## 纵向联邦学习的参与方样本生成方法

摘要：

关键字：

# 1 引言

当今，人工智能的发展要求机器学习需要面对来自多参与方的相同对象不同属性的数据。例如，在开展信用风险评估时，银行、电商企业都有大量的历史数据可供信用风险评估，而银行和电商企业所拥有的同一对象的数据又有不同的属性维度，如果结合二者的不同属性来开展共同对象的信用风险评估，无疑会获得更为有价值的结果。然而，因参与各方受安全隐私保护的约束，又难于为对方提供己方数据以训练机器学习模型。所以，纵向联邦学习[ ]应运而生。在纵向联邦学习过程中，需要将不同参与方拥有的ID相同的样本对象进行加密样本对齐。在加密样本对齐后，我们将具有相同ID的多方样本称为纵向联邦下的联合样本。联合样本集利用来自多方的更多属性帮助纵向联邦机器学习模型训练。在很多实际应用中，当多个参与方进行纵向联邦学习时，各参与者所拥有的可用于样本对齐的样本对象并不完全一致。如图1（a）所示，B方相对于A方，样本对象数量较少。又如图1（b）所示，A方相对于B方，或B方相对于A方，都各自有一部份样本对象缺少。此时，在对多个参与者的样本对象进行加密样本对齐后，联合样本量会远远少于各参与方独立数据集中的完整样本量。众所周知，样本量是影响机器学习模型训练效果的重要因素之一。因此，在纵向联邦学习中，为解决多方样本对齐后联合样本量不足的问题，研究获得更多的多方联合训练样本的科学方法，是一个有价值的研究方向。

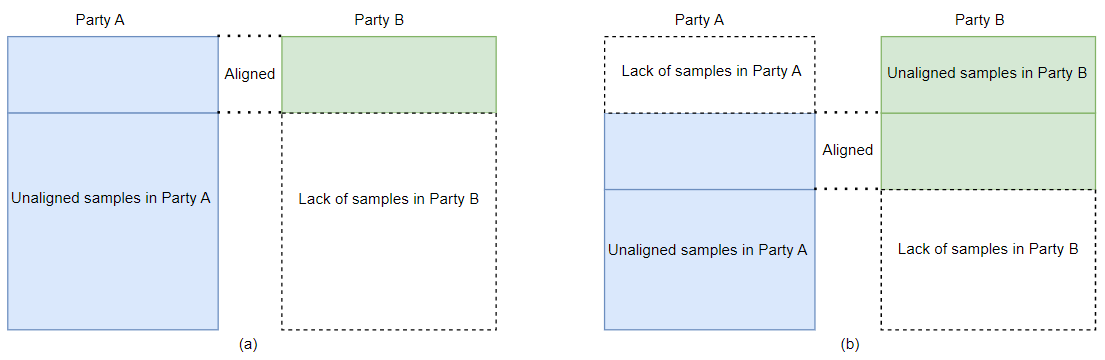


图1 参与方对齐样本不足的图例

解决多方样本对齐后联合样本不足的问题，有两种办法：（1）生成全新的联合样本：舍弃参与方未对齐的样本，基于对齐的联合样本，进行多方联合样本生成，扩充联合样本集的样本；（2）为缺少样本的参与方生成样本，以构建完整联合样本：保留参与方未对齐的样本，为缺少样本的参与方生成样本，使其参与多方样本对齐，为未对齐的样本构建完整联合样本，以此获得更多联合样本。

第一种方法：生成全新的联合样本。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：基于生成对抗网络(GAN)的纵向联邦学习数据生成方法：FedDA[73]、VertiGAN框架 [78 ] 、VFLGAN [79] 、GTV[74]，和基于马尔可夫随机场（MRFs）的合成表格数据生成框架：VertiMRF[75]。然而，对于机器学习模型的训练来说，不仅与样本数量有关，更与样本数据的质量等多种因素都有关。在第一种方法中，由联邦生成方法得到的整个联合样本的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时，用于训练联邦生成模型的对齐联合样本也很少，无法训练出表现优秀的纵向联邦生成模型。因此，无法生成高质量的联合样本。此外，其他参与方未对齐的样本因为无法进行样本对齐，而不能加入到联合样本集中，无法用于后续的纵向联邦机器学习任务。这将极大地减少了纵向联邦机器学习模型训练中的真实数据信息。因此，第一种‘生成全新的联合样本’的方法不是最佳的确决方法。

第二种方法：生成缺少的参与方样本。目前，有多种生成方法可用于为样本缺少的参与方生成样本，其中，表现较为优秀的是最新的一些基于深度学习的生成方法，例如，生成对抗网络（GAN）[]、自动编码器（AE）[]和去噪扩散概率模型（DDPM）[]。这些方法能够通过学习数据的分布，然后使用该分布生成高质量样本数据。 GAN 通过生成器和判别器之间的对抗过程生成样本数据，例如使用条件 GAN 进行表格数据建模（CTGAN）[ ]和基于生成对抗网络的数据合成（TableGAN）[ ]。自动编码器（AE）通过学习数据的概率表示来创建样本，例如，TVAE[ ] 和变分自动编码器（VAE）[ ] 等模型。 DDPM 使用降噪扩散过程来迭代细化噪声数据并生成高质量样本，比如 TabDDPM[ ] 等模型。在为样本缺少的参与者生成样本时，这些方法仅从自己一方的数据中学习，不考虑其他参与者的影响。事实上，当通过组合来自多方的数据创建联合样本时，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。尽管上述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。为了给样本缺少的参与方生成更高质量的样本，我们需考虑结合其他方数据对该方协作地进行数据生成，以多方数据的相关性提升样本缺少参与方的生成效果。此时，‘为缺少样本的参与方生成样本’的问题，将在纵向联邦下，转化为‘在联合样本中针对缺少数据进行填补’的问题。同样地，参与协作生成的各方都需要数据安全隐私保护，无法直接汇总各方数据对联合样本中的缺少数据进行数据填补模型训练和计算。我们需要一种非常有效的方法来为这种情景下的这些数据进行生成和填补，以获得完整的联合样本。

针对上述问题，本文提出了一种基于纵向联合学习的参与方样本生成方法（VF-PSG），旨在为多方协作学习提供更多高质量的联合训练样本。通过纵向联合学习将关联规则和数据填补模型结合起来，该方法可以用于为样本缺少的参与方生成合成样本，同时确保多方的数据隐私。我们的主要贡献如下：

（1）我们提出了一种新的基于纵向联合学习（Vertical Federated Learning，VF-PSG）的参与方样本生成方法，可为多方协作场景下的机器学习提供更多高质量的联合训练样本。

（2）为了给样本缺少的参与方生成数据，本文提出了一种基于多方的关联规则属性生成方法。该方法通过学习每个参与者属性之间的相关性，在安全隐私保护下，为样本缺少的参与者生成重要属性值。

(3) 基于多方关联规则生成方法生成的属性值，我们提出了一种多方数据填补方法，为样本缺少的参与者生成剩余属性值。这种方法强调数据填补过程中其他参与方数据的影响，最大限度地挖掘了多方协作学习的潜力。

本文的组织结构如下： 第 2 节讨论了数据生成和数据填补方法的相关工作。第 3 节介绍了我们解决参与者样本生成问题的方法。第 4 节介绍实验结果，第 5 节给出结论。

# 2 相关工作

2.1 数据生成和数据填补方法

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，旨在通过已有的样本来生成新的样本数据。

基于自编码的生成模型，如：自编码器AE[35]、变分自编码器VAE[37]等。L. Xu [38]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[39]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[40]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。M. Arjovsky 等人[42, 43]引入Wasserstein距离来替代JS散度和KL散度，并将其作为优化目标，从而提出了WGAN（Wasserstein GAN），从根本上解决了原始GAN的梯度消失问题。L. Xu 等人[44]提出了一种生成对抗网络Tabular GAN (TGAN)，使用LSTM和MLP分别作为生成器和判别器，生成如病例或教育记录等表格数据。2019年，L. Xu 等人[38]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。Y. Yu 等人[45]提出了一种基于改进的条件生成对抗网络CWGAN(Conditional Wasserstein Generative Adversarial Nets)来学习滚动轴承故障的时频谱特征，并根据输入类别生成相应故障类别的时频谱。J. Lee 等人[47]提出了一个广义的GAN框架的表格合成，它结合了GAN的对抗训练和可逆神经网络的负对数密度正则化，以提高生成数据的综合质量。2021年，M. Esmaeilpour 等人[48]提出了一种用于合成包含连续列、二进制列和离散列的表数据集的双判别器GAN。S. Singh 等人[49]提出MeTGAN，使用稀疏线性层来克服CTGAN的内存瓶颈，大大减少了训练的内存使用。Z. Zhao 等人[50]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。J. Engelmann 等人[52]提出了一种基于条件Wasserstein GAN的方法，对具有数值和分类变量的表格数据集进行建模，并通过一个辅助分类器来特别关注下游分类任务。

DDPM[55]是 J. Ho 等人在2020年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。该方法通过多步马尔科夫链精确建模生成过程，生成质量高但效率相对较低。在此基础上，A. Kotelnikov 等人[56]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。

伴随着人工智能和大数据技术的快速发展和应用，对基于生成对抗网络和扩散模型的数据生成方法为近几年的重要方法。然而，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。因此，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简单替换。均值填补（Mean Imputation）[57]用特征均值替代缺失值，简洁快速，但易损失数据变异性。K近邻填补（K-Nearest Neighbors Imputation, KNNI）[33]利用KNN方法寻找与缺失样本最相似的K个邻居，以其均值或众数填补缺失值，仅有助于保持局部数据结构。

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）[59]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）[34]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）[61]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受到深度生成模型强大能力的启发，通过隐式或显式分布建模实现高质量填补。自编码器（Autoencoders, AEs）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。S. Ryu 等人[36]采用自动编码器方法对智能电表的缺失数据进行填补。生成对抗网络（GANs）利用生成器和判别器的对抗训练生成符合真实分布的合成样本和填补值，常用于复杂高维场景下的缺失数据重构。J. Yoon 等人[68]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[69]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[70]提出了伪标签条件生成对抗网络（Pseudo-label Conditional Generative Adversarial Imputation Networks, PC-GAIN）。该方法在传统GAIN模型的基础上，通过引入伪标签作为条件信息，增强了生成模型对数据分布的捕捉能力，提高了对不同模式缺失数据的适应性。X. Miao 等人[71]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，通过在生成器中引入VAE的潜变量正则和重构损失来加强表示学习的稳健性，从而有效避免模式崩塌并提高数据缺失值填补的质量。基于扩散模型的填补方法（以TabCSDI为例）[72]，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。

