差异。因此，Cnum=5不一定是FedPSG-CAG的最佳设定。但即便如此，基于GANs纵向联邦填补模型的FedPSG-CAG都优于所有直接用生成模型为B方生成样本的RMSE。

FedPSG-CAG 的优秀表现在于它将相关属性生成与纵向联邦填补模型相结合，有效捕捉B方属性之间的内在关系，确保强相关属性生成的数据分布与原始数据更一致。此外，在纵向联邦学习框架下，FedPSG-CAG利用来自多个参与者的数据，在保护数据隐私的同时提高了生成效果。相比之下，其他基线方法（如 CTGAN ， TabDDPM等）只注重通过学习参与方本地数据的整体分布来生成数据，一方面没有考虑相关属性在生成数据中更有效的数据分布学习，也没有充分利用其他参与方的数据信息，从而导致准确性和一致性降低，尤其是在高缺失率的情况下。

**表9 不同生成方法的B方缺少样本生成RMSE结果（Bank and Credit datasets）**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset(LackR-B)  Methods | Bank | | | Credit | | |
| 0.2 | 0.5 | 0.8 | 0.2 | 0.5 | 0.8 |
| CTGAN[ ] | 0.5099 | 0.5213 | 0.7554 | 0.5456 | 0.6385 | 0.66 |
| TableGAN[ ] | 0.5951 | 0.6865 | 0.7312 | 0.6008 | 0.6975 | 0.7838 |
| CTAB-GAN[ ] | 0.4773 | 0.5644 | 0.6454 | 0.5533 | 0.6674 | 0.6976 |
| TVAE[ ] | 0.4265 | 0.4862 | 0.697 | 0.4305 | 0.5534 | 0.674 |
| TabDDPM[ ] | 0.4227 | 0.4728 | 0.5236 | 0.4478 | 0.5363 | 0.575 |
| FedPSG-CAG-0（VF-GAIN） | 0.4615 | 0.4981 | 0.519 | 0.5048 | 0.5354 | 0.5518 |
| FedPSG-CAG-5（VF-GAIN） | 0.2518 | 0.3378 | 0.3468 | 0.3509 | 0.3815 | 0.4994 |
| FedPSG-CAG-5（VF-VGAIN） | 0.2372 | 0.2997 | 0.3537 | 0.3495 | 0.3610 | 0.4621 |

**（注：红色字体表示最优，蓝色字体表示次优，橙色字体表示再次优）**

**表10 不同生成方法的B方缺少样本生成RMSE结果（Letter and News datasets）**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset(LackR-B)  Methods | Letter | | | News | | |
| 0.2 | 0.5 | 0.8 | 0.2 | 0.5 | 0.8 |
| CTGAN[ ] | 0.5328 | 0.5664 | 0.6218 | 0.5209 | 0.5626 | 0.6034 |
| TableGAN[ ] | 0.5398 | 0.6148 | 0.7068 | 0.4722 | 0.5204 | 0.6234 |
| CTAB-GAN[ ] | 0.5226 | 0.561 | 0.6559 | 0.4815 | 0.522 | 0.6572 |
| TVAE[ ] | 0.5203 | 0.5614 | 0.6634 | 0.5018 | 0.5492 | 0.6348 |
| TabDDPM[ ] | 0.4921 | 0.5413 | 0.5833 | 0.455 | 0.5024 | 0.5671 |
| FedPSG-CAG-5（VF-GAIN） | 0.4049 | 0.4468 | 0.4897 | 0.4112 | 0.463 | 0.5133 |
| FedPSG-CAG-5（VF-VGAIN） | 0.3698 | 0.3823 | 0.4768 | 0.4341 | 0.4579 | 0.5021 |

**（注：红色字体表示最优，蓝色字体表示次优）**

## 4.4 实验三：参与方缺少样本的不同处理方式对联邦分类模型训练的效果对比

为了进一步验证 FedPSG-CAG 为样本缺少参与方生成样本的有效性，我们将生成后的B方样本与A方样本构建联合样本集，用于纵向联邦分类任务机器学习模型的训练。验证和对比采取不同的方式处理B方缺少样本后，所构建的联合样本集用于支撑不同的纵向联邦分类模型的训练效果。

