## 一种纵向联邦关联规则生成和数据填补的参与方样本生成方法

## 基于纵向联邦学习的参与方样本生成方法

Participants sample generation method based on association rules and data imputation within vertical federated learning

Xin Liua, Hangxuan Hea, Ying Qiana,\*, Feng Chena, Qingjie Weia, Hongjun Zhua

a *School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China*

Abstract: When multiple parties align their samples for vertical federation learning, some participants may lack certain samples that other participants possess, resulting in insufficient joint samples and ultimately impairing performance. To address this issue, we propose a novel Participants Sample Generation method based on Association Rules and data imputation within vertical federated learning, abbreviated as FedPSG-AR. Overall, to generate the missing samples, FedPSG-AR first generates partial attributes for those samples by vertical federated association rules, and then impute the remaining attributes by vertical federated imputation models. Specifically, an attribute generation method is proposed based on vertical federated association rules, which includes computing multi-party attribute correlations, establishing relationship of attribute values, generating attribute values by vertical federated association rules. It can generate the attributes in the missing samples of participants that are highly correlated with those of other parties under data secure privacy protection. Then, a vertical federated imputation framework based on GANs is constructed to generate the remaining attributes of these missing samples. In this federated imputation framework, we redesign the model structure, loss function and training process based on GANs. Experiments on multiple public datasets have thoroughly validated that our method outperforms the state-of-the-art baseline models currently available for participants sample generation in vertical federation learning.

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，属性相关性，关联规则，数据填补

Keywords: Participants sample generation, Vertical federated learning, Attribute correlation, Association rule, Data imputation

**1 引言**

Nowadays, with the development of AI, machine learning often needs to deal with data from multiple participants, where the same object has different attributes. For example, when conducting credit risk assessment, banks and e-commerce companies have a large amount of historical data available for evaluation. However, the data they hold on the same entity have different attributes. By combining the different attributes from both sources to assess the credit risk of the common entity, more valuable results can undoubtedly be obtained. However, due to security and privacy protection constraints, it is difficult for participants to share their data with others for training machine learning models. As a result, vertical federated learning (VFL) has emerged. In the process of VFL, sample entities with the same ID but held by different participants need to undergo encrypted sample alignment. After this alignment, we refer to the multi-party samples with the same ID as joint samples under vertical federated learning. The joint sample set leverages additional attributes from multiple parties to enhance the training of VFL models. In many real-world applications, when multiple participants engage in VFL, the sample entities available for alignment are not always completely identical across all participants. As shown in Figure 1(a), Party B has fewer samples compared to Party A. Similarly, as shown in Figure 1(b), either Party A relative to Party B or Party B relative to Party A has a portion of missing samples. After performing encrypted sample alignment across multiple participants, the number of joint samples will be significantly smaller than the complete sample size in each participant's independent dataset. As is well known, sample size is a crucial factor affecting the performance of machine learning model training. Therefore, in vertical federated learning, generating samples for participants with missing data to address the issue of insufficient joint samples after multi-party alignment, and exploring scientific methods to obtain more joint training samples, is a valuable research direction.

当今，人工智能的发展要求机器学习需要面对来自多个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，为样本缺少的参与方生成样本，以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。

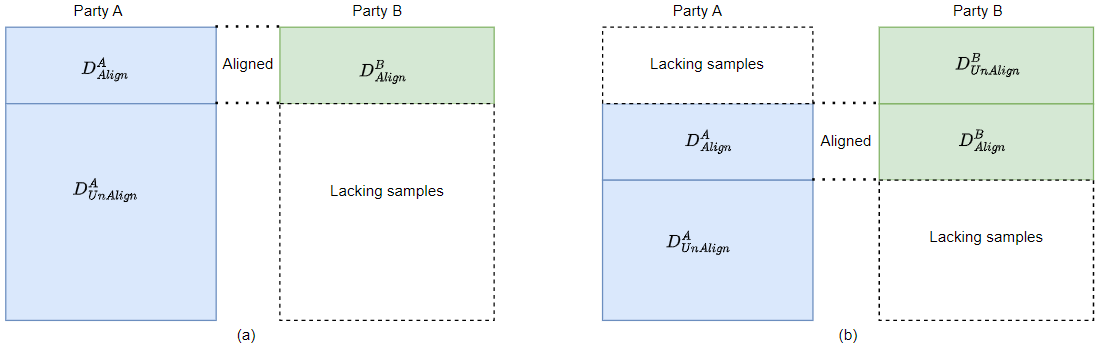


图1 参与方对齐样本不足的图例

Figure 1: Illustration of participants with insufficient aligned samples

There are two approaches to addressing the issue of insufficient joint samples after multi-party sample alignment: ① Generating new joint samples. The unaligned samples will be discarded. New joint samples will be generated based on aligned joint samples to expand the joint sample set. ② Generating samples for participants with missing samples to construct complete joint samples. Retain the unaligned samples from participants and generate samples for the participants with missing samples, enabling them to participate in multi-party sample alignment. This process constructs complete joint samples for the unaligned samples, thereby obtaining more joint samples.

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。

The first approach involves generating new joint samples. Currently, several sample generation methods based on Vertical Federated Learning (VFL) can used to generate the joint samples from multiple participants. For instance, vertical federated data generation methods based on Generative Adversarial Network (GANs): FedDA[73], VertiGAN [78], VFLGAN [79], GTV[74], and tabular data generation methods based on Markov Random Fields (MRFs) within vertical federated framework: VertiMRF[75]. However, in machine learning model training, not only is the sample size important, but the quality of the sample data is also crucial. Vertical federation generation methods yield completely new joint samples, and the data across all parties is synthesized and non-real. Moreover, when the missing sample ratio of some participants is high, the additional unaligned samples of other participants cannot be incorporated into the joint sample set. As a result, there are few joint samples available for training the vertical federation generation model, making it impossible to train an excellent vertical federation generation model. Therefore, methods for generating joint samples do not guarantee the acquisition of high - quality joint samples. The first approach of ‘generating new joint samples’ is not the optimal solution.

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，其他参与方的未对齐样本因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。

The second approach involves generating samples for participants with missing samples. Currently, various methods are available for generating these missing samples, such as Generative Adversarial Networks (GANs) (GANs)[ ] , Auto-Encoders (AEs)[ ], and Denoising Diffusion Probabilistic Models (DDPMs)[ ]. These methods are capable of generating high-quality data by learning the underlying data distribution and then creating synthetic samples. When generating samples for participants with missing samples, these methods learn locally from the data of such participant and do not consider the influence of other parties. However, in vertical federated learning applications, there are always certain associations among the data of all participants within the joint sample. Although some deep learning-based generation methods, such as the examples mentioned above, perform well, the models applied locally within each party overlook the role and impact of multi-party data associations on the generation results. To generate high-quality samples for participants with missing data, we need to collaboratively integrate data from other parties. This approach leverages multi-party data relevance to enhance sample generation. In this case, the problem of 'generating samples for participants with missing samples' transforms into the problem of 'imputing missing data in joint samples' within a vertical federation. Furthermore, to ensure data security and privacy protection, all parties involved in the collaborative generation process must prevent their data from being directly aggregated for training or computing the imputation model for participants with missing samples. Therefore, we need a highly effective method to generate and impute these missing sample data in this scenario to obtain a complete joint sample.

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，这些方法仅对本地数据进行学习，不考虑其他参与者对生成过程带来的影响。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，但仅在本方本地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协作地对该方进行数据生成，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。

To address these challenges，this paper proposes a Participants Sample Generation method based on association rules and data imputation within vertical federated learning, referred to FedPSG-AR. It aims to generate samples for participants with missing samples while ensuring the privacy of multi-party data., thereby providing more high-quality joint samples for federated machine learning. The contributions of this paper are as follows:

1. To generate the missing samples, this paper proposes a attribute generation method based on vertical federated association rules. The method generates the attributes in the missing samples of participants that are highly correlated with those of other parties under secure privacy protection.
2. Based on the attributes generated by vertical federated association rules, we construct a framework of vertical federated imputation to generate the remaining attributes. In this framework, we redesign the model structure, loss function and training process based on GANs. This method enhances the influence of other participants' data in the imputation process, maximizing the potential of multi-party collaborative learning.
3. This paper proposes a novel Participants Sample Generation method based on vertical federated learning, which design an effective strategy for generating the missing samples of participants. We first generates partial attributes of the missing samples by vertical federated association rules, and then impute the remaining attributes by vertical federated imputation models.

The paper is organized as follows: Section 2 discusses related work on data generation, and data imputation. In Section 3, we introduce our method for addressing the problem of participants samples generation. Experimental results are presented in Section 4, followed by conclusions in Section 5.

针对上述问题，本文提出了一种纵向联邦关联规则生成和数据填补的参与方样本生成方法（FedPSG-AR），通过纵向联合学习将关联规则和数据填补模型结合起来，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。本文的主要贡献如下：

1. 为了给样本缺少的参与方生成数据，本文提出了一种基于纵向联邦关联规则的属性生成方法。该方法通过学习不同参与者属性之间的相关性，在安全隐私保护下，为样本缺少的参与方生成高关联属性。
2. 基于联邦关联规则生成的属性值，我们构建了一种基于GANs的纵向联邦数据填补框架，用于生成样本缺少的参与方的剩余属性值。并对该框架下的填补模型结构、损夫函数、训练过程进行了重新设计。该方法增加了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。
3. 我们提出了一种新的基于纵向联合学习的参与方样本生成方法，该方法设计了一种“先在样本缺少方通过关联规则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。同时，基于本文方法研究，我们验证、对比和探讨了四种基于纵向联合学习的不同参与方样本生成方法，为解决参与方缺少样本生成问题提供创新且有效的思路。

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 Related Word**

2.1 Data generation methods

In addition to early machine learning-based data generation methods, state-of-the-art approaches that have demonstrated exceptional performance include autoencoder-based generative models, Generative Adversarial Networks (GANs), and Diffusion Probabilistic Models (DDPMs), among others. These methods are designed to generate new sample data by learning underlying patterns from existing samples.

Autoencoder-based generative models include Autoencoder (AE) [35], Variational Auto encoder (VAE) [37], and others. L. Xu et al. [38] proposed an improved VAE that explicitly models the joint distribution of hidden variables and tabular data features, enabling effective generation and reconstruction of both continuous and discrete features.

Proposed by I. J. Goodfellow et al. in 2014, Generative Adversarial Networks (GAN) is inspired by Game Theory. Internally, it consists of a generator (G) and a discriminator (D) network, which engage in an adversarial game against each other to efficiently generate data. M. Mirza et al. [40] proposed Conditional Generative Adversarial Nets (CGANs) to introduce conditional variables into the modeling of both the generative and discriminative models, to transform GANs from unsupervised to supervised learning. L. Xu et al. [38] proposed CTGAN, which is based on Conditional GAN, to generate tabular data for modeling tabular data distributions and samples. J. Lee et al. [47] proposed a generalized GAN framework for tabular data synthesis, integrating adversarial training with the negative log-density regularization of invertible neural networks to enhance the overall quality of the generated data. S. Singh et al. [49] introduced MeTGAN, which employs sparse linear layers to address the memory bottleneck of CTGAN, significantly reducing memory consumption during training. Z. Zhao et al. [50] proposed CTAB - GAN based on CTGAN for modeling different data types, including continuous variables, categorical variables, and mixed - feature variables. J. Engelmann et al. [52] proposed a conditional Wasserstein GAN-based approach for modeling tabular datasets with both numerical and categorical variables, incorporating an auxiliary classifier to specifically enhance performance on classification tasks.

The Diffusion Probabilistic Model (DDPM) [55] was proposed by J. Ho et al. It transforms the data distribution into a standard normal distribution via a forward step-by-step noise-addition process and then learns an inverse denoising process to generate high-quality target data. A. Kotelnikov et al. [56] proposed TabDDPM, a diffusion-model-based tabular data generation method. This method can efficiently capture the complex relationships between numerical and categorical features through a specific noise-addition and denoising process.

**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，它们旨在通过对已有样本的学习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[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。

2.2 Data imputation methods

With the rapid development and application of artificial intelligence and big data technologies, data generation methods based on Generative Adversarial Networks (GANs) and Diffusion Probabilistic Models (DDPMs) have become prominent in recent years. However, when certain data elements are missing in a sample, these methods fail to provide high-quality training samples for machine learning tasks such as classification and prediction. Consequently, generating and imputing missing data elements has emerged as a critical research direction. Currently, the main data imputation methods include statistical-based imputation methods, traditional machine-learning-based imputation methods, and deep-learning-based imputation methods.

Statistical-based imputation methods primarily replace missing values using statistical characteristics. Mean Imputation [57] substitutes missing values with the feature mean. K-Nearest Neighbors Imputation (KNNI) [33] identifies the K most similar neighbors to the missing sample using the KNN method and imputes the missing values with their mean or mode, which helps preserve the local data structure.

