尊敬的各位老师：

上午好！我是软件工程专业的研究生。今天我向各位老师汇报的论文题目是《联邦半监督学习方法及其在样本生成中的应用研究》。下面，我将从研究背景与意义、研究现状、研究内容、结论与展望四个方面进行汇报。

**研究背景与意义**

随着大数据时代的到来和人工智能技术的迅猛发展，数据资源已成为驱动各行各业决策能力和服务质量提升的战略要素。然而，数据隐私保护法规日益严格，使得不同机构之间的数据协作面临着前所未有的挑战，极大地限制了数据价值的深度挖掘。在此背景下，联邦学习作为一种新兴的分布式机器学习方法应运而生。它允许多个数据拥有方在不共享原始数据的前提下，通过交换模型参数或梯度信息实现模型的联合训练，成功解决了数据隐私保护与联合建模之间的矛盾。在现实应用中，VFL场景中常常只有一方拥有标签信息，这进一步加剧了联合模型训练的复杂性。如何高效利用未完全标记数据和未对齐样本成为当前VFL领域亟待解决的难题。

**国内外研究现状**

在联邦半监督学习的应用中，根据标记数据的位置不同，通常可分为两种情况：标签存储在客户端和标签存储在服务器端。对于前者，研究者们提出了多种算法，如Liang等人提出的随机采样共识联邦半监督学习（RSCFed），Fan等人提出的私有联邦半监督学习（FedSSL），以及Lin等人提出的联邦PU学习（FedPU）方法。对于后者，Zheng等人提出了一种不相干的学习方案，Diao等人提出了半监督联邦学习方法（SemiFL）。在数据生成方法领域，现有技术主要集中在生成模型上，包括基于自编码器的方法，如VAE；基于生成对抗网络（GANs）的方法，如CTGAN、TableGAN；以及针对表格数据的特定模型，如CTAB-GAN、MeTGAN等。这些方法通过学习丰富的表示从现有数据集中生成新样本，但在联邦学习环境下，它们往往忽视了不同参与方之间的数据相关性。

**论文研究的主要内容**

本论文针对纵向联邦中未标记样本缺失以及对齐样本不足的问题开展研究，具体的研究内容包括：一是研究未标记样本缺失的PU学习问题，针对VFL场景中仅含正样本和未标记样本的半监督学习问题，特别关注了未标记数据缺失的PU学习（UDD-PU）挑战，研究了一种结合VFL与PU学习的方法——VFPU；二是研究对齐样本不足的问题，针对VFL中普遍存在的样本对齐不足的问题，研究了一种基于联邦半监督学习的参与方样本生成方法——FedPSG-PUM。该方法利用半监督学习策略，结合特征相关性，提出了一种创新的多任务框架，以有效提升联邦学习的性能。

**相关理论的方法与应用**

本研究涉及联邦学习、半监督学习及表格数据生成三大核心理论。联邦学习是一种分布式训练方法，采用"数据在本地，模型在中央"的理念，根据特征空间和样本空间的重叠情况分为横向联邦学习（HFL）、纵向联邦学习（VFL）和联邦迁移学习（FTL）。半监督学习通过同时利用少量标注数据与大量未标注数据进行联合训练，提升模型泛化能力，其方法包括伪标签生成和PU学习。表格数据生成技术则致力于在保持数据特征分布与关联关系的同时，提升所生成数据的真实性与多样性，主要分为基于统计模型和基于深度生成模型两大类，为解决联邦学习中的样本不足与特征缺失问题提供了思路。

**基于多方联邦的半监督学习方法**

联邦学习（FL）为多方协作数据隐私保护问题提供了一种解决方案，允许多个数据所有者在不共享原始数据的前提下，通过交换模型参数或梯度实现联合训练。在VFL场景下，我们面临一种特殊的PU学习问题：各方只能访问部分重叠样本ID的特征空间，且目标方仅持有标记为正类的数据，其他参与方则仅持有未标记数据，这导致了一个全新的挑战，即未标记数据缺失的PU学习问题（UDD-PU）。

**问题分析与定义**

在UDD-PU问题中，我们考虑一个多方数据环境，其中涉及A方、B方和C方三个数据拥有者。A方持有一个只包含正类样本的数据集P，而B方和C方共同持有一个未标注数据集U，且U不包含P中的任何样本。这三方的目标是通过合作共同训练一个推荐模型，该模型能够从U中识别出可靠的正类样本R。这一问题设置使得传统的VFL算法在其标准形式下无法适用，因为各方都不拥有完整的标签数据，而传统PU学习方法假设学习者可以同时访问P和U，这在本联邦学习场景中不成立。