2.2 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[1]提出。联邦学习的方法包括在本地数据上训练具有一致模型架构的本地模型，各参与方独立进行模型训练，利用各自的数据资源优化模型参数。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[19][20]（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。在这种设置下，传统的集中式机器学习方法由于隐私和数据传输的限制难以应用，而VFL通过分布式计算和隐私保护技术，为跨组织的特征级模型协同训练提供了一种有效的解决方案。

在纵向联邦学习的应用场景中，当训练数据集中样本量不足时，通过样本生成的方式来进行数据增强是很有必要的。目前，有一些基于纵向联邦学习（VFL）的样本生成方法，主要包括基于GAN的方法和基于马尔可夫随机场（MRFs）的方法。基于生成对抗网络(GAN)的方法：J. Zhang 等人[73]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 [78 ] 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。X. Yuan 等人 [79] 提出了另一个基于纵向联邦学习（VFL）的生成对抗网络——VFLGAN。GTV[74]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。基于马尔可夫随机场（MRFs）的方法：W. Lin 等人[75]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，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[77]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。作者提出了一种隐私保护下的纵向联邦KNN特征填补方法，通过在多方数据不出本地、仅共享安全加密的特征相似度信息，实现缺失数据的KNN填补。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，例如，本文引言中所提出的问题。

# 3 纵向联邦的参与方样本生成方法

本文提出了一种名为参与方对齐样本生成的方法（VF-PSG），该方法将纵向联邦学习（VFL）框架与属性相关性、关联规则以及数据生成与填补技术相结合。在确保数据隐私的前提下，该方法的核心思想是利用各方特征属性之间的相关性，为样本缺少的参与方生成样本中的重要属性值，再采用联邦填补策略生成的剩余属性值，从而结合多方数据达到缺少样本生成的目的。该方法的优势在于，在构建多方机器学习的联合样本集时，能保留未对齐样本参与方的样本，从而保留更多真实数据。此方法充分考虑了各方数据相关性，同时保护了各方的数据隐私，增强了样本缺少参与方的样本生成的真实性，最终生成高质量的联合样本集，以提升多方机器学习模型的训练效果。

如图2所示，整个过程由服务器协调，参与方通过加密通信交换公钥和生成的信息。VF-PSG方法主要包括两个阶段。第一阶段是基于多方的关联规则生成方法，如图1(a)所示。在数据隐私保护下，基于对齐样本中各方特征属性的相关性，为样本缺少的参与方生成样本重要属性值。具体来说，这一阶段基于多方的关联规则生成方法包括以下三个过程：多方属性相关性计算、属性对应关系建立和关联规则属性生成。首先，利用各方的可对齐样本集，在多个参与方之间计算属性之间的相关性，并构建多方属性相关性矩阵。然后，从该矩阵中识别出具有强相关性的属性对，为该属性对中的属性列中的所有值建立对应关系。最后，基于这些对应关系，针对两列属性的值建立关联规则，为样本缺少的参与方生成样本重要属性值。第二阶段是基于VFL的联邦数据填补方法，如图1(b)所示。利用基于多方的关联规则生成方法生成的属性值，并结合其他参与方的样本数据，多方联邦数据填补方法将为样本缺少的参与方生成这些样本的剩余属性值。

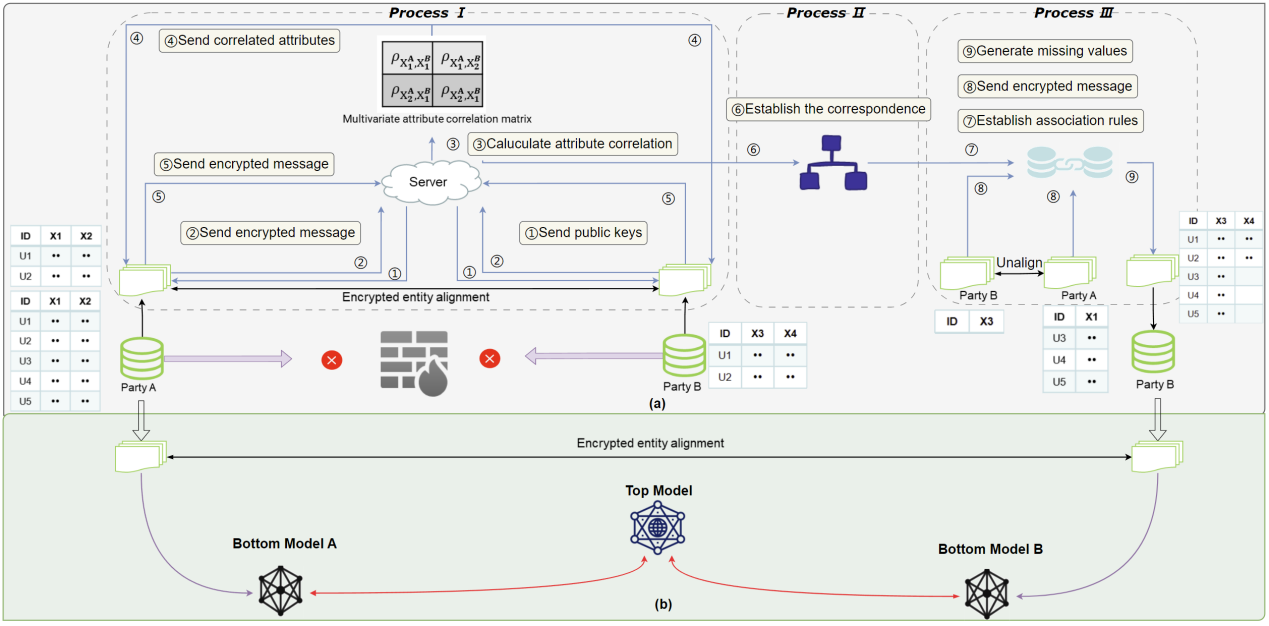


图2：参与方对齐样本生成方法架构图：（a）基于多方的关联规则生成方法；（b）基于纵向联邦的填补方法

3.1 数据预处理和数据描述

假设在一个给定的场景中，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的方法，我们以两个参与方为例：A方和B方，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，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，双方安全地执行样本对齐过程。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[ ]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。在本文方法的整个纵向联邦学习过程中，我们使用全同态加密[]（Cheon-Kim-Kim-Song, CKKS）进行数据加密运算。

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，，， 表示A方的样本数量，a表示A方样本的属性个数。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， ， ， 表示B方的样本数量，b表示B方样本的属性个数。A方和B方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。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。

3.2 多方关联规则生成方法

与其他参与方合作为样本缺少的参与方生成样本时，首先要生成样本的重要属性值，这些属性值是由A、B两方对齐样本中相关性确定的。因此，我们需要在数据安全保护下计算它们两方属性之间的相关性，将高相关的属性设定为重要属性。根据属性相关性建立属性对应关系，通过关联规则为这些重要属性赋值。

（1）多方属性相关性计算

计算A方第m个属性和B方第n个属性之间相关性系数的具体过程如下：

协调方C方生成密钥对，并将公钥密钥发送给A方和B方，用于整个多方关联规则生成方法。对于A方第m个属性列，，，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，将加密后发送给B放。对于B方第n个属性列，，，循环遍历中的每个属性值，并对其赋排名（排名方式与相同）。得到每个属性值的排名，加密。在B方计算加密后的与之间的差值：

根据差值进行皮尔曼相关性系数[ ]计算，得到相关性系数，计算公式如下：

其中，为对齐的样本数，即。

根据公式（2）求出A方和B方每个属性之间的相关性系数，得到多方属性相关性矩阵M：

特别说明，如果多方协作学习过程中，不只A、B两方，可将未对齐参与方的所有属性列与样本缺少参与方的所有属性列之间进行属性相关性矩阵计算。

（2）属性对应关系建立

基于多方属性相关性矩阵，该过程首先识别对齐样本数据集中各方之间具有强相关性的属性列，建立A、B两方具有强相关性的属性对。为每个属性对建立它们所有取值之间的对应关系，具体过程如下：

① 循环遍历相关性矩阵M中的每一个值，找出当前矩阵中最大的相关性系数，并据此确定具有强相关性的属性对)。

② B方初始化一个空的对应关系表，用于存储属性中的每个取值与属性的取值之间的对应关系。

③ 根据属性对)，循环遍历和中所有取值。假设属性列总共包含s个取值，值的集合为，其中是属性列m中的第k个属性值；属性列共包含个取值，值的集合为，其中是属性列n中的第q个属性值。A方和B方分别查找和中第i个样本的属性和的取值和，并加密得到和。A方将加密后的发送给B方，B方记录当取值为时，对应的属性取值，和的对应关系为表示为（）。遍历查找和中所有样本属性取值的对应关系，并将这些对应关系存储到关系表中。

④ 将的值重置为 0。

重复上述过程，根据相关性矩阵M，为A、B两方间所有属性列建立属性对应关系表矩阵R：

R

（3）相关属性值生成

根据属性相关性建立的关系表矩阵R，使用关联规则为样本缺少的参与方生成重要属性值。关联规则[]用于发现数据项之间的隐含关系，广泛应用于数据挖掘和市场篮分析等领域。它们通过分析数据项的共现模式来识别项之间的强关联。例如，在零售分析中，购买商品A的客户很可能也会购买商品B。关联规则通常以“如果-那么”（if-then）格式表示。在为样本缺少的参与方生成样本时，关联规则帮助建立两个属性列之间的值映射，这些映射用于推断并生成数据。

