重 庆 邮 电 大 学

CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

博士学位论文

DOCTORAL DISSERTATION

论文题目	**重庆邮电大学学位论文格式模板**
学科专业	**电子科学与技术**
学　　号	**S20202222**
作者姓名	**张三**
指导教师	**李四　教授**
学　　院	**光电工程学院/重庆国际半导体学院**

学校代码	10617	UDC	xxxxxx
分 类 号	xxxxxx	密级	

学 位 论 文

重庆邮电大学学位论文格式模板

某 某

指导教师 **某某某 教 授**

某 某 副教授

申请学位级别	**博士**	学科专业	**XXXX**
专业学位领域		**XXXXXX**	
答辩委员会主席	**某某某 教授**	论文答辩日期	**2021 年 5 月 20 日**
学位授予单位和日期	**重庆邮电大学**	**2021 年 6 月**	

Dissertation Template for Doctoral Degree of Engineering in CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

A Doctoral Dissertation Submitted to
Chongqing University of Posts and Telecommunications

Discipline	**XXXX**
Student ID	**XXXX**
Author	**XXXX**
Supervisor	**XXXX**
School	**XXXX**

重庆邮电大学

学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文中不包含其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在论文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

重庆邮电大学

学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于 ：

□**公开论文**

□**涉密论文**，保密____年，过保密期后适用本授权书。

（请在以上方框内选择打“√”）

作者签名：　　导师签名：

日期：　　年　　月　　日

摘　要

学位论文是研究生从事科研工作的成果的主要表现，集中表明了作者在研究工作中获得的新发明、新理论或新见解，是研究生申请硕士或博士学位的重要依据，也是科研领域中的重要文献资料和社会的宝贵财富。

为进一步规范我校研究生学位论文撰写格式，提高研究生学位论文质量，参照国家标准《学位论文编写规则》（GB/T 7713.1-2006），结合我校实际，制定本模板。

关键词：学位论文，撰写规范，论文模板，重庆邮电大学

ABSTRACT

Dissertation /Thesis is postgraduate’s main academic performance to display her/his works of scientific research, which shows the author’s new invention, new theory or new opinion in her/his research. It is the crucial document for the graduate students to apply for degree, and it is also the important scientific research literature and the valuable wealth of society.

In order to further standardize the format of dissertation/thesis writing and improve graduate dissertation/thesis quality, this temolate is formulated with reference to the national standard ”Rules for Dissertation Writing” (GB/T 7713.1-2006) and the reality of CQUPT.

Keywords： Dissertation/Thesis, Writing Specification, Thesis Template, Chongqing University of Posts and Telecommunications

目 录

图目录

表目录

主要符号表

符号	说明	页码
c	电磁波的相平面速度	10

缩略词表

英文缩写	英文全称	中文全称
CQUPT	Chongqing University of Posts Telecommunications	重庆邮电大学

第 1 章 绪论

1.1 研究背景及意义

1.2 国内外研究现状

1.2.1 联邦半监督学习

当前的联邦半监督学习方法主要基于横向联邦架构（即参与方共享特征，但样本 ID 不同）。根据标注数据的位置分为标签在客户端和标签在服务器端两种情况[1]。

(1) 标签在客户端

标注数据存于客户端，而服务器只能获取未标注数据。例如，某家公司欲利用智能手机拍摄的图片训练一个物体检测的联邦学习模型，但无法直接访问用户的本地数据，只能依赖用户的标注，如图 1-1 所示。首先，对于不同的参与方（用户），样本 ID 是不同的，但参与方（用户）手机里的图片特征是一样的，所以构成了横向联邦的设置。其次，用户通常不愿为每张照片标注，给基于横向联邦的半监督学习创造了一个“客户端标注”（label-at-client）的环境。

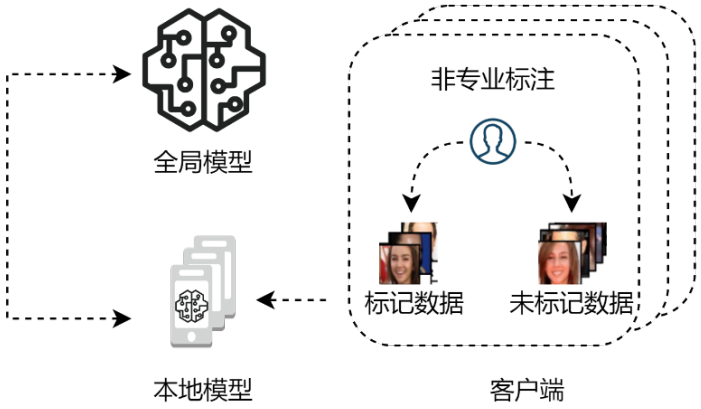

图 1-1 标记数据在客户端的情况

Fig. 1-1 Labeled data on the client side

RSCFed[2] 主要关注联邦半监督学习中的标签隔离问题和数据异质性问题。在局部训练中，采用师生模型[3] 对无标签数据进行训练。为进一步解决数据异质性问题，RSCFed 提出了子共识抽样法和距离加权聚合法。在每一轮中，通过对所有参与者的多个子集进行独立抽样，从而聚合出多个子共识模型，这样每个子共识模型都有望包含拥有标记数据的参与者。此外，本地模型会根据它们与子共识模型

的距离进行加权，这样偏差模型就会得到较低的权重，其影响也会降到最低。

FedSSL [4] 解决了标签隔离问题、数据隐私问题和数据异构问题。为了便于对未标记的客户端进行本地训练，FedSSL 利用了伪标记技术。此外，为解决数据异构问题，FedSSL 学习全局生成模型，从统一的特征空间生成数据，从而通过生成的数据缓解数据异构问题。最后，为了防止生成模型造成的隐私泄露，FedSSL 利用差分隐私（DP）来限制生成模型中训练数据的信息泄露。

FedMatch [5] 提出了一种客户端间一致性损失来解决数据异质性问题。具体来说，对每个客户端的前 k 个最近客户端进行采样，在每个数据样本上，本地模型的输出与前 k 个客户端模型的输出进行正则化，以确保一致性。此外，FedMatch 还提出了分离式学习方法，将标注数据和未标注数据的参数分开，未标注数据的参数是稀疏的。更新时，只有未标记数据的客户端会上传稀疏张量，从而降低通信成本。

FedPU [6] 研究了半监督学习中更具挑战性的环境–正向和无标签学习，在这种环境中，每个客户端只有类的子集标签。在这种情况下，客户端只掌握所有类别中的一部分信息，从而导致严重的标签隔离问题。为了解决这个问题，FedPU 提出了一个新颖的目标函数，即把学习客户端负类的任务交给拥有负类标签数据的其他客户端。这样，每个客户端只负责学习正类，并可自行进行局部训练。根据经验，在正类和无标签学习设置中，所提出的 FedPU 优于 FedMatch。

AdaFedSemi [7] 提出了一种系统，可在联合半监督学习中利用服务器端无标记数据实现效率与模型准确性之间的权衡。在每一轮学习中，模型都是通过客户端的标注数据进行训练，并在服务器端进行汇总。服务器端未标注数据通过伪标注纳入训练过程。AdaFedSemi 确定了平衡效率和性能的两个关键参数，即客户端参与率 P 和伪标签的置信度阈值 τ。较低的 P 可以降低通信成本和模型准确性，而较高的 τ 可以降低服务器端计算成本，同时也限制了未标记数据的使用。。实验表明，AdaFedSemi 通过动态调整 P 和 τ 在不同的训练阶段实现了效率和准确性之间的良好平衡。

DS-FL [8] 解决了与 AdaFedSemi 类似的问题，即客户端拥有标签数据，而服务器拥有非标签数据。它提出了一种集合伪标签解决方案来利用服务器端的非标签数据。具体来说，它不是对数据样本使用单一的伪标签，而是对所有客户端生成的伪标签进行平均，。这将创建一个客户端模型集合，并提供更好的性能。此外，由于只传输伪标签而不是模型参数，通信成本可以大大降低。此外，DS-FL 发现在伪标签上进行训练会导致预测熵增大。因此，DS-FL 提出了一种减少熵的聚合方法，即在聚合之前使局部输出。

(2) 标签在服务器端

标注数据存于服务器，而客户端只有未标注数据。例如，一家可穿戴设备公司

希望利用联邦学习训练健康监测模型图，如图 1-2 所示。在这种情况下，由于用户通常缺乏专业知识，无法标注健康相关数据，因此客户端的数据是未标注的。标签数据在客户端的情况比标签在服务器端设置更复杂，原因是所有客户端都只拥有未标记数据，无法为联邦模型提供额外的监督信号, 仅使用无标签数据进行训练可能会导致遗忘从有标签数据中学到的知识，从而影响模型性能[5,9]。

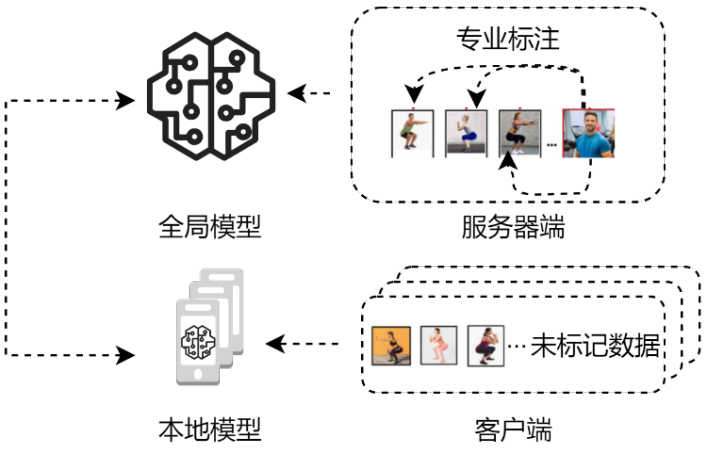

图 1-2 标记数据在服务端的情况

Fig. 1-2 Labeled data on the server side

为了解决标签数据和非标签数据之间的隔离问题，FedMatch[5] 提出了一种不相干的学习方案，即分别为标签数据和非标签数据设置两套参数。在非标签数据上进行训练时，标签数据的参数是固定的，反之亦然，以防止知识被覆盖。非标记数据的参数会在参与者和服务器之间传输，而非标记数据的参数设置为稀疏的，这为通信效率带来了额外的好处。此外，为了解决不同客户端持有的异构数据问题，FedMatch 提出了客户端间一致性损失，这样不同参与者的本地模型就能在相同数据上产生相似的输出。

SemiFL[9] 采用另一种方法来解决这些挑战。它建议使用标注数据对全局模型进行微调，以提高其质量，并减轻客户端无监督训练所造成的遗忘。此外，SemiFL 建议最大限度地提高客户端模型与全局模型之间的一致性，而不是在客户端之间对模型输出进行正则化。具体来说，全局模型为客户端的未标记数据生成伪标签，而客户端的局部模型则根据伪标签进行训练。实证结果表明，与 FedMatch 相比，SemiFL 能产生更有竞争力的结果。

