通过上述技术的有机结合，FedPSG-PUM方法构建了一个全面的纵向联邦样本生成框架，能够在保护数据隐私的同时，高效解决纵向联邦学习中的特征缺失问题，为构建高质量的联合样本集提供了坚实的理论基础和技术支持。

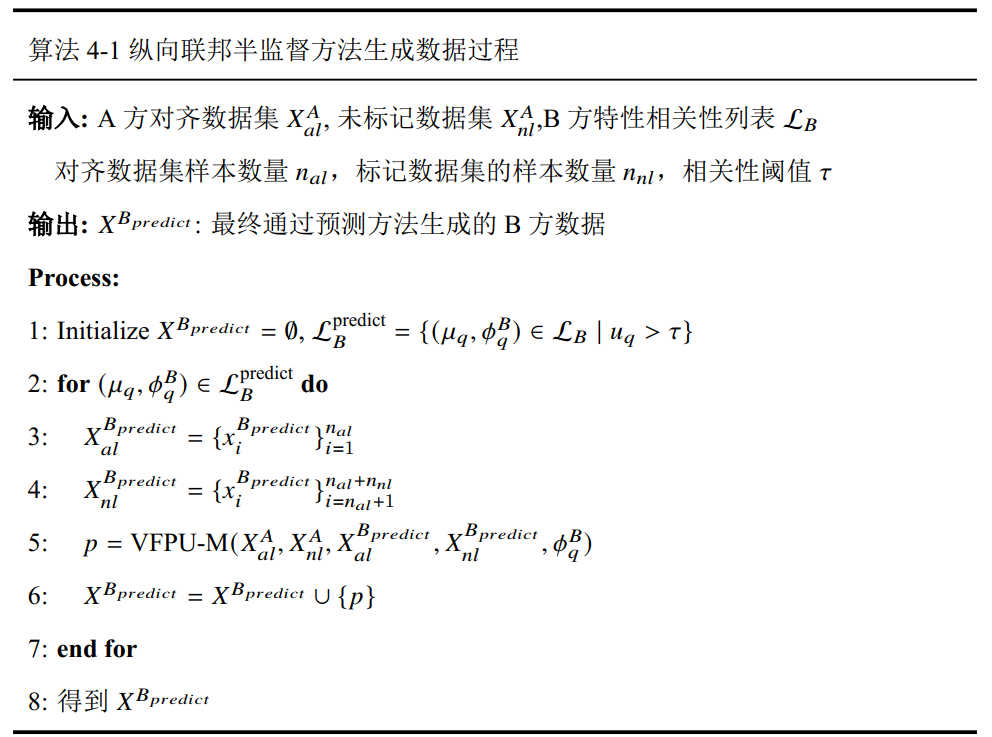
在该框架下，A方和B方的数据仍然保持隐私保护，即A方无法直接访问B方的原始数据，B方也无法直接获取A方的数据。然而，通过联邦学习的协作训练机制，A方可以利用自身数据和部分推测得到的B方数据进行模型训练，以优化对B方特征的预测能力。与此同时，B方的高相关性特征被逐列拆分，使得每一列特征都可以单独作为监督信号，从而有效地提升模型的学习能力。

在本研究中，该问题被重新表述为一个纵向联邦半监督学习（Vertical Federated Semi-Supervised Learning, VFSSL）问题，其中特征集 和 的数据被划分为有标签部分和无标签部分。Liu等人\textsuperscript{\cite{liu2023multi}}提出了VFPU（Vertical Federated Positive-Unlabeled Learning）算法来解决纵向联邦半监督学习问题。然而，VFPU 方法主要适用于PU（Positive-Unlabeled）学习，即仅包含正样本和未标记样本的情况，而在本研究中， 特征列可能涉及二分类、多分类甚至回归任务，VFPU 方法无法直接适用。因此，本文在 VFPU 方法的基础上进行了改进，提出了一种新的方法——VFPU-M（Multi-task VFPU），使其能够适用于多种任务类型。VFPU-M 主要通过以下五个步骤实现纵向联邦半监督学习，如图 4-2 所示。