Traditional machine-learning-based imputation methods attempt to predict missing values by learning complex feature relationships. XGBoost Imputation (XGBI) [59] constructs a prediction model with XGBoost to predict the missing values either in a regression or classification manner. MissForest Imputation (MissFI) [34] iteratively predicts the missing values based on the Random Forest algorithm, and it can maintain high imputation accuracy under nonlinear relationships. Multiple Imputation by Chained Equations (MICE) [61] iteratively creates an imputation model for each variable via linear regression or generalized linear models and is suitable for multiple - imputation scenarios.

Deep-learning-based imputation algorithms, inspired by the powerful capabilities of deep generative models, achieve high-quality imputation through implicit or explicit distribution modeling. Auto - encoders (AEs), which embed data in the latent space through an encoding - decoding structure, preserve key data features in low - dimensional representations and impute the missing values during the decoding process. S. Ryu et al. [36] employed an autoencoder-based approach to impute missing data in smart meters. Generative Adversarial Networks (GANs) leverage adversarial training between a generator and a discriminator to produce synthetic samples and imputed values that align with the real distribution. They are commonly used for missing data reconstruction in complex, high-dimensional scenarios. J. Yoon et al. [68] proposed GAIN, which utilizes GAN to impute missing data. S. E. Awan et al. [69] proposed Conditional Generative Adversarial Imputation Network (CGAIN), inspired by CGAN. CGAIN takes the sample classification labels as inputs for generator and discriminator to impute the missing values in samples, simultaneously, addresses the problem of sample imbalance. Y. Wang et al. [70] proposed Pseudo-label Conditional Generative Adversarial Imputation Networks (PC - GAIN). PC - GAIN pre-trains on low-missing-rate data to learn category information, employs pseudo-labels to determine an auxiliary classifier, and integrates it into a GAN framework to enhance data imputation. X. Miao et al. [71] proposed VGAIN, an imputation model that integrates the concept of Variational Autoencoders (VAEs) into the GAN framework. By incorporating VAE's latent variable regularization and reconstruction loss into the generator, VGAIN enhances the robustness of representation learning, effectively preventing mode collapse and improving the quality of missing data imputation. Diffusion model-based imputation methods, exemplified by TabCSDI [72], reconstruct missing values through a forward noise addition process followed by a reverse denoising process. This approach enables a more flexible and smooth approximation of the data distribution.

2.2 数据填补方法

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，对样本中的缺失数据元素进行生成填补成为一个重要的研究方向。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。

2.3 Vertical federated data generation and data imputation methods

Federated Learning (FL), first proposed by B. McMahan et al. [1], is a distributed machine learning framework that enables multiple participants to collaboratively train a global model while preserving data privacy and security, without sharing raw data. Depending on the application scenario, FL can be categorized into different types, such as horizontal federated learning and vertical federated learning. Among them, Vertical Federated Learning (VFL) [19][20] is suited for scenarios where different participants possess data with the same sample space but different feature attributes. By leveraging distributed computing and privacy-preserving techniques, VFL provides an effective solution for feature-level collaborative model training across organizations.

In the application scenario of vertical federated learning, when the size of the joint samples in the training dataset is insufficient, data augmentation can be achieved through sample generation. Currently, there are several sample generation methods based on Vertical Federated Learning (VFL), mainly including GAN-based methods and Markov Random Field (MRF)-based methods. J. Zhang et al. [73] proposed FedDA, a vertical federated learning data augmentation method based on the Generative Adversarial Network. X. Jiang et al. [78] proposed VertiGAN, which utilizes a GAN model consisting of a multi-output global generator and multiple local discriminators to generate a high-utility synthetic integrated dataset. X. Yuan et al. [79] proposed VFLGAN, another vertical federated learning method based on generative adversarial networks. GTV [74] is a VFL framework specialized in generating high-fidelity synthetic tabular data. W. Lin et al. [75] proposed VertiMRF for generating synthetic data in the vertical federated learning setting and providing differential privacy protection for all shared information.

In practical applications, when some data elements are missing in samples from multiple local datasets, it becomes challenging to provide high-quality samples for training models in vertical federated learning tasks (e.g., federated classification, prediction, etc.). The data from different parties typically show some associations. Therefore, vertical federated data imputation models are essential for imputing these missing data elements. W. Du et al. [77] stated in their paper that research on missing-data imputation in vertical federated application scenarios is still relatively unexplored. They also proposed a privacy-protected vertical federated KNN feature imputation method. However, the KNN computation method could not meet the requirements for high-quality sample data imputation.

In summary, there are many excellent data generation and imputation methods. However, data generation and imputation methods in vertical federated scenarios are uncommon. Although existing research has demonstrated the feasibility of generating data in Vertical Federated Learning using models such as Generative Adversarial Networks , there are still numerous challenges in various practical application scenarios. For example, the challenging problem introduced in the introduction section of this paper.

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[1]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[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。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。

**3 Participants sample generation method based on association rules and data imputation under VFL**

This paper proposes a participants sample generation method based on association rules and data imputation under vertical federated learning, abbreviated as FedPSG-AR, which integrates with attribute correlations, association rules, and imputation techniques within the VFL framework. The core idea of our method is to generate highly correlated attributes with other parties for participants with missing samples while ensuring data privacy. Subsequently, a federated imputation based on GANs is then employed to generate the remaining attributes of these missing samples. The FedPSG method primarily consists of two stages. The first stage involves an attribute generation method based on vertical federated association rules. As shown in Figure 2(a), it includes the following three processes implemented under vertical federated learning: Computing multi-party attribute correlations; Establishing relationship of attribute values; Generating attribute values by vertical federated association rules. First, based on the aligned sample set from each party, attribute correlations between two parties are computed to construct a multi-party attribute correlation matrix. Next, a strongly correlated attribute pair is identified from the matrix, establishing correspondences between all values in the respective attribute columns. Finally, based on these correspondences, association rules are established between the values of the two attribute columns, to generate correlated attributes of the missing samples for the participant. The second stage involves a vertical federated imputation based on GANs, as shown in Fig.2(b). Utilizing the attribute values generated by vertical federated association rules and combining sample data from multiple participants, the vertical federated imputation method is used to generate the remaining attribute values of these missing samples.

**3 基于纵向联邦关联规则和数据填补的参与方样本生成方法**

本文提出了一种基于纵向联邦关联规则生成和数据填补的参与方样本生成方法（FedPSG-AR），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相关属性值，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。如图2所示，FedPSG方法主要包括两个阶段。第一阶段是基于纵向联邦的关联规则生成方法，如图2(a)所示，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，为该属性对涉及的属性列中的所有值建立对应关系。最后，基于两个属性列间的对应关系，建立关联规则，为样本缺少的参与方生成高相关属性值。第二阶段是基于GANs的纵向联邦填补方法，如图2(b)所示。利用纵向联邦关联规则方法生成的属性值，并结合其他参与方的样本信息，纵向联邦填补方法将为参与方生成这些缺少样本的剩余属性值。

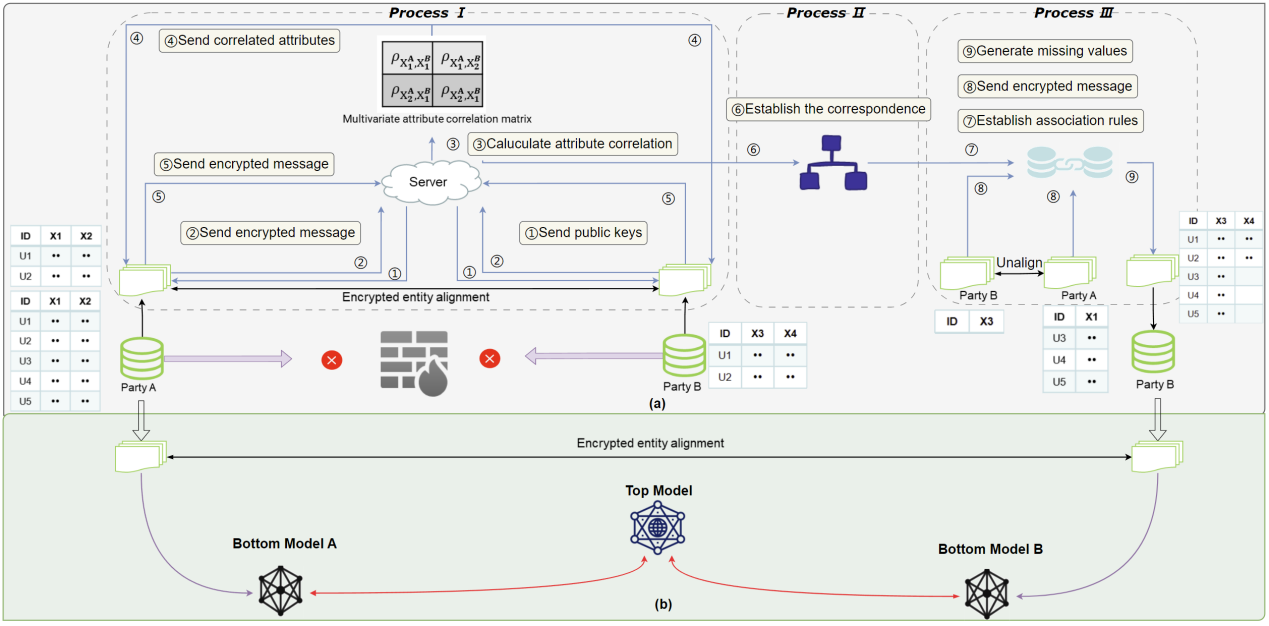


图2：参与方对齐样本生成方法架构图：（a）基于纵向联邦的关联规则生成方法；（b）基于纵向联邦的填补方法

Fig.2：Overall Schematic of Participants Sample Generation Method proposed in this paper：(a) Association rule generation based on vertical federated; (b) Vertical federated imputation based on GANs

3.1 Data Preprocessing and Data Description

Suppose there are N data owners (N participants) and a central server in a given vertical federated learning scenario. To clearly illustrate the methodology of this paper, we take the example of two participants: Party A and Party B, with their trusted collaborator C serving as the central server. Party A and Party B possess sensitive data and are required to safeguard data privacy during their collaborative model-training process. Assume that the label column of the joint samples is held by Party B.

We apply various preprocessing techniques to the data held by Parties A and B, including data cleaning, normalization, and feature encoding. Categorical features are handled using one-hot encoding, while numerical features are normalized via standard scaling. After preprocessing, the two parties securely align their samples based on ID spaces. To protect data privacy during this alignment, we employ the Blind RSA-based Private Set Intersection (PSI) protocol [ ]. This protocol allows both parties to securely compute the intersection of their datasets without revealing any additional information about the samples they hold. Throughout the vertical federated learning process in the proposed method, fully homomorphic encryption (Cheon-Kim-Kim-Song, CKKS [ ]) is employed for secure data computation.

Suppose the sample set in Party A as , , , where is the -th sample of Party A, is the -th attribute of the-th sample, , a. denotes the number of samples in Party A, and *a* denotes the number of attributes in Party A's samples. In Party B, the sample set is represented as ， , , where is the -th sample of Party B, is the -th attribute of the -th sample, , . denotes the number of samples in Party B, and *b* denotes the number of attributes in Party B's samples. Parties A and B align their datasets, and , based on their sample IDs through encrypted matching, resulting in aligned samples. The - samples in Party A cannot find corresponding aligned samples in Party B. The - samples in Party B cannot find corresponding aligned samples in Party A. In Party A, the aligned sample set is represented as , and the unaligned sample set is . In Party B, the aligned sample set is represented as , and the unaligned sample set is . As illustrated in Fig. 1, the number of samples in equals that in , and they have identical sample IDs.

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。假设联合样本的标签列在B方。

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。的样本数量和是相同的，且它们的样本ID是一致的。

3.2 Attribute generation method based on vertical federated association rules

Attribute generation method based on vertical federated association rules (i.e., an attribute generation method through VF-AR) utilizes the correlations between the attributes of the participant with missing samples and those of other parties. The attributes of these missing samples are generated using association rules derived from these correlations within the vertical federation framework.

(1) Computing multi-party attribute correlations

The specific process for calculating the correlation coefficient between the m-th attribute of Party A and the n-th attribute of Party B is as follows:

Party C, as the coordinator, generates a key pair and sends the public key to Parties A and B to be used in the entire multi - party association rule generation method. For the *m*-th attribute columns of the samples from Party A, iterate over each attribute value in , and assign ranking to them, where . The rankings start from 1. For values with the same ranking, the average ranking is used (that is, the rankings of the same values are averaged). The ranking of each attribute value is obtained, encrypted, and then sent to Party B. For the n-th attribute columns of the samples from Party B, iterate over each attribute value in , and assign ranking to them, where . The ranking of each attribute value is obtained, encrypted. The difference between encrypted and is calculated as follows:

The Spearman's association coefficient[ ], , is then calculated based on the differences , using the following formula:

where, num is the number of aligned samples, i.e. .