(3) 联邦半监督学习方法总结

表 1-1 总结了前文介绍的基于横向联邦的半监督学习方法，按照标签在客户端/服务器端划分。

表 1-1 联邦半监督学习方法总结

Table 1-1 Summary of federated semi-supervised learning methods

类别	方法	半监督学习算法	隐私保护方案	数据异质问题	性能
标签在客户端	RSCFed	教师学生模型	无	加权距离聚合	无
	FedSSL	伪标记	差分隐私	全局生成模型	无
	FedMatch	伪标记	无	客户端间一致性	分散学习和稀疏学习
	FedPU	PU Learning	无	客户端间一致性	无
	AdaFedSemi	伪标记	无	无	调整置信度阈值和参与率
	DS-FL	集成未标记	无	无	传输日志、无参数
标签在服务器端	SemiFL	伪标记	无	减少熵的平均值	
	FedMatch	伪标记	无	客户端间一致性	分散学习和稀疏学习

1.2.2 数据生成方法

生成合成数据的主要策略集中在生成模型上，这些模型旨在从现有数据集中学习丰富的表示，并随后生成新的样本。这些方法在多个领域中得到了应用，如医学研究、金融、教育和各种工业应用，其中高质量合成数据的生成对于隐私保护和有限数据集的增强都至关重要，以下将对常见的数据生成方法进行详细介绍。生成合成数据的主要策略集中在生成模型上，这些模型旨在从现有数据集中学习丰富的表示，并随后生成新的样本。这些方法在多个领域中得到了应用，如医学研究、金融、教育以及各种工业应用，在这些领域中，生成高质量的合成数据对隐私保护和有限数据集的增强至关重要。

基于自编码器的方法：这一类别中的先驱性工作是自编码器（Autoencoder，AE）[10]，其训练目的是将高维输入映射到低维潜在编码中，然后从这些编码中重构原始数据。通过强制隐藏（潜在）层降维，AE 有效地学习到了压缩表示。然而，纯粹的 AE 是确定性的，缺乏灵活采样的直接机制。变分自编码器（Variational Autoencoder，VAE）[11] 通过引入概率潜在变量框架解决了这一问题，从而可以从学习到的潜在分布中采样出新的、未见过的数据点。这一扩展极大地拓宽了基于自编码器模型在合成数据生成中的潜在应用范围。在 AE 和 VAE 的基础上，Xu,Lei 等人[12] 提出了一种针对表格数据生成和重构的改进方法。其方法更准确地建模了潜在变量与表格特征的联合分布，而后者可能包含连续和离散属性。通过关注不同类型变量之间的相互作用，该方法确保生成的合成数据忠实地反映了真实世界表格数据集（如医疗和教育数据）中存在的复杂依赖关系。

基于生成对抗网络 (Generative Adversarial Networks，GANs) [13] 的方法：自 2014 年引入以来，GANs 对生成建模领域产生了重大影响。其基本概念基于博弈论：两

个网络——生成器（Generator，G）和判别器（Discriminator，D）——在对抗循环中训练。判别器学习区分真实数据和合成数据，而生成器则试图生成能够欺骗判别器的样本。当判别器无法再区分真实数据与生成数据时，这个极小极大博弈就告一段落，这表明生成器已经捕捉到了底层分布的统计特征。在 GANs 出现后不久，[14] 提出了条件生成对抗网络（Conditional Generative Adversarial Networks，CGANs），其中生成器和判别器均基于辅助信息（如类别标签或特定输入变量）进行条件化。这一扩展框架使得生成能够针对特定类别或属性进行定向，实际上将原本的无监督设置转变为有监督或半监督范式。尽管取得了这些进展，传统 GANs 往往仍然面临梯度消失或模式崩溃的问题, 为了解决这些问题，Banach 等人用 Wasserstein 距离替换了 Jensen–Shannon (JS) 和 Kullback–Leibler (KL) 散度，从而诞生了 Wasserstein GAN (WGAN) [15,16]。

针对表格数据的 GANs：虽然 GANs 最初因图像合成而受到欢迎，但它们在处理通常具有异构特征类型、不平衡和复杂依赖关系的表格数据集时，同样证明既具有挑战性又极具价值。针对这些挑战，[17] 提出了表格 GAN（Table GAN，TGAN），该方法在生成器网络中应用了长短期记忆单元，并在判别器中采用了多层感知器，从而使深度架构适应于表格数据生成（例如健康记录或学生成绩数据）。2019 年，Xu, Lei 等人 [12] 提出了 CTGAN（Conditional Table GAN），一种专门为处理不平衡离散列、多模态连续列以及表格数据固有复杂性而设计的改进型条件 GAN 架构。通过利用条件采样策略，CTGAN 在建模详细分布和生成与真实数据密切对应的表格行方面表现出色。表格 GANs 的进一步发展，Lee 等人 [18] 探索了一种可逆的表格 GAN 框架，该框架将对抗训练与来自可逆神经网络的负对数密度正则化相结合。该方法在训练过程中增加或降低真实样本的对数密度，从而根据隐私和质量目标生成更接近或更远离真实数据流形的合成样本。为了解决模式崩溃并提高样本多样性，Nguyen 等人 [19] 提出了一种双判别器 GAN，该方法结合了 KL 散度和反向 KL 散度，以更好地捕捉真实世界数据分布的多模态特性。Singh 等人 [20] 针对内存限制问题提出了 MeTGAN，在生成器和判别器中采用稀疏线性层，从而显著降低了内存开销，而对合成质量影响不大，这在处理具有高基数类别变量的表格数据集时尤为有利。在扩展 CTGAN 功能方面，Zhao 等人 [21] 提出了 CTAB-GAN，能够对连续、类别和混合类型的特征进行建模。该方法能够有效处理偏态分布和多样化的特征类型，通常在与真实数据的相似性和下游任务准确性方面优于其他基线方法。Engelmann 等人 [22] 提出了一种针对同时包含数值和类别变量的表格数据的条件 Wasserstein GAN，通过引入辅助分类器，使模型在生成逼真合成数据的同时提升了下游任务（如信用风险评估）中分类器的性能。

基于扩散概率模型 (Denoising Diffusion Probabilistic Models，DDPMs) [23] 的方

法：与 GANs 一样，DDPMs 因其生成高保真样本的能力而受到关注。DDPMs 在多个前向步骤中系统地向数据添加高斯噪声，将原始数据分布转化为一个可处理的先验分布（例如标准正态分布）。随后，该模型通过多步马尔可夫链学习逆向（去噪）过程，逐步去除噪声以恢复数据分布。尽管 DDPMs 在某些领域（尤其是计算机视觉）生成了最先进的合成样本，但其多步特性可能会带来较高的计算开销。TabDDPM 结合了针对表格特征混合性质精心设计的噪声添加与去除调度，同时考虑数值和类别变量的分布，展现了学习多列之间复杂关系的能力，并在生成结构化数据时缓解了纯图像中心扩散方法的缺陷 [24]。

1.3 论文研究的主要内容

学位论文……

1.4 论文组织结构

本文……

第 2 章 相关理论介绍

2.1 本章引言

本章引言……

2.2 联邦学习

在过去几年中，深度学习取得了迅速进展，尤其在人工智能和智能制造等领域 [1]。然而，一些障碍依然存在，包括数据治理方面的挑战以及敏感信息的保护。联邦学习作为一种有前景的方法浮现出来，用以应对这些问题。接下来的部分将对联邦学习进行概述，探讨其概念、架构框架和训练方法，从而加深我们对其演变过程和潜在应用的理解。

2.2.1 联邦学习概念

在传统的集中式机器学习模式中，所有数据通常被集中存储在同一个中心服务器上进行训练与推理。然而，这种“数据集中化”的方式在实际应用中面临着多重挑战。首先是“数据孤岛”现象：由于不同机构或系统的数据缺乏互联互通，数据共享或合并分析变得困难[25]。例如，医院之间常因合规与隐私限制而无法共享病人数据，金融机构也难以交换客户交易信息。其次是隐私与安全隐患：在大数据时代，用户产生的大量敏感信息（如医疗记录、财务数据等）被集中保存在少数服务器上，一旦发生数据泄露，将给用户和机构带来巨大损失[26]。最后，随着诸如欧盟《通用数据保护条例》（General Data Protection Regulation，GDPR）等越来越严格的数据保护法规不断出台，跨境或跨区域的数据传输受到严苛限制，集中化的数据处理模式在许多场景下已难以适用。

在这种背景下，如何既能充分利用分散的海量数据，又能有效保护数据隐私、符合合规要求，成为了机器学习领域亟待解决的关键问题。分布式数据的异构性、地理分散性以及敏感性，使得传统方法难以直接在统一服务器上进行整合和建模。针对这一问题而提出的联邦学习，为分布式数据管理与隐私保护提供了新的解决思路。

联邦学习（Federated Learning，FL）最早由 Google 研究团队于 2016 年提出[27]，并在 2017 年的工作中进一步完善[26]。它的核心思想是：在不集中原始数据的情况下，由多方（如分布于不同地理位置或不同组织的设备、服务器等）各自在本地对模型进行训练，并将训练得到的模型参数或梯度发送到中心服务器进行聚合，从而

形成一个全局模型[28]。在整个过程中，原始数据始终保留在各自的本地存储，不会被外传，极大地降低了数据泄露的风险。

联邦学习，有时也被称为协同学习，是一种机器学习方法，其中多个参与方——通常称为客户端——在不将数据集中存储的前提下，共同开发一个模型[29]。该方法的一个关键特点是数据的内在异质性；由于数据仍然分布在各个客户端，各地点可用的样本往往不符合独立同分布（i.i.d.）的模式。联邦学习主要是由数据隐私、数据最小化和访问权限等问题所驱动。其应用涵盖了多个领域，包括国防、电信、物联网以及制药。联邦学习旨在开发机器学习算法——例如深度神经网络——以便在不同节点上存储的各个本地数据集上进行训练，而无需直接交换原始数据。相反，每个节点在自身数据上训练本地模型，并在定期的时间间隔内互相共享模型参数（例如权重和偏置），其系统架构如图 2-1 所示。

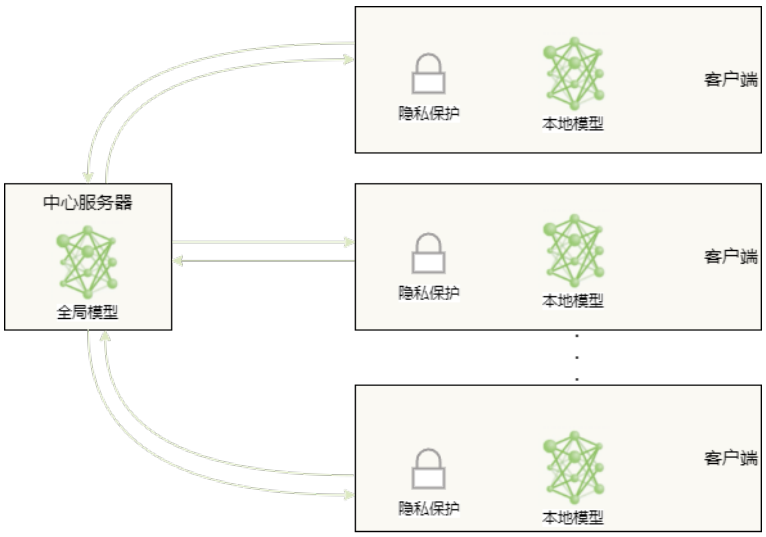

图 2-1 联邦学习系统架构

Fig. 2-1 Federated learning system architecture

在整个工作流程中，中心服务器首先会初始化全局模型并将模型参数广播给所有客户端。各客户端在接收到这些参数后，利用本地数据进行训练，然后将更新后的本地模型（不包含原始数据）发送回中心服务器。聚合器再对所有客户端的模型参数进行加权平均等聚合操作，以生成新的全局模型。这个过程反复进行，直至模型收敛或达到预先设定的最大迭代次数。