当A方样本缺少时：

步骤1：B方遍历属性，计算每个属性取值的样本个数；同时，B方向A方发送关系表。A方遍历属性，根据的关系表，记录与中每个属性取值的样本数，并发送给B方。

步骤2：B方根据接收到的数据，计算相对于的比值：

B方建立属性与的关联规则，表示如下：

其中，表示，当某个样本在中属性取值为时，该样本在中属性取值为的概率为，。

步骤3：A方属性列中的样本值由该样本B方属性值的关联规则决定。在B方，当属性值为时，B方查找，从选择最大的所对应的发送给A方。A方初始化随机掩码，计算，并将其发送给C方，C方解密得到+，重新发送回A方，A方去除随机掩码得到并赋值给，其中j。

当B方样本缺少时：

步骤1：A方遍历属性，记录每个属性取值的样本个数；同时，B方遍历属性，根据的关系表，记录与对应的中每个属性取值的样本数，并发送给A方。

步骤2：A方根据接收到的数据，计算相对于的比值：

A方建立属性与的关联规则，表示如下：

其中，表示，当某个样本在中属性取值为时，该样本在中属性取值为的概率为，。

步骤3：B方属性列中的样本值由该样本A方属性值的关联规则决定。在A方，当属性值为时，A方查找，从选择最大的所对应的发送给B方。B方初始化随机掩码，计算，并将其发送给C方，C方解密得到+，重新发送回B方，B方去除随机掩码得到并赋值给，其中j。

基于各方间的高相关性属性，多方关联规则可以为样本缺少的参与方生成重要属性值。除了通过本文提出的多方关联规则生成重要属性值外，我们还可以使用生成模型为样本缺少方生成这些属性，如CTGAN、TableGAN、CTABGAN、VAE和TabDDPM等。但它们仅在样本缺少方的数据上进行生成模型的训练。尽管这些生成模型可以仅使用样本缺少方的数据生成这些重要属性值，但由于生成过程中的噪声和随机性，生成的数据质量通常较低。此外，生成数据的过程并未考虑来自其他参与方的信息的影响。相比之下，使用多方关联规则能为样本缺少方生成表现各方之间强相关性的属性值，通过以推演的方式结合其他参与方的信息，获得更准确的样本属性值。生成这些重要属性值后，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3.3节所述。基于多方关联规则和基于仅单方生成模型生成这些重要属性值后，再用联邦填补模型填补剩余属性值，实验4.1详细比较了它们针对参与方样本生成的最终效果。

## 3.3 纵向联邦数据填补方法

当我们通过联邦关联规则为样本缺少参与方生成样本重要属性值后，还需要对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3所示。

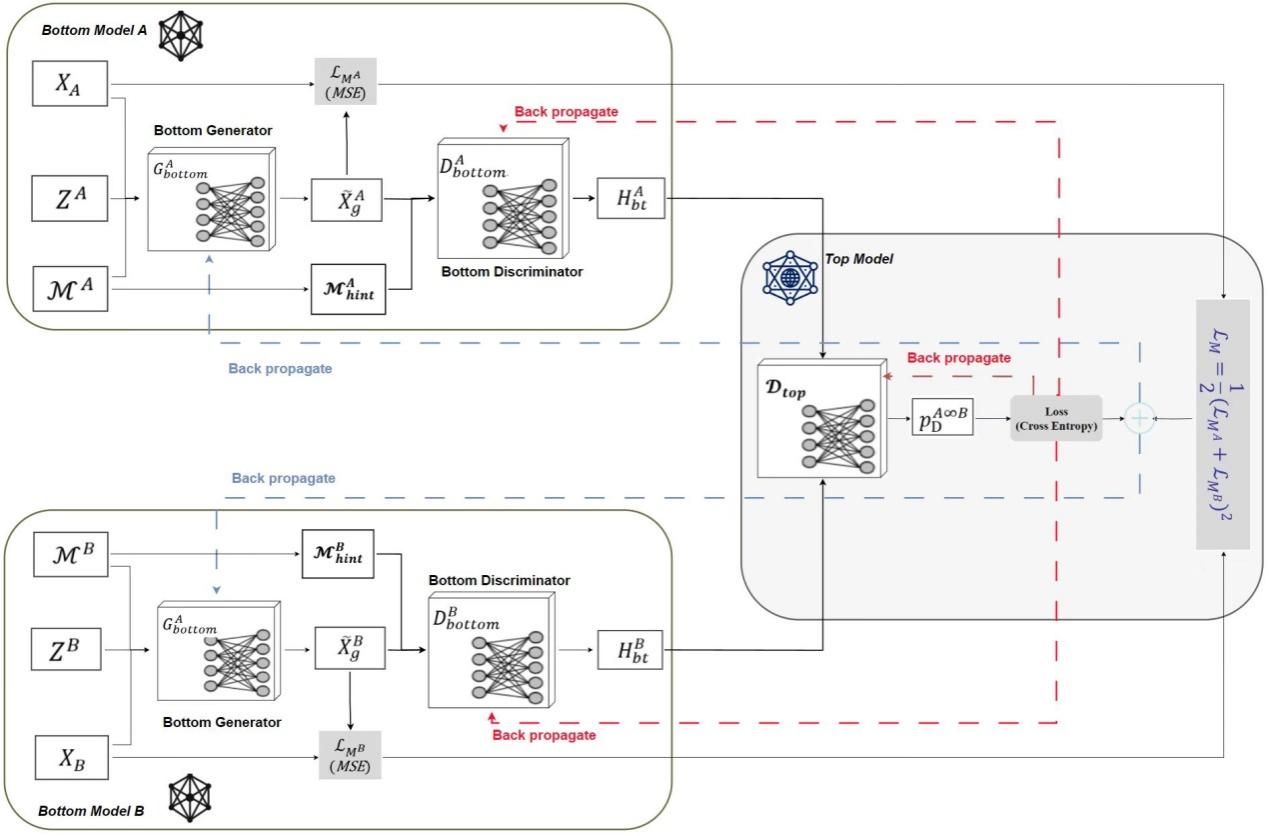


Fig.3：The model framework of Collaborative Data Imputation Based on VFL（以VF-GAIN为例）

（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方剩余需填补的样本数据表示为A方的缺失数据，用向量表示：

即

将B方剩余需填补的样本数据表示为B方的缺失数据，用向量表示：

即

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。以下以基于GAIN的联系联邦填补为例，详述各方底部模型的结构和计算过程。

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两方的生成器网络主要包括2个ResNet全连接模块（ResNet-FC）和1个全连接网络模块（FC）。前两个ResNet-FC模块中使用Relu激活函数，并在其中采用批量规范化操作（Batch Normalization）稳定和加快学习过程，最后一个FC模块的输出中使用Sigmoid激活函数。

A、B方底部生成器的生成结果为：

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，如公式(XXX)所示:

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：

由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：

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(XXX)所示:

和B方生成的中的部份，将用于联邦学习过程中损失函数和的运算，详细损失函数公式（XXXX）。其中，。

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[加引用]。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[加引用]。

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：

在上述底部模型中，我们基于 GAIN 提出了一个纵向联邦学习的底部模型框架。需要注意的是，不同填补模型的联邦底部模型有所差异。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-PC-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。它们的各单方网络结构与原本的填补模型（GAIN、PC-GAIN、VGAIN、CGAIN）的结构一致。也就是说，VF-GAIN 的各方底层模型与 GAIN 的结构一致，VF-PC-GAIN 与 PC-GAIN 对应，VF-VGAIN 与 VGAIN 对应，VF-CGAIN 则与 CGAIN 对应。GAIN、PC-GAIN、VGAIN、CGAIN的具体模型结构和实现细节可以参考这些论文 [参考文献1]、[参考文献2]、[参考文献3]、[参考文献4]。

对于VF-CGAIN，A方、B方底部生成器的输入，除了包括VF-GAIN的输入外，还需要在A方或B方其中一方中加入联合样本的条件向量C，用于指示联合样本所属样本类别。A、B两方底部生成器网络的结构与VF-GAIN相同。设作为条件向量C的标签列在B方，其各方生成器的生成结果为：

A、B两方底部判别器公式为：

对于VF-PC-GAIN，其底部模型结构PC-GAIN是结合了生成器、判别器以及辅助分类器三部分，以提升数据填补质量。生成器的输入与VF-GAIN的输入一致，判别器接收与VF-GAIN判别器一样的输入作为输入。此外，辅助分类器利用低缺失率数据中的伪标签进行预训练，并在对抗训练阶段固定不变，用以约束生成器生成具有明确类别特征的数据，同时通过熵损失提高生成数据的类别一致性与特征保真度。辅助分类器公式为：

其中，表示辅助分类器。

对于VF-VGAIN，A方和B方的部生成器和判别器的输入、输出和VF-GAIN完全一样，不同之处在于，A方和B方的生成器网络结构包括编码器、嵌入层和解码器，与VAE结构相同。

（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：

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

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：

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：

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。

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、、。生成器损失包括底部生成器损失函数和服务器端损失函数。

生成器的总损失函数为：

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。

相比于VF-GAIN的损失函数，由于VF-CGAIN，VF-PC-GAIN, VF-VGAIN的底部模型的不同，导致其底部模型损失函数也略有不同，详细如下：

对于VF-CGAIN，由于其底部模型采用的是CGAIN填补模型，需要加入类别标签，即

其中，表示第条样本的标签向量，其余部分的损失函数与上述VF-GAIN一致。

对于VF-PC-GAIN，由于其模型中加入了辅助分类器，故需要加入辅助分类器的损失函数：

其中，和分别表示A方和B方第条样本的辅助分类器的输出。因此，VF-PC-GIAN的生成器总损失为：

其中，为超参数，在本文实验中超参数取值参照PC-GAIN方法中的超参数设定。而其判别器损失函数则和VF-GAIN一致。

对于VF-VGAIN，由于A、B两方底部生成器结构和VAE结构一致，即包含了编码器、嵌入层和解码器。编码器将输入向量映射到潜在表示 ，并通过在潜在分布上施加先验正则化来进行正则化，即先验正则化项​。在训练生成器时，不仅确保填补的缺失值能够成功地欺骗判别器，还需要最小化编码器和解码器的重建误差和先验正则化项的总和，因此，VF-VGAIN的生成器总损失为：