（1）实验三的实验设计

该A、B两方联合样本集由采用不同生成方法得到的B方生成数据和A方、B方原有数据组成，分类模型训练时双方数据都受安全隐私保护。本实验中，用于处理B方缺少样本的方式包括：FedPSG-CAG，本实验采用的GM为TabDDPM，Cnum=5，联邦填补模型为VF-VGAIN；在B方本地用TabDDPM生成缺少样本，本实验记为”TabDDPM”。TabDDPM由实验二验证为目前数据生成文献中的SOTA方法。此外，在A、B两方联合样本集构建时，还有两种方式。第一种方式：不对B方缺少样本进行样本生成，而直接基于A方、B方原有数据进行样本对齐建立联合样本集。在本实验中，我们将这种处理B方缺少样本的方式记为“N-GM”。第二种方式：在第一种方式构建的联合数据集基础上，采用目前文献中SOTA的纵向联邦样本生成方法VertiGAN[ 3 ]，为A、B两方生成全新的联合样本，扩充第一种方式的联合样本集样本量。其中，生成的联合样本数量跟B方缺少样本数量相同。在本实验中，我们将这种处理方式记为“A∞B-GM”。

本实验的纵向联邦分类模型包括：纵向联邦逻辑回归（VF-LR）[37]、纵向联邦支持向量机（VF-SVM）[38 ]、纵向联邦梯度提升决策树（VF-GBDT）[39 ]、纵向联邦随机森林（VF-RF）[ 40]、和FinalNet[ 41 ]，它们代表了不同类型和规模的分类学习模型。评估纵向联邦机器学习分类模型训练效果的评估指标为：准确率（ACC）[ 42]、AUC[ 42] 和 F1 分数[42 ]。实验中，我们先将基于A方、B方原有数据进行样本对齐建立的联合样本集，随机划分其中30%的样本作为联邦分类模型的测试集。剩余70%的样本，再加上针对B方缺少样本采用的各种不同方式获得的联合数据集，一起作为分类模型的训练集。本实验中所采用的数据集为‘Bank’和‘Credit’，数据集中B方的缺少比仍设置为0.2、0.5和0.8。本实验中，逻辑回归（LR）：学习率设为0.01，迭代次数为1000次，批次大小为64；支持向量机（SVM）：正则化参数取值为1；梯度提升决策树（GBDT）：树的数量设为20，学习率为0.1，树的最大深度为6，样本采样率为20%；随机森林（RF）：树的数量设为100，树的最大深度为3，最小样本分割数为2，最小样本叶子数为1；FinalNet：学习率为0.001，迭代次数为1000次，批次大小为64，隐藏层维度为128，正则化参数λ为0.0001。

（2）实验三的结果分析

我们对FedPSG-CAG、TabDDPM、A∞B-GM、N-GM这4种处理方式得到的联合样本量进行统计和分析，有以下结论：

① FedPSG-CAG、TabDDPM这两种方式是为B方相对于A方缺少的样本进行生成，使其可以与A方未对齐样本构建完整的联合样本。而A∞B-GM生成的联合样本跟B方缺少样本量相同。因此，由FedPSG-CAG，TabDDPM，A∞B-GM构建的联合样本集的样本量都是一样的，它们由A方样本数量确定。