For all attributes from Party A and all attributes from Party B, the association coefficient between each attribute pair can be calculated using Equation (2), constructing the multi-party correlation coefficient matrix , which is represented as follows:

In particular, when there are more than two parties involved in the multi-party collaborative learning process, the correlation coefficient matrix can be computed between all attribute columns of participants with unaligned samples and those of participants with missing samples.

3.2 纵向联邦关联规则生成方法

纵向联邦关联规则生成方法利用样本缺少参与方与其他参与方之间属性的相关性，通过关联规则为缺少样本的属性赋值。在数据安全保护下，该方法的实现过程包括：计算A、B两方属性之间的相关性，建立属性对应关系、基于联邦关联规则的属性值生成。

（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。

1. Establishing relationship of attribute values

Based on the multi-party correlation coefficient matrix, this process first identifies attribute columns with high correlations between the parties in the aligned sample dataset, and then establishes attribute pairs that show strong correlations between Parties A and B. For each attribute pair, a correspondence relationship is established for all of their values. The detailed procedure is as follows:

1. Iterate through each value in the correlation coefficient matrix M to identify the largest correlation coefficient and determine the attribute pairs with strong correlations ).
2. Party B initializes an empty table to store the correspondence relationship between each value from and the values from .
3. Loop through all the values in and corresponding to the attribute pair ).

Assume that the attribute columns contain a total of *s* distinct values, represented as , where is the -th value. Similarly, attribute columns contains a total of distinct values, represented as , where is the -th value. Party A and Party B respectively look up the values and of the attributes and for the i-th sample in and , and then encrypt them to obtain and . Parties A sent the encrypted value to Party B, and Party B records the attribute value corresponding to when takes the value of .

Parties A and B then process the values from attributes and ​in a loop traversal manner, respectively. Each party encrypts their corresponding attribute values and using the public key, and sends the encrypted data to coordinator C. Coordinator C records that when takes the value , corresponds to the attribute with value , and stores these relationships in the table. The correspondence between and is represented as ().

Traverse and search for the correspondence between the attribute values of all samples in and , and store these correspondences in the relationship table .

④ Resets to 0.

Repeat this process from step ① to step ④ until the total number of attributes in the participant with miss samples that are highly correlated with those of other parties meets the specified requirement. The number of selected correlated attributes also represents the degree of correlation between the attributes of Party A and Party B to be determined in the experiment. These attributes are need to generate by vertical federation association rule.

（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

1. Generating attribute values by vertical federated association rule

According to the correspondence table established based on attribute correlations, association rules are utilized to generate these attributes of missing samples for the participants. Association rules[ ] are used to discover hidden relationships between data items and are widely applied in fields such as data mining and market basket analysis. They identify strong associations between items by analyzing their co-occurrence patterns. For example, in retail analysis, customers who buy product A are likely to also purchase product B. Association rules are typically expressed in an 'if-then' format. During the process of for participants sample generation, association rules assist in establishing value mappings between two attribute columns, which are then used to infer and generate relevant data.

When Party A's samples are missing:

Step 1: Party B traverses all attribute values in and compute the sample counts for each attribute value , denoted as , where . Simultaneously, Party B send the correspondence table to Party A. Party A traverses all attributes values in . Based on between and , Party A records the sample counts for the attribute values in that correspond to the value and send it to Party B.

Step 2: Based on the received data, Party B calculates the ratio , where corresponds to

Party B create an association rule for attribute columns and , denoted as follows:

where, represents the probability that a sample with attribute value in also has attribute value in , denoted as , and .

Step 3: The sample value in the attribute column of Party A is determined by the association rule corresponding to the attribute value of in Party B for that sample. In Party B, when the attribute value of is , Party B looks up , and selects the value corresponding to the largest from , then sends it to Party A. Party A initializes a random mask , computes , and sends it to Party C. Party C decrypts to obtain +, and sends it back to Party A. Party A removes the random mask to obtain and assigns it to , where, j.

When Party B's samples are missing:

Step 1: Party A traverses all attribute values in and compute the sample counts for each attribute value , denoted as , where . Simultaneously, Party B traverses all attributes values in . Based on between and , Party B records the sample counts for the attribute values in that correspond to the value and send it to Party A.

Step 2: Based on the received data, Party A calculates the ratio , where corresponds to

Party A create an association rule for attribute columns and , denoted as follows:

where, represents the probability that a sample with attribute value in also has attribute value in , denoted as , and .

Step 3: The sample value in the attribute column of Party B is determined by the association rule corresponding to the attribute value of in Party A for that sample. In Party A, when the attribute value of is , Party A looks up , and selects the value corresponding to the largest from , then sends it to Party B. Party B initializes a random mask , computes , and sends it to Party C. Party C decrypts to obtain +, and sends it back to Party B. Party B removes the random mask to obtain and assigns it to , where, j.

（3）基于纵向联邦关联规则的属性值生成

根据属性相关性建立的关系表矩阵R，纵向联邦关联规则可以为样本缺少的参与方生成这些高相关的属性值。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。

3.3 Vertical federated imputation based on GANs

After the correlated attributes with other parties are generated for the missing samples, we also need to generate the remaining attributes in order to obtain the complete sample information. This paper propose a Vertical federated imputation framework based on Generative Adversarial Networks (GANs) to impute the remaining attributes of the missing samples. This framework will use GANs to perform collaborative imputation with multi-party data, while ensuring data security and privacy protection. This vertical federated imputation framework includes the bottom models of each party and the top model, as shown in Fig. 3.

## 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高相关属性值后，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。

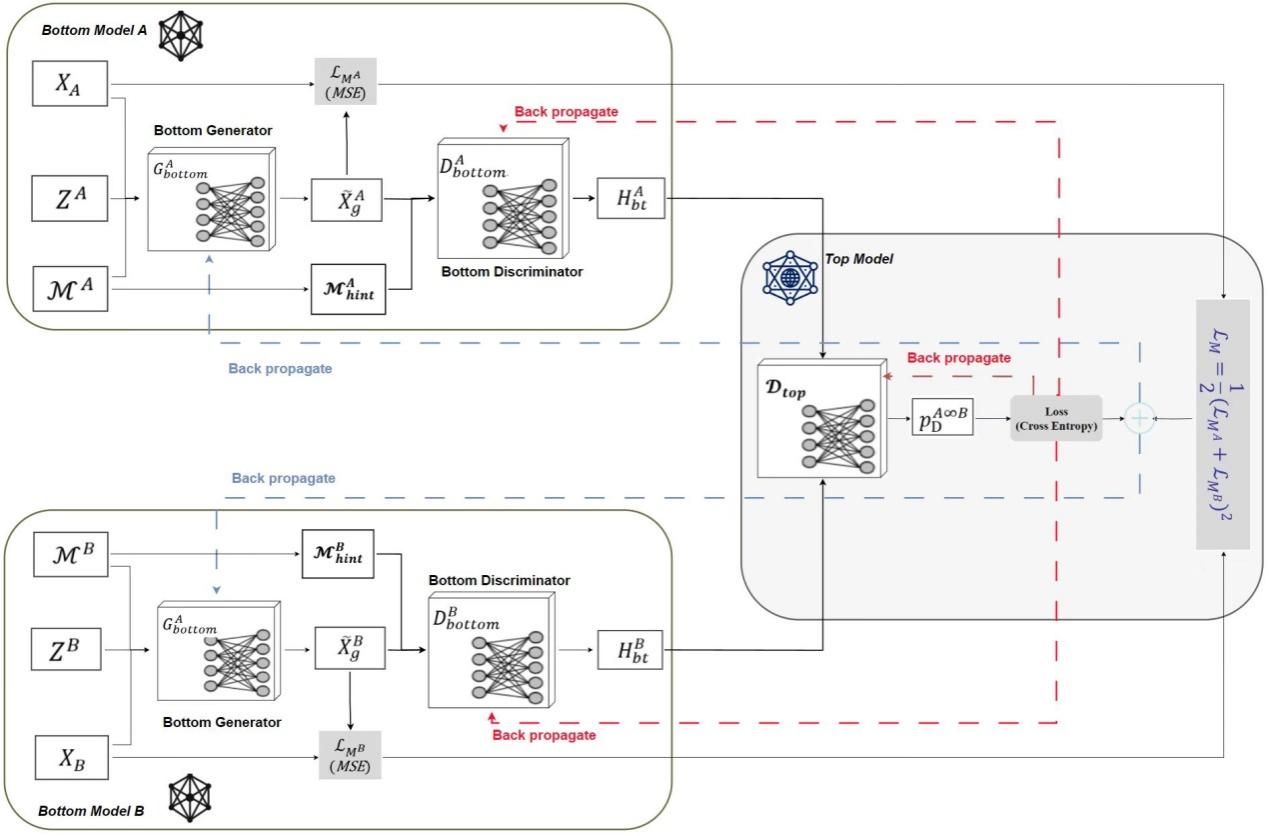


Fig.3：The framework of vertical federated imputation Based on GANs

1. Mask vector

The mask vector plays a crucial role in the data imputation process. It indicates the locations of the remaining attributes in the sample that need to be imputed. During the imputation process, the generator from each party uses this information to impute the remaining attributes values based on the observed data from the parties in the joint sample.

Set the encoding vectors of the feature attributes as for Party A's sample and as  for Party B's sample. In this paper, the federated imputation models based on GANs normalize numerical variables using Min-Max normalization and encode categorical variables with One-Hot encoding. Set the mask vector of Party A as ，where, ， represents the dimensionality of the sample vector of Party A. And the mask vector of Party B as ，where ， represents the dimensionality of the sample vector of Party B.



The remaining sample data to be imputed for Party A and Party B is referred to as the missing data from Party A and Party B, and is represented by the vectors , , respectively:

， （4）

that is,

，（5）

Where, denotes element-wise multiplication between vectors.

1. Bottom models

In the Bottom model, we design separate generators and discriminators for each party (e.g., Party A and Party B, as shown in Fig. 3). The main task of each local generator is to generate reasonable data for the missing values in participants with missing samples, based on federated encrypted collaborative learning from each party's local data. The discriminator, on the other hand, is responsible for determining whether the generated data is close to the real data, providing feedback to optimize the generator. It should be noted that we mention a variety of vertical federation imputation models based on GANs including VF-GAIN, VF-CGAIN and VF-VGAIN, in the experiment in Section 4. These vertical federated imputation models are similar in overall architecture, and the differences are the network structure of their respective bottom models. In other words, the bottom models of each party in VF-GAIN, VF-CGAIN and VF-VGAIN have the same structure as that of GAIN[ ] CGAIN[ ],VGAIN[ ].

In the federated learning process, the inputs of Party A's bottom generator include: , and random noise vector . And, the inputs of Party B's bottom generator include: , and random noise vector . The outputs of the bottom generators of Parties A and B are:

（6）

The joint sample , comprising the sample data generated by Party A and the sample data generated by Party B, is represented as follows:

（7）

where, because the data of both parties are subject to data security and privacy requirements, denotes the concatenation under vertical federated learning, which is not a direct concatenation of the data from Parties A and B. It is only for the convenience of representing the joint sample.

The imputed samples , are obtained from the generated vectors of Party A, of Party B, and the observed data from the original missing sample vectors , , respectively:

， （8）

Under such circumstances, the imputation operation can be seen as being performed based on joint samples consisting of Parties A and B. The imputed joint sample is denoted as , which consists of data from party A and data from party B, as shown in Equation (9):

（9）

in generated from Party A and in generated from Party B will apply in calculating the generator loss function in the federated learning process.

In this paper, to focus the discriminator on evaluating the imputation quality of the missing data and ensure a more accurate distinction between observed and imputed data, the hint mask of GAIN is introduced to assist in training the model [add citation]. The hint mask depends on . We define a hyperparameter as the probability of randomly sampling from , such that the elements of the mask vector with a value of 1 are set to 0 with probability . In this paper, the hyperparameter is set to 0.9, which is the value referenced from the GAIN model [add citation].

The bottom network of discriminator uses a fully connected neural network structure, which primarily consists of two hidden layers. Each hidden layer employs the LeakyReLU activation function, and dropout is applied to randomly ignore some neurons with a certain probability to reduce model overfitting. In this paper, the negative slope of the LeakyReLU is set to 0.2, and the dropout rate is set to 0.5. The output of the bottom discriminator in Party A is a vector obtained by and through its network, and the output of the bottom discriminator in Party B is a vector obtained by and through its network:

（10）

1. Top model

In the vertical federated imputation framework, the top model is primarily utilized to coordinate the data imputation process between Party A and Party B. Based on the bottom discriminators of Parties A and B, the quality of the generated and imputed data is further evaluated and optimized by a top discriminator. The top discriminator uses a fully connected network to concatenate the output hidden vectors from the bottom discriminators of Parties A and B. It contains two hidden layers, with each layer using the LeakyReLU activation function. In this paper, the negative slope of the LeakyReLU function is set to 0.2, and the Dropout ratio is set to 0.5. The vectors and output from the bottom discriminators of Parties A and B are input into the top discriminator to obtain its output probability vector:

(11)

where,

denotes the concatenation of vectors and under vertical federated learning, obtained from the bottom discriminators of Parties A and B. denotes the concatenation of hint mask vectors and under vertical federated learning from Parties A and B. The output is the probability vector of the discriminator distinguish the imputed and real data of elements in the joint samples.