其中

式中，表示KL散度，表示A、B两方解码器的联邦输出，为标准正态分布。在本文实验中超参数取值参照VGAIN方法中的超参数设定。

1. 训练过程

设作为条件向量C的标签列在B方，顶部模型在B方。基于GANs的纵向联邦填补方法的训练过程如下：

Step1：A方初始化底部生成器权重参数，底部判别器权重参数，B方初始化底部生成器权重参数，底部判别器权重参数，以及顶部判别器权重参数。协调者C方创建密钥对，并将公共密钥发送给A方和B方，通过同态加密技术确保在后续交换信息过程中，无需暴露原始数据和隐私信息，其中同态加密操作用表示。A方将自己拥有的样本数据放入A方底部生成器中，B方将自己拥有的样本数据放入B方底部生成器中。将、和输入A方底部生成器网络。将、和输入B方底部生成器网络。

Step2：A方底部生成器网络进行正向传播得到和，并计算得到A方的损失。由B方底部生成器进行正向传播得到和，并计算得到B方的损失。A方由、在底部判别器网络进行正向传播得到。B方由、在底部判别器网络进行正向传播得到。A方对、的每一个元素值进行加密，得到和加密元素向量，并将、和发送B方。B方对、的每一个元素值进行加密，得到和加密元素向量。B方收到A方的、、，将和联合成为加密元素向量，将和联合成为，输入顶层判别器网络进行正向传播计算，计算判别器的总损失。同时，在B方计算生成器损失，由和传入的计算，以及由、计算生成器总损失。 将判别器的总损失和生成器总损失发送给C方。

Step3：A方、B方初始化加密随机掩码、，这些加密随机掩码可以保证在传输过程中中间结果无法被逆向还原，从而保证了数据的隐私性。

在判别器训练的反向传播过程中：在B方，计算和+，并由和传入的将B方顶层判别器网络第一层的权重梯度拆分成为和，以便A、B两方计算和。B方将，用于计算+。将+和+发送C方，将发送给A方。在A方，计算+，并将其发送C方。

在生成器训练的反向传播过程中：B方计算+，并将其发送C方。A方计算+，并将其发送C方。

C方，解密和、+、+、+、+、+，将+，+发送给A方，将+、+、+发送给B方。

Step4：A方、B方解除收到的梯度信息上的加密随机掩码，并根据这些梯度信息分别在各方更新底部生成器、判别器网络权重参数、、，以及更新顶部判别器网络权重参数。

Step5：重复迭代Step2-Step4，直到训练结束。

# 4 实验

## 4.1 数据集和数据准备

在实验中，我们使用了四个数据集来评估所提出的方法：银行营销数据集 [参考文献] 和德国信贷数据集 [参考文献]、**Letter Recognition**数据集[ ]和**Online News Popularity 数据集**[ ]。

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。这些营销活动是通过电话进行的，产生的数据集包含 45,211 个实例和 16 个属性，以及一个 ID 列和一个标签列。该数据集按时间顺序排列，从 2008 年 5 月至 2010 年 11 月，目的是对客户是否会认购（是/否）定期存款（以变量 y 表示）进行分类。

② 德国信贷数据集广泛应用于机器学习和数据科学，尤其是分类和信贷风险建模任务。该数据集源自信用评分系统，可用于训练贷款违约预测和推荐系统的模型。一共有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。我们的目标是将申请人划分为信用 “良好 ”或 “不良”。

**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。该数据集旨在将每个样本分类为26个英文字母之一（A-Z），广泛应用于模式识别和图像分类研究。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，帮助训练和评估字符分类模型的性能。

**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，目标是预测文章将获得的分享次数或浏览量。

为简单起见，本文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，我们针对图1(a)情况进行了充分的实验论证。图1(b)情况可根据图1(a)情况进行有效性的合理推理。在实验中，本文为图1(a)情况中的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。

## 4.2 实验一的设计与结果分析

为了结合多方信息对样本缺少的参与方进行样本生成，在本文中提及了VF-PSG的两种方法：（1）方法一：用生成模型独立地为该方生成样本的重要属性值（GM表示生成模型），再由纵向联邦填补模型生成剩余的属性值，称为VF-PSG-GM。（2）方法二：通过联邦关联规则生成样本的重要属性值（VF-AR表示联邦关联规则），再由纵向填补模型生成剩余的属性值，称为VF-PSG-AR。这两种方法的主要区别在于样本的重要属性值生成方法不同。为了验证这两种方法的优劣，B方在相对于A方的不同缺少比下，本文做了以下三种实验设定。

（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，设置不同的重要属性值的生成列数，以验证在生成不同的属性列数下，两种方法为样本缺少的参与方生成数据的最终效果。其中，VF-PSG-GM中的GM采用TabDDPM生成B方缺少样本的重要属性值，TabDDPM[ ]是目前文献中生成效果表现SOTA的生成模型。VF-PSG-AR中的VF-AR采用基于Aprior算法的联邦关联规则方法生成B方缺少样本的重要属性值。两种方法采用的生成B方样本剩余属性值的填补模型都是VF-GAIN。

实验设定二：固定方法中重要属性值的生成列数，设置剩余属性值的不同填补模型，以验证在不同的填补模型下，两种方法为样本缺少的参与方生成数据的最终效果。其中，固定的重要属性值的生成列数是由实验设定一实验结果中的最佳列数确定。生成B方样本的重要属性值时，两种方法采用的生成模型与实验设定一相同。用于对比的填补模型为目前效果较好的几种模型：Mean[ ]、MissFI（其中随机森林中树的棵数设为100）[ ]、MICE[ ]、GAIN[ ]、PC-GAIN[ ]、VGAIN[ ]、CGAIN[ ]。实验中，将它们在纵向联邦框架下进行填补，被表示为VF-MissFI，VF- MICE，VF- GAIN， VF-PV-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：固定方法中重要属性值的生成列数，验证多方纵向联邦填补和仅单方填补对B方剩余属性列的生成效果。一种方式为非联邦填补：基于B方原有数据和使用本文两种方法生成获得的B方重要属性列，对B方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方原有数据和使用本文两种方法生成获得的B方重要属性列，纵向联邦地对B方剩余属性列进行数据填补。生成B方样本的重要属性值时，两种方法采用的生成模型与实验设定一相同。填补模型采用了实验设定二的实验结果中较优的VF-GAIN、VF-CGAIN、VF-PC-GAIN、VF-VGAIN的联邦和非联邦实现。

本实验用于评估数据生成效果的评估指标均为RMSE[ ]，用于表示生成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。实验中，重要属性值的生成列数用‘Cnum’表示。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。其中，PC-GAIN中聚类过程使用的邻居数量设为4，聚类方法采用KM（K-means）方法。

（2）实验一的结果分析

表1、表2和图4展示了实验设定一的实验结果。LackR-B表示B方相对于A方的样本缺少率，Cnum表示不同的重要属性值的生成列数。当Cnum=0时，表示没有用GM或者VF-AR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，表示用GM或者VF-AR生成B方缺少样本的1个属性列，剩余样本属性采用联邦填补模型生成；以此类推...。如表1、表2和图4所示，在B 方不同的样本缺少率下，当Cnum为4、5、6时，VF-PSG-GM和VF-PSG-AR两种方法的RMSE值表现更优，如表1和表2中的红色和紫色字体表示的结果。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“先生成一些重要属性值，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A方和B方的属性相关性，能更多的生成一些相关性高的属性，再进行填补，生成获得的样本的RMSE更低，更利于得到接近真实的样本集。而随着生成重要属性值的列数的增加，例如表1和图4中**Bank Dataset**的7列和8列，表2和图4中**Credit Dataset**的7列至11列，生成样本的RMSE也会有所增加，这说明并非生成的重要属性列数越多越好，一些非重要和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，可能会导致关联规则或生成模型在推理和学习过程产生误差，降低生成数据的质量。而采用联邦填补方法生成非重要属性列，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，通过生成部份重要且高相关性属性列，结合联邦填补方法生成填补剩余属性列，能够在保证生成数据的质量。同时，将VF-PSG-GM和VF-PSG-AR的实验结果相比，表1和表2中RMSE值更低的实验结果用加粗字体进行表示。从加粗的实验结果可以看出，在B 方不同的样本缺少率下，在实验设定一中，在重要属性的不同生成列数下，VF-PSG-AR总体上来说优于VF-PSG-GM。这表明通过联邦关联规则生成重要属性的方式比直接在B方采用生成模型去生成重要属性的方式更佳。特别地，即使B方中的GM采用了目前文献表现最优的TabDDPM进行数据生成，其实验结果也无法超越联邦关联规则的实验结果。

表3展示了实验设定二的实验结果。在此，我们固定的生成重要属性的列数是由表1和表2中实验结果相对较优的列数---5列。在B 方不同的样本缺少率下，在实验设定二中，VF-PSG-GM和VF-PSG-AR的最优和次优RSME结果都出现在当填补模型为VF-GAIN、VF-PC-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-GAIN、VF-CGAIN时；而Credit数据集的最优结果出现在填补模型为VF-VGAIN时。这些结果也表明，基于GAN改进的纵向联邦填补模型的有效性。两种方法中的最差RSME结果都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。同时，如表3所示，将VF-PSG-GM和VF-PSG-AR的实验结果相比，RMSE值更低的实验结果用加粗字体进行表示。从加粗的实验结果可以看出，在B 方不同的样本缺少率下，在实验设定二中，不管选择用什么填补模型对缺少样本的剩余属性进行生成，VF-PSG-AR总体上来说优于VF-PSG-GM。这更进一步表明通过联邦关联规则生成重要属性的方式的优势，它能适应不同填补模型对剩余属性的生成，确保VF-PSG-AR的鲁棒性。

表4展示了实验设定二的实验结果。在此，我们将Cnum分别设置为1列、5列、8列。在使用本文方法的GM和VF-AR为B方生成缺少样本的重要属性列后，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4中的联邦填补方式是指利用A、B两方原有和生成数据进行基于VF-GAIN模型的填补生成，非联邦填补方式是指仅利用B方原有和生成数据进行基于单方GAIN模型的填补生成。如表4的实验结果所示，采用本文方法VF-PSG生成B方缺少样本的重要属性列后，采用基于VF-GAIN的联邦填补方式为B方生成剩余属性，全面优于基于单方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，当生成重要属性值的列数为5列时，在B 方不同的样本缺少率下，VF-PSG-GM和VF-PSG-AR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。而且VF-PSG-AR中联邦和非联邦填补方式的实验结果都优于VF-PSG-GM中联邦和非联邦填补方式，如表4的加粗实验结果所示，这与实验设定一和二的结论是一致的。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，我们在生成重要属性值的列数为5列的情况下，采用GAIN、PC-GAIN、VGAIN、CGAIN进行了联邦和非联邦填补的对比。如表5所示，VF-PSG-GM和VF-PSG-AR两种方法下，基于VF-GAIN、VF-PC-GAIN、VF-VGAIN、VF-CGAIN的联邦填补方式均全面优于基于单方GAIN、PC-GAIN、VGAIN、CGAIN的非联邦填补方式，这于表4的实验结论一致。