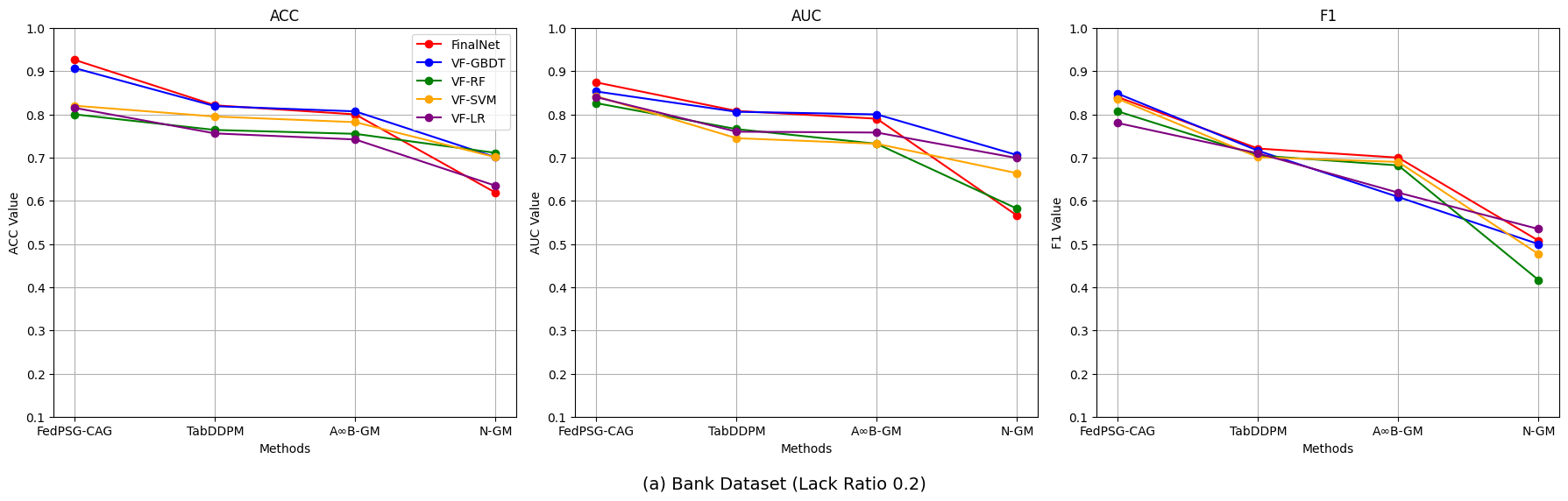
② N-GM这种处理方式构建的联合数据集中的样本量最少。因为，在N-GM方式下，当A方样本与法B方缺少的样本对齐时，在构建联合样本集时，A方未对齐样本将被丢弃。因此，N-GM方式下，联合样本集的样本数量由缺少样本生成前B方原有样本数量确定。

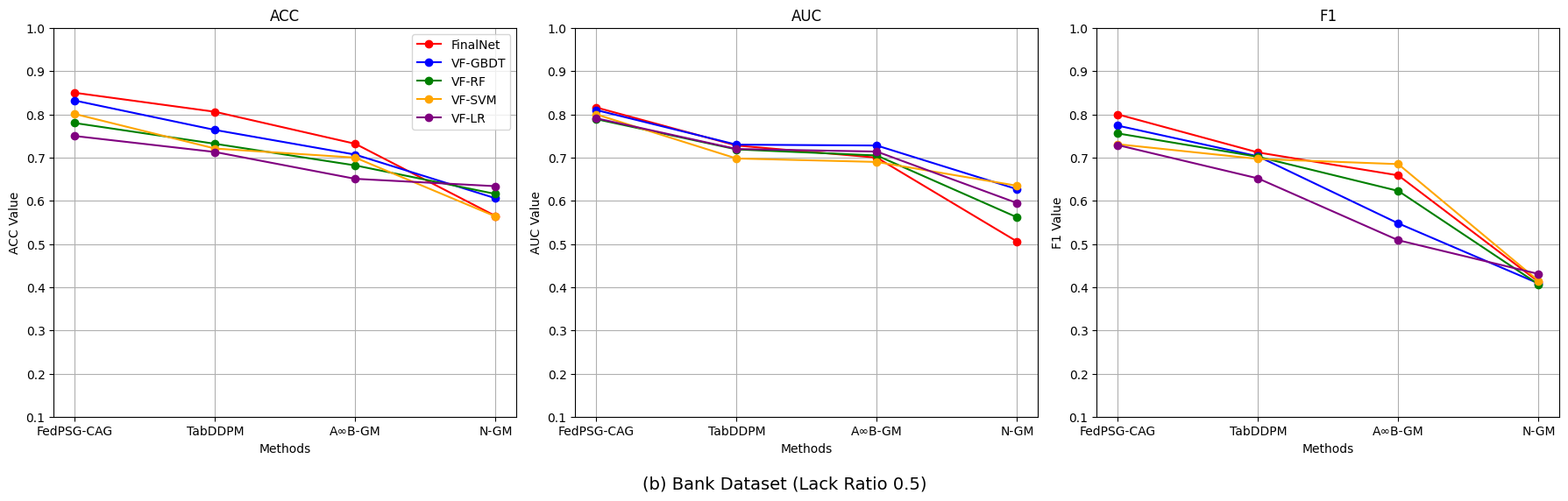
实验结果如图4所示，图4**（a）-（f）分别表示了Bank和Credit数据集在B方样本缺少比为0.2、0.5、0.8下，构建联合样本集时不同缺少样本处理方式下，不同纵向联邦分类模型的**ACC、AUC、F1 分数**结果曲线。**图4**（a）-（f）每个子图中不同颜色的曲线表示不同的纵向联邦分类模型，横坐标表示针对B方缺少样本不同的处理方式，即：**FedPSG-CAG、TabDDPM、A∞B-GM、N-GM，纵坐标表示**纵向联邦分类模型在基于不同处理方式得到的联合数据集上训练后，模型在测试集上得到的**ACC、AUC、F1 分数值。