联邦学习与分布式学习的主要区别在于对本地数据集特性所作假设不同[30]。分布式学习旨在利用并行计算能力，通常假设每个本地数据集均为独立同分布（i.i.d.）且规模大致相同。而联邦学习则不做这些假设，它设计用于处理异构数据，数据集的规模可能存在显著差异。此外，联邦学习中的客户端往往较为不可靠，

因其通常依赖于如 Wi-Fi 等较不稳定的通信方式，并运行在电池供电的设备（如智能手机或物联网设备）上，因此面临更高的失败或中途退出风险。相比之下，分布式学习通常依赖于数据中心内的节点，这些节点具备强大的计算能力，并通过高速网络互联[29]。联邦学习的核心思想是在不共享数据的情况下训练模型，仅交换模型参数。

2.2.2 联邦学习分类

联邦学习可根据特征空间和样本空间的重叠情况，主要划分为横向联邦学习（Horizontal Federated Learning）、纵向联邦学习（Vertical Federated Learning）以及联邦迁移学习（Federated Transfer Learning）[25,28]，如图 **??** 所示。这种划分方式能够帮助研究人员与工程实践者在不同场景下选择适宜的联邦学习方案，从而更高效地利用分布式数据进行模型训练。

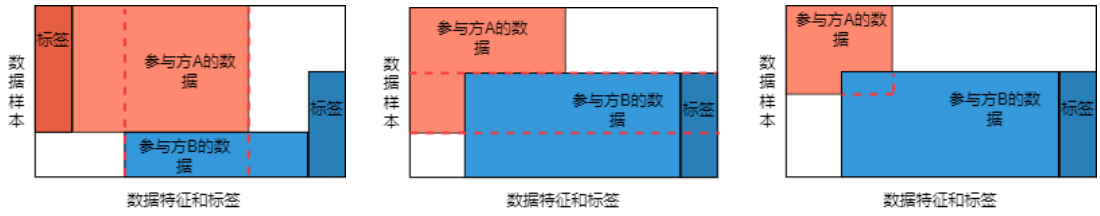

图 2-2 联邦学习分类。(a) 横向联邦学习；(b) 纵向联邦学习；(c) 迁移联邦学习

Fig. 2-2 Federated Learning Classification. (a) Horizontal Federated Learning; (b) Vertical Federated Learning; (c) Transfer Federated Learning

在横向联邦学习（HFL）中，各数据拥有方在特征空间上较为相似，但用户或样本并不重叠[25,31]，如图 2-2 (a) 所示。举例来说，若多家银行都想联合训练一个信用风险评估模型，但它们各自的用户群体并无明显重叠，且每家银行都有较为类似的特征字段（如年龄、职业、收入等），这种场景便适合采用横向联邦学习。通过让各银行保留本地数据，仅传递模型参数或梯度进行聚合，既能提升模型的泛化能力，又在最大程度上保护了用户隐私。

与之相对，纵向联邦学习（VFL）则适用于样本空间高度重叠，但特征集差异较大的场景[32,33]，如图 2-2 (b) 所示。例如，某些银行与电商平台在用户群体上可能有很大交集，却掌握了不同维度的用户信息：银行侧有用户财务信用数据，电商侧则有用户消费行为数据。在这种情形下，联合建模可以充分利用双方不同来源的特征信息，从而获得更全面、更准确的刻画。当各方在训练过程中只需对对齐后的公共用户进行联合梯度更新时，无需共享原始数据，依然能够保证个人隐私与数

据安全。

第三种形式是联邦迁移学习（FTL），其特点在于样本空间与特征空间均不完全重叠[25,33]，如图 2-2 (c) 所示。在部分数据量较少或者特征不足的场景里，可以通过迁移学习的方法来补充数据或特征不足的问题。此时，不同组织或机构拥有的用户群体与特征集合可能仅存在少量重叠，但依然可通过联邦迁移学习的思想，将部分预训练模型的知识进行迁移，从而提升任务的性能或模型的泛化能力。这种方法特别适用于小样本或垂域知识不足的情景，有效地拓展了联邦学习在更多应用场景下的可行性。

综上所述，横向联邦学习、纵向联邦学习及联邦迁移学习分别应对了特征空间或样本空间上的不同重叠情况。它们在数据类型、业务场景和技术实现上各有侧重，却都遵循了联邦学习的核心理念：在不直接交换数据的前提下，实现多方协作与知识共享。这种多元化的联邦学习形式为不同行业和应用需求提供了灵活的选择，也在隐私保护和跨域协作方面展现出巨大的潜力。

2.3 半监督学习

2.3.1 半监督学习概念

在实际任务中，有标注数据往往十分稀缺，且标注过程需要投入大量的人力与时间成本[34]. 然而，大量的未标注数据通常容易获取，如果能够加以有效利用，便能显著提升模型的学习性能[35].

半监督学习的核心思想在于同时利用少量标注数据与大量未标注数据进行联合训练，从而提升模型的泛化能力[36]. 通过对未标注数据的“挖掘”或“自学习”，半监督学习有效弥补了标注样本不足所带来的数据缺陷[34].

半监督学习通常建立在以下三个基本假设之上：首先，平滑性假设（Smoothness Assumption）认为在特征空间距离较近的样本，其标签也应相近；其次，聚类假设（Cluster Assumption）指出同一簇中的样本倾向于具有相同的标签；最后，低密度分隔假设（Low-density Separation Assumption）认为不同类别被低密度区域所分隔[35].

与监督学习相比，半监督学习更加强调对未标注数据的有效利用；而与无监督学习相比，半监督学习则依靠少量标注样本对模型进行指导，使模型能够在庞大的未标注数据中学习到更具区分度的特征[36].

2.3.2 半监督学习方法

(1) 伪标签生成方法

该方法基于自训练（Self-training）假设[35,37]，通过为未标记数据生成伪标签

来实现半监督学习，并将这些伪标签作为额外的训练数据。未标记数据对应的损失项通常可以表示为

$$L_u = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} R\Big(y_i^j,\ p(x_i \mid \theta)\Big), \tag{2-1}$$

其中，$p(x_i \mid \theta) = \mathrm{softmax}\big(f(\theta, x_i)\big)$ 为模型 $f(\theta)$ 在样本 x_i 上的预测概率分布，而 y_i^j 则指生成的伪标签，可以是一热编码或概率向量。为了实现熵最小化，通常采用一热编码的形式，使模型在预测时更加自信，从而将决策边界放置在低密度区域[34]. 当使用软标签时，为了达到类似的熵最小化效果，往往会对伪标签进行锐化处理，例如采用 0-1 阈值化：

$$y_i^j = \mathrm{argmax}\Big(p(\theta', x_i)\Big), \tag{2-2}$$

即直接以当前模型的预测作为伪标签。

在上述表达式中，$R(\cdot,\cdot)$ 表示用于度量差异的距离函数，常用的包括均方误差（MSE）或交叉熵等[36]. 为了进一步保留那些置信度较高的伪标签样本，可对公式 (2-1) 进行更深层次的优化，通过引入一个筛选函数 $\mathrm{Filter}(\cdot)$ 来实现：

$$L_u = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} \mathrm{Filter}\Big(R\big(y_i^j,\ p(x_i \mid \theta)\big)\Big), \tag{2-3}$$

使得模型仅聚焦于高可信度的伪标签样本。

此外，为了构建总的目标函数，常将带标签数据上的监督损失与未标记数据上的伪标签损失结合，得到

$$L = L_s + \alpha\, L_u, \tag{2-4}$$

其中 α 用于平衡两部分损失，既可视为一个固定超参数，也可以随训练迭代而动态调整（如 $\alpha = \alpha(t)$）。通过对 L 的梯度进行反向传播并更新模型参数，模型会在每次迭代中不断重新预测未标记数据的伪标签，从而持续提升模型对未标记数据的刻画能力，直至收敛或满足预设的终止条件。

(2) PU（Positive and Unlabeled）学习

正例与未标注（Positive and Unlabeled，PU）学习是一种有效的半监督学习策略，旨在仅依赖少量确证的正例样本与大量未标注数据来构建分类模型，而无需明确的负例样本[38,39]. 在传统的二分类场景中，研究人员通常需要完整且平衡的正负例数据集以训练模型。然而，随着数据规模的指数级增长，人工标注的成本与难度显著上升，大量真实世界数据缺乏可靠标记。为应对这一挑战，PU 学习通过将未知标签样本纳入训练过程，减少对全面标注的依赖，同时利用少量的正例样本来学习准确的决策边界。

近年来，PU 学习方法在多个领域均展现出卓越潜力。与传统方法相比，PU 学习无需显式负例，也能有效定位负类分布或异常模式，并对类别不平衡问题进行修正[38]. 此外，PU 学习对大规模未标注数据的需求大幅降低了标注成本，尤其适合工业监测与异常检测等场景：在此类应用中异常事件极为稀少，未标注数据往往代表了正常状态，因此仅使用少量正例便可使算法逐步“发掘”潜在负类特征，从而显著提升分类器的稳健性。

一种直接而高效的方法是将已标记的正例样本视为正类，将未标注样本简单视为负类进行训练[38]. 尽管此策略在理论上似乎过于简化，但有研究表明其依然能取得令人满意的实验效果。根据 Elkan 与 Noto[38] 的理论，当满足一定假设（如正例样本是随机抽取的标注数据）时，由正例与未标注数据训练的分类器，其输出分数与同时拥有完整正负例的分类器之间存在正比关系。因此，在对样本进行排序或度量其成为正例可能性时，该方法的排名性能几乎可以与使用完整标注数据的模型相媲美。

Mordelet 与 Vert[39] 提出了针对 PU 学习的改进袋装法（PU Bagging）。该方法的主要流程如下：

1. 自助采样（Bootstrap）：在构造训练集时，保留所有正例样本，同时从未标注样本池中采用有放回的方式随机抽取子集，并将其视为伪负例。

2. 分类器训练：使用“正例 + 伪负例”的自助样本来训练分类器。

3. 袋外评估（Out-of-Bag）：将训练好的分类模型应用于未参与当前自助抽样的未标注样本，并记录预测得分。

4. 迭代与聚合（Ensemble）：重复上述采样与训练过程，每个未标注样本会累计多个袋外预测分数，最后以平均分作为最终结果。

实验证明，当正例样本十分稀少或未标注数据中的负类所占比例本就较低时，PU Bagging 往往表现优异，甚至可超越一些更为复杂的 PU 学习算法。此外，该方法在处理大规模未标注数据时展现了较高的计算效率，更适合实际工业场景中的在线或分布式训练需求。