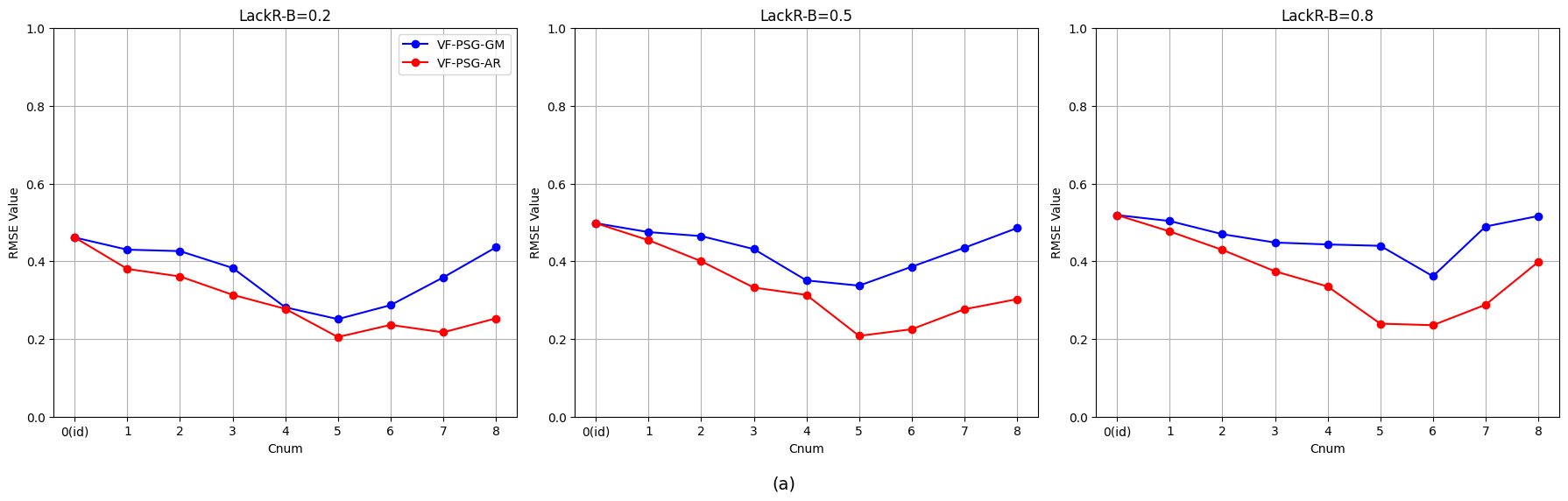
**表1** 重要属性值的不同生成列数的生成效果by VF-PSG（RMSE**，on Bank Dataset**）

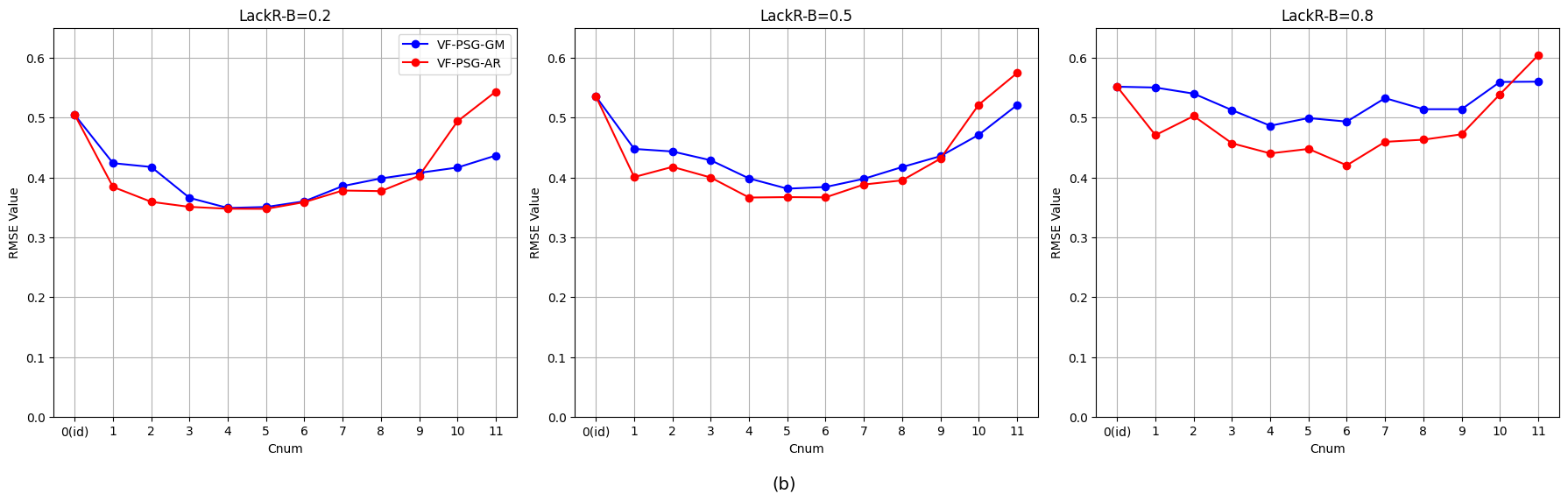
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cnum  LackR-B | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.2 | VF-PSG-GM | 0.4615 | 0.4302 | 0.4265 | 0.3832 | 0.2821 | 0.2518 | 0.2876 | 0.3588 | 0.4360 |
| VF-PSG-AR | **0.4615** | **0.3807** | **0.3614** | **0.3142** | **0.2782** | **0.2057** | **0.2369** | **0.2177** | **0.2536** |
| 0.5 | VF-PSG-GM | 0.4981 | 0.4753 | 0.4651 | 0.4317 | 0.3512 | 0.3378 | 0.3864 | 0.4351 | 0.4856 |
| VF-PSG-AR | **0.4981** | **0.4545** | **0.4005** | **0.3328** | **0.3139** | **0.2087** | **0.2258** | **0.2771** | **0.3029** |
| 0.8 | VF-PSG-GM | 0.519 | 0.5038 | 0.4700 | 0.4485 | 0.4436 | 0.4401 | 0.3961 | 0.4897 | 0.5166 |
| VF-PSG-AR | **0.519** | **0.4770** | **0.4296** | **0.3746** | **0.3356** | **0.2401** | **0.2362** | **0.2887** | **0.3986** |

**（注：紫色字体表示次优，红色字体表示最优）**

**表2** 重要属性值的不同生成列数的生成效果by VF-PSG（RMSE**，on Credit Dataset**）

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cnum  LackR-B | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 0.2 | VF-PSG-GM | 0.5048 | 0.4241 | 0.4177 | 0.3661 | 0.3493 | 0.3509 | 0.3604 | 0.3858 | 0.3986 | 0.4079 | **0.4168** | **0.4368** |
| VF-PSG-AR | **0.5048** | **0.3844** | **0.3595** | **0.3511** | **0.3480** | **0.3478** | **0.3590** | **0.3784** | **0.3774** | **0.4031** | 0.494 | 0.5436 |
| 0.5 | VF-PSG-GM | 0.5354 | 0.4479 | 0.4436 | 0.4289 | 0.3986 | 0.3815 | 0.3843 | 0.3980 | 0.4176 | 0.4358 | **0.4714** | **0.521** |
| VF-PSG-AR | **0.5354** | **0.4009** | **0.4179** | **0.4001** | **0.3667** | **0.3674** | **0.3670** | **0.3884** | **0.3954** | **0.4316** | 0.5271 | 0.5745 |
| 0.8 | VF-PSG-GM | 0.5518 | 0.5504 | 0.5402 | 0.5127 | 0.4866 | 0.4994 | 0.4935 | 0.5327 | 0.5142 | 0.5346 | 0.5597 | **0.5602** |
| VF-PSG-AR | **0.5518** | **0.4709** | **0.5028** | **0.4573** | **0.4404** | **0.4479** | **0.4205** | **0.4597** | **0.4633** | **0.4723** | **0.5391** | 0.6043 |

****

****

**图4** 重要属性值的不同生成列数的生成效果折线图by VF-PSG: (a) Bank Dataset; (b) Credit Dataset

**表3** 不同填补模型的生成效果by VF-PSG（RMSE**，on Bank and Credit Dataset**）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Imputation  Dataset  （LackR-B） | | | Mean | VF-MissFI | VF-MICE | VF-GAIN | VF-PC-GAIN | VF-VGAIN | VF-CGAIN |
| Bank | 0.2 | VF-PSG-GM | 0.4077 | 0.4036 | 0.4214 | 0.2518 | 0.2947 | 0.2372 | 0.2691 |
| VF-PSG-AR | **0.4075** | **0.3333** | **0.3233** | **0.2057** | **0.2728** | **0.2291** | **0.1691** |
| 0.5 | VF-PSG-GM | 0.4744 | 0.4563 | 0.4047 | 0.3378 | 0.3229 | 0.2997 | 0.2722 |
| VF-PSG-AR | **0.4578** | **0.3501** | **0.3743** | **0.2087** | **0.2890** | **0.2335** | **0.2332** |
| 0.8 | VF-PSG-GM | 0.5287 | 0.4776 | 0.4520 | 0.4401 | 0.3719 | 0.3537 | 0.3832 |
| VF-PSG-AR | **0.4989** | **0.4006** | **0.3567** | **0.2401** | **0.3011** | **0.3024** | **0.2533** |
| Credit | 0.2 | VF-PSG-GM | 0.4815 | 0.4481 | 0.4512 | 0.3509 | 0.4139 | 0.3495 | 0.4376 |
| VF-PSG-AR | **0.4567** | **0.4372** | **0.4471** | **0.3478** | **0.4083** | **0.3373** | **0.4029** |
| 0.5 | VF-PSG-GM | 0.5006 | 0.4516 | 0.4911 | 0.3815 | 0.4309 | 0.3610 | 0.4522 |
| VF-PSG-AR | **0.4773** | **0.4359** | **0.4789** | **0.3674** | **0.4112** | **0.3503** | **0.4242** |
| 0.8 | VF-PSG-GM | 0.5437 | 0.4928 | 0.5210 | 0.4994 | 0.4832 | 0.4621 | 0.4657 |
| VF-PSG-AR | **0.4910** | **0.4701** | **0.4982** | **0.4479** | **0.4373** | **0.4030** | **0.4433** |