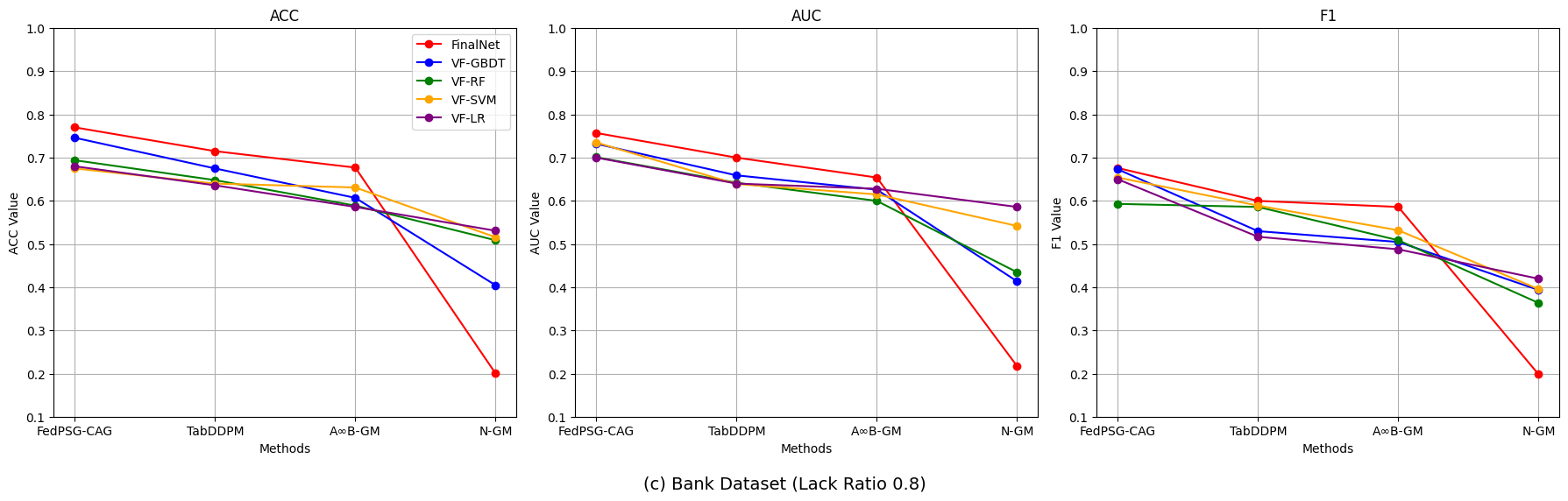
如图4所示，在两个数据集的不管在何种缺少比下，采用FedPSG-CAG方法生成B方缺少样本所构建的联合样本集，不管用于VF-LR、VF-SVM、VF-GBDT、VF-RF、FinalNet哪种联邦机器学习模型的训练，其测试集的ACC、AUC、F1评估指标都是表现最为出色的。然后依次是TabDDPM、A∞B-GM、N-GM。分析出现这种结果的原因：