目前，多数 PU 学习算法通常可以归纳为“两步法”，其主要步骤为首先进行可靠负样本识别，从未标注数据中筛选出高置信度的负例子集，尽管其规模可能较小，但标签质量较高，为后续训练提供基准。接着进行迭代分类器训练，将已知正例与前一步识别的可靠负例合并训练初始分类器，然后将该分类器应用到剩余的未标注数据。这个过程通常会进行多次迭代，随着更多可靠负例被标识并加入训练集，分类器的决策边界不断得到精炼，直至满足停止准则或取得收敛。

虽然两步法在机制上与 Shubham Jain 的“伪标签”策略相似，但其针对 PU 问题进行了专门设计，通过动态更新和扩充可靠负集来逐步提升分类器性能。相比于一般的半监督学习，两步法在 PU 场景下更强调对未知类别分布的主动探索，因而能在缺乏负例注释的环境中持续改进模型的判别能力。

2.4 表格数据生成

表格数据生成（Tabular Data Generation）是指通过一定的算法或模型，基于现有真实数据或先验知识，合成出具有与真实数据相似统计分布和特征的虚拟表格数据。随着大数据与隐私保护需求的不断增长，表格数据生成技术在医疗、金融、电商、社交媒体等领域具有越来越重要的应用价值。该技术不仅可以用于数据扩增（Data Augmentation）和算法测试，还能够在保证隐私安全的前提下，为数据分析和机器学习模型提供更多的训练样本 [40]。

表格数据生成技术的核心目标在于在保持数据特征分布与关联关系的同时，最大程度地提升所生成数据的真实性与多样性，其典型生成过程需要综合考虑多个方面。生成数据需与真实数据在数值特征和类别特征上保持相似的分布特征，以避免重要特征的偏差或缺失，同时在多列或多特征维度的表格数据中捕捉并保留不同字段间复杂的相关性 [41]。此外，生成过程还需兼顾隐私保护，既要避免暴露个体敏感信息，又要保证数据的可用性和真实性，而为了满足大规模数据需求，生成技术的效率和可扩展性也成为重要的考量指标。

当前，针对表格数据的生成方法主要分为基于统计模型和基于深度生成模型两大类，前者通过参数化或非参数化的统计分布对真实数据进行建模，后者则更多依赖神经网络，尤其是生成对抗网络（GAN）和变分自动编码器（VAE）等框架来生成高维度和多样化的数据 [42]。基于统计模型的方法通过对单变量分布及变量间相关关系建模，能够生成近似真实分布的数据，实现简单且具备一定可解释性，但在数据相关性复杂或维度较高时容易出现失配。而近年来，GAN 和 VAE 等深度生成模型在图像、文本等领域取得显著成果后，也逐渐应用于表格数据生成，通过生成器与判别器间的对抗训练，能捕捉多维特征关系并生成高保真合成数据，然而这类方法通常需要大量样本训练，且在超参数选择与模型可解释性方面仍面临较大挑战。

2.5 本章小结

本章介绍了……

第 3 章 基于多方联邦的半监督学习方法研究

3.1 本章引言

在当今数据驱动的世界中，机器学习模型在众多现实世界的应用领域中发挥着至关重要的作用，例如金融风控[43,44]、医疗诊断[45,46]、智能制造[47,48]以及智能交通[49,50]等。这些模型的训练通常依赖于大规模数据集，其中包含敏感信息，例如个人身份信息、行为数据、社交关系以及上下文数据（如地理位置、时间戳和环境状态）。为了提升模型的预测性能，传统方法通常采用集中式数据存储和处理方式，即将所有数据汇总至中央服务器进行训练。然而，这种集中式存储方式带来了显著的隐私风险，包括数据泄露、未经授权的访问以及潜在的数据滥用。此外，由于数据通常由多个机构或组织分别持有，并受到严格的隐私保护法规（如 GDPR 和 CCPA）的约束，跨多个数据所有者整合数据以训练高效的机器学习模型变得越来越困难，甚至在某些情况下是不可能的。

为了解决这一问题，联邦学习（Federated Learning, FL）作为一种分布式机器学习方法[26]，提供了一种隐私保护的解决方案。联邦学习允许多个数据所有者（客户端）在不共享原始数据的情况下，通过交换中间参数（如模型梯度或模型权重）来协同训练机器学习模型。现有的联邦学习研究主要集中于监督学习场景，即假设所有客户端都拥有完全标记的数据。然而，在许多实际应用中，数据通常是不完全标记的，这可能是由于标注成本高昂、缺乏领域专家、资源受限或标注工具不足等因素所导致的。因此，近年来，一些研究开始探索半监督学习（Semi-Supervised Learning, SSL）在联邦学习中的应用，特别是仅涉及正样本和未标记数据的问题，这类问题被称为 PU（Positive and Unlabeled）学习问题。针对 PU 问题以及其他有限标记数据场景，研究人员提出了一些方法，例如：FedPU 算法[6]：针对联邦学习环境中的 PU 问题，每个客户端仅对其数据集的一小部分进行标记，并利用 PU 学习策略提升模型性能。RSCFed 算法[2]：通过聚合多个子共识模型来更新全局模型，以解决非独立同分布（Non-IID）本地客户端的不均匀可靠性问题。FedMatch 算法[5]：通过优化客户端间的一致性和参数分解，提高有限标记数据下的联邦学习效果。AdaFedSemi 算法[7]：利用设备上的标记数据和云端的未标记数据，基于多臂赌博机算法优化客户端参与度和伪标签质量。

然而，这些研究主要集中在所有数据所有者共享相同特征空间的场景，即横向联邦学习（Horizontal Federated Learning, HFL）（见文献[25]）。在 HFL 场景下，每个数据所有者拥有相同的特征集，但不同的样本。然而，在许多实际应用中，数据所有者可能共享相同的样本 ID 空间，但其特征空间不同，例如：金融与医疗数据

融合：银行可能拥有用户的金融交易记录，而医院则掌握用户的健康数据。两者希望联合训练一个信用风险评估或健康预测模型，但无法直接共享数据。智能制造与供应链优化：制造商可能拥有生产数据，而供应商则掌握物流信息。两者希望协同优化供应链效率，但数据属于不同的企业，难以直接整合。多机构联合风控：不同的金融机构可能分别持有部分用户的信用信息，但由于竞争关系和隐私法规的限制，无法直接共享数据。

上述场景属于纵向联邦学习（Vertical Federated Learning, VFL）[25]，即数据所有者共享相同的样本 ID 空间，但特征空间不同。尽管 VFL 在隐私保护机器学习领域具有重要应用价值，但目前针对 VFL 进行半监督学习的研究仍然较少，尤其是在 PU 学习问题上的研究几乎为空白。在 VFL 场景下，研究了一种特定的 PU 学习问题，其特征如下：

1. 数据分布：多个数据所有者（机构）持有部分重叠的样本 ID，但特征空间不同，且数据具有专有性。

2. 数据类型：某一方（通常是目标方）仅拥有正样本数据，而其他方仅持有未标记数据，且无法直接访问彼此的数据。

3. 目标：所有方希望联合训练一个机器学习模型，以从未标记数据中识别可靠的正样本，同时保护各方数据的隐私。

上述第二个特征对应于 PU 学习问题，即从正样本和未标记数据中学习有效的分类模型。然而，传统的 PU 学习方法 [51–54] 通常假设正样本和未标记数据都可供训练使用。然而，在 VFL 场景下，正样本和未标记数据分布在不同的数据所有者之间，且无法直接共享，这引入了一个新的挑战。将此问题称为未标记数据缺失的 PU 学习问题（Unlabeled-Data-Deficient PU, UDD-PU）。传统的 PU 学习框架无法直接解决这一问题，因为它们通常假设训练过程中可以同时访问正样本和未标记数据。

为应对这一挑战，本章提出了一种新的方法——纵向联邦学习与正样本和未标记数据（Vertical Federated Learning with Positive and Unlabeled data, VFPU）。VFPU 允许多个数据所有者在不共享原始数据的情况下，协同训练一个机器学习模型，以从未标记数据中识别可靠的正样本。VFPU 主要具有以下特点：

- 隐私保护：通过联邦学习框架，各方仅交换加密的中间计算结果，而不直接共享原始数据，确保数据隐私性。

- PU 学习优化：结合 PU 学习策略，在 VFL 场景下有效利用未标记数据，提高模型的分类性能。

- 适应多种应用场景：VFPU 可广泛应用于金融、医疗、智能制造等多个领域，解决数据孤岛问题，同时满足隐私保护要求。

本章的结构安排如下：3.2 节将对 UDD-PU 问题进行分析定义，3.3 节将系统性地阐述多方联邦半监督学习的问题定义与方法框架，并详细介绍 VFPU 的执行流程和算法设计。3.3 节在多个真实数据集上进行实验，验证 VFPU 方法的有效性，并对实验结果进行深入分析。最后，3.4 节对本章的研究工作进行总结。

3.2 未标记数据缺失问题（UDD-PU）的分析与定义

在本章节讨论的语境中，首先探讨一个分布式数据场景：K 个独立的数据所有者各自掌控着大型数据集的不同部分，并存在一个中央服务器作为协作协调者。每个数据所有者持有的数据被系统地组织成矩阵形式，具体记作 $\mathcal{D}_k$（下标 k 表示第 k 个数据所有者）。该矩阵中，每行对应一个独立样本（即具体的数据实例或观测值），每列对应特定特征（即样本的可测量属性或特性）。从全局视角，整个分布式数据集可抽象为结构化三元组 $(\mathcal{I},\mathcal{X},\mathcal{Y})$，其中：$\mathcal{I}$ 为样本 ID 空间，包含所有样本的唯一标识符；$\mathcal{X}$ 为特征空间，涵盖所有样本的全部特征维度；$\mathcal{Y}$ 为标签空间：包含所有可能的分类标签或类别。

在传统纵向联邦学习（Vertical Federated Learning, VFL）框架中，通常存在一个关键假设：至少有一个参与方持有其数据部分的完整标签集，从而能够实现有监督的预测模型训练。然而，这一假设在大量现实场景中往往难以成立，因为获取全量标注数据集面临严峻挑战。这种挑战主要源于实际商业环境中的多重约束，包括但不限于：隐私保护法规对数据使用的限制、高昂的人工标注成本、行业竞争导致的数据孤岛现象、敏感信息共享的法律风险以及动态数据更新带来的标注滞后问题。

为了更清晰、更具体地理解这一概念，现在考虑一个具体的示例场景，其中涉及三位不同的数据拥有者，将其分别称为 A 方、B 方和 C 方。这三方各自持有敏感数据——这些信息对其业务至关重要，但由于隐私和安全方面的考虑，他们不愿直接公开。因此，他们需要在确保数据隐私严格保护的前提下，以安全的方式进行合作，充分利用各自的数据资源。