* 基于对齐数据进行纵向联邦训练
* 利用训练好的基学习器对未对齐数据进行预测
* 计算预测结果的置信度
* 选择高置信度样本加入对齐数据集
* 重复上述过程，直至满足终止条件

通过 VFPU-M 方法，本研究能够在纵向联邦学习框架下有效利用无标签数据，从而提升模型的泛化能力，并适用于多种任务类型（如二分类、多分类和回归任务）。实验结果表明，VFPU-M 在不同任务场景下均能取得优于传统 VFPU 方法的性能，进一步验证了其有效性和适用性。

在本阶段的研究工作中，主要需要执行两个核心算法，以确保数据处理和模型训练的有效性。这两个算法分别为：算法1——基于纵向联邦半监督学习（Vertical Federated Semi-Supervised Learning, VFSSL）的方法用于数据生成的过程，以及算法2——VFPU-M（Vertical Federated PU Learning with Model Adaptation）算法。在后续章节中，我们将对这两个算法的理论基础、实现细节及其在本研究中的具体应用进行详细介绍。



(1) 算法 1：基于纵向联邦半监督学习的数据生成过程

本小节介绍了一种基于纵向联邦半监督学习（Vertical Federated Semi-supervised Learning, VFSL）的方法，用于生成 B 方缺失的数据。该方法的核心思想是利用 A 方与 B 方对齐数据之间的统计相关性，结合联邦学习框架，在保证数据隐私的前提下，对 B 方未对齐样本进行特征补全。算法 4-1 详细描述了该数据生成过程。

算法的输入包括以下几个关键元素：A 方对齐数据：，表示 A 方与 B 方样本空间对齐的特征矩阵，其中 为对齐样本的数量，

为 A 方特征维度。A 方未对齐数据：，表示 A 方未对齐部分的特征矩阵，其中 为未对齐样本的数量。B 方相关系数列表：，其中 表示 B 方特征列 与 A 方数据的皮尔逊相关系数， 为 B 方特征维度。相关性阈值：，用于筛选与 A 方数据具有显著相关性的 B 方特征列。该阈值的选取通常基于统计显著性检验，以确保筛选出的特征在统计上具有可靠性。