1. Loss function

The loss function of the whole vertical federated imputation consists of both generator loss and discriminator loss. The optimization process follows the adversarial training mechanism of Generative Adversarial Network. Specifically, the generator aims to improve the quality and realism of the generated data by minimizing the difference between the generated data and the real data. Meanwhile, the discriminator enhances its ability to distinguish real data from generated data by maximizing its classification accuracy. This training objective can be formally expressed as:

（12）

Firstly, we fix the generator and train the discriminator. The minimum batch size is selected in each round from the joint sample set . Each sample in the minimum batch includes and in Party A, and in Party B.

The total loss function of the discriminator is:

（13）

， is the j-th element in the hint mask vector for the i-th sample, and is the j-th element in for the i-th sample. It should be noted that, for VF - CGAIN, the category labels need to be added because its bottom model is consistent with that of CGAIN; i.e., , where, denotes the label vector of the i-th sample.

Subsequently, we fix the discriminator and train the generator. The minimum batch size is selected in each round from the joint sample set . Each sample in the minimum batch includes , and in Party A, , and in Party B.

The total loss function of the generator is:

（14）

Where,

The loss function in Party A is：

The loss function in Party B is：

Where, i refers to the i-th sample in the joint sample set of the batch. In party A, *j* is the j-th element in the i-th sample of Party A, is the observed data of Party A, is the generated data of Party A, and and represent the j-th elements in the i-th samples and , respectively. In party B, *j* is the j-th element in the i-th sample of Party B, is the observed data of Party B, is the generated data of Party B, and and represent the j-th elements in the i-th samples and , respectively. The parameter in the loss functions of both Parties A and B is a hyperparameter. In the experiments presented in this paper, the value of is chosen based on the hyperparameter setting used in the GAIN method.

（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

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。

1. Training process

According to the assumption, the label column of the joint samples is held by Party B. Therefore, the top model resides in Party B. The training process of vertical federated imputation dased on GANs is as follows.

Step 1: Party A initializes the bottom generator's weight parameters and the bottom discriminator's weight parameters . Party B initializes the bottom generator's weight parameters , the bottom discriminator's weight parameters , and the top discriminator's weight parameters . The coordinator, Party C, creates an encryption key pair and distributes the public key to Parties A and B. Homomorphic encryption enables secure information exchange without revealing raw data or private information. represents the fully homomorphic encryption, CKKS. Input , and into Party A's bottom generator network, and input , and into Party B's bottom generator network.

Step 2: Party A performs forward propagation of the input information through the bottom generator network to obtain and , and computes its corresponding loss . Party B performs forward propagation of the input information through the bottom generator network to obtain and , and computes its corresponding loss . Party A performs forward propagation of and through the bottom discriminator network to obtain . Party B performs forward propagation of and through the bottom discriminator network to obtain . Party A encrypts and each element value of , obtaining and the encrypted element vector , and sends , and to Party B. After receiving , and from Party A, Party B combines and into an encrypted element vector , combines and into , and performs forward propagation of , through the top discriminator network, and computes the total loss of the discriminator . Simultaneously, Party B computes the generator loss , computes from and the received , and the total generator loss from and . Party B sends the total discriminator loss and the total generator loss to Party C.

Step 3: Party A and Party B initialize encrypted random masks , , respectively. These masks ensure that intermediate results cannot be reverse-engineered during transmission, thereby guaranteeing data privacy.

The back-propagation process of the discriminator training: Party B computes and +, and uses and the received to split the weight gradient of the first layer of Party B's top-level discriminator network into and , enabling Party A and Party B to compute and . Party B uses to compute +. Then, Party B sends + and + to Party C, and sends to Party A. Party A computes +, and sends it to Party C.

The back-propagation process of the generator training: Party B computes +, and sends it to Party C. Party A computes +, and sends it to Party C.

Party C decrypts , , +, +, +, + and +, sends +, + to Party A, and sends +, + and + to Party B.

Step 4: Party A and Party B remove the encrypted random masks from the received gradient information. Based on these gradients, Party A and Party B update the weights parameters , , and of their respective bottom generator and discriminator networks, and update the weights parameters of the top discriminator network.

Step 5: Repeat Steps 2 to 4 iteratively until the training is complete.

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。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。

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。‘Bank’ and ‘Credit’数据集代表了样本规模相对较大和较小的两种情况。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)，即当=∅的情况进行了充分的实验论证。图1(b)，即当≠∅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，MisR-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 Experiments**

4.1 Datasets and Data Preparation

In our experiments, we utilized four datasets to evaluate the proposed method: the Bank Marketing Dataset, the German Credit Dataset, the Letter Recognition Dataset and Online News Popularity Dataset:

①The Bank Marketing Dataset [reference] pertains to the direct marketing campaigns of a Portuguese banking institution, which contains 45,211 examples and 16 attribute features, along with an ID column and a label column. It aims to classify whether a client will subscribe to a term deposit.

②The German Credit Dataset [reference] originates from a credit scoring system, and its objective is to classify applicants as having either "good" or "bad" credit. The dataset includes 1000 samples, and each sample includes 21 attributes covering financial, demographic, and social characteristics: 13 categorical attribute features, 7 numerical attribute features, and a label column.

③The Letter Recognition dataset [ ] is used for character recognition tasks and contains 20,000 instances and 16 attribute features. Each instance includes statistical features of handwritten letter images, such as size, shape, and contour. This dataset aims to classify each sample into one of the 26 English letters.

④The Online News Popularity dataset [ ] is used to predict the popularity of online news articles and contains 39,644 instances and 60 attribute features. The dataset includes information such as article content, social media interactions, and publishing time. For simplicity, we refer to the four datasets as 'Bank' , 'Credit', ’Letter’ and ‘News’, respectively.

In the experiments on the 'Bank' and 'Credit' datasets, the attributes are vertically partitioned between Parties A and B based on feature ownership. Specifically, Party A contains customer information, while Party B contains bank information. In the Bank dataset, Party A has 8 attributes while Party B has 8 attributes excluding the ID column excluding the ID column. In the Credit dataset, Party A has 9 attributes and Party B has 11 attributes excluding the ID column. 'Bank' and 'Credit' datasets represent two scenarios: one with a relatively large sample size and the other with a relatively small sample size. 'Letter' and 'News' datasets are used to simulate a vertical - federated learning scenario. And, we divide the original attribute columns equally between Party A and Party B according to their order. Excluding the financial domain, ‘Letter’ and ‘News’ datasets represent different application scenarios as well as different sample sizes and feature numbers.

To verify the validity of the proposed method, we thoroughly conducted sufficient experimental demonstrations for the case where =∅ in Fig. 1. The case when ≠∅ can be inferred in the same way. In the experiments of this paper, we set different missing sample ratios for Party B when =∅. For example, a missing sample ratio of 0.2 indicates that 80% of the samples in Party B can be aligned with those in Party A, while 20% of the samples are missing in Party B, relative to Party A. In the experiments, Cnum represents the number of attributes generated by VF-AR in Party B. And, we denote the missing sample ratio of Party B relative to Party A as MisR-B.

## 4.2 实验一的设计与结果分析

## 4.2 Experiment 1：Different settings in FedPSG-AR

FedPSG-AR first generates partial attributes for the missing samples of Party B, then imputes the remaining attributes of these missing samples. Initially generated partial attributes have strong correlations with the attributes of Party A. In Experiment 1, we will evaluate the effectiveness of generating complete data for the missing samples in Party B by using different partial attribute generation approaches or various remaining attribute imputation methods to determine the optimal setting for FedPSG-AR. This experiment consists of three groups, each with a different missing rate (MisR-B) for Party B's samples. The three groups of experiments are conducted on the two datasets, 'Bank' and 'Credit'. The evaluation metric employed to assess the effectiveness of sample generation is RMSE. RMSE [ ] is used to indicate the average degree of deviation between the data in the generated sample and the real data. In Experiment 1, for the federated or non-federated imputation models based on GANs and the generation models involved in the comparative experiment, the total number of training iterations is set to 10,000, the number of epochs is set to 10, the learning rate is set to 0.001, and the optimizer is set to Adam.

FedPSG-AR先为B方缺失样本生成部分属性，再填补这些缺失样本的剩余属性。先生成的部分属性均与A方属性高相关。我们将验证采用不同的部份属性生成方式，或不同的剩余属性填补方法，获得B方缺失样本完整数据的效果，以确定FedPSG-AR的最佳设定。实验一共设置了三组实验，均设置了不同的B方样本缺失率MisR-B。实验所采用的数据集为Bank和Credit. 本实验用于评估样本生成效果的评估指标均为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。实验一中，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。

(1) The experimental design of Experiment 1

Group 1 in Experiment 1: We fix remaining attribute imputation methods, and compare different partial attribute generation approaches. In this group of experiment, there are two approaches used to generate the attributes in the missing samples of Party B that are highly correlated with those of Party A: ①The local generation model learns the data distribution of these attributes in the observed samples of Party B to generate attribute values for the missing samples of Party B. We refer to this as AG. Its complete process of sample generation is referred to as FedPSG-AG. ② Combining the observed samples from Party A, these attributes of the missing samples in Party B are generated by vertical federated association rule, which is VF-AR in our proposed method FedPSG-AR. In FedPSG-AG, TabDDPM [ ] is used as its local generation model AG, which is a state-of-the-art model in the data generation field. And both the remaining attribute imputation methods in FedPSG-AG and FedPSG-AR use VF-GAIN. This group of experiment presents a comparison of experimental results for different values of Cnum. The number of attributes represented by Cnum also reflects different degrees of attribute correlation.

Group 2 in Experiment 1: We fix partial attribute generation approaches, and compare different remaining attribute imputation methods. In this group of experiment, the imputation models used for comparison include Mean[ ], MissFI (The number of trees in the Random Forest is set to 100)[ ], MICE[ ], GAIN[ ], CGAIN[ ], VGAIN[ ]. In this experiment, they are deployed and implemented within a vertical federated framework, which are represented as VF-MissFI, VF-MICE, VF-GAIN, VF-CGAIN, VF-VGAIN. Mean imputation fills the missing values by calculating the mean of the current column, and this process does not require any vertical federated computation. This group of experiment adopts the optimal results from Group 1, i.e., the partial attribute generation approach adopts VF-AR, with Cnum set to 5.

Group 3 in Experiment 1: We fix partial attribute generation approaches, and compare different remaining attribute imputation approaches. In this group of experiment, the imputation approaches adopted include multi-party vertical federated imputation and local non-federated imputation. Multi-party vertical federated imputation: Combing the observed samples from Party A and Party B, and the partial attributes generated by VF-AR in the missing samples of Party B, the remaining attributes of Party B are imputed through the implementation of vertical federated imputation methods. Local non-federated imputation: Combing the observed samples from Party B and the partial attributes generated by VF-AR in the missing samples of Party B, the remaining attributes of Party B are imputed through the implementation of local imputation methods. The imputation methods adopt the federated and non-federated implementation of GAIN, CGAIN, VGAIN, which are the outstanding models in Group 2. Cnum is set to 1, 5, and 8.

（1）实验一的实验设计

第一组实验：固定剩余属性填补方法，对比不同的部分属性生成方式。本组实验中，用于为B方缺失样本生成与A方高相关属性的方式有两种：①采用本地生成模型学习B方观测样本中这些属性的数据分布，为B方缺失样本生成这些属性值，我们称之为AG。样本完整生成过程称之为FedPSG-AG。② 联合A方观测样本，基于纵向联邦关联规则生成B方缺失样本的这些属性值，即本文方法FedPSG-AR中的VF-AR。在FedPSG-AG中采用的本地生成模型为目前表现SOTA的TabDDPM。而FedPSG-AG和FedPSG-AR中的剩余属性填补模型都为VF-GAIN。本组实验展示了不同Cnum的实验结果对比。Cnum所表示的属性个数，也代表了A、B方不同的属性相关性程度。

第二组实验：固定部份属性生成方式，对比不同的剩余属性填补模型。本组实验中，用于对比的填补模型包括：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，它们将在纵向联邦框架下进行实现，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。本组实验采用了第一组实验中的最佳结果，即部份属性生成方式采用VF-AR，Cnum设置为5.