在该场景中，所涉及的数据样本可以被划分为两大基本类别：正类（positive）和负类（negative），这可以对应于某些应用场景中的理想（或期望）行为与非理想（或不期望）行为。一个值得注意的特点是，这三方在样本标识符（ID）上存在一定程度的重叠，即某些样本可能同时出现在多个参与方的数据集中，但每一方所收集或观察到的特征可能有所不同。

具体而言，A 方持有一个样本集合，记作 P，该集合的独特之处在于它仅包含正类样本——即被明确标记为正类的数据实例，不包含任何负类或不确定的样本。而 B 方和 C 方共同持有一个未标注数据集，记作 U，其中的样本类别（正类或负类）未知。需要特别强调的是，U 不包含 A 方数据集 P 中的任何样本，即 U 是一个独立的数据集，不与 A 方的正类样本重叠。

这三方的主要目标是通过合作共同训练一个推荐模型，该模型的作用是分析未标注数据集 U，并从中识别出可以被可靠归类为正类的样本。这一过程的输出结果记作 R，即从 U 中成功提取出的可靠正类样本集合。一旦这些样本被识别出来，R 将被提供给 A 方，使其能够基于这些推断出的正类实例向客户或用户提供精准的产品推荐，从而提升其业务能力和市场竞争力。

该问题设置引入了一个重大挑战，使得传统的纵向联邦学习（VFL）算法在其标准形式下无法适用。核心问题在于，参与的各方——A 方、B 方和 C 方——都不拥有完整的标签数据。在传统的 VFL 方案中，至少需要有一方持有完整的标注数据，以便模型能够从监督信号中学习，并在联邦系统中传播。然而，在本问题中，A 方仅拥有正类样本的标签（即数据集 P），而 B 方和 C 方的数据集 U 完全没有标签，这导致了一个关键的缺口，使得标准的 VFL 技术无法直接应用。

为了解决这一挑战，可以考虑采用半监督学习（semi-supervised learning）策略，该策略专门用于处理包含部分标注数据和未标注数据的场景。在本问题背景下，一个特别相关的方法是 PU 学习（Positive and Unlabeled learning，正类与未标注学习）。PU 学习是一种特殊的半监督学习技术，它利用一组带有正类标签的样本（P）和一组未标注样本（U）进行训练，目标是在未标注数据集中识别出正类实例。PU 学习特别适用于缺少负类标签或负类标签不可靠的情况，因此在本问题中具有较大的应用潜力。

然而，PU 学习在本场景下的直接应用存在一个关键限制：传统的 PU 学习方法假设学习者可以同时访问 P 和 U，从而能够直接对比正类样本和未标注样本，并进行有效的训练。然而，在本联邦学习场景中，这一假设并不成立，因为 A 方仅持有 P，而 U 分布在 B 方和 C 方手中。由于隐私保护的要求，各方无法简单地将数据汇总到一个集中存储库，也不能自由共享数据给 A 方，这使得传统 PU 学习方法难以直接应用，成为本问题的一个重要挑战。

在这种情况下，A 方希望将推荐服务集成到其业务运营中，但面临一个全新的复杂挑战，将其称为“未标记数据缺失的 PU 学习问题”（Unlabeled-Data-Deficient PU, UDD-PU）。该问题的核心特征在于，A 方仅能访问正类样本集 P，但无法直接获取未标注数据 U，这使得传统的 PU 学习方法无法直接应用。此外，由于数据在联邦学习环境下分布——即 U 分散存储在 B 方和 C 方手中——并且所有参与方都

施加了严格的隐私保护要求，因此现有的 PU 学习方法若不进行重大改进，无法直接适用于该场景。

因此，UDD-PU 问题代表了联邦学习、半监督学习和隐私保护计算的独特交汇点，需要创新性的解决方案，以便在这些约束条件下实现有效的协作和模型训练。

总的来说，这一扩展解释突出了该场景的复杂性、传统方法的局限性，以及新问题的提出，为进一步的研究探索或方法论发展奠定了详细的基础，可用于撰写相关研究论文。

3.3 基于多方联邦的半监督学习方法

在推荐系统中，有效利用多个参与方的数据，同时应对特定挑战至关重要。在某些场景下，会出现 UDD-PU 学习问题，即某一方（在本例中为 A 方）缺乏足够的未标记样本，导致无法使用传统的正例-未标记（PU）学习技术来有效训练推荐模型。这种未标记数据的不足会影响模型的泛化能力，导致性能下降和推荐结果不够准确。为了解决这一问题，本章提出了 VFPU 算法，这是一种结合了纵向联邦学习框架与 PU 学习技术的新方法。VFPU 旨在解决 A 方未标记样本不足的问题，通过更好地利用分布式数据资源，提升推荐模型的性能。

在本章中，基于 VFPU 的推荐过程主要包括三个核心步骤：数据预处理、加密样本对齐以及 VFPU 算法的执行。这些步骤共同构成了一个稳健的流程，以在隐私保护的协作环境中提升推荐的准确性。VFPU 的主要目标是在多个参与方持有的未标记数据集中识别可靠的正例样本。通过精准定位这些可靠的正例样本，模型能够更好地区分正例和负例，即使在缺乏明确标注的负例数据的情况下——这是 PU 学习场景中的常见挑战。

这一能力在推荐系统中尤为关键，因为用户偏好通常是通过交互隐式表达的，而非通过显式标签标注。因此，模型需要从有限或噪声较大的数据中推断用户偏好。通过识别这些可靠的正例样本，模型能够更深入地理解正例实例的特征，例如用户偏好或产品相关性，从而最终生成更准确、更个性化的产品推荐，以满足 A 方的需求。

对于 A 方而言，由于数据资源有限，其推荐模型的性能可能受到约束。然而，通过这一联邦学习方法，A 方可以利用 B 方和 C 方的丰富数据集，从而显著增强推荐模型的稳健性。同时，确保数据隐私的保护，并促进多个参与方之间的无缝协作，严格遵循联邦学习的核心原则。在整个过程中，数据隐私得到了精心维护，每个参与方的原始数据始终保持本地化，仅共享模型更新或聚合后的信息，而不会泄露敏感的个人记录。

这种隐私保护的协作方式在图 3-1 中得到了展示，该图详细说明了推荐过程的具体流程。本章的后续章节将对这一过程中的各个步骤进行深入探讨，以提供对该方法的全面理解。

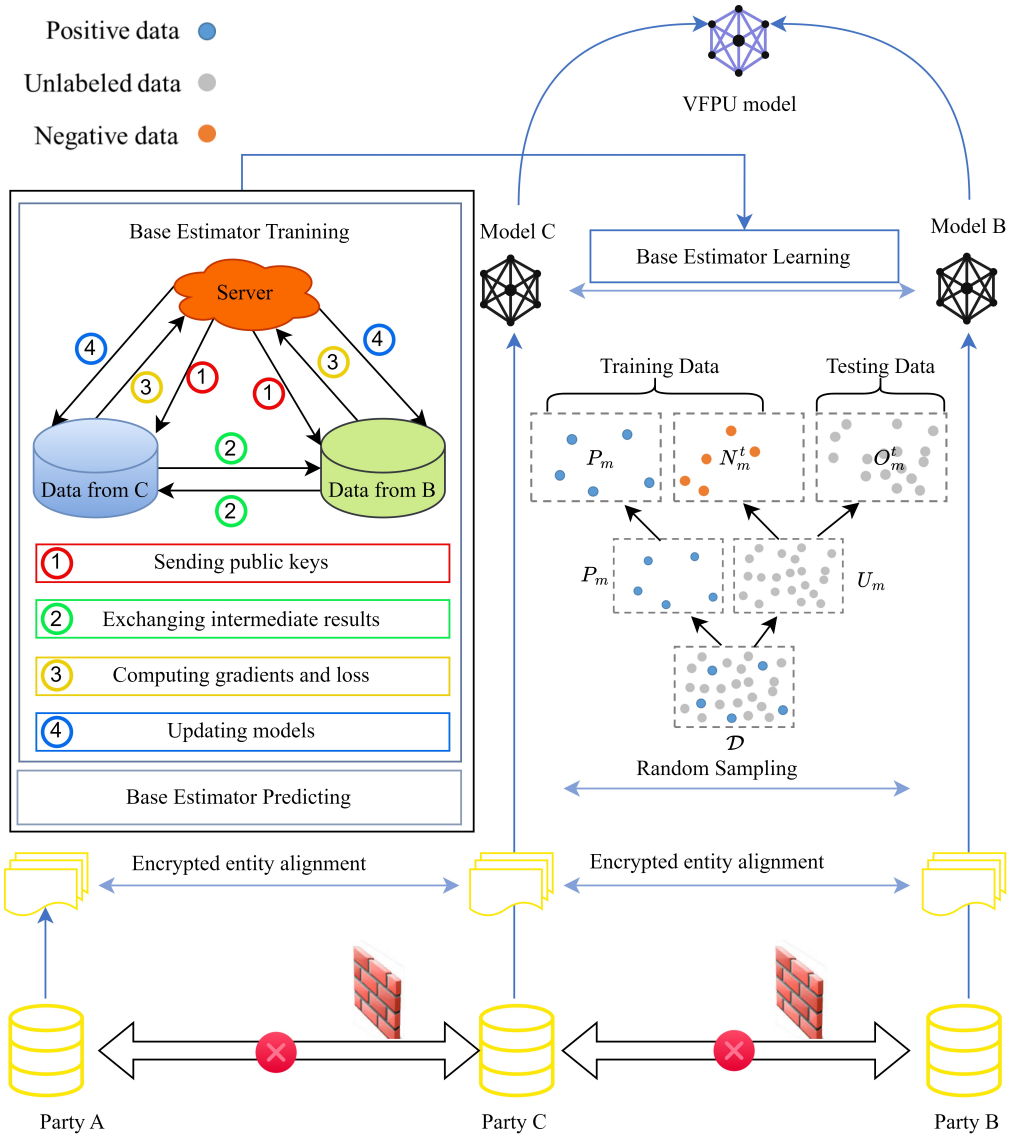

图 3-1 VFPU 算法总体流程

Fig. 3-1 The overall process of the proposed VFPU algorithm

3.3.1 数据预处理与加密样本对齐

为了有效地训练模型，VFPU 框架包含了一个两部分的预处理阶段：数据预处理和加密样本对齐。这些过程确保了来自 A、B 和 C 方的数据集在不损害隐私的情况下进行协调和安全对齐。

(1) 数据预处理

对 A、B 和 C 方持有的异构数据应用各种预处理技术，包括数据清洗、规范化和特征编码。数据清洗是第一步，它涉及解决常见的数据质量问题，例如通过插补或删除处理缺失值，消除重复项，并解决数据收集过程中可能出现的错误。这些努力确保数据集可靠，并且没有可能扭曲模型性能的噪声。规范化紧随其后，确保各方所有特征都达到可比的规模，这对许多机器学习算法的最佳运行至关重要。具体来说，数值特征使用标准化缩放进行规范化，这是一种将数据转换为零均值和单位方差的技术，从而防止范围较大的特征不成比例地影响模型。同时，分类特征通常表示定性属性，例如产品类别或用户人口统计信息，使用独热编码进行处理。此方法通过为每个类别创建二进制向量来将分类变量转换为数字格式，确保模型解释这些特征，而不假设从其他编码方案（如标签编码）可能产生的任何意外的序数关系。总之，这些预处理步骤创建了一个标准化、高质量的数据集，为后续处理做好准备。