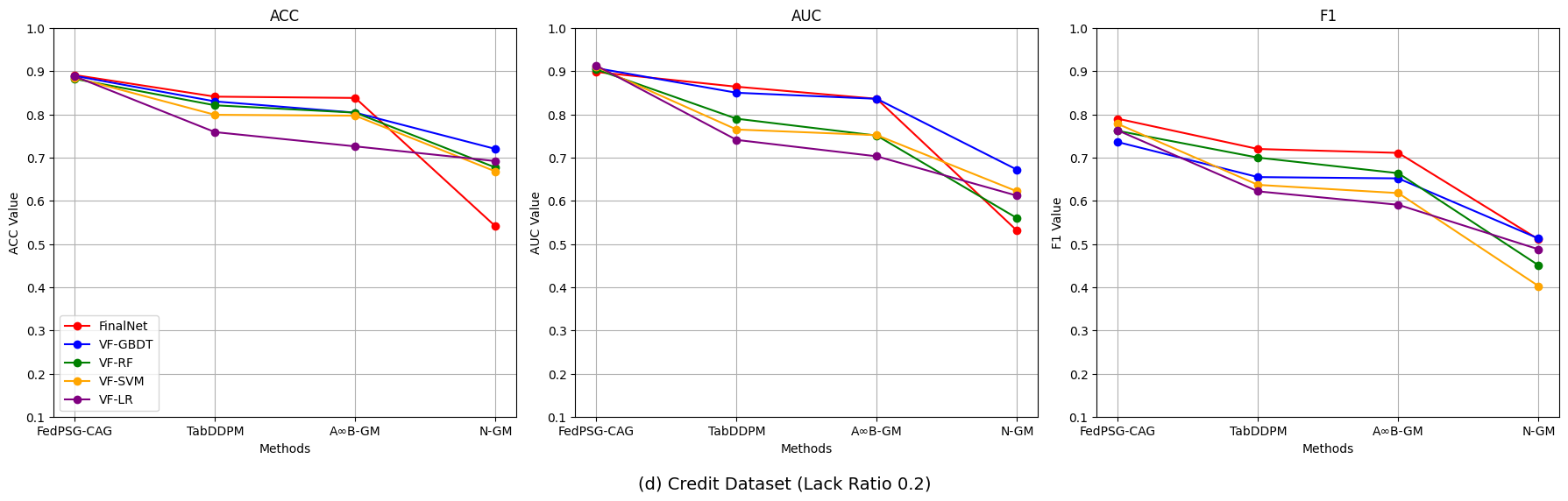
① 从样本数量的角度来分析，FedPSG-CAG、TabDDPM、A∞B-GM三种方式构建的联合数据集样本量相同，N-GM最少。如图4所示，在两个数据集的不管在何种缺少比下，样本量最少的N-GM方式下的联合样本集训练的纵向联邦分类模型的评估指标结果最差。这表明了联合样本量对纵向联邦机器学习模型训练的重要性。其他因素相同的情况下，样本量较大的联合样本集，训练的纵向联邦机器学习模型分类性能更好。如图4所示，我们用不同的联邦机器学习模型验证了这一结论。在同一个数据集中，随着B方缺少比的增加，N-GM方式获得的联合样本集样本量将大幅下降。当联合样本集样本量下降后，五种纵向联邦机器学习模型的评估指标值都下降了，尤其，FinalNet相对于VF-LR、VF-SVM、VF-GBDT、VF-RF的评估指标值有了更大幅度的下降。FinalNet是一种纵向联邦学习的深度学习框架。这同时也表明对于深度神经网络等复杂模型，用更多的样本量进行训练是有必要的。而本文方法提供了一种更有效解决参与方样本缺少时的联合样本量问题。