算法的核心目标是利用 A 方数据预测 B 方未对齐样本的特征值，并生成完整的 B 方数据矩阵 。

整个过程可分为以下三个主要步骤：

初始化阶段：目标数据集 为空集，表示尚未生成任何 B 方数据。通过相关性筛选，从 中选取所有相关性系数大于阈值 $\tau$ 的特征列，构造预测特征集合：

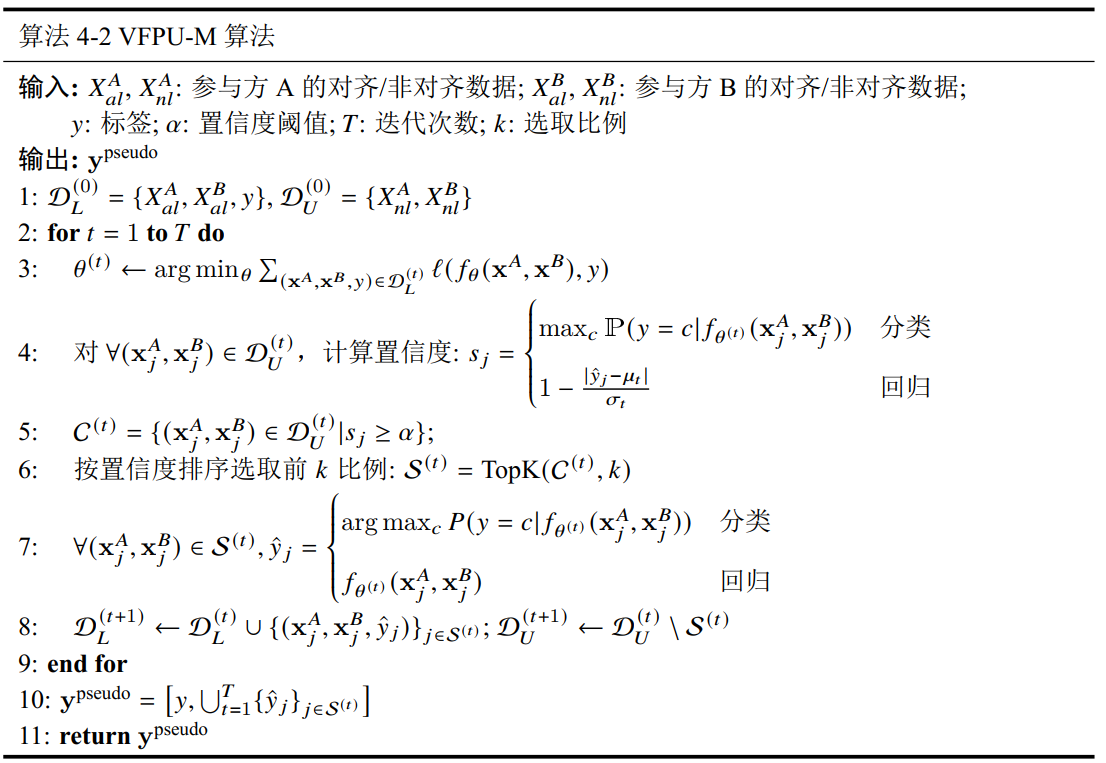


该筛选过程通常采用假设检验方法，以确保仅保留统计上显著相关的特征，从而提高数据生成的可靠性。特征级联邦数据生成：对于每个满足相关性筛选条件的特征列 ，执行以下步骤，数据分区 将 B 方特征列 的预测数据划分为：对齐部分：，对应于 A 方对齐样本的 B 方特征值。未对齐部分：，需要通过联邦学习方法进行预测。这种数据分区方式与 A 方数据结构保持一致，有助于后续联邦建模的执行。联邦预测建模：采用 VFPU-M（Vertical Federated Prediction with Unlabeled Missing data）算法进行特征预测，该算法将在后续章节详细介绍。其函数形式如下：

其中：对齐部分 直接使用 B 方已有的特征值。未对齐部分 通过 VFPU-M 进行预测，生成伪标签数据。该过程确保了 B 方数据的补全，同时符合联邦学习的隐私保护要求。

数据合成：将 VFPU-M 预测得到的伪标签向量 合并至目标数据集 $X^{B\_{predict}}$，完成当前特征列的数据生成。迭代执行：通过遍历所有满足相关性筛选条件的特征列，最终生成完整的 B 方数据矩阵：

该矩阵不仅保留了与 A 方数据的统计相关性，同时符合纵向联邦学习的隐私保护约束。



(2) 算法2：VFPU-M算法

本小节将介绍VFPU-M算法，如算法 4-2 所示，这是一种面向半监督联邦学习的创新方法，旨在处理这样一种场景：在两个参与方中，只有一部分数据（“对齐数据”）具有可靠的标签信息，而另一部分（“非对齐数据”）则缺乏标签，需要通过伪标签（Pseudo-Labeling）的方式来挖掘潜在信息。VFPU-M 在维护数据私密性的同时，通过迭代地对高置信度无标签样本进行伪标签生成并加入后续训练，有效提升了联邦模型在异构数据上的学习能力。

VFPU-M算法的核心思想是将拥有有限可靠标签的对齐数据和大量无标签的非对齐数据同时纳入训练过程，并在保证原始数据隐私的基础上，不断对高置信度的无标签样本赋予伪标签，使其得以融入有标签数据集中，从而迭代式地壮大可监督训练集。在具体实现中，算法首先对对齐数据构成的有标签数据集记为$\mathcal{D}\_L^{(0)}$，并将各参与方非对齐的剩余数据聚合至无标签数据集$\mathcal{D}\_U^{(0)}$。接下来，VFPU-M初始化一个联邦模型$f\_{\theta}$，模型的参数$\theta$在分布式环境下进行更新，参与方只共享必要的模型梯度或参数信息，不会直接交换任何原始数据。每一轮迭代中，算法都在现有的有标签数据集上训练模型，最小化以下损失函数：