(2) 加密样本对齐

在数据预处理之后，三方参与一个安全的样本对齐过程，该过程分两个不同的步骤进行，以同步其数据集，同时保护隐私。这种对齐对于实现纵向联邦学习至关重要，在纵向联邦学习中，不同的方持有重叠样本集的互补特征。

步骤 1: B 方和 C 方对齐其样本 ID 空间，只保留两方都共享的样本，丢弃未对齐的样本。此步骤确保 B 方和 C 方在一个共同的样本集上操作，这是纵向联邦学习的先决条件，在纵向联邦学习中，各方为相同的实体贡献不同的特征集，例如用户或项目。通过关注其样本集的交集，建立了一个一致的基础，其中 B 方和 C 方的特征对应于相同的人或项目，允许模型有效地从组合特征空间中学习。因此，B 方和 C 方现在共享相同的样本，但保持独特的、互补的特征，为协作训练奠定了基础。

步骤 2: A 方和 C 方对齐其样本 ID 空间，不删除任何样本，与步骤 1 相比采用了一种更具包容性的方法。对齐的样本定义为存在于 A 方和 C 方数据集中的样本，而未对齐的样本仅存在于 C 方。这种对齐过程利用了来自 A 方的可用标签信息来丰富 C 方的数据集。具体来说，出现在 A 方和 C 方中的样本在 C 方中被分配一个标签 1，表示它们是正样本，因为它们对应于 A 方数据中已知的正实例，例如与确认的产品交互的用户。相反，C 方中缺少 A 方对应物的样本被分配一个标签 -1，将其标记为可能包含正负实例混合的未标记样本。这种标记策略将 C 方的数据集转换为适合 PU 学习的数据集，其中挑战在于区分未标记集中真正的正样本。通过不丢弃任何样本，此步骤最大限度地利用了可用于训练的数据，同时利用 A 方的正样本来指导该过程。

在完成加密样本对齐后，C 方获得了一个包含正样本和无标签样本的数据集。

这种转换有效地将原始的 UDD-PU 推荐问题——其特点是 A 方缺乏无标签数据——转变为一个纵向联邦训练场景。在这个场景中，PU 学习问题由 B 方和 C 方协作解决，其中 B 方提供额外的特征，C 方提供有标签和无标签的样本。这种合作框架充分利用了所有方的优势，克服了 A 方在数据方面的限制，最终使其推荐系统受益。

为了在样本对齐过程中保护数据隐私，采用了基于盲 RSA 的私有集交集（PSI）协议 [55]。这种密码学技术使所有方能够安全地计算其数据集的交集，而不会暴露除共享样本 ID 之外的任何信息。具体来说，PSI 确保 B 方和 C 方能够识别它们的共同样本，同时 A 方和 C 方能够确定它们的重叠样本，而不会泄露各自数据集的全部内容或关于未对齐样本的任何敏感细节。这种保护隐私的机制是联邦学习范式的核心，促进了各方之间的信任，并确保遵守数据保护标准。随着数据预处理和加密样本对齐的顺利完成，B 方和 C 方的数据集现已完全准备就绪——经过对齐、标记和保护——可以用于执行 VFPU 算法。该算法及其训练过程的细节将在后续章节中详细阐述，基于此处奠定的基础。

3.3.2 基于正样本与未标记数据的纵向联邦学习

纵向联邦 PU 学习（VFPU）算法的目标是在纵向划分的数据环境中，安全且高效地从未标记数据中识别可靠的正例样本。这种场景在现实应用中经常出现，其中不同的组织持有相同数据主体的不同特征，但由于隐私问题或法规限制，无法直接共享原始数据。例如，银行可能拥有金融交易记录，而电子商务平台则掌握在线购物历史，这些数据都与相同的客户相关。识别正例样本（如可能违约的贷款客户或可能响应特定营销活动的客户）至关重要，但由于缺乏完整的标注数据以及特征的分布式存储，这一任务极具挑战性。

VFPU 通过巧妙结合既有的正例与未标记（PU）学习技术和纵向联邦学习框架来应对这一挑战。具体而言，它利用了 Liu 等人 [52] 提出的两步技术，以及 Mordelet 和 Vert [51] 提出的 PU bagging 方法的鲁棒性。这些方法经过调整和整合，形成了一种安全协议，使得各方能够在不泄露数据隐私的情况下进行协作训练。

算法 3-1 详细描述了 VFPU 过程。第一阶段采用两步技术，首先从未标记数据中识别出一组可靠的负例样本。这一过程利用了正例样本，并假设未标记数据中同时包含正例和负例。通过仔细分析特征分布，可以提取出可靠的负例集合，从而更准确地表示数据的真实分布。随后，利用这些可靠的负例样本以及原始的正例样本训练一个初步分类模型，该模型作为第二阶段的基础。

第二阶段引入 PU bagging 方法，以进一步增强 VFPU 的鲁棒性和性能。通过从正例和可靠负例集合中生成多个自助采样（bootstrap）样本，并在每个样本上训

算法 3-1 VFPU 算法训练过程

输入: 参与方 B, C。对齐的数据集 $\mathcal{D}_B$,$\mathcal{D}_C$ 和 ID 集合 $\mathcal{I}_B, \mathcal{I}_C$。最大迭代次数 M，随机采样迭代次数 T 和 θ，其中 θ 是可靠正样本的采样率。

输出: R，可靠正样本的集合。

参与方 C 执行:

for $m = 1, 2, \ldots, M$ **do**
  $P_m = \{i | \mathcal{Y}_i^C = 1,\ i \in \mathcal{I}_C\}$
  $U_m = \{i | \mathcal{Y}_i^C = -1,\ i \in \mathcal{I}_C\}$
  for $t = 1, 2, \ldots, T$ **do**
    $N_m^t = \{U_m | P_m|\}$
    $O_m^t = U_m - N_m^t$
    加密并发送 N_m^t、P_m 和 O_m^t 给其他参与方。
    通知参与方设置训练数据和测试数据。
    S_m^t = Base_Estimator_Learning()
  end for
  $\mathcal{P}_m(u) = \sum_{t=1}^{T} S_m^t(u) / \sum_{t=1}^{T} I(u \in O_m^t),\ \forall \mathrm{u} \in U_m$
  $R_m = \{\mathcal{P}_m | U_m| \times \theta ID\}$
  $\mathcal{Y}_r^C = 1,\ \forall r \in R_m$
end for
$R = \bigcup_{m=1}^{M} R_m$

函数 Base_Estimator_Learning():
  服务器创建加密密钥对，将公钥发送给 B 和 C
  B 和 C 加密、交换梯度和损失。
  B 和 C 添加掩码，将加密值发送给服务器。
  服务器解密并回传值。B 和 C 去除掩码，更新模型。
  返回测试数据上正类的预测概率。

练独立的分类器，PU bagging 有效地缓解了初始可靠负例选择可能带来的偏差。最终预测结果通过集成这些独立分类器的预测结果获得，从而在未标记数据中更稳定、准确地识别正例样本。这种集成方法还增强了 VFPU 在处理噪声数据或不完整数据时的适应能力，而这些问题在现实数据集中普遍存在。

此外，纵向联邦学习框架确保了整个过程中的数据隐私。每个参与方都保留对自身数据的控制权，并且在训练过程中仅共享中间结果（如模型参数或加密梯度）。这种去中心化的方法避免了集中式数据存储的需求，并最大程度地降低了数据泄露的风险。VFPU 结合了强大的 PU 学习技术和纵向联邦学习的隐私保护特性，为

分布式环境下涉及正例和未标记数据的各种现实应用提供了一种有前景的解决方案。这一方法为更高效的协作和知识发现铺平了道路，同时遵守数据隐私标准。

(1) 建立初始样本集

VFPU 算法采用迭代方式运行，并通过双层方法增强正样本识别过程的稳健性和可靠性。外层循环包含 M 次迭代，为优化可靠正样本的选择提供了多次机会。在每次迭代 $m \in \{1, ..., M\}$ 中，内层循环执行 T 轮随机采样、训练和预测。这种嵌套结构有助于提高算法的稳定性和准确性，尤其是在处理现实场景中常见的噪声数据或不平衡数据时。

在每次迭代 m 开始时，算法基于 C 方提供的标签建立两个基本样本集。在纵向联邦学习框架中，C 方被指定为持有训练数据标签或部分标签的一方。这一指定至关重要，因为它指导了数据的初始划分。这两个集合定义如下：

$$\begin{aligned} P_m &= \{i | \mathcal{Y}_i^C = 1,\ i \in \mathcal{I}_C\}; \\ U_m &= \{i | \mathcal{Y}_i^C = -1,\ i \in \mathcal{I}_C\}, \end{aligned} \tag{3-1}$$

其中，$\mathcal{I}_C$ 表示 C 方的 ID 空间，即 C 方可用的所有样本标识符的集合。$\mathcal{Y}^C$ 表示 C 方的标签空间，包含每个样本 ID 对应的标签。i 代表 ID 空间中的特定样本 ID。因此，P_m 代表迭代 m 中的正样本集，具体而言，即 C 方提供正标签（$\mathcal{Y}_i^C = 1$）的样本集合。相反，U_m 代表迭代 m 中的未标记样本集，包含 C 方提供负标签或未标记（$\mathcal{Y}_i^C = -1$）的样本。

需要注意的是，在 PU 学习（Positive-Unlabeled Learning）的背景下，未标记集合 U_m 假设包含真实的正样本和真实的负样本的混合体。VFPU 旨在有效地区分这些样本，并在 U_m 中识别出可靠的正样本。该算法的迭代特性，以及后续涉及 PU 学习技术的步骤，有助于完成这一去歧义过程。通过在多个迭代和采样轮次中不断优化正样本的选择，VFPU 旨在逐步收敛到更准确、更稳健的真实正样本识别结果。这种精细的划分和迭代优化对于在未标记数据的情况下实现高性能至关重要。

(2) 采样、训练与预测

如 3-1 所示，在第 m 次迭代的第 t-th ($t \in \{1, 2, ..., T\}$) 轮采样过程中，使用自助法（bootstrapping）[51] 从 U_m 生成伪负样本集 N_m^t。数学上可以表示为：

$$N_m^t = \{\text{Randomly select } |P_m| \text{ elements from } U_m\}, \tag{3-2}$$

其中，$|P_m|$ 是 P_m 中包含的样本数量。在此过程中，随机选择是有放回的，这意味着 U_m 中的同一元素可能会被多次选中。这种策略不仅在样本生成过程中引入了随机性，还在不确定 U_m 内真实标签分布的情况下，有助于构建一个平衡的数据集。

由于未标记样本的实际类别未知，N_m^t 被视为一组伪负样本，可能同时包含真正的负样本和正样本。通过从 U_m 中抽取 $|P_m|$ 个元素，可以构造出与 P_m 规模相同的 N_m^t，从而为分类任务提供一个可比且平衡的数据集。这种平衡对于机器学习至关重要，因为它试图减轻类别不平衡可能带来的不利影响，并为后续的训练过程提供一个稳健的样本空间。