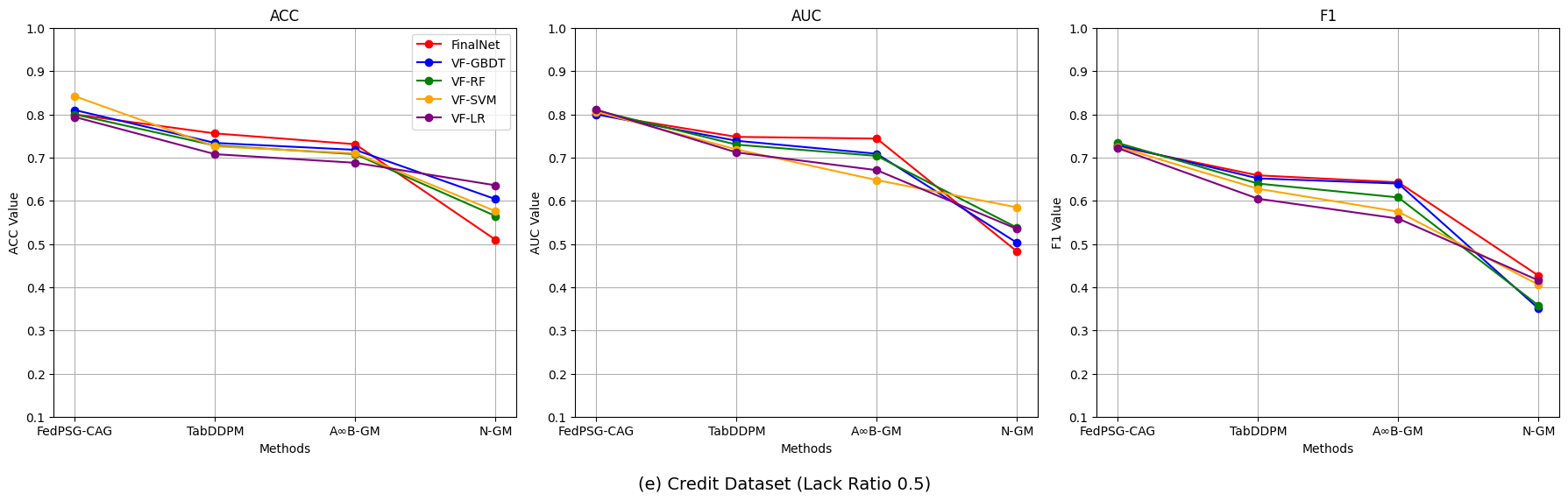
② 从样本质量的角度分析，虽然FedPSG-CAG、TabDDPM、A∞B-GM三种方式的联合样本集样本量相同，但他们的联合样本的质量却依次降低。FedPSG-CAG、TabDDPM这两种方式，A方的数据全部被保留了。比起A∞B-GM方式生成全新的联合样本，前两种方式构建的联合样本集中有更多的真实数据用于模型训练，联合样本质量更高。前两种方式中，FedPSG-CAG能够有效地学习缺少样本参与方本地高相关属性之间的数据分布，以及利用多方数据之间的关联性进行缺少样本的生成。而TabDDPM方式，缺少样本参与方的数据是由TabDDPM在B方本地生成的。这说明本文方法是解决参与方样本缺少问题的有效方法。如图4所示的实验结果验证了这一结论，也同时表示了模型的训练效果不仅仅依赖于样本量，还跟样本质量等其他因素有关。即使在B方高缺少比的情况下，本文方法也能为联合样本集提供更高质量的训练样本，更好地支持纵向联邦机器学习模型训练。如图4所示，这一结论适用于不同的机器学习模型。