**（注：紫色字体表示次优，红色字体表示最优）**

**表4 基于GAIN联邦和非联邦填补的生成效果by** VF-PSG when num=1,5,8 （RMSE**，on Bank and Credit Dataset**）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VF-Imputed(Y/N)  Dataset(LackR-B) | | | 1 | | 5 | | | 8 | |
| VF-GAIN | GAIN | VF-GAIN | GAIN | GAIN | VF-GAIN | GAIN |
| Bank | 0.2 | VF-PSG-GM | 0.4302 | 0.4477 | 0.2518 | | 0.3305 | 0.4360 | 0.4372 |
| VF-PSG-AR | **0.3807** | **0.4041** | **0.2057** | | **0.2779** | **0.2536** | **0.3468** |
| 0.5 | VF-PSG-GM | 0.4753 | 0.491 | 0.3378 | | 0.3524 | 0.4856 | 0.5072 |
| VF-PSG-AR | **0.4545** | **0.4744** | **0.2087** | | **0.295** | **0.3029** | **0.3437** |
| 0.8 | VF-PSG-GM | 0.5038 | 0.5149 | 0.4401 | | 0.3468 | 0.5166 | 0.5418 |
| VF-PSG-AR | **0.4770** | **0.4951** | **0.2401** | | **0.314** | **0.2986** | **0.3577** |
| Credit | 0.2 | VF-PSG-GM | 0.4241 | 0.5542 | 0.3509 | | 0.5095 | 0.3986 | 0.5513 |
| VF-PSG-AR | **0.3844** | **0.5317** | **0.3478** | | **0.4945** | **0.3774** | **0.5017** |
| 0.5 | VF-PSG-GM | 0.4479 | 0.5923 | 0.3815 | | 0.5108 | 0.4176 | 0.5467 |
| VF-PSG-AR | **0.4009** | **0.5748** | **0.3674** | | **0.5044** | **0.3954** | **0.4456** |
| 0.8 | VF-PSG-GM | 0.5504 | 0.5791 | 0.4994 | | 0.5314 | 0.5142 | 0.5551 |
| VF-PSG-AR | **0.4709** | **0.5631** | **0.4479** | | **0.5158** | **0.4633** | **0.4615** |

**表5 基于不同填补模型联邦和非联邦的生成效果by** VF-PSG when num=5（RMSE**，on Bank and Credit Dataset**）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VF-Imputioant(Y/N)  Dataset(LackR-B) | | | 5列 | | 5列 | | 5列 | | 5列 | |
| VF-GAIN | GAIN | VF-PC-GAIN | PC-GAIN | VF-VGAIN | VGAIN | VF-CGAIN | CGAIN |
| Bank | 0.2 | VF-PSG-GM | 0.2518 | 0.3305 | 0.2947 | 0.3653 | 0.2372 | 0.2764 | 0.2691 | 0.3531 |
| VF-PSG-AR | **0.2057** | **0.2779** | **0.2728** | **0.3329** | **0.2291** | **0.2378** | **0.1691** | **0.2198** |
| 0.5 | VF-PSG-GM | 0.3378 | 0.3524 | 0.3229 | 0.4245 | 0.2997 | 0.3236 | 0.2722 | 0.3336 |
| VF-PSG-AR | **0.2087** | **0.295** | **0.2890** | **0.3234** | **0.2335** | **0.2854** | **0.2332** | **0.2775** |
| 0.8 | VF-PSG-GM | 0.3468 | 0.4401 | 0.3719 | 0.4479 | 0.3537 | 0.386 | 0.3832 | 0.4018 |
| VF-PSG-AR | **0.2401** | **0.314** | **0.3011** | **0.4096** | **0.3024** | **0.3364** | **0.2533** | **0.3333** |
| Credit | 0.2 | VF-PSG-GM | 0.3509 | 0.5095 | 0.4139 | 0.4835 | 0.3495 | 0.4289 | 0.4376 | 0.4618 |
| VF-PSG-AR | **0.3478** | **0.4945** | **0.4083** | **0.4267** | **0.3373** | **0.4086** | **0.4029** | **0.4273** |
| 0.5 | VF-PSG-GM | 0.3815 | 0.5108 | 0.4309 | 0.5108 | 0.3610 | 0.4573 | 0.4522 | 0.5003 |
| VF-PSG-AR | **0.3674** | **0.5044** | **0.4112** | **0.476** | **0.3503** | **0.4354** | **0.4242** | **0.467** |
| 0.8 | VF-PSG-GM | 0.4994 | 0.5314 | 0.4832 | 0.532 | 0.4621 | 0.4846 | 0.4657 | 0.5059 |
| VF-PSG-AR | **0.4479** | **0.5158** | **0.4373** | **0.492** | **0.4030** | **0.4217** | **0.4433** | **0.4651** |

## 4.3 实验二的设计与结果分析

为了验证本文方法的有效性，在实验一的基础上，我们将VF-PSG的两种方法VF-PSG-GM和VF-PSG-AR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[ ]、TableGAN[ ]、CTAB-GAN[ ]、VAE[ ] 和 TabDDPM[ ]。实验也设置了B方在相对于A方的不同缺少比。本实验数据生成效果的评估指标均也为RMSE，用于表示生成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VF-PSG-GM和VF-PSG-AR的两种方法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 方不同的缺少率下，当Cnum=1时，VF-PSG两种方法的RMSE值都较差，而列数为5时，两种方法的RMSE值都表现较优。又根据实验一中实验设定二的实验结果表3 得知，在B 方不同的缺少率下，联邦填补模型为VF-MissFI时，两种方法的RMSE值都较差，而联邦填补模型为基于GAN的系列模型时，两种方法的RMSE值都表现较优。 因此，在本组实验的VF-PSG两种方法中，Cnum分别设置为1和5，联邦填补模型分别包括VF-MissFI、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。在本组实验中，对于VF-PSG-GM和VF-PSG-AR的两种方法实验设置有以下说明：根据实验二第一组实验结果，在本组实验的VF-PSG两种方法中，Cnum分别设置为5，且我们将从基于GAN的系列模型中任意选取一个VF-GAIN作为联邦填补模型。

在本实验中，VF-PSG-GM 设置的生成模型为仍为TabDDPM，VF-PSG-AR 使用 Apriori 算法进行联邦关联规则生成。当Cum设置为1 和 5时，分别用VF-PSG-GM-1/ VF-PSG-GM-5，和VF-PSG-AR-1/ VF-PSG- AR -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6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，CTAB-GAN和TabDDPM生成数据的RMSE值是较优的，甚至其大部份RMSE值优于了当Cnum=1且填补模型为VF-MissFI的VF-PSG-GM的结果。虽然 MissFI 依赖于随机森林的强大预测能力，通过已观测的特征可以预测缺失的特征值。但当样本数量较少或数据中存在大量缺失值时，MissFI 的性能会下降。在处理高维度、非线性和特征复杂交互的数据时，MissFI 的填补效果不理想，无法有效捕捉特征之间的非线性依赖关系，导致填补准确性下降。

而如表6结果所示，当缺少比较高时，不管Cnum为1还是5，基于GAN的纵向联邦填补模型的VF-PSG-GM和VF-PSG-AR都优于直接用生成模型为B方生成样本的RMSE。尤其是，当Cnum为5时，VF-PSG-GM和VF-PSG-AR取得了非常优异的生成效果。

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7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，CTAB-GAN和TabDDPM生成数据的RMSE值是较优的。但是，基于GAN的纵向联邦填补模型的VF-PSG-GM和VF-PSG-AR都全面优于所有直接用生成模型为B方生成样本的RMSE。尤其是，VF-PSG-AR取得了最优的生成效果。

VF-PSG 的卓越性能归功于其创新的模型设计，它将关联规则算法、生成模型与纵向联邦填补模型整合在一起，特别是将关联规则与填补模型相结合。这种组合能有效捕捉属性之间的内在关系，确保生成的数据与原始数据在逻辑和统计上保持一致--尤其是当属性之间存在强相关性时。此外，在纵向联邦学习框架下，VF-PSG 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 和 TableGAN）只注重通过学习参与者自己的数据来生成数据，而没有充分利用其他参与者的信息，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。

需要强调的是，在我们的方法中，纵向联邦填补模型的选择并不是讨论的重点。本文方法关注于为样本缺少的参与方进行样本生成的一种创新思路。当然，在通常场景下数据填补表现越好的填补模型，在我们的联邦学习框架中也会有更好的表现，表2和表3的实验结果也遵循了这一结论。对于本文方法中填补模型的选择，我们不推荐使用基于统计的填补方法，例如均值填补。原因在于，这类方法简单地用特征的平均值、中位数或其他统计特征替代缺失值，无法捕捉属性列数据之间的复杂关系，也无法联合利用其他参与方的数据。在这种情况下，联邦填补并没有实施，且如均值填补等基于统计的填补方法可能低估数据的方差和变异性，导致填补后的数据分布偏离真实情况。这可能引入偏差，影响后续机器学习模型的训练效果，尤其在数据缺失率较高或数据分布不均匀的情况下。