**数据预处理与加密样本对齐**

针对UDD-PU问题，我们提出了基于正样本和未标记数据的纵向联邦学习方法（VFPU）。该方法包括数据预处理与加密样本对齐、基于正样本与未标记数据的纵向联邦学习两个主要环节。在数据预处理阶段，对各方异构数据应用清洗、规范化和编码技术；在加密样本对齐阶段，采用基于盲RSA的私有集交集（PSI）协议安全计算数据集交集。VFPU算法通过反复从未标记数据中随机抽样，将被抽样的数据暂时视作负样本，形成训练集（正负样本比例平衡）与测试集，在联邦环境下迭代训练基学习器，并基于测试集预测计算每个未标记样本为正例的概率，选择高概率样本作为可靠正样本。

**数据集介绍与实验准备**

在实验中，我们利用了三个多样的数据集：Bank Marketing Dataset（41,188个样本，20个特征），Default of Credit Card Clients Dataset（30,000名客户，24个变量）以及Adult Census Dataset（1994年美国人口普查数据）。数据预处理包括分类特征的独热编码和数值特征的标准化缩放。将训练数据纵向划分给三个参与方，并为模拟标签稀缺情况，从方B和C数据中移除标签，从方A中随机选择10%正样本保留标签。在模型参数方面，最大迭代次数M为5，特征采样率θ为0.02，随机采样迭代次数T为10，采用密钥长度为512位的Paillier加密系统。

**实验设计与基线模型介绍**

我们设计了三组实验：一是联邦与非联邦PU学习性能比较，使用四种基学习器（LR、RF、GBDT、LGB）在三个数据集上评估联邦学习对分类质量的影响；二是分析不同基学习器对VFPU性能的影响，重点关注精确度、召回率、F-score和精确度-召回率曲线随可靠正样本数量的变化趋势；三是与其他半监督学习方法的对比，包括GBDT、Bagging\_GBDT、2Step\_GBDT、Pseudo-labeling、MixMatch、FixMatch、CoMatch、AdaMatch和SoftMatch等九种半监督方法。

**实验结果分析**

实验结果表明：一、联邦与非联邦PU学习性能比较中，联邦变体在准确率、召回率、精确率和AUC方面略低于非联邦版本，但差异极小（平均差异仅为0.0174）；二、不同基学习器对比显示，VFPU\_GBDT在所有数据集上性能最优，如在Bank数据集利用1500个可靠正样本时，VFPU\_GBDT的精确度达51.94%，召回率17.28%，F-score为0.26，明显优于其他基学习器；三、在与其他半监督方法比较中，No\_Fed\_VFPU\_GBDT在所有测试数据集和不同样本数量条件下均表现最佳，如在Census数据集上，当num=1300时实现了94.57%的推荐准确率，远超其他方法。

**基于联邦半监督学习的参与方样本生成方法**

随着隐私保护法律的日益严格，联邦学习成为解决多方协作问题的关键技术之一。在VFL中，多个数据持有方可以在不泄露原始数据的前提下，通过对齐样本的特征协作训练模型。然而，VFL要求所有参与方共享足够的联合样本，这在实际应用中往往难以实现。针对样本对齐不足导致的联合样本稀缺问题，我们提出了一种基于联邦半监督学习的参与方样本生成方法（FedPSG-PUM），通过估计缺失特征的表示来生成样本，有效提升VFL性能。

**问题分析与定义**

在典型的两方VFL场景中，参与方A和B通过各自独立的特征集进行合作学习，其中只有一方拥有标签。传统VFL方法仅使用对齐样本来构建联邦模型，而将非对齐样本弃置不用，这在对齐样本数量较少时，限制了模型性能。FedPSG-PUM方法结合了纵向联邦半监督学习和表格数据生成技术，通过将对齐样本视为有标签数据，非对齐样本视为无标签数据，利用半监督学习从对齐样本中学习，同时采用表格数据生成技术填补相关性较弱的特征缺失值，显著提高了数据利用率和模型泛化能力。