****









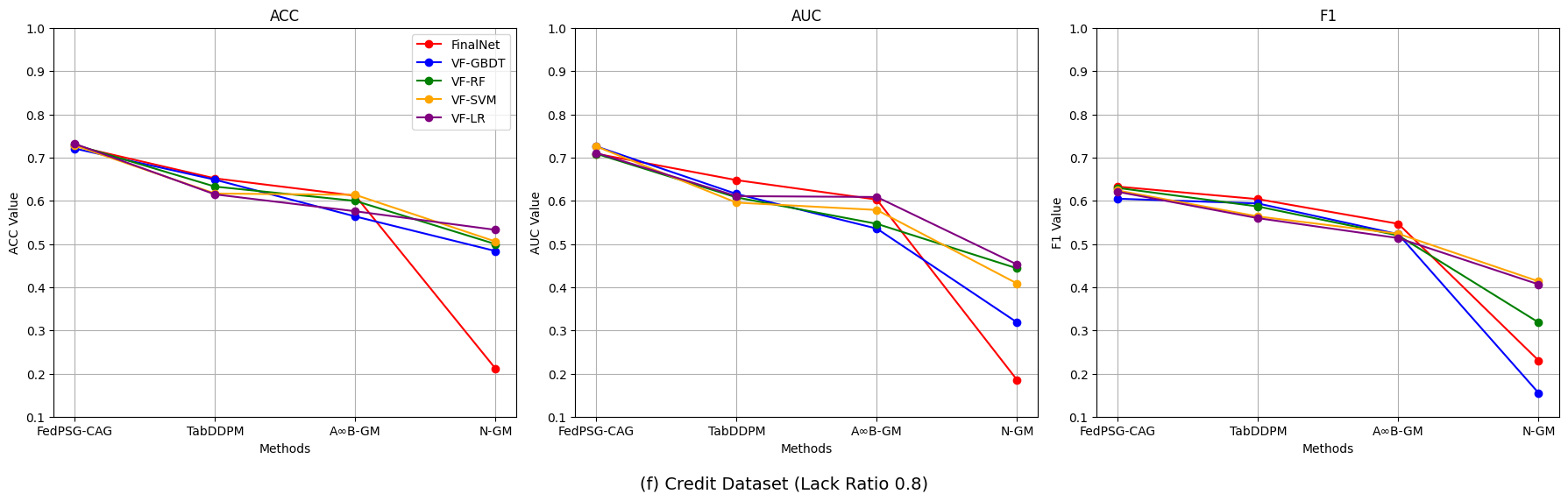


图4 基于不同B方缺少样本处理方式而获得的联合样本集，不同纵向联邦分类模型的验证结果:（a）缺少比0.2下的Bank数据集；（b）缺少比0.5下的Bank数据集；（c）缺少比0.8下的Bank数据集；（d）缺少比0.2下的Credit数据集；（e）缺少比0.2下的Credit数据集；（f）缺少比0.2下的Credit数据集

# 5 Conclusion

在纵向联邦学习场景，当某个或某些参与方相对其他方缺少样本时，加密样本对齐后，联合样本较少，不利于纵向联邦机器学习模型的训练。In this paper, we propose a novel Participants Sample Generation Method Based on Related attributes generation and Vertical Federated imputation (FedPSG-CAG) to address the challenge of data generation for participant with lack of samples in multi-party collaborative scenarios. 本文采取一种“先生成一些高相关属性值，再用纵向联邦填补方法进行填补”的方式进行参与方缺少样本生成。FedPSG-CAG首先计算和获取样本缺少参与方的对齐样本集合中具有高相关性的属性列，用于在样本缺少参与方进行本地生成模型的训练，并将训练好的生成模型用于生成缺少样本中的相关属性值。然后，本文构建了纵向联邦学习的填补模型框架，用于生成参与方缺少样本中的剩余属性值。通过一系列的实验表明：在多方纵向联合学习场景下，当某些参与方样本缺少时，是非常有必要为其进行样本生成的。这可以使多方在联邦机器学习模型训练过程获得更多联合样本，为联合训练样本集保留更多真实数据。FedPSG-CAG 更好地解决相对样本缺少的参与方的样本生成问题。它将相关属性生成与纵向联邦填补模型相结合，有效捕捉样本缺少参与方属性之间的内在关系，并在纵向联邦学习框架下利用来自多个参与方的数据关联性，在保护数据隐私的同时提高了生成效果。

# 参考文献：

[1] Liu Y, Kang Y, Zou T, et al. Vertical federated learning: Concepts, advances, and challenges[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3615-3634.

[2] J. Zhang, Y. Jiang, "A Data Augmentation Method for Vertical Federated Learning," Wireless Communications and Mobile Computing, 2022. https://doi.org/10.1155/2022/6596925.

[3]Jiang X, Zhang Y, Zhou X, et al. Distributed GAN-Based Privacy-Preserving Publication of Vertically-Partitioned Data[J]. Proceedings on Privacy Enhancing Technologies, 2023.

[4] Z. Zhao, H. Wu, A. Van Moorsel, et al. "GTV: generating tabular data via vertical federated learning," arXiv preprint arXiv:2302.01706, 2023.

[5] Yuan X, Yang Y, Gope P, et al. VFLGAN: Vertical Federated Learning-based Generative Adversarial Network for Vertically Partitioned Data Publication[J]. arXiv preprint arXiv:2404.09722, 2024.

[6] F. Zhao, Z. Li, X. Ren, et al. "VertiMRF: Differentially Private Vertical Federated Data Synthesis," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024:4431-4442.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[8] Michelucci U. An introduction to autoencoders[J]. arXiv preprint arXiv:2201.03898, 2022.

[9] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[10] Bank D, Koenigstein N, Giryes R. Autoencoders[J]. Machine learning for data science handbook: data mining and knowledge discovery handbook, 2023: 353-374.

[11] Kingma D P. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[12] L. Xu, M. Skoularidou, A. Cuesta-Infante, et al. "Modeling tabular data using conditional gan," in Advances in Neural Information Processing Systems, 2019. arXiv:1907.00503v2 [cs.LG].

[13] M. Mirza, S. Osindero, "Conditional Generative Adversarial Nets," Computer Science, 2014:2672-2680.

[14] Z. Zhao, A. Kunar, H. Van der Scheer, et al. "CTAB-GAN: Effective Table Data Synthesizing," arXiv:2102.08369v1 [cs.LG], 2021. https://arxiv.org/pdf/2102.08369v1.pdf.