**表6 基于不同生成方法的B方缺少样本生成效果对比（RMSE，on Bank and Credit Dataset）**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |
| VF-PSG-GM-1（VF-MissFI） | 0.4363 | 0.4862 | 0.5296 | 0.5017 | 0.5345 | 0.5756 |
| VF-PSG-AR-1（VF-MissFI） | 0.4333 | 0.469 | 0.4826 | 0.4942 | 0.5312 | 0.5333 |
| VF-PSG-GM-1（VF-GAIN） | 0.4302 | 0.4753 | 0.5038 | 0.4241 | 0.4479 | 0.5504 |
| VF-PSG-AR-1（VF-GAIN） | 0.3807 | 0.4545 | 0.4470 | 0.3844 | 0.4009 | 0.4709 |
| VF-PSG-GM-1（VF-VGAIN） | 0.4254 | 0.4666 | 0.5024 | 0.42 | 0.4266 | 0.5367 |
| VF-PSG-AR-1（VF-VGAIN） | 0.3856 | 0.4322 | 0.4229 | 0.3745 | 0.4162 | 0.453 |
| VF-PSG-GM-5（VF-MissFI） | 0.4036 | 0.4563 | 0.4776 | 0.4481 | 0.4516 | 0.4928 |
| VF-PSG-AR-5（VF-MissFI） | 0.3333 | 0.3501 | 0.4006 | 0.4372 | 0.4359 | 0.4701 |
| VF-PSG-GM-5（VF-GAIN） | 0.2518 | 0.3378 | 0.4401 | 0.3509 | 0.3815 | 0.4994 |
| VF-PSG-AR-5（VF-GAIN） | 0.2057 | 0.2087 | 0.2401 | 0.3478 | 0.3674 | 0.4479 |
| VF-PSG-GM-5（VF-VGAIN） | 0.2372 | 0.2997 | 0.3537 | 0.3495 | 0.3610 | 0.4621 |
| VF-PSG-AR-5（VF-VGAIN） | 0.2291 | 0.2335 | 0.3024 | 0.3373 | 0.3503 | 0.4030 |

**（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）**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |
| VF-PSG-GM-1（VF-MissFI） | 0.4363 | 0.4862 | 0.5296 | 0.5017 | 0.5345 | 0.5756 |
| VF-PSG-AR-1（VF-MissFI） | 0.4333 | 0.469 | 0.4826 | 0.4942 | 0.5312 | 0.5333 |
| VF-PSG-GM-1（VF-GAIN） | 0.4302 | 0.4753 | 0.5038 | 0.4241 | 0.4479 | 0.5504 |
| VF-PSG-AR-1（VF-GAIN） | 0.3807 | 0.4545 | 0.4470 | 0.3844 | 0.4009 | 0.4709 |
| VF-PSG-GM-1（VF-VGAIN） | 0.4254 | 0.4666 | 0.5024 | 0.42 | 0.4266 | 0.5367 |
| VF-PSG-AR-1（VF-VGAIN） | 0.3856 | 0.4322 | 0.4229 | 0.3745 | 0.4162 | 0.453 |
| VF-PSG-GM-5（VF-MissFI） | 0.4036 | 0.4563 | 0.4776 | 0.4481 | 0.4516 | 0.4928 |
| VF-PSG-AR-5（VF-MissFI） | 0.3333 | 0.3501 | 0.4006 | 0.4372 | 0.4359 | 0.4701 |
| VF-PSG-GM-5（VF-GAIN） | 0.2518 | 0.3378 | 0.4401 | 0.3509 | 0.3815 | 0.4994 |
| VF-PSG-AR-5（VF-GAIN） | 0.2057 | 0.2087 | 0.2401 | 0.3478 | 0.3674 | 0.4479 |
| VF-PSG-GM-5（VF-VGAIN） | 0.2372 | 0.2997 | 0.3537 | 0.3495 | 0.3610 | 0.4621 |
| VF-PSG-AR-5（VF-VGAIN） | 0.2291 | 0.2335 | 0.3024 | 0.3373 | 0.3503 | 0.4030 |

**表7 基于不同生成方法的B方缺少样本生成效果对比（RMSE，on Letter and News Dataset）**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |
| VF-PSG-GM-5（VF-GAIN） | 0.4049 | 0.4468 | 0.4897 | 0.4112 | 0.463 | 0.5133 |
| VF-PSG-AR-5（VF-GAIN） | 0.3667 | 0.4297 | 0.465 | 0.3125 | 0.4138 | 0.4689 |

## 4.4 实验三的设计与结果分析

为了进一步验证 VF-PSG 为样本缺少参与方生成数据的有效性，我们将生成后的B方数据与A方数据构建联合样本集，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。而本实验中，用于处理B方缺少样本的方式包括：VF-PSG-GM；VF-PSG-AR；用TabDDPM独立地为B方生成缺少样本的数据，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，采用目前文献中较优的纵向联邦样本生成方法VertiGAN[ ]，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”。

不同的纵向联邦分类模型包括：纵向联邦逻辑回归（VF-LR）[ ]、纵向联邦支持向量机（VF-SVM）[ ]、纵向联邦梯度提升决策树（VF-GBDT）[ ]、纵向联邦随机森林（VF-RF）[ ]、和FinalNet[ ]。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）[ ]、AUC[ ] 和 F1 分数[ ]。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VF-PSG-GM、VF-PSG-AR、TabDDPM、A∞B-GM、N-GM这5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 VF-PSG-GM、VF-PSG-AR、TabDDPM这三种方式是为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VF-PSG-GM，VF-PSG-AR，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B方原有样本数量确定。

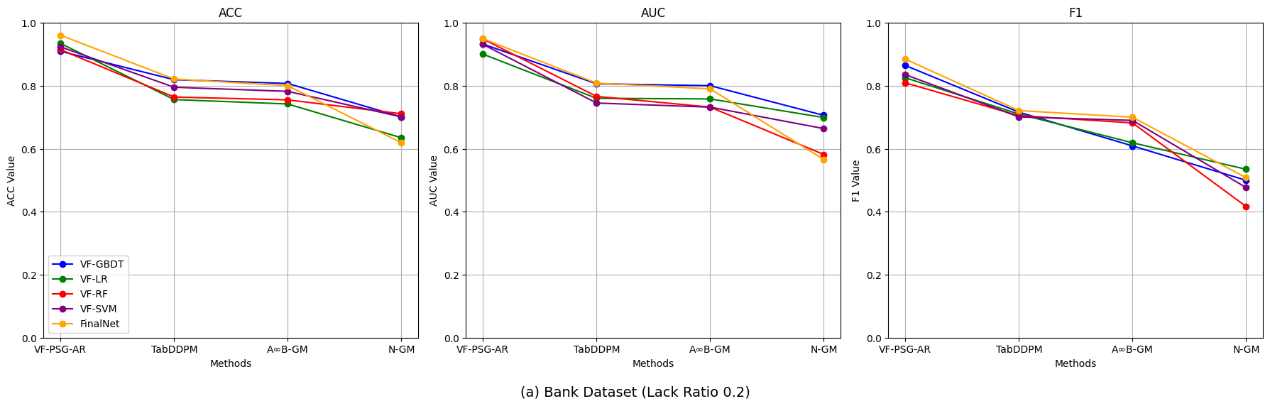
实验结果如图5所示，图5**（a）-（f）分别表示了Bank和Credit数据集在B方样本缺少率为0.2、0.5、0.8下，构建联合样本集时不同缺少样本处理方式下，不同纵向联邦分类模型的**ACC、AUC、F1 分数**结果曲线。**图5**（a）-（f）每个子图中不同颜色的曲线表示不同的纵向联邦分类模型，横坐标表示针对B方缺少样本不同的处理方式，即：**VF-PSG-GM、VF-PSG-AR、TabDDPM、A∞B-GM、N-GM，纵坐标表示**纵向联邦分类模型在基于不同处理方式得到的联合数据集上训练后，模型在测试集上得到的**ACC、AUC、F1 分数值。

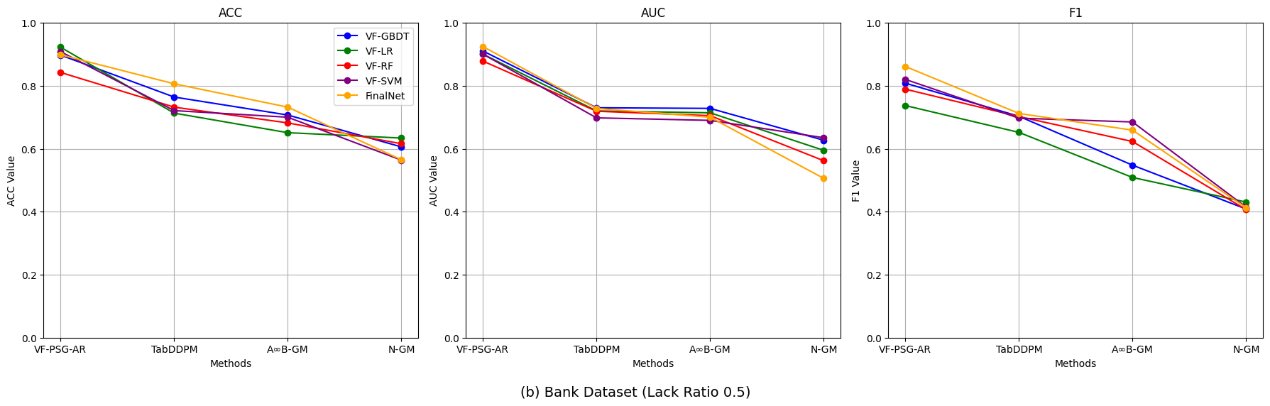
如图5所示，在两个数据集的不管在何种缺少率下，采用VF-PSG两种方法生成B方缺少样本所构建的联合样本集，不管用于VF-LR、VF-SVM、VF-GBDT、VF-RF、FinalNet哪种联邦机器学习模型的训练，其测试集的ACC、AUC、F1评估指标都是表现最为出色的，而其中，VF-PSG-AR更优。然后依次是TabDDPM、A∞B-GM、N-GM。分析出现这种结果的原因：