在训练过程中，P_m 和 N_m^t 被合并为一个二元分类训练集。该训练集用于训练纵向联邦学习模型，使其能够区分正样本和负样本，并将此知识应用于未来的预测任务。这两个集合的结合确保了学习算法能够接触到多样化的示例，在优化训练模型的泛化性能方面发挥着至关重要的作用。此外，训练过程同时利用了真实的正样本和伪负样本，即使伪负样本集中可能包含被错误标记的数据点，也能促进模型学习到稳健的决策边界。

自助法（bootstrapping）是一种从数据集中随机选择样本（有放回）的技术。这种统计方法在机器学习中被广泛采用，因为它有助于减少过拟合，并提供对模型性能更精确的估计。采用该技术使 VFPU 能够创建多样化且平衡的训练集，从而提高模型的泛化能力，减少潜在偏差，并增强推荐模型的整体性能。此外，在难以获取真实负样本的情况下，该方法尤为有用，因为它通过合成一个近似于真实负样本分布的伪负样本集，弥补了标注数据不足的问题。

在自助法过程中未被选中的样本称为袋外样本（out-of-bag samples）。这些样本在验证模型性能方面发挥着重要作用，因为它们提供了对分类误差的无偏估计。袋外评分（out-of-bag score）表示袋外样本被分类为正样本的预测概率。因此，为了获得袋外样本集 O_m^t，需要从 U_m 中排除 N_m^t 中的样本，数学表达如下：

$$O_m^t = U_m - N_m^t. \tag{3-3}$$

这种策略提供了一种内部模型评估机制，而无需单独划分验证集，从而充分利用所有可用数据进行模型开发和性能估计。

然后，C 方对 N_m^t、P_m 和 O_m^t 进行加密，并将其发送给其他方。在示例中，另一方是 B 方。随后，B 方和 C 方基于交换的三组样本 ID 共同建立各自的训练和测试数据集。这一联合过程在联邦学习环境中至关重要，因为隐私和数据安全是首要考虑因素。具体来说，按照以下公式进行数据集构建：

$$\begin{aligned} \mathcal{D}_{train}^K &= \{(i, x_i, y_i) \mid i \in P_m \ or \ i \in N_m^t\}; \\ \mathcal{D}_{test}^K &= \{(i, x_i, y_i) \mid i \in O_m^t\}, \end{aligned} \tag{3-4}$$

其中，$\mathcal{D}_{train}^K$ 代表二元分类训练数据，$\mathcal{D}_{test}^K$ 代表测试数据，$K \in \{B, C\}$。这里，

$x_i \in \mathcal{X}$ 表示样本 i 关联的特征向量来自特征空间 $\mathcal{X}$，而 $y_i \in \mathcal{Y}$ 表示相应的标签属于标签空间 $\mathcal{Y}$。这种加密样本 ID 与相应特征的无缝集成，促进了一个安全且协作的训练环境，在需要高隐私标准的场景中至关重要，同时确保了高质量的模型输出。

总之，通过结合自助法、袋外评估机制以及联邦学习方之间的安全样本共享机制，方法有效应对了处理不平衡和部分标注数据集的常见挑战。这种方法不仅保证了多样化的训练过程，还显著增强了模型在实际应用中的预测结果的完整性和可靠性。

一旦 B 方和 C 方准备好了各自的训练和测试数据集，二分类问题就转变为一个垂直联邦训练和预测任务。在这种情况下，一个基础估计器——代表每个参与实体的传统机器学习模型——被调整为可在垂直联邦学习（VFL）框架内使用。理解 VFL 的一般训练过程至关重要，如 [25] 中所描述。总体而言，这个过程包括四个关键步骤，这些步骤共同展示了如何在多方训练数据上训练基础估计器，同时确保在整个合作过程中严格保护各方数据集的隐私。四个步骤的描述如下：

- 第一步：服务器通过生成一系列加密密钥对来启动该过程。在此过程中，服务器安全地创建这些加密材料，并将相应的公钥发送给 B 方和 C 方。这一步为后续的所有隐私保护操作奠定了基础，建立了一个安全的信息交换通道。

- 第二步：在收到公钥后，B 方和 C 方对中间计算结果进行加密并交换这些结果。这些结果至关重要，因为它们涉及到梯度和损失的计算，这些计算对于学习过程至关重要。通过仅共享加密数据，每方确保不会直接透露或传输原始的敏感数据值，从而遵守了 VFL 范式中固有的隐私约束。

- 第三步：在交换加密的中间结果后，B 方和 C 方计算必要的加密梯度，用于模型参数的更新。除了计算这些梯度外，双方还引入了额外的掩码机制，以进一步隐藏计算出的梯度值。同时，每方还计算了加密版本的模型损失。应用这一额外的掩码步骤是为了防止在后续传输过程中可能发生的数据泄露。一旦这些涉及梯度和损失的加密值生成，它们会被安全地传输到服务器进行进一步处理。

- 第四步：在收到加密数据后，服务器负责解密接收到的梯度和损失。在完成解密过程后，服务器将解密后的梯度和损失发送回 B 方和 C 方。收到后，双方将移除之前应用的额外掩码。这个去掩码步骤至关重要，因为它恢复了真实的梯度信息，这对于更新模型参数是必需的。只有在这些梯度成功去掩码后，B 方和 C 方才会对基础估计器执行实际的参数更新，从而以同步和隐私保护的方式推进训练过程。

为了支持在 VFL 框架内的一般训练过程，已提出多种隐私保护的机器学习算法[25]。其中一些著名的例子包括逻辑回归（LR）[56,57]、随机森林（RF）[58]、梯度提升决策树（GBDT）[56]、XGBoost（XGB）[59,60] 和 LightGBM（LGB）[61]。这些算法旨在促进安全和高效的训练，同时在多个数据所有者之间保持数据隐私。在本文中，利用不同的基础估计器——每个算法对应其中之一——来全面评估推荐模型的整体性能。这种多估计器评估不仅突显了 VFL 框架的灵活性和广泛适用性，还提供了关于不同模型在隐私保护设置下行为的深入见解。包含多种模型架构强化了方法在处理复杂且敏感数据集时的鲁棒性和适应性，适用于跨不同数据源的应用。

通过仔细遵循上述四个步骤，并结合成熟的隐私保护技术，所提出的方法确保了模型性能与数据安全之间的强平衡。加密数据的交换、协作计算以及随后的解密和去掩码过程，展示了防止任何无意暴露敏感信息的严格措施。因此，这一过程促进了一个安全的训练环境，各方可以共同从集体数据洞察中受益，同时严格遵守隐私协议，使其特别适用于涉及敏感或受监管数据的应用。

(3) 确定可靠正样本

在第 m-th 次迭代中，采样、训练和预测的整体过程被重复执行 T 轮。完成这 T 轮后，会积累大量信息，从而形成一组概率，记为 $\mathcal{P}_m$，它与 U_m 中所有未标记样本相关联。这些概率反映了 U_m 中每个样本属于正类的可能性或置信度。这些概率评分在后续决策过程中起着关键作用，例如识别一部分高度可靠的正样本，并迭代更新训练集以提升模型性能。

为了计算完整的集合 $\mathcal{P}_m$，有必要为每个未标记样本 u（其中 $u \in U_m$）确定概率 $\mathcal{P}_m(u)$。该概率 $\mathcal{P}_m(u)$ 是通过将样本 u 在第 m-th 次迭代中所有 T 轮的袋外评分相加，然后除以 u 在这些轮中作为袋外样本出现的总次数来归一化得到的。该过程在数学上可表示为以下公式：

$$\mathcal{P}_m(u) = \frac{\sum_{t=1}^{T} S_m^t(u)}{\sum_{t=1}^{T} I(u \in O_m^t)}. \tag{3-5}$$

在该表达式中，$S_m^t(u)$ 表示在第 m-th 次迭代中第 t-th 轮为样本 u 分配的袋外评分。值得注意的是，如果 u 在某一轮中未作为袋外样本出现，则定义 $S_m^t(u)$ 为 0。指示函数 $I(u \in O_m^t)$ 在样本 u 包含在袋外集合 O_m^t 中时取值为 1，否则取值为 0。因此，分母 $\sum_{t=1}^{T} I(u \in O_m^t)$ 有效地计数了 u 参与袋外估计的轮数。此归一化过程确保分配给每个样本的概率反映了其在参与预测的各轮中的平均表现，从而提高了所得概率估计的鲁棒性和可靠性。

基于计算得到的概率 $\mathcal{P}_m(u)$，接下来对 U_m 中的所有未标记样本进行排序。该

排序以非递增顺序执行，因此具有较高概率的样本（即更可能为真正正样本的样本）会排在列表的顶部。利用这一排序好的列表，下一步便是选择一组可靠正样本。正式地，令 R_m 表示在第 m-th 次迭代中识别出的可靠正样本集合。在方法中，通过选择排名前 $|U_m| \times \theta$ 的样本来进行选择，其中 θ 是手动设置的比例，代表可靠正样本的采样率。该选择过程可简洁地表述为：

$$R_m = \{Chose\ Top\ |U_m| \times \theta\ IDs\ from\ \mathcal{P}_m\}. \tag{3-6}$$

这一两步选择过程可以具体描述为：

- 第一步：将 $\mathcal{P}_m$ 中的所有样本按照计算得到的概率以非递增顺序进行排序。此排序确保最有可能为正样本的样本获得优先考虑。
- 第二步：从排序结果中选择前 $|U_m| \times \theta$ 个样本，其中 θ 作为可调参数决定了未标记数据集中被视为可靠正样本的样本比例。

通过执行这一过程，最有信心且可能真正属于正类的样本便从其他样本中被区分出来。因此，集合 R_m 包含了这些被选中的样本，并随后被用来更新由 Party C 维护的训练标签。

在识别过程之后，这些新识别出的可靠正样本会被纳入 Party C 维护的训练集中。此更新过程包括为 R_m 中的每个样本分配一个正标签，从而逐步减少未标记数据集 U_m 的规模。从形式上表示，该更新过程表达为：

$$\mathcal{Y}_r^C = 1,\ \ r \in R_m, \tag{3-7}$$

其中集合 R_m 中的每个样本 r 被明确赋予标签 1，表明其被归类为正类。该标签更新在训练数据的迭代细化过程中至关重要，因为它利用了从袋外预测集合中获得的置信度评分来提高后续迭代中训练集的整体质量和可靠性。

总之，识别可靠正样本的过程是一种精心设计的机制，它将概率预测转换为一组高置信度的正实例。该机制不仅提升了训练数据的质量，还为改进整体推荐模型性能提供了一个迭代途径。通过基于样本的袋外表现不断更新样本标签，算法有效地缓解了标签不确定性和数据稀疏性相关的问题。此外，这种丰富而系统的方法为后续的模型训练和预测任务奠定了坚实的基础，确保不断发展的数据集逐步趋向更高的准确性和一致性。

(4) 最终结果及其在推荐中的应用

完成算法的所有 M 次迭代后，最终的可靠正样本集合 R 是通过取每次迭代中所有 R_m 集合的并集得到的，其中 $m = 1, 2, \ldots, M$ 表示在整个过程中所执行的各个