[15] A. Kotelnikov, D. Baranchuk, I. Rubachev, et al. "Tabddpm: Modelling tabular data with diffusion models," in International Conference on Machine Learning. PMLR, 2023:17564-17579.

[16] A. Farhangfar, L. A. Kurgan, W. Pedrycz, "A novel framework for imputation of missing values in databases," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 37(5):692-709.

[17] F. You, D. Li, H. Zhang, et al. "A Random Forest Approach for Missing Data Imputation based on Normalized KNNI," Journal of Chengdu University of Information Technology, 2021, 36(1):32-40.

[18] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016:785-794.

[19] J. Meng, C. Li, "Missing Data Imputation for Categorical Data Based on Random Forest Model," Statistics and Information Forum, 2014, 29(9):86-90.

[20] P. Royston, I. R. White, "Multiple imputation by chained equations (MICE): implementation in Stata," Journal of statistical software, 2011, 45:1-20.

[21] S. Ryu, M. Kim, H. Kim, "Denoising autoencoder-based missing value imputation for smart meters," IEEE Access, 2020, 8:40656-40666.

[22] J. Yoon, J. Jordon, M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in International conference on machine learning, PMLR, 2018:5689-5698. https://arxiv.org/pdf/1806.02920.pdf.

[23] S. E. Awan, M. Bennamoun, F. A. Sohel, et al. "Imputation of missing data with class imbalance using conditional generative adversarial networks," Neurocomputing, 2021, 453:164-171.

[24] Y. Wang, D. Li, X. Li, et al. "PC-GAIN: Pseudo-label conditional generative adversarial imputation networks for incomplete data," Neural Networks, 2021, 141:395-403.

[25] X. Miao, Y. Wu, L. Chen, et al. "An experimental survey of missing data imputation algorithms," IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7):6630-6650.

[26] S. Zheng, N. Charoenphakdee, "Diffusion models for missing value imputation in tabular data," arXiv preprint arXiv:2210.17128, 2022.

[27] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in Artificial Intelligence and Statistics. PMLR, 2017, pp. 1273–1282.

[28] Q. Yang, Y. Liu, T. Chen, Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–19, 2019.

[29] W. Du, Y. Wang, G. Meng, et al. "Privacy-Preserving Vertical Federated KNN Feature Imputation Method," Electronics, 2024, 13(2):381.

[30]De Cristofaro E, Tsudik G. Practical private set intersection protocols with linear complexity[C]//International Conference on Financial Cryptography and Data Security. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010: 143-159.

[31] Zar J H. Spearman rank correlation[J]. Encyclopedia of biostatistics, 2005, 7.

[32] Cheon J H, Kim A, Kim M, et al. Homomorphic encryption for arithmetic of approximate numbers[C]//Advances in cryptology–ASIACRYPT 2017: 23rd international conference on the theory and applications of cryptology and information security, Hong kong, China, December 3-7, 2017, proceedings, part i 23. Springer International Publishing, 2017: 409-437.

[33]Zhou X, Liu X, Lan G, et al. Federated conditional generative adversarial nets imputation method for air quality missing data[J]. Knowledge-Based Systems, 2021, 228: 107261.

[34]Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.

uci.edu/ml

[35]Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)[J]. Geoscientific model development discussions, 2014, 7(1): 1525-1534.

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

[37] Yang S, Ren B, Zhou X, et al. Parallel distributed logistic regression for vertical federated learning without third-party coordinator[J]. arXiv preprint arXiv:1911.09824, 2019.

[38]Hartmann V, Modi K, Pujol J M, et al. Privacy-preserving classification with secret vector machines[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 475-484.

[39]Tian Z, Zhang R, Hou X, et al. ${\sf FederBoost} $: Private Federated Learning for GBDT[J]. IEEE Transactions on Dependable and Secure Computing, 2023.

[40] Wu Y, Cai S, Xiao X, et al. Privacy preserving vertical federated learning for tree-based models[J]. arXiv preprint arXiv:2008.06170, 2020.

[41] Zhu J, Jia Q, Cai G, et al. Final: Factorized interaction layer for ctr prediction[C]//Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval. 2023: 2006-2010.

[42] Murphy K P. Machine learning: a probabilistic perspective[M]. MIT press, 2012.