**方法框架**

FedPSG-PUM方法包含三个核心流程：一是计算跨方特征相关性，通过隐私保护的Spearman秩相关分析，安全高效地量化不同参与方特征间的统计关联；二是纵向联邦半监督方法预测数据，基于"列级特征分解"策略和"多列对一列模型"方法，设计多任务联邦半监督学习算法VFPU-M，通过迭代式伪标签生成和高置信度样本筛选，准确预测高相关性特征；三是生成模型与纵向联邦半监督学习的协同框架，对于低相关性特征，采用如VF-GAIN、TabDDPM等生成模型技术进行填补，构建一个半监督学习与生成模型协同的混合架构。

**数据集介绍与实验准备**

实验选取了UCI机器学习库中的四个数据集：银行营销数据集（Bank，45,211个样本），德国信贷数据集（Credit，1,000条贷款记录），字母识别数据集（Letter，20,000个手写英文字母样本）和在线新闻传播度数据集（News，39,644篇新闻）。各参与方按照特征所有权进行了纵向数据划分，如在Bank数据集中，参与方A和B各自持有8个属性列。实验重点研究了B方样本缺失的场景，设置了不同级别的样本缺失率（MisR-B）和相关性阈值（τ）。

**实验设计与基线模型介绍**

实验设计了三组对比：一是调节相关性系数阈值τ的取值，评估FedPSG-PUM性能；二是固定τ=5，比较不同基学习器（VFPU\_LR、VFPU\_GBDT、VFPU\_RF、VFPU\_LGB）的表现；三是探究置信度阈值α对模型性能的影响，α取值区间为[0.5, 0.9]，步长0.05。此外，还设计了基线对比实验，将FedPSG-PUM与CTGAN、TableGAN、CTAB-GAN、TVAE和TabDDPM等单方本地生成方法进行性能比较，并评估生成样本在多种纵向联邦分类模型（VF-LR、VF-SVM、VF-GBDT、VF-RF、FinalNet）训练中的效果。

**实验结果分析**

实验结果显示：一、当τ∈[0.5, 0.6]时，FedPSG-PUM的RMSE指标达到最优水平，表明联邦半监督学习预测高相关属性列数据与联邦填补生成剩余属性的混合策略具有显著优势；二、基学习器对比中，VFPU\_GBDT在所有缺失率条件下均表现最优，如在Bank数据集的MisR-B=0.2场景下，RMSE仅为0.2097，较VFPU\_RF（0.2640）降低20.6%；三、置信度阈值α=0.7时模型性能最佳，表现出"U型"分布特征；四、与单方本地生成模型相比，FedPSG-PUM(VF-GAIN)在所有条件下均显著优于单方模型，如在Bank数据集MisR-B=0.2时，RMSE仅为0.2097，比最佳基线模型TabDDPM（0.4227）降低50.4%；五、在样本生成效果评估中，基于FedPSG-PUM构建的联合样本集在ACC、AUC和F1指标上全面超越其他方法。

**总结与展望**

本研究提出了两种创新性框架VFPU和FedPSG-PUM，分别针对正样本与未标记数据学习问题以及纵向联邦学习中的样本未对齐问题提出了解决方案。VFPU首次将PU学习引入纵向联邦学习场景，通过加密样本对齐和协同训练实现了与传统集中式方法相近的效果；FedPSG-PUM结合了联邦半监督学习和生成模型技术，在保护数据隐私的前提下有效解决了样本未对齐问题，显著提升了模型性能。未来工作将进一步扩展现有框架至多方参与的复杂生态系统，增强隐私保护机制，结合最新生成模型技术，优化计算与通信成本，并推广到医疗健康、金融风控等多个应用领域。

以上是我的论文汇报，恳请各位老师批评指正，谢谢！

**问题1：论文中的创新点是什么？**

**回答1：** 该论文的主要创新点在于提出了两个新颖框架来解决联邦学习中的关键挑战。首先，VFPU方法首次将PU学习(仅使用正样本和未标记数据的学习)引入到纵向联邦学习场景中，通过设计加密样本对齐和协同训练流程，有效解决了未标记数据缺失的PU学习问题。其次，FedPSG-PUM方法通过结合联邦半监督学习与表格数据生成技术，创新性地提出了基于特征相关性的分层处理策略，为高相关特征采用半监督预测，为低相关特征采用生成模型补全，显著提升了对齐样本不足情况下的模型性能。这两个框架共同为隐私保护环境下的数据协作提供了新思路。