第三组实验：固定部份属性生成方式，对比不同的剩余属性填补方式。本组实验中，采用的填补方式包括多方纵向联邦填补和仅单方非联邦填补。联邦填补：基于A方、B方观测样本和使用VF-AR生成的B方缺失样本的部分属性，对B方剩余属性进行纵向联邦填补。非联邦填补：基于B方观测样本和使用VF-AR生成获得的B方缺失样本的部分属性，对B方剩余属性进行本地填补。填补模型采用了第二组实验结果中较优的VF-GAIN、VF-CGAIN、VF-VGAIN的联邦和非联邦实现。Cnum设置为1,5,8.

(2) Results and Analysis of Experiment 1

Table 1, Table 2 and Fig. 4 demonstrate the experimental results for Group 1 in Experimental 1. In Table 1, Table 2, the generation effects of FedPSG-AG and FedPSG-AR are compared, with the results showing lower RMSE values highlighted in bold. Meanwhile, the optimal and suboptimal results are highlighted in red and blue, respectively. When Cnum=1, one attribute column of the missing samples in Party B is generated by AG or VF-AR, and the remaining attributes are generated by VF-GAIN; and so forth…. As shown in Table 1 and Table 2, from the bolded experimental results, it can be observed that under different B values, for each set of Cnum, comparing different RMSE values, FedPSG-AR generally outperforms FedPSG-AG. This indicates that when remaining attribute imputation methods is fixed, VF-AR is superior to AG in generating the attributes in the missing samples of Party B that are highly correlated with those of Party A. Specifically, for AG, even when the local generation model of Party B employs TabDDPM, which achieves the best performance in the current literature, its experimental results still do not surpass those of vertical federated association rule. Because, for these attributes of Party B that are highly correlated with Party A, VF-AR can learn the data distribution rules better than the local generation models by leveraging data from Party A. From the experimental results highlighted in red and blue in Tables 1 and Table 2, FedPSG-AR demonstrates a lower RMSE value when Cnum=4, 5, 6. According to the attribute correlation between Party A and Party B, vertical federated association rule generates more highly correlated attributes, then perform federated imputation for the remaining attributes, resulting in a lower RMSE for the generated samples. It contributes to obtaining a sample set that more closely approximates the real one. However, it is worth noting that the RMSE will also increase as the Cnum of Party B increases as shown in Fig. 4. These results indicate that a larger number of attributes in Party B generated by VF-AR is not necessarily better. In fact, some attributes that are non-highly correlated with those of Party A are more suitable for generation through federated imputation. Because attributes with low correlation to Party A may introduce errors in the learning and inference process of VF-AR, thereby reducing the quality of the generated data. Federated imputation can fully leverage the data collaboration among multiple participants to improve the accuracy of generated data. Consequently, generating partial highly correlated attributes and combining federated imputation to impute the remaining attributes, can help ensure the quality of the generated samples. The value of Cnum needs to be determined based on the specific dataset. Generally, when the correlations between attributes of Party A and Party B are not less than 0.5, the number of selected attributes is sufficient to meet the requirements for participants sample generation.

（2）实验一的结果分析

**表1** A、B方不同的属性相关性阈值的生成效果by FedPSG（RMSE**，on Bank Dataset**）

Table 1 RMSE obtained by FedPSG-AG/AR with different Cum of Party B for Bank Dataset (The optimal and suboptimal results are highlighted in red and blue, respectively.)

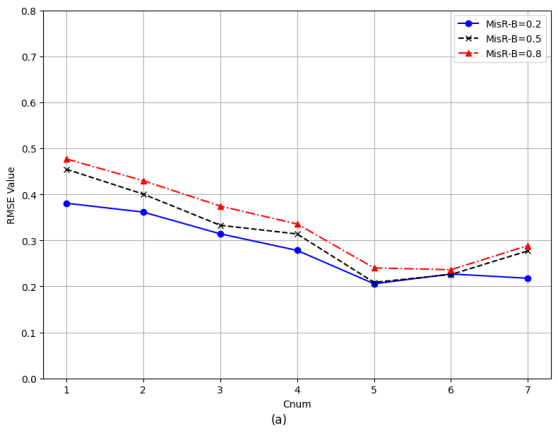
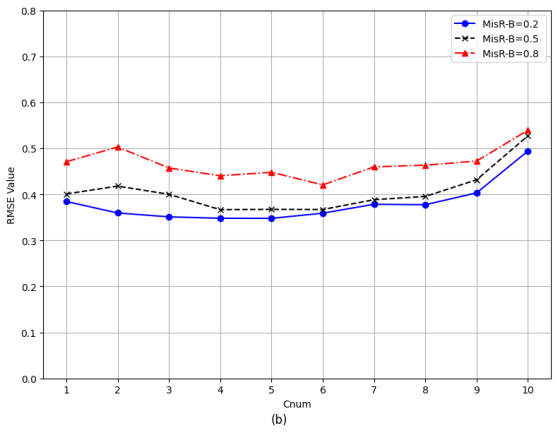
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cnum  MisR-B | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.2 | FedPSG-AG | 0.4615 | 0.434 | 0.4273 | 0.4048 | 0.3067 | 0.2820 | 0.3055 | 0.3529 | 0.4226 |
| FedPSG-AR | **0.4615** | **0.3807** | **0.3614** | **0.3142** | **0.2782** | **0.2057** | **0.2369** | **0.2177** | **0.2536** |
| 0.5 | FedPSG-AG | 0.4981 | 0.4826 | 0.4517 | 0.4444 | 0.3585 | 0.3119 | 0.3451 | 0.4585 | 0.4708 |
| FedPSG-AR | **0.4981** | **0.4545** | **0.4005** | **0.3328** | **0.3139** | **0.2087** | **0.2258** | **0.2771** | **0.3029** |
| 0.8 | FedPSG-AG | 0.519 | 0.5007 | 0.4536 | 0.4521 | 0.4273 | 0.4217 | 0.4464 | 0.474 | 0.5301 |
| FedPSG-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 FedPSG（RMSE**，on Credit Dataset**）

Table 2 RMSE obtained by FedPSG-AG/AR with different Cum of Party B for Credit Dataset (The optimal and suboptimal results are highlighted in red and blue, respectively.)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cnum  MisR-B | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 0.2 | FedPSG-AG | 0.5048 | 0.4962 | 0.4959 | 0.5126 | 0.471 | 0.4637 | 0.4353 | 0.4218 | 0.3946 | 0.4074 | 0.4282 | 0.4462 |
| FedPSG-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 | FedPSG-AG | 0.5354 | 0.4985 | 0.5051 | 0.5373 | 0.4825 | 0.4745 | 0.4549 | 0.4403 | 0.4247 | 0.4771 | 0.4863 | 0.535 |
| FedPSG-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 | FedPSG-AG | 0.5518 | 0.508 | 0.5445 | 0.5859 | 0.5232 | 0.5085 | 0.4725 | 0.4726 | 0.4891 | 0.5122 | 0.5715 | 0.5699 |
| FedPSG-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** A、B方不同的属性相关性阈值的生成效果折线图by FedPSG: (a) Bank Dataset; (b) Credit Dataset

Fig. 4 RMSE line graph obtained by FedPSG-AR with different Cum of Party B: (a) Bank Dataset; (b) Credit Dataset

表1、表2和图4展示了第一组实验的实验结果。在表1、表2中，将FedPSG-AG和FedPSG-AR生成效果对比，RMSE值更低的实验结果用加粗字体进行表示。同时，将FedPSG-AR所得最优和次优的RMSE值用红色和蓝色字体表示。当Cnum=1时，表示用AG或者VF-AR生成B方缺少样本的1个属性列，剩余属性均采用VF-GAIN生成；以此类推...。如表1和图4所示，从加粗的实验结果可以看出，在不同的MisR-B下，对比不同Cnum的每一组RMSE，FedPSG-AR总体上来说优于FedPSG-AG。这表明：在填补模型固定的情况下，VF-AR用于生成B方缺失样本中与A方高相关的属性时，优于AG。特别地，对于AG，即使B方本地生成模型采用了目前文献表现最优的TabDDPM，其实验结果也无法超越联邦关联规则的实验结果。因为，对于与A方高相关的这些B方属性来说，VF-AR比本地生成模型能更好地学习数据分布规则。从表1 和表2的红色和蓝色字体表示的实验结果看，当Cnum为4、5、6时，FedPSG-AR的RMSE值表现较优。这说明：根据A方和B方的属性相关性，联邦关联规则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。但随着Cnum的增加，生成样本的RMSE也会有所增加，这说明并非采用联邦关联规则生成的属性列数越多越好，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，与A方相关性较低的属性，可能会导致关联规则在学习和推理过程产生误差，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作作用，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，再结合联邦填补生成剩余属性，能够在保证生成样本的质量。Cnum的取值需要根据具体的数据集来确定，一般来说，当A、B两方属性间的相关性大于0.5时，所选择出的属性的个数能满足样本生成的要求。

Table 3 demonstrates the experimental results for Group 2 in Experimental 1. When Cnum=5, under the different values of MisR-B, the optimal and suboptimal RMSE results are obtained by FedPSG-AR when the remaining attribute imputation method is VF-GAIN, VF-VGAIN, or VF-CGAIN. Particularly, the optimal RMSE values of Bank dataset occur when imputation model is VF-VGAIN or VF-CGAIN, and the optimal RMSE values of Credit dataset occur when imputation model is VF-VGAIN. The results demonstrate the effectiveness of vertical federated imputation models based on GANs. The worst RMSE values of the two datasets occur when imputation model is Mean. Because Mean imputation is a mean calculation on all the values of the current attribute column, and during the imputation process, it neither references nor combines the data from other parties.

表3展示了第二组实验的实验结果。当Cnum=5时，在不同的MisR-B下，FedPSG-AR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-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。这些结果表明了基于GANs的纵向联邦填补模型的有效性。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，得到的填补值误差较大。

**表3** 不同填补模型的生成效果by FedPSG（RMSE**，on Bank and Credit Dataset**）MisR-B=0.2, 0.5. 0.8

Table 3 RMSE obtained by FedPSG-AR with different imputation models for Bank and Credit Dataset (The optimal and suboptimal results are highlighted in red and blue, respectively. MisR-B=0.2, 0.5. 0.8)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Imputation  Dataset | | Mean | VF-MissFI | VF-MICE | VF-GAIN | VF-PC-GAIN | VF-VGAIN | VF-CGAIN |
| Bank | 0.2 | **0.4075** | **0.3333** | **0.3233** | **0.2057** | **0.2728** | **0.2291** | **0.1691** |
| 0.5 | **0.4578** | **0.3501** | **0.3743** | **0.2087** | **0.2890** | **0.2335** | **0.2332** |
| 0.8 | **0.4989** | **0.4006** | **0.3567** | **0.2401** | **0.3011** | **0.3024** | **0.2533** |
| Credit | 0.2 | **0.4567** | **0.4372** | **0.4471** | **0.3478** | **0.4083** | **0.3373** | **0.4029** |
| 0.5 | **0.4773** | **0.4359** | **0.4789** | **0.3674** | **0.4112** | **0.3503** | **0.4242** |
| 0.8 | **0.4910** | **0.4701** | **0.4982** | **0.4479** | **0.4373** | **0.4030** | **0.4433** |

Table 4 and Table 5 demonstrate the experimental results for Group 3 in Experimental 1. Firstly, Cum is set to 1, 5, 8, respectively. In table 4, the model adopted by federated imputation approach is VF-GAIN, and the model adopted by non-federated imputation approach is local GAIN in Party B. As shown in Table 4, for the remaining attributes in FedPSG-AR, federated imputation based on VF-GAIN overall outperform non-federated imputation based on local GAIN. Because federated imputation introduces more multi-party data for collaborative learning, the incorporation of additional data features improves the quality of the missing data generated for Party B in the joint samples. Meanwhile, under the different values of MisR-B, the experimental results obtained by federated imputation and non-federated imputation in FedPSG-AR when Cnum=5 are superior to the results obtained when Cnum=1 and Cnum=8, which is consistent with the conclusion of Group 1 in Experiment 1. Furthermore, we compare the federated imputation and non-federated imputation of GAIN, CGAIN, VGAIN when Cum = 5 in FedPSG-AR. As shown in Table 5, the federated imputations based on VF-GAIN, VF-CGAIN, VF-VGAIN overall outperform the non-federated imputations based on local GAIN, CGAIN, VGAIN, which is consistent with the conclusion of Table 4.