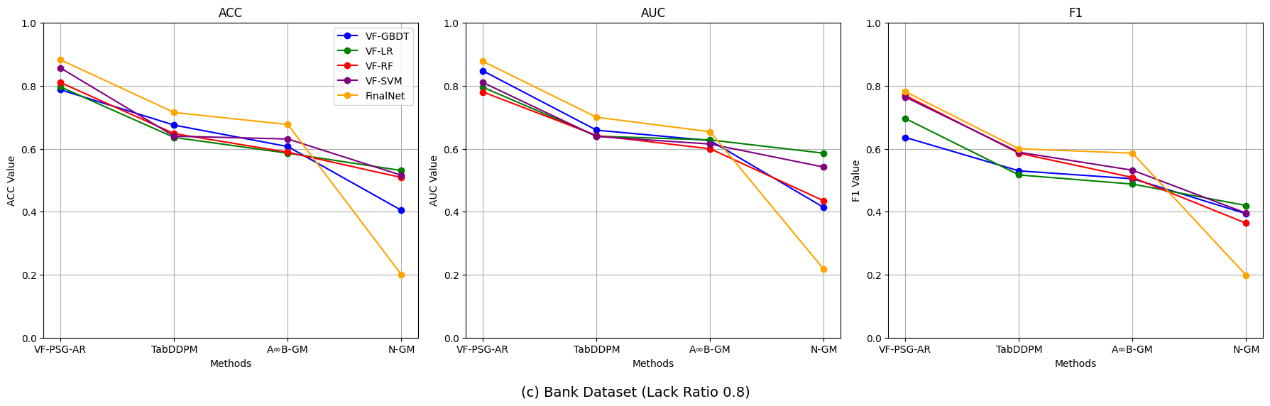
① 从样本数量的角度来分析，VF-PSG-AR、VF-PSG-GM、TabDDPM、A∞B-GM四种方式的联合数据集样本量相同，N-GM最少。如图5所示，在两个数据集的不管在何种缺少率下，样本量最少的N-GM方式下的联合样本集训练的纵向联邦分类模型的评估指标结果最差。这表明了联合样本量对纵向联邦机器学习模型训练的重要性。其他因素相同的情况下，样本量较大的联合样本集，训练的纵向联邦机器学习模型分类性能更好。如图5所示，我们用不同的联邦机器学习模型验证了这一结论。在同一个数据集中，随着B方缺少比的增加，N-GM方式获得的联合样本集样本量将大幅下降。当联合样本集样本量下降后，五种纵向联邦机器学习模型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。

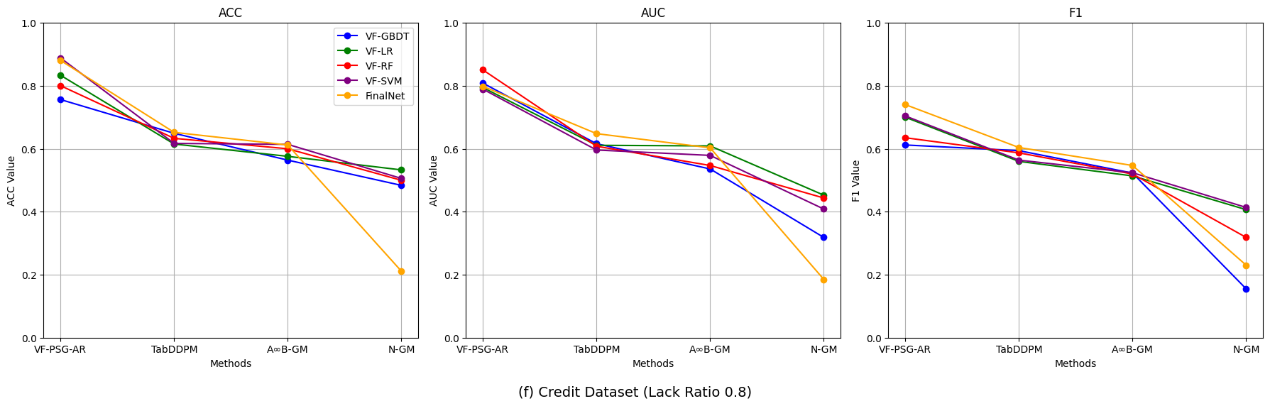
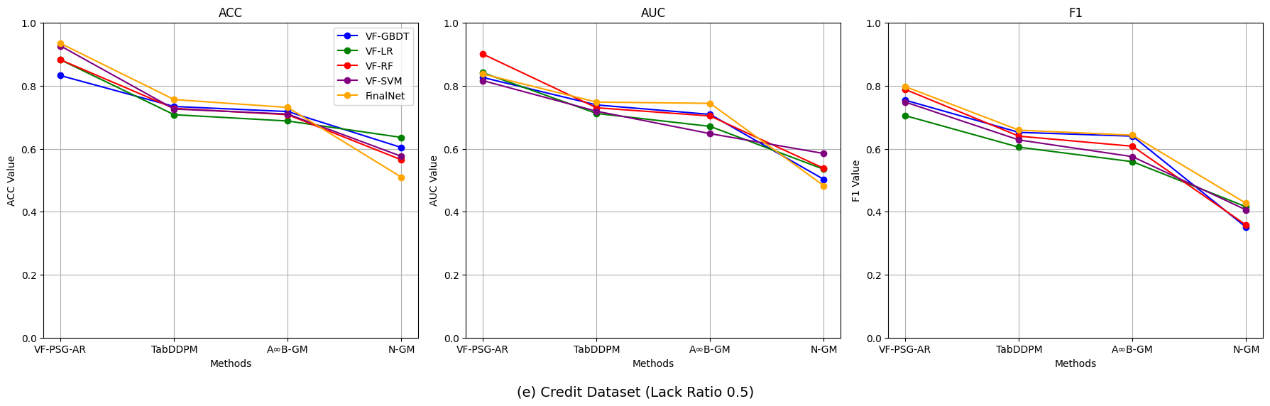
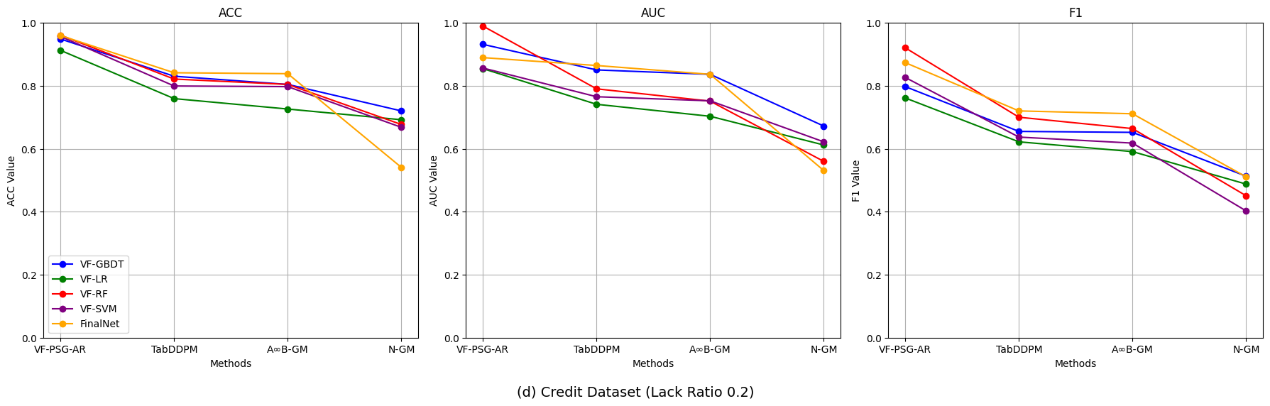
② 从样本质量的角度分析，虽然VF-PSG-AR、VF-PSG-GM、TabDDPM、A∞B-GM四种方式的联合样本集样本量相同，但他们的联合样本的质量却依次降低。VF-PSG-AR、VF-PSG-GM、TabDDPM这三种方式，A方的数据全部被保留了。比起A∞B-GM方式生成全新的联合样本，前三种方式构建的联合样本集中有更多的真实数据用于模型训练，联合样本质量更高。前三种方式中，VF-PSG尤其是基于关联规则的VF-PSG能够有效地利用多方数据之间的相关性进行B方缺少样本的生成和填补。而TabDDPM方式，B方缺少样本的数据是由TabDDPM在B方独立地生成的。这说明本文方法是解决参与方样本缺少问题的非常有效的方法。如图5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5所示，这一结论适用于不同的机器学习模型。

综上所述，通过实验3进一步表明：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。更重要是的，VF-PSG尤其是结合关联规则的VF-PSG能充分利用多方数据之间的相关性，联合多方数据为样本缺少参与方生成样本，且生成效果优于基于单方表现最优的生成模型，从而帮助联邦机器学习分类模型获得更优异的训练效果。

****

****

****

****

**图6 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）缺少率0.2下的Bank数据集；（b）缺少率0.5下的Bank数据集；（c）缺少率0.8下的Bank数据集；（d）缺少率0.2下的Credit数据集；（e）缺少率0.2下的Credit数据集；（f）缺少率0.2下的Credit数据集**

# 5 Conclusion（需要修改）

In this paper, we propose a novel Participants Sample Generation Method Based on Vertical Federated Learning (VF-PSG) to address the challenge of data generation for participant with fewer samples in multi-party collaborative scenarios. This problem occurs when certain participants have significantly fewer data samples, which limits the effective use of data in vertically federated learning settings. VF-PSG integrates attribute correlation analysis, association rule mining, and data generation through imputation techniques, enabling the generation of high-quality aligned samples for participants with fewer samples. Our proposed method is designed to ensure that important attributes are generated for participants with fewer samples, which are then used alongside a federated imputation process to generation the lack samples. This approach effectively expands the joint dataset, leading to improvements in overall model performance. We evaluated the method through a series of experiments comparing VF-PSG with other state-of-the-art generative models such as CTGAN, TableGAN, CTABGAN, VAE, and TabDDPM. Metrics such as RMSE, ACC, AUC, and F1-score were employed to quantify the effectiveness of the proposed approach.

Experimental results demonstrate that VF-PSG consistently outperforms existing methods across different lack ratios for Party B, especially in scenarios with high lack ratios. Specifically, when the lack ratios of Party B reached 80%, VF-PSG achieved significantly better RMSE results on both datasets than other generation models like TableGAN and CTGAN. This underscores the capability of VF-PSG to provide a highly effective solution for sample generation in parties with fewer samples.

The superior performance of VF-PSG can be attributed to its innovative model design, which integrates an association rule algorithm with generation and imputation models under the vertical federated learning framework. This combination effectively captures intrinsic relationships between attributes, ensuring that the generated data maintains logical and statistical consistency with the original data—particularly when attributes are strongly correlated. Furthermore, VF-PSG leverages information from multiple participants while preserving data privacy, thus making it suitable for applications in privacy-sensitive environments like healthcare and finance.