**问题2：你选题的缘由是什么？研究具有何种现实指导意义？**

**回答2：** 选择本课题的缘由主要基于联邦学习实际应用中面临的两个关键挑战：标记数据稀缺和样本对齐不足。随着数据隐私保护法规日益严格，金融、医疗等行业的机构间数据协作面临前所未有的挑战，而现有联邦学习方法往往过度依赖于大量标记数据和完全对齐样本，无法充分利用丰富的未标记数据和未对齐样本。本研究的现实指导意义在于：一方面，提出的方法能够在保护数据隐私的前提下，提高联邦学习中数据的利用效率，降低对标记数据的依赖；另一方面，通过生成高质量的样本数据，解决了样本对齐不足问题，为金融风控、医疗诊断、智能推荐等实际应用场景提供了可行的技术方案。

**问题3：请详细解释VFPU算法的工作原理及其解决UDD-PU问题的关键步骤。**

**回答3：** VFPU算法的工作原理主要基于迭代式伪标签生成与筛选。其关键步骤首先是数据预处理与加密样本对齐，通过盲RSA协议安全地确定各方共有样本，形成对齐数据集。第二步是反复从未标记数据中随机抽样构建平衡训练集，将未被抽取的样本作为测试集，基于对齐样本在联邦环境下训练基学习器。第三步是对测试集样本生成预测得分，计算每个未标记样本为正例的概率。第四步是根据预测概率选择高置信度样本作为可靠正样本，加入正样本集并移出未标记集。最后通过多轮迭代，逐步扩充正样本集，实现对未标记数据中隐藏正样本的有效识别，解决了UDD-PU问题中多方数据无法集中访问的核心挑战。

**问题4：在实验结果中，VFPU\_GBDT表现优于其他基学习器的原因是什么？这对实际应用有何启示？**

**回答4：** VFPU\_GBDT表现优于其他基学习器的原因主要有三点：首先，GBDT作为一种集成学习算法，通过迭代训练多个弱学习器并纠正前序模型的错误，具有较强的泛化能力和鲁棒性，特别适合处理UDD-PU问题中的数据不平衡和标签噪声。其次，GBDT能自动处理特征之间的复杂非线性关系，这对于捕捉联邦环境中多方异构数据的潜在模式至关重要。第三，GBDT对过拟合有较好的抵抗能力，在有限标记数据条件下依然能保持稳定性能。这一发现对实际应用的启示是：在设计联邦半监督学习系统时，基学习器的选择应当考虑数据特性和任务需求，并优先考虑具有较强自适应能力和鲁棒性的集成方法；同时在资源允许的情况下，应当针对不同应用场景验证多种基学习器的表现，以选择最适合的算法组合。

**问题5：FedPSG-PUM方法中，相关性阈值τ和置信度阈值α对模型性能有何影响？如何在实际应用中确定这些参数的最优值？**

**回答5：** 根据实验结果，相关性阈值τ和置信度阈值α对FedPSG-PUM方法的性能具有显著影响。当τ∈[0.5, 0.6]时，模型RMSE达到最优，说明适当平衡联邦半监督预测与生成模型填补的特征比例更有效。置信度阈值α则表现出"U型"分布特征，α=0.7时性能最佳，过高或过低都会导致性能下降。这是因为α过低时会引入大量低质量伪标签造成噪声，而α过高则筛选条件过严，可能导致有用样本的损失。在实际应用中确定这些参数的最优值，应当结合数据特性和业务需求采用交叉验证方法，在代表性数据子集上系统性地测试参数组合，通过网格搜索或贝叶斯优化等技术寻找最优配置。此外，还可以根据特征相关性分布和模型预测置信度的统计特性，为不同应用场景定制参数选择策略。

**问题6：联邦与非联邦PU学习性能比较实验中，平均性能差异仅为0.0174，这对联邦学习隐私保护与模型性能的权衡有何启示？**

**回答6：** 联邦与非联邦PU学习性能比较实验中得到的微小差异（平均仅0.0174）具有重要启示意义。这表明通过精心设计的联邦学习框架，可以在确保数据隐私的同时，几乎不损失模型性能。这一发现打破了传统认为"隐私保护必然导致显著性能下降"的观念。对于金融、医疗等高度重视隐私保护的行业，这意味着机构可以放心采用联邦学习技术进行数据协作，无需担忧模型质量的大幅下降。从更广泛的角度看，这为数据隐私与模型性能的权衡提供了新思路，证明通过优化算法设计和协作机制，可以实现"鱼与熊掌兼得"的效果。在实际部署中，组织应更加关注联邦架构的设计与优化，而非过度担忧隐私保护带来的性能损失。