**表4 基于GAIN联邦和非联邦填补的生成效果by** FedPSG when num=1,5,8 （RMSE**，on Bank and Credit Dataset**）MisR-B=0.2, 0.5. 0.8

Table 4 After generating partial attributes, RMSE obtained by federated and non-federated imputation when Cnum=1,5,8 for Bank and Credit Dataset, MisR-B=0.2, 0.5. 0.8

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VFImp(Y/N)  Dataset | | Cnum=1 | | Cnum=5 | | | Cnum=8 | |
| VF-GAIN | GAIN | VF-GAIN | GAIN | GAIN | VF-GAIN | GAIN |
| Bank | 0.2 | **0.3807** | 0.4041 | **0.2057** | | 0.2779 | **0.2536** | 0.3468 |
| 0.5 | **0.4545** | 0.4744 | **0.2087** | | 0.295 | **0.3029** | 0.3437 |
| 0.8 | **0.4770** | 0.4951 | **0.2401** | | 0.314 | **0.2986** | 0.3577 |
| Credit | 0.2 | **0.3844** | 0.5317 | **0.3478** | | 0.4945 | **0.3774** | 0.5017 |
| 0.5 | **0.4009** | 0.5748 | **0.3674** | | 0.5044 | **0.3954** | 0.4456 |
| 0.8 | **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** FedPSG when num=5（RMSE**，on Bank and Credit Dataset**）MisR-B=0.2, 0.5. 0.8

Table 5 After generating partial attributes, RMSE obtained by federated and non-federated imputation based on different GANs when Cnum=5 for Bank and Credit Dataset (The optimal and suboptimal results are highlighted in red and blue, respectively. MisR-B=0.2, 0.5. 0.8)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VFImp(Y/N)  Dataset | | 5列 | | 5列 | | 5列 | |
| VF-GAIN | GAIN | VF-VGAIN | VGAIN | VF-CGAIN | CGAIN |
| Bank | 0.2 | **0.2057** | 0.2779 | **0.2291** | 0.2378 | **0.1691** | 0.2198 |
| 0.5 | **0.2087** | 0.295 | **0.2335** | 0.2854 | **0.2332** | 0.2775 |
| 0.8 | **0.2401** | 0.314 | **0.3024** | 0.3364 | **0.2533** | 0.3333 |
| Credit | 0.2 | **0.3478** | 0.4945 | **0.3373** | 0.4086 | **0.4029** | 0.4273 |
| 0.5 | **0.3674** | 0.5044 | **0.3503** | 0.4354 | **0.4242** | 0.467 |
| 0.8 | **0.4479** | 0.5158 | **0.4030** | 0.4217 | **0.4433** | 0.4651 |

## 4.3 Experiment 2：Comparison between FedPSG-AR and the SOTA baseline models

To validate the effectiveness of our method for participants sample generation, based on Experiment 1, we comprehensively compare FedPSG-AR with other baseline models. This experiment is also conducted with varying missing sample ratios of Party B relative to Party A (MisR-B). The datasets used in the experiment are Bank, Credit, Letter and News. The evaluation metric for this experiment is also RMSE, which indicates the average deviation between the data in the generated samples and the real data. In Experiment 2, the total number of training iterations for each generation method is set to 10,000, with 10 epochs, a learning rate of 0.001, and the Adam optimizer.

1. The experimental design of Experiment 2

In this experiment, for generation the missing samples of Party B, the methods used for validation and comparison are categorized into three main types:

1. The entire missing sample is completely generated locally in Party B, which is abbreviated as ‘Entirely generated locally’. The baseline models used for comparison are the current state-of-the-art models in the field of data generation: CTGAN[ ], TableGAN[ ], CTAB-GAN[ ], VAE[ ], and TabDDPM[ ].
2. Partial attributes of the missing sample are generated locally in Party B, and the remaining attributes are generated through federated imputation, which combines the data from Party A and Party B. This method is abbreviated as ‘Partially generated locally, partially generated federally’. As the first approach to researching this issue of participants samples generation, our team proposed a method called FedPSG-CAG, which first generates partial highly correlated attributes within Party B by local generation models, then generates the remaining attributes by federated imputation. In FedPSG-CAG, the remaining attributes imputation combines the observed samples from Party A and partial attributes of the missing samples in Party B generated by local generation models.

Note: Here, we do not select FedPSG-AG, which was used for comparison with FedPSG-AR in the Group 1 in Experiment 1. FedPSG-AG is also a mothod of ‘Partially generated locally, partially generated federally’. But, the initially generated partial attributes in FedPSG-AG are selected for comparison with VF-AR in FedPSG-AR and are determined by the correlation with the attributes of Party A. They themselves do not necessarily have a strong correlation. However, the initially generated partial attributes in FedPSG-CAG are selected based on their correlation with attributes within Party B. The generation model executed locally, can more easily learn the data distribution of attributes with strong correlation. Therefore, the overall generation performance of FedPSG-CAG is superior to that of FedPSG-AG. This can be observed from the experimental comparison results of FedPSG-AG-5 in Tables 1 and Tables 2 with FedPSG-CAG-5 in Table 6 on Bank and Credit datasets. Hence, in Experiment 2, FedPSG-AG is no longer used for comparison.

1. The entire missing sample is generated by combining the data of Party A and Party B. This method is abbreviated as ‘Entirely generated federally’ including the following two methods:

One method is used the vertical federated imputation based on GANs constructed in Section 3.3 of this paper. This method combines the observed samples from Party A and Party B to generate all attributes of the missing samples in Party B, which is abbreviated as ‘Completely imputed federally’. When Cnum=0, FedPSG-AR, FedPSG-CAG and FedPSG-AG are all the approaches of ‘Completely imputed’, denoted as FedPSG-AR-0 or FedPSG-CAG-0，FedPSG-AG-0.

The other method is the whole process of FedPSG-AR proposed in this paper. Combining the observed samples from Party A and Party B, this method firstly generates partial attributes of the missing samples in Party B by VF-AR, then generates the remaining attributes by federated imputation based on GANs.

According to the conclusions of Experiment 1, for consistency in comparison, Cnum in both FedPSG-CAG and FedPSG-AR is set to 5, and the federated imputation model adopts VF-GAIN and VF-VGAIN. The local generation model of Party B in FedPSG-CAG uses TabDDPM. When Cnum is set to 5, FedPSG-CAG-5 and FedPSG-AR-5 are used to denote the corresponding settings.

## 4.3 实验二的设计与结果分析

为了验证本文方法在生成B方缺失样本的有效性，在实验一的基础上，我们将FedPSG-AR与其他基线模型进行了充分的比较。本实验设置了不同的B方样本缺失率MisR-B。实验所采用的数据集为Bank，Credit，‘Letter’ and ‘News’. 样本生成效果的评估指标也为RMSE，用于表示生成数据与真实数据之间的平均偏差程度。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。

（1）实验二的实验设计

本实验中，用于验证比较的方法分为三大类：

①B方本地生成。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[ ]。

②部分本地生成，部份结合他方联邦填补。本团队在研究解决参与方样本生成问题时，最先提出一种“先本地生成B方高相关属性，再用纵向联邦填补生成剩余属性”的方法，称为FedPSG-CAG。在该方法中，剩余属性的填补联合了A方观测样本的数据。

特别说明：在此，我们没有选择实验一第一组实验中用于与FedPSG-AR比较的FedPSG-AG。FedPSG-AG也是一种“部分本地生成，部份结合他方联邦填补”的方法。但是FedPSG-AG中先生成的部份属性是为了与FedPSG-AR的VF-AR进行比较而选择的，是由与A方属性间的相关性决定的，它们之间本身并不一定具有很强的相关性。而FedPSG-CAG中的部分属性是由B方内部属性间的相关性决定的。在本地生成时，生成模型更容易学习到具有相关性的属性之间的数据分布。因此，FedPSG-CAG的生成效果总体上来说优于FedPSG-AG。这可以从表1、表2中FedPSG-AG-5与表6中FedPSG-CAG-5关于Bank，Credit数据集的实验对比结果看出。所以，在实验2中，我们不再将FedPSG-AG用于实验对比。

③完全结合他方生成。该类方法包括两种方式：

一种方式是采用本文3.3小节构建的基于GANs的纵向联邦填补，结合A方观测样本对B方缺失样本的所有属性进行生成，记为FedPSG-AR-0或FedPSG-CAG-0，FedPSG-AG-0。

另一种方式是本文方法FedPSG-AR的完整过程，联合A方观测样本，先用VF-AR生成B方缺失样本的部分属性，再联邦填补剩余属性。

根据实验一的实验结论：为了比较的统一性，FedPSG-CAG和FedPSG-AR中的Cnum都设置为5，填补模型都采用了VF-GAIN和VF-VGAIN。FedPSG-CAG中的B方本地生成模型采用了TabDDPM。当Cum设置为 5时，分别用FedPSG-CAG-5和FedPSG-AR-5表示。

(2) Results and Analysis of Experiment 2

Table 6 and Table 7 demonstrate the experimental results for Experimental 2. Under the different values of MisR-B, the RMSE values of TabDDPM are superior when ‘Entirely generated locally’ using for the missing samples of Party B. This is also the reason why FedPSG-AG and FedPSG-CAG chose TabDDPM as the local model for generating partial attributes of the missing samples in the experiments of this paper.

As shown in the results of Table 6 and Table 7, regardless of whether the MisR-B is high or low, the following conclusions can be drawn by analyzing the experimental results of FedPSG-CAG-5: The RMSE values of FedPSG-CAG-5 are all superior to ‘Entirely generated locally’. And, The RMSE values of FedPSG-CAG-5 are all superior to FedPSG-CAG/AG/AR-0, the method ‘Completely imputed federally’. These results indicate that generating the missing samples of Party B using the approach is effective which first generates partial highly correlated attributes within Party B by local generation models, then generates the remaining attributes by federated imputation.

FedPSG-CAG/AG/AR-0, the method ‘Completely imputed federally’, when compared with TabDDPM, the optimal ‘Entirely generated locally’, leads to the following conclusions: When MisR-B is relatively high, FedPSG-CAG/AG/AR-0 is superior to TabDDPM. However, when MisR-B is relatively low, TabDDPM is better learn the distribution of local data for generating the missing samples.

Furthermore, as shown in the results of Table 6 and Table 7, regardless of whether the MisR-B is high or low, our method FedPSG-CAG-5 generally outperforms all the other aforementioned methods. Based on FedPSG-CAG which first generates partial highly correlated attributes within Party B by local generation models, the method, which first generates partial attributes that are highly correlated with those of Party A by VF-AR, is superior.

In particular, 'Letter' and 'News' datasets equally divide the attribute columns between Parties A and B to simulate the vertical federated scenario. The data correlation, sample size, and number of attribute columns differ from those in 'Bank' and 'Credit' datasets. Therefore, Cnum=5 is not necessarily the optimal setting for FedPSG-CAG/FedPSG-AR on 'Letter' and 'News'. However, even so, the RMSE values of FedPSG-CAG-5/FedPSG-AR-5 based on VF-GANs outperforms all methods ‘Entirely generated locally’ and ‘Completely imputed federally’. Moreover, FedPSG-AR-5 is superior to FedPSG-CAG-5.

The outstanding performance of FedPSG-AR can be attributed to its innovative model design, which combines association rule generation, data imputation within vertical federated learning model. This combination effectively captures correlations between attributes from each party, ensuring that the generated data maintains logical and statistical consistency with the observed data—especially when the attributes are strongly correlated. Additionally, FedPSG leverages data associations from multiple parties under a vertical federated learning framework, enhancing generation performances while preserving data privacy. In contrast, these baseline methods (e.g., CTGAN, TabDDPM, etc.) focus solely on generating data by locally learning the overall distribution of the participant's data. They do not fully exploit the data associations with other participants. As a result, accuracy and consistency are compromised, particularly in cases with a high MisR-B.