其中可以是交叉熵损失、均方误差或其他根据任务需求设定的目标函数。通过这一过程，模型的预测能力在隐私保护的分布式框架下得到稳定提升。

在完成训练后，算法会利用更新后的模型对无标签数据集中的所有样本进行推断，并计算每个样本对应的置信度分数。对于分类任务，常用的做法是取模型输出的所有类别预测概率中最大的那一项作为置信度，这一数值越高表明模型对该样本的预测越有把握。而在回归任务中，则可以根据当前预测输出的整体分布情况，以

来衡量某个样本的预测距离是否接近分布中心，从而得到相应的置信度度量。随后，VFPU-M会预先设定一个置信度阈值，并从高于该阈值的候选样本中进一步挑选前$k$比例的高置信度样本，以降低噪声标签引入的风险。对于这些被选中的无标签样本，算法在本地生成对应的伪标签：如果是分类任务，则将预测概率最高的类别视为伪标签；若为回归任务，则直接采用当前模型的数值输出。通过这种方式，算法兼顾了可靠性与实用性，不会大规模地将未知真实性的数据强行纳入模型训练，而是倾向于选择那些与模型预测高度一致的样本来生成伪标签，从而减少噪声标签的累积效应。

在得到高置信度样本的伪标签后，这些样本便会从无标签数据集中移除，并被加入到新的有标签数据集中，以便在下一轮训练中继续完善模型。由此可见，VFPU-M在每轮迭代中都实现了有标签数据集的动态扩充，并不断强化模型的半监督学习能力。随着迭代进行，越来越多的高置信度样本得到有效利用，使联邦模型逐步学习到更全面、更丰富的数据模式，对于数据异构性和标签稀缺性都具有较好的适应能力。

与其他半监督联邦学习方法相比，VFPU-M具有以下突出的优势。其一，在标签缺乏的场景下，算法能够借助高置信度伪标签循序渐进地增强对无标签数据的利用效率，有效提升模型性能的同时不会过度牺牲鲁棒性。其二，VFPU-M既支持分类任务也支持回归任务，通过简单调整置信度度量与伪标签生成策略即可轻松适配多种应用需求。其三，该算法在任何分布式或跨机构场景中都能显著降低数据泄露风险，各参与方只需交换必要的模型更新信息而无需分享原始数据，切合当下对隐私与安全的严苛要求。其四，由于VFPU-M采用迭代式的联邦训练与筛选机制，当数据中存在较大异质性时，算法可以逐步利用对齐与非对齐数据的混合信息，从而平衡不同分布、不同来源数据对模型的综合贡献。

总的来说，VFPU-M算法为半监督学习与联邦学习的紧密结合提供了一种新颖而高效的范式。它在吸收对齐数据中已有标签信息的同时，充分利用非对齐无标签数据所蕴含的潜在价值，借助高置信度伪标签的迭代式筛选与生成，达到在分布式环境中同步提升模型精度与鲁棒性的目标。在未来的研究中，可以继续探讨如何动态更新置信度阈值，引入更先进的自适应策略，以及在多方恶意节点存在或标签噪声更为复杂的条件下增强算法的稳定性。通过与其他联邦学习或半监督学习技术有机结合，VFPU-M有望进一步拓展应用广度，在更多跨领域合作中发挥关键作用，为数据隐私与机器学习性能的双重需求提供行之有效的解决方案。

为更好地应对A方特征与目标特征间低相关性（即）给半监督学习带来的预测性能下降问题，VFPU-M-Syn引入了生成模型技术，通过捕捉低相关性特征的分布规律来合成数据，从而实现对缺失特征的合理补全。将其在B方进行本地生成模型的训练，并将训练好的生成模型用于生成B方缺少样本中的相关属性值。可采用的生成模型例如：CTGAN、TableGAN、CTABGAN、VAE和TabDDPM等。