**问题7：FedPSG-PUM方法相比单方本地生成方法(如TabDDPM)在RMSE指标上有显著提升，这种提升的内在机制是什么？**

**回答7：** FedPSG-PUM方法相比单方本地生成方法的显著提升源于其独特的混合架构设计。首先，FedPSG-PUM通过跨方特征相关性分析，能够量化不同特征间的依赖关系，从而实现更有针对性的特征处理策略，而单方本地生成方法无法获取这种跨参与方的信息。其次，对于高相关性特征，FedPSG-PUM采用VFPU-M算法进行半监督预测，能够充分利用对齐样本中的监督信号和结构信息，比单纯依靠生成模型更准确。第三，FedPSG-PUM的迭代式伪标签生成机制能够逐步提高预测质量，随着高置信度样本的累积，预测准确率不断提升。最后，针对低相关性特征，FedPSG-PUM选择性地应用专业生成模型进行补全，保证了整体生成数据的质量和一致性。这种"分而治之"的综合策略使得FedPSG-PUM能够在Bank数据集上将RMSE从TabDDPM的0.4227显著降低至0.2097，降幅达50.4%。

**问题8：在处理样本未对齐问题时，FedPSG-PUM方法如何保证数据隐私的安全性？是否存在潜在的隐私泄露风险？**

**回答8：** FedPSG-PUM方法通过多层隐私保护机制确保数据安全。首先，在跨方特征相关性计算过程中，采用同态加密技术对秩向量进行加密，参与方只能访问加密后的数据，无法获取原始信息。其次，在联邦半监督学习阶段，各方保留原始数据本地，仅交换必要的模型参数或加密梯度，符合联邦学习"数据不出域"的核心原则。第三，生成模型阶段在B方本地执行，不涉及原始数据的跨方传输。尽管如此，仍存在一些潜在风险，如通过模型参数反向推导原始数据的模型反演攻击，或多轮交互过程中的信息累积泄露。为进一步加强安全性，可考虑引入差分隐私机制为梯度添加噪声，或结合可信执行环境(TEE)技术确保计算过程的安全性。总体而言，FedPSG-PUM在保证功能实现的同时，通过精心设计的协议和加密机制，有效降低了隐私泄露风险。

**问题9：论文提出的方法在高缺失率(如MisR-B=0.8)情况下依然保持良好性能，这一特性在实际业务场景中有何应用价值？**

**回答9：** 论文方法在高缺失率下保持良好性能的特性具有显著的实际应用价值。在金融风控领域，不同金融机构间的客户重叠率通常较低，往往不足20%，传统联邦学习方法因对齐样本太少而性能欠佳，而本方法能够有效利用未对齐样本，大幅提升反欺诈模型的准确率。在医疗健康领域，不同医院之间的患者重叠度同样有限，通过本方法可以构建更全面的患者画像和诊断模型，促进罕见病研究和精准医疗发展。在跨境电商场景，各地区用户行为数据难以直接整合，本方法能够帮助企业在保护用户隐私的前提下，构建高质量的全球用户推荐系统。此外，本方法还能降低联邦学习的部署门槛，使更多中小机构能够参与数据协作，无需担心对齐样本不足的问题，促进数据要素更广泛的流通与价值释放。

**问题10：如何评价论文提出的VFPU-M算法中多任务学习框架的设计？这种设计对其他联邦学习场景是否有借鉴意义？**

**回答10：** VFPU-M算法中的多任务学习框架设计具有显著的创新性和实用价值。该设计采用"列级特征分解"策略和"多列对一列模型"方法，将复杂的特征预测问题分解为多个相对独立的子任务，每个任务专注于预测单一特征列，这种分解策略降低了模型复杂度，减少了过拟合风险。同时，通过动态伪标签生成和置信度筛选机制，实现了半监督学习的渐进式优化，特别适合联邦环境下的异构数据处理。这种设计对其他联邦学习场景具有重要借鉴意义：首先，任务分解思想可应用于联邦迁移学习中的特征选择与映射；其次，渐进式伪标签策略可用于解决联邦学习中的数据不平衡问题；第三，基于特征相关性的差异化处理机制可推广到多方联邦学习的模型个性化优化。未来研究可考虑将此框架扩展到更复杂的异构联邦学习和垂直领域知识迁移场景中。