（2）实验二的结果分析

实验二的实验结果如表6和表7所示。在B方不同的MisR-B下，B方本地生成B方缺失样本时，TabDDPM生成数据的RMSE值是较优的。这也是本文实验中FedPSG-AG和FedPSG-CAG选择TabDDPM作为本地模型生成缺失样本部分属性的原因。

而如表6和表7结果所示，无论B方缺少比高低，分析FedPSG-CAG-5的实验结果可得出以下结论：一方面，FedPSG-CAG-5的RMSE都优于B方本地生成的方法。另一方面，FedPSG-CAG-5的RMSE都优于FedPSG-CAG/AG/AR-0。这表明采取“先本地生成B方高相关属性值，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。

对于FedPSG-CAG/AG/AR-0来说，与B方本地生成中最优的TabDDPM相比，有以下结论：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。

进一步地，如表6和表7结果所示，无论B方缺少比高低，FedPSG-AR-5整体上优于上述其他所有方法。在FedPSG-CAG“先本地生成一些高相关属性值，再用纵向联邦填补方法进行填补”的基础上，采用VF-AR先生成与A方高相关的部份属性，再联邦填补剩余属性，是一种更佳的方式。

特别地，‘Letter’ and ‘News’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Bank，Credit有差异。因此，Cnum=5不一定是FedPSG-CAG/FedPSG-AR关于‘Letter’ and ‘News’数据集的最佳设定。但即便如此，基于VF-GANs的FedPSG-CAG/FedPSG-AR都优于所有B方本地生成，以及完全结合他方全填补生成的RMSE。

FedPSG-AR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各方属性之间的相关性，确保生成的数据与观测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，FedPSG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。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，一方面没有考虑属性相关性对于数据分布学习的重要作用，也没有充分利用其他参与方的数据关联信息，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。

Table 6 RMSE obtained by different methods when generating the missing samples of Party B for Bank and Credit Dataset (The optimal, suboptimal and third-best results are highlighted in red, blue, purple, respectively. MisR-B=0.2, 0.5, 0.8)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset  Methods | | Bank | | | Credit | | |
| 0.2 | 0.5 | 0.8 | 0.2 | 0.5 | 0.8 |
| Entirely generated locally | 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 |
| Partially generated locally, partially generated federally | 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 |
| Entirely generated federally (Completely imputed federally) | FedPSG-CAG/AR-0 (VF-GAIN) | 0.4615 | 0.4981 | 0.519 | 0.5048 | 0.5354 | 0.5518 |
| FedPSG-CAG/AR-0 (VF-VGAIN) |  |  |  |  |  |  |
| Entirely generated federally (FedPSG-AR) | FedPSG-AR-5 (VF-GAIN) | 0.2057 | 0.2087 | 0.2401 | 0.3478 | 0.3674 | 0.4479 |
| FedPSG-AR-5 (VF-VGAIN) | 0.2291 | 0.2335 | 0.3024 | 0.3373 | 0.3503 | 0.4030 |

Table 7 RMSE obtained by different methods when generating the missing samples of Party B for Letter and News Dataset (The optimal, suboptimal and third-best results are highlighted in red, blue, purple, respectively. MisR-B=0.2, 0.5, 0.8)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset  Methods | | Letter | | | News | | |
| 0.2 | 0.5 | 0.8 | 0.2 | 0.5 | 0.8 |
| Entirely generated locally | 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 |
| Partially generated locally, partially generated federally | 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 |
| Entirely generated federally (Completely imputed federally) | FedPSG-CAG/AR-0 (VF-GAIN) | 0.47 | 0.4981 | 0.5541 | 0.4602 | 0.5046 | 0.5566 |
| FedPSG-CAG/AR-0 (VF-VGAIN) |  |  |  |  |  |  |
| Entirely generated federally (FedPSG-AR) | FedPSG-AR-5 (VF-GAIN) | 0.3667 | 0.4297 | 0.465 | 0.3125 | 0.4138 | 0.4689 |
| FedPSG-AR-5 (VF-VGAIN) |  |  |  |  |  |  |

## 4.4 Experiment 3：Constructed joint sample sets for training federated classification models

To further validate the effectiveness of FedPSG-AR, we combine the generated samples from Party B with the observed samples from Party A to obtain new joint samples. Then, a joint sample set, constructed from these new joint samples and original aligned joint samples, is used for training the vertical federated classification models. Experiment 3 will validate and compare the training performance of various vertical federated classification models using the joint sample set constructed after addressing the missing samples in Party B with different methods. The vertical federated classification models adopted in Experiment 3 include VF-LR[ ], VF-SVM[ ], VF-GBDT[ ], VF-RF[ ], and FinalNet[ ]. These models represent classification learning models of different types and scales. The data from Parties A and B are under security and privacy protection during the training of the vertical federated classification models. The evaluation metrics for assessing the training performance of these classification models are ACC [ ], AUC [ ], and F1 score [ ]. In the experiment, we first construct a joint sample set based on the original aligned samples from Party A and Party B, then 30% of the joint samples are randomly divided as the test set for these federated classification models. The remaining 70% of the joint samples, together with the joint samples obtained by mentioned four approaches, are served as the training set of these federated classification models. Experiment 3 is conducted on the two datasets, 'Bank' and 'Credit'. The MisR-B is also set to 0.2, 0.5, and 0.8. In Experiment 3, the model configurations are as follows: VF-LR：Learning rate set to 0.01, 1000 iterations, batch size of 64; VF-SVM: Regularization parameter set to 1; VF-GBDT: Number of trees set to 20, learning rate of 0.1, maximum tree depth of 6, sample sampling rate of 20%; VF-RF: Number of trees set to 100, maximum tree depth of 3, minimum samples for splitting set to 2, minimum samples for leaf nodes set to 1; FinalNet: Learning rate set to 0.001, 1000 iterations, batch size of 64, hidden layer dimension of 128, regularization parameter λ set to 0.0001.

## 4.4 实验三的设计与结果分析

为了进一步验证本文方法FedPSG-AR 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。本实验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[ ]，它们代表了不同类型和规模的分类学习模型。纵向联邦分类模型训练时双方数据都受安全隐私保护。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方观测样本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。

(1) The experimental design of Experiment 3

In Experiment 3, five approaches are used to address the missing samples in Party B and construct the joint sample set:

1. FedPSG-AR, a method of ‘Entirely generated federally’ proposed in this paper. We first generate partial attributes in the missing samples of Party B that are highly correlated with those of Party A by VF-AR, then federally impute the remaining attributes. In this method, Cnum=5, and vertical federated imputation model is VF-VGAIN/VF-GAIN.
2. FedPSG-CAG, a method of ‘Partially generated locally, partially generated federally’. We first generate partial highly correlated attributes of the missing samples within Party B using local generation models, then federally impute the remaining attributes. In this method, the highly correlated attributes generation model adopted is TabDDPM, with Cnum=5, and vertical federated imputation model is VF-VGAIN/VF-GAIN.

③ TabDDPM, a state-of-the-art model for data generation, is directly employed to locally generate the missing samples in Party B.

④ No sample generation is performed for the missing samples in Party B, and a joint sample set is constructed directly by aligning the original samples from Party A and Party B. In Experiment 3, we refer to this approach as N-GM.

⑤Based on the joint sample constructed using the approach N-GM, a vertical federated sample generation method, VertiGAN [ ], is applied to generate new joint samples for Parties A and B, thereby expanding the sample size of the joint sample set from the approach N-GM. The number of generated joint samples is the same as the number of missing samples in Party B. In Experiment 3, we refer to this approach as A∞B-GM.

（1）实验三的实验设计

In Experiment 3, five approaches are used to handle the missing samples in Party B and construct the joint sample set:

1. FedPSG-AR, 本文方法的完整过程，联合A方观测样本，先用VF-AR生成B方缺失样本的部分属性，再联邦填补剩余属性。In this method, Cnum=5, and vertical federated imputation model is VF-VGAIN/VF-GAIN.
2. FedPSG-CAG, 先本地生成B方缺失样本的高相关属性，再联邦填补剩余属性。 In this method, the highly correlated attributes generation model adopted is TabDDPM, with Cnum=5, and vertical federated imputation model is VF-VGAIN/VF-GAIN.

③ TabDDPM, a state-of-the-art model for data generation, is directly employed to locally generate the missing samples in Party B.

④ No sample generation is performed for the missing samples in Party B, and a joint sample set is constructed directly by aligning the original samples from Party A and Party B. In Experiment 3, we refer to this approach as N-GM.

⑤ Based on the joint sample constructed using the approach N-GM, the state-of-the-art vertical federated sample generation method, VertiGAN [ ], is applied to generate new joint samples for Parties A and B, thereby expanding the sample size of the joint sample set from the approach N-GM. The number of generated joint samples is the same as the number of missing samples in Party B. In Experiment 3, we refer to this approach as A∞B-GM.

1. Results and Analysis of Experiment 3

We statistically analyze the joint sample sizes obtained from these five approaches, FedPSG-AR, FedPSG-CAG, TabDDPM, A∞B-GM, N-GM, and draw the following conclusions:

1. FedPSG-AR, FedPSG-CAG and TabDDPM is to generate the missing samples in Party B relative to Party A. Then, the generated samples of Party B can be aligned with the unaligned samples in Party A to construct complete joint samples. A∞B-GM is to generate joint samples, and the number of generated joint samples is the same as the number of missing samples in Party B. Therefore, the sample sizes of the joint sample sets constructed by FedPSG-AR, FedPSG-CAG，TabDDPM，A∞B-GM are the same, and they are determined by the number of observed samples from Party A.
2. The joint sample set constructed by N-GM has the least number of samples. Because, under the N-GM approach, if the samples from Party A can not be aligned with the missing samples from party B, the unaligned samples from Party A will be discarded when constructing the joint sample set. Therefore, the number of samples in the joint sample set under the N-GM approach is determined by the original observed number of samples from Party B before the missing samples are generated.

The experimental results are shown in Fig. 5. Fig.5(a)–(f) represent the ACC, AUC, and F1 score curves of different vertical federated classification models on the Bank and Credit datasets, with MisR-B = 0.2, 0.5, and 0.8, under various missing sample addressing methods for constructing the joint sample set. In each subplot of Fig.5(a)–(f), the curves in different colors represent different vertical federated classification models. The x-axis indicates the different missing sample addressing methods for constructing the joint sample set, while the y-axis shows the ACC, AUC, and F1 scores values obtained by training these federated classification models on the joint datasets.

As shown in Fig. 5, for both datasets, regardless of the value of MisR-B, the joint sample set constructed by using the FedPSG-AR, FedPSG-CAG approaches to generate the missing samples for Party B consistently yields the best performance in terms of ACC, AUC, and F1 scores on the test set, across all federated classification models, VF-LR, VF-SVM, VF-GBDT, VF-RF and FinalNet. The next best performance is observed with the TabDDPM, A∞B-GM, N-GM approaches, in that order. The reasons for these results are as follows:

1. From the perspective of sample size, the joint sample sets constructed using the four approaches, FedPSG-AR, FedPSG-CAG, TabDDPM, and A∞B-GM, have the same sample size, with N-GM having the smallest. As shown in Fig. 5, for both datasets, regardless of the values of MisR-B, the vertical federated classification models trained on the joint sample set constructed using the N-GM approach, which has the smallest sample size, yield the worst values of evaluation metrics. It indicates the importance of joint sample size in training vertical federated machine learning models. With other factors held constant, a larger joint sample set leads to better classification performance in the trained vertical federated machine learning models. As shown in Fig. 5, we verified this conclusion using different federated machine learning models. In the same dataset, as the sample missing ratio of Party B increases, the sample size of the joint sample set obtained by the N-GM approach decreases significantly. As the joint sample set size decreases, the evaluation metrics of all five vertical federated machine learning models decline. Notably, the evaluation metrics of FinalNet show a more significant decrease compared to VF-GBDT, VF-RF, VF-SVM, VF-LR. FinalNet is a vertical federated deep learning framework. The results also indicate that for complex models like deep neural networks, it is necessary to train with a larger sample size. Our method FedPSG-AR provides a effective solution to obtain more joint samples when one or some participants have missing samples.
2. From the perspective of sample quality, although the sample sizes of the joint sample sets obtained by the four approaches, FedPSG-AR, FedPSG-CAG, TabDDPM, A∞B-GM, are the same, the quality of their joint samples decreases successively. For the three approaches, FedPSG-AR, FedPSG-CAG, TabDDPM, all the data from Party A has been retained. Compared to the A∞B-GM approach, which generates entirely new joint samples, the joint sample sets constructed using the approaches, FedPSG-AR, FedPSG-CAG, TabDDPM, contain more real data for model training, resulting in higher quality joint samples. FedPSG-AR and FedPSG-CAG, especially FedPSG-AR, can effectively and sufficiently leverage the data associations between multiple parties to generate and impute the missing samples of Party B. For the TabDDPM approach, the data of participant with missing samples is generated locally in Party B by TabDDPM. The experimental results shown in Fig. 5 demonstrate that our method FedPSG-AR effectively addresses the issue of sample generation for participants with missing samples. Additionally, they indicate that model training performance depends not only on sample size but also on other factors, such as sample quality. Even in cases where MisR-B is high, our method FedPSG-AR can generate higher-quality training samples for the joint sample set, which better supports the training of vertical federated machine learning models. As shown in Fig. 5, this conclusion holds true for different machine learning models.

（2）实验三的结果分析

我们对FedPSG-AR、FedPSG-CAG、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 FedPSG-AR、FedPSG-CAG、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FedPSG-AR、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。

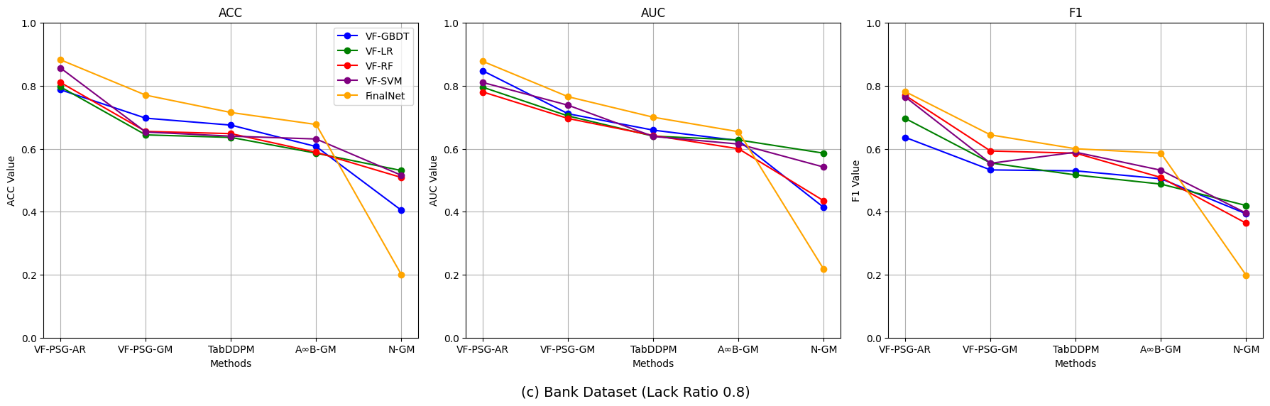
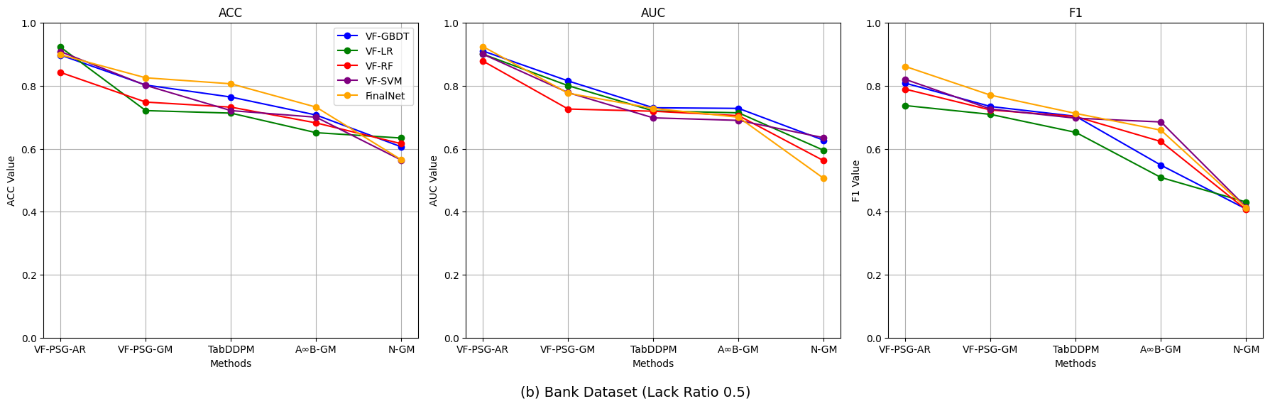
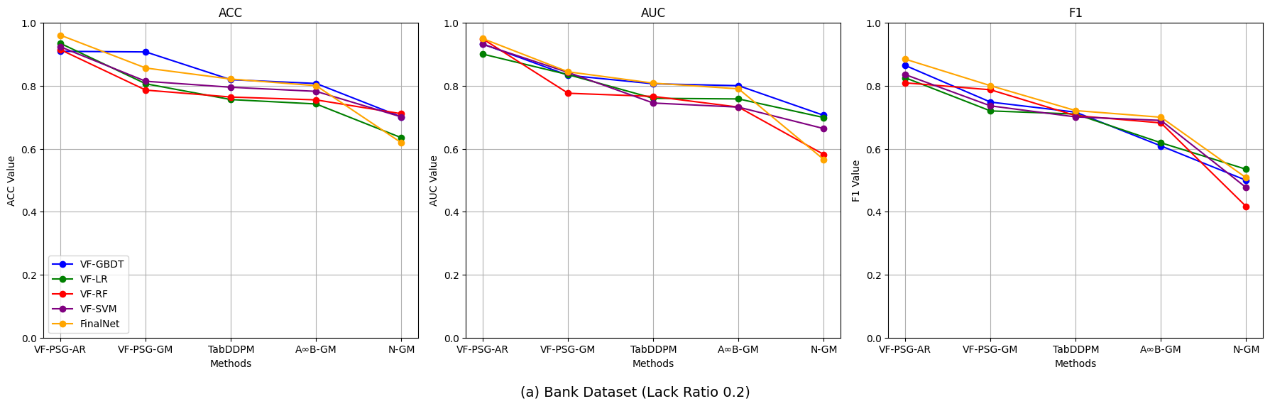
② N-GM这种处理方式构建的联合数据集中的样本量最少。因为，在N-GM方式下，当A方样本与法B方缺少的样本对齐时，在构建联合样本集时，A方未对齐样本将被丢弃。因此，N-GM方式下，联合样本集的样本数量由B方观测样本数量确定。

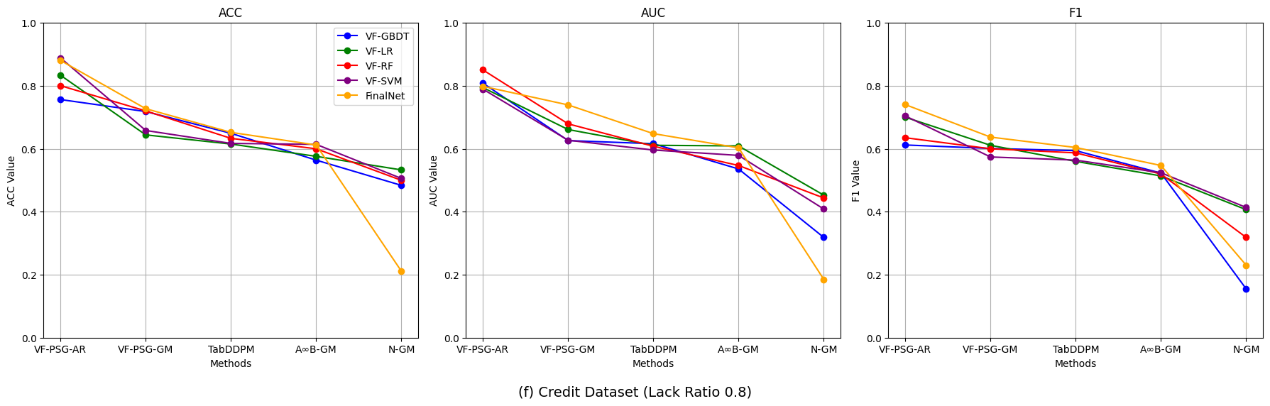
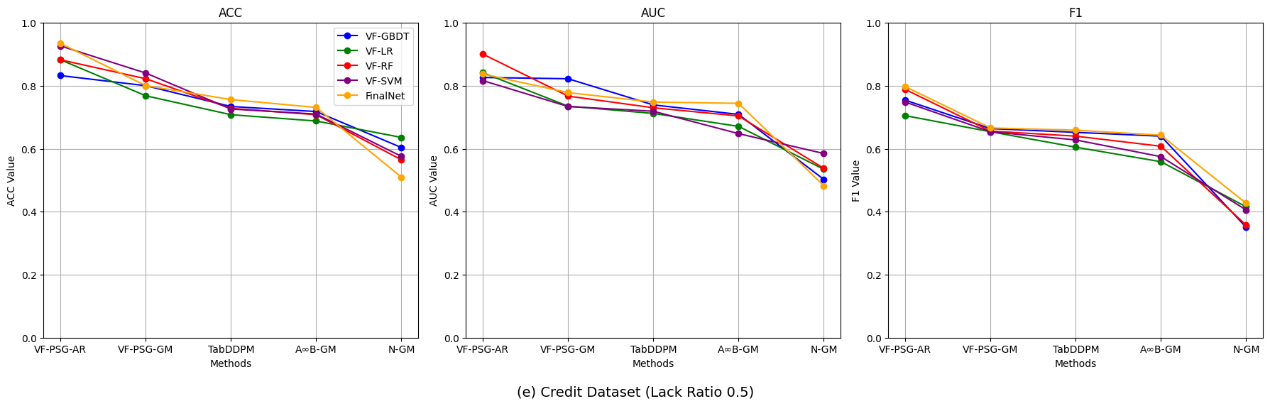
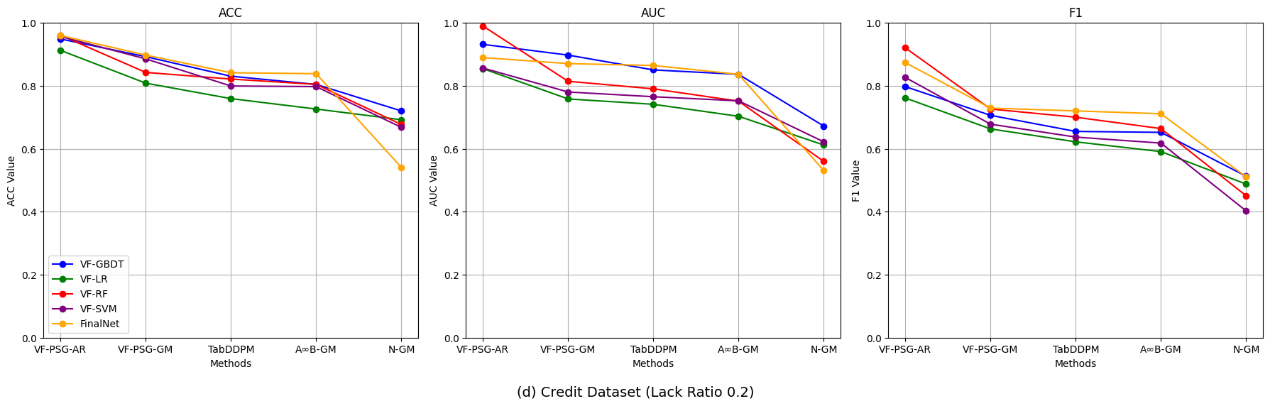
实验结果如图5所示，图5**（a）-（f）分别表示了Bank和Credit数据集在B方样本缺少率为0.2、0.5、0.8下，构建联合样本集时不同缺少样本处理方式下，不同纵向联邦分类模型的**ACC、AUC、F1 分数**结果曲线。**图5**（a）-（f）每个子图中不同颜色的曲线表示不同的纵向联邦分类模型，横坐标表示针对B方缺少样本不同的处理方式，即：**FedPSG-AR、FedPSG-CAG、TabDDPM、A∞B-GM、N-GM，纵坐标表示**纵向联邦分类模型在基于不同处理方式得到的联合数据集上训练后，模型在测试集上得到的**ACC、AUC、F1 分数值。

如图5所示，在两个数据集的不管在何种缺少率下，采用FedPSG-AR、FedPSG-CAG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，FedPSG-AR表现更为优秀。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-AR、FedPSG-CAG、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FedPSG-AR是一种非常有效的解决参与方样本缺少时的联合样本量问题的方法。

② 从样本质量的角度分析，虽然FedPSG-AR、FedPSG-CAG、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。FedPSG-AR、FedPSG-CAG、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，FedPSG-AR、FedPSG-CAG，尤其是FedPSG-AR，能够有效地利用多方数据之间的关联性进行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方本地生成的。这也能说明本文方法FedPSG-AR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。

****

****

**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**

Fig. 5 The verification results of different vertical federated classification models (需要重画该图，一些模型及颜色排序要修改，可参照会议论文，另，方法名称要修改，还有就是VF-GAIN、VF-VGAIN需要确定)

**5 Conclusion**

In multi-party collaborative scenarios, this paper proposes a novel Participants Sample Generation method based on association rules and data imputation within vertical federated learning, referred to FedPSG-AR, to address the challenge of data generation for participant with missing samples. When one or some participants may lack certain samples that other participants possess, the number of joint samples becomes limited after encrypted sample alignment, which is unfavorable for training vertical federated machine learning models. It is essential to generate samples for participants with missing data. FedPSG-VR integrates attribute correlations among multi-parties, association rules, and data imputation techniques with vertical federated learning, to generate high-quality samples for those participants. Attribute generation method based on vertical federated association rules generates the highly correlated attributes for the missing samples of participants by learning attribute correlations among different parties under secure privacy protection. Then, vertical federated imputation based on GANs are construct to generate the remaining attributes, maximizing the potential of multi-party collaborative learning. Experiments on various public datasets demonstrate that FedPSG-AR consistently outperforms existing methods across different sample missing ratios for Party B, especially in scenarios with high missing ratios. As future work, we would focus on improving the generation methods of highly correlated attributes and the federated imputation methods.

# Acknowledgments

This work was supported by the Chongqing Returned Overseas Scholars Entrepreneurship and Innovation Support Program No. CX2024086; the National Social Science Foundation of China under Grant No. 20BXW097; the Chongqing Technology Innovation and Application Development Special Major Project under Grant No. CSTB2023TIAD-STX0034.

# References