单独迭代。该迭代方法确保算法在多个轮次中系统地细化正样本的选择，逐步提升最终集合 R 的质量和可靠性。具体来说，每个 R_m 包含在第 m 次迭代中识别出的正样本，而并集操作则将这些中间结果整合为一个全面的最终集合。该步骤在计算上虽简单但至关重要，因为它将所有迭代的结果聚合成一个统一的集合，反映出算法在从整个数据集中区分可靠正样本方面所做的累计努力。

借助这一精心筛选出的可靠正样本集合 R，Party A（可能代表服务提供商或推荐系统运营者等相关方）现在可以利用这些信息为 R 中的每个样本量身定制推荐。该个性化过程涉及对每个样本的特征或偏好进行分析，并设计与之匹配的推荐策略。通过这种方式，Party A 能显著提高推荐系统的准确性和相关性，确保向最终用户提供的建议——无论是产品推荐、内容建议还是其他个性化输出——既精准又符合情境。基于一组可靠正样本对推荐进行微调的能力，凸显了该算法的实际效用，将理论计算与实际应用相结合。

利用 R 来增强推荐系统的这一步骤不仅仅是一个程序上的形式问题，而是整个算法有效性的基石。推荐系统的成功取决于其识别和优先考虑最相关样本的能力，而这一任务正是该迭代算法通过其结构化设计所完成的。通过聚焦于可靠正样本，系统最大程度地减少了噪声或无关数据的引入，这些数据本可能降低推荐质量。此外，R 的应用使得 Party A 能够优化资源分配——将计算和操作上的努力集中在最有可能带来高用户满意度和参与度的样本上。

为说明这一过程的重要性，可以考虑推荐系统在各个领域的广泛应用，如电子商务平台、流媒体服务和社交网络，其中建议的准确性直接影响用户体验和业务成果。通过取 R_m 集合的并集来创建 R，确保了这些应用的坚实基础，提供了一种可扩展且适应性强的解决方案，能够容纳不同规模和复杂度的数据集。此外，算法的迭代特性使其能够适应用户偏好或数据分布可能随时间变化的动态环境，从而成为长期部署的多功能工具。

在实践中，可以通过精确度、召回率或用户满意度等指标来评估这一最终步骤的有效性，这些指标量化了由于精心筛选出的集合 R 而提升的推荐质量。例如，更高的精确度表明推荐项中相关的比例更大，而召回率的提升则反映了系统从数据集中检索出更广泛相关项的能力。这些可衡量的结果突显了算法在提升推荐系统性能方面所作的贡献，验证了迭代过程及随后 R 应用的重要性。因此，这一收尾阶段不仅完成了算法工作流程，也确立了其作为解决现实推荐挑战的实际方案的价值。

3.4 实验结果与分析

本节描述了 VFPU 算法的实验设计和评估。首先，提供了数据集和实验设置的概述。随后，提出了一组用于指导实验的研究问题，然后根据每个研究问题展示并讨论实验结果。

3.4.1 数据集

在实验中，利用了三个多样且成熟的数据集来评估所提出方法的有效性：Bank Marketing Datasetg [62]、Default of Credit Card Clients Dataset [63] 以及 Adult Census Dataset [64]。每个数据集的选择都经过精心考虑，以代表不同的真实世界领域和预测挑战，从而使能够在多样化的场景下对方法的鲁棒性进行评估。

Bank Marketing Dataset 源自 UCI 机器学习库，数据来自葡萄牙某银行针对定期存款产品进行的直销电话营销活动。这些活动通过电话直接与客户沟通，旨在说服他们订购定期存款产品。完整的数据集包含四个版本，但在本研究中，专门采用了“bank-additional-full”版本，该版本包含 41,188 个样本和 20 个输入特征。这些特征涵盖了客户的各种属性，包括年龄、工作类型、婚姻状况、教育程度以及诸如是否有住房贷款或个人贷款等金融细节，同时也包括了营销活动相关变量，如联系次数和以往互动的结果。数据按照日期顺序排列，涵盖了从 2008 年 5 月到 2010 年 11 月的时间段，反映了此期间营销工作的时间演变。该数据集的主要分类目标是预测客户是否会订购定期存款，即由变量“y”（是/否）表示的二元结果。此数据集在客户关系管理方面具有实际意义，并且由于大多数客户并未订购，从而带来了类别不平衡的问题，这为预测建模带来了实际挑战。

Default of Credit Card Clients Dataset 来自 Kaggle，提供了 30,000 名信用卡客户财务行为的详尽数据，数据集中包含 24 个变量。这些变量融合了丰富的人口统计信息（如性别、教育、婚姻状况和年龄）和详细的还款历史数据，涵盖了每月账单、前期还款金额以及多个月份内的付款延迟记录。该数据集为构建预测模型以评估客户下月违约风险提供了基础，其目标变量为表示违约状态（是/否）的二元指标。人口统计和交易特征的结合为信用风险提供了细致的视角，使该数据集在金融分析中极具价值。其庞大的数据量和详细的特征集进一步提升了其在训练和评估旨在信用风险预测的复杂机器学习模型中的适用性。

Adult Census Dataset 同样源自 UCI 机器学习库，是一个广泛认可的人口统计和社会经济数据集，数据来自 1994 年美国人口普查。该数据集包含各种属性，如年龄、工作类别（例如私人部门、联邦政府）、教育程度、职业、种族、性别、资本利得、资本损失、每周工作小时数以及原籍国，为个人特征提供了详尽的快照。

数据集的分类任务是预测个人年收入是否超过 50,000 美元，这一二元结果已成为收入预测模型的基准。除了在机器学习中的应用之外，该数据集还经常被研究人员和从业者用于探讨与收入分配、就业趋势和经济差异相关的社会模式。其多样的特征集和现实意义使其成为评估预测模型的理想候选者，同时也应对了诸如缺失值和类别数据处理等挑战。为简洁起见并在本文中保持一致性，分别将这些数据集称为“Bank”、“Credit”和“Census”。

在数据预处理阶段，采用了系统化技术以确保数据集适用于机器学习算法。所有三个数据集中普遍存在的分类特征（例如“Bank”中的工作类型、“Credit”中的教育程度以及“Census”中的职业）均采用独热编码（one-hot encoding）进行编码。该方法将分类变量转换为一系列二元列，每个列对应一个类别，从而防止模型错误地假设类别之间存在序数关系，并能有效地进行数值处理。数值特征，如年龄或账单金额，则使用标准化缩放方法进行归一化，该方法将数据中心化为均值为零并缩放为单位方差。这种归一化确保了不同尺度的特征在模型学习过程中具有同等的贡献，尤其对于依赖梯度优化的算法尤为关键。

各数据集均由各自数据提供方预先划分为训练集和测试集，确保模型训练与在未见数据上的评估有初步的分离。为了与实验设计保持一致，该设计模拟了分布式学习环境，进一步将训练数据纵向划分为三个不同部分，分别分配给三个假想方：A、B 和 C。该纵向划分意味着每个方持有整个样本集中的一部分独特特征，这反映了数据在多个实体间分布的场景，正如联邦学习框架中常见的情况。为模拟标签数据稀缺或分布不均的情况，从方 B 和 C 持有的数据中移除了数据标签，从而生成了统称为 U 的未标记数据。对于方 A，在其训练数据子集中识别出所有属于正类的样本，并随机选择其中 10% 的样本保留其标签，构成正样本集 P。而对于其余 90% 的正样本以及所有负样本，其标签均被忽略，在方 A 的数据中这些样本实际上被视为未标记。因此，实验设置由仅供方 A 使用的一小部分标记正样本集 P，以及分布于三个方的更大未标记数据集 U 组成，从而使能够探讨方法在分布式环境下利用有限标记数据的能力。

3.4.2 实验环境及参数设置

在垂直联邦预测不确定性（VFPU）算法的联邦训练和预测阶段，通过集成针对垂直联邦学习（VFL）场景量身定制的多样化基学习器，对其性能进行了严格评估。具体而言，垂直联邦逻辑回归（VFPU_LR）的实现改编自 Aono 等人的基础性工作 [65]，该工作利用安全多方计算（SMC）协议在梯度交换过程中保护数据隐私。对于基于树的方法，则使用 FedTree 库 [66] 开发了垂直联邦梯度提升决策树（VFPU_GBDT）和垂直联邦随机森林（VFPU_RF），该库是一个针对分布式树构建

并支持同态加密的最先进框架。此外，垂直联邦 LightGBM（VFPU_LGB）的实现则基于 FATE（Federated AI Technology Enabler）框架 [67]，该框架为跨异构数据分区的安全联邦模型训练和推断提供了强大的基础设施。

为了建立全面的基线比较，传统的集中式机器学习模型在聚合数据集上进行了训练，以模拟非联邦环境。这些模型包括：(i) 逻辑回归（LR），(ii) 梯度提升决策树（GBDT），(iii) 随机森林（RF），(iv) XGBoost（XGB）[68]，以及 (v) LightGBM（LGB）[69]，均使用 scikit-learn [70] 和原生库实现。这种双重方法使能够在相同评估指标下，直接比较联邦与非联邦范式的性能。

实验中统一了超参数配置以确保公平性。为平衡计算效率与模型收敛性，最大迭代次数 M 固定为 5，而特征采样率 θ 被设置为 0.02，以缓解高维数据场景中的过拟合问题。为确保随机优化的稳定性，随机采样迭代次数 T 设定为 10。针对 VFL 中固有的数据隐私问题，在跨方计算中采用了密钥长度为 512 位的 Paillier 加密系统 [71] 对中间输出进行加密。此密钥长度在加密安全性与计算开销之间提供了务实的折衷。所有实验均假定默认的两方设置 [66]，反映了现实中拥有互补特征空间的组织间常见的 VFL 协作模式。

评估指标: 采用了多维度评估框架以全面评估模型性能。准确率（acc）和 F 分数（F）[72] 分别量化了整体预测正确性和类别特定的平衡性。精确率和召回率 [52] 被用来评估第一类/第二类错误的权衡，这在不平衡分类任务中尤为关键。ROC 曲线下面积（AUC）[72] 提供了一个与阈值无关的正负类别可分性度量。

AUC 指标来源于接收器操作特征（ROC）曲线，对于在不同决策阈值下评估模型鲁棒性尤其具有信息量，因为它量化了所有分类阈值下真正率（灵敏度）与假正率（1-特异性）之间的关系。与此同时，F 分数被定义为精确率与召回率的调和平均值，在类别分布偏斜或误分类成本不对称的场景下，作为模型稳定性的统一指标。这些指标均通过 5 折交叉验证在独立测试集上计算，以确保统计可靠性。

3.4.3 实验一：比较联邦和非联邦的 PU 学习的性能

在本小节中，深入研究了联邦学习对正类和无标签（PU）学习性能的影响。具体来说，比较了使用 VFPU 算法的联邦 PU 学习与非联邦 PU 学习在多种基础估计器和数据集上的表现。核心目标是评估在其他算法参数和实验配置保持不变的情况下，联邦学习对分类质量的影响程度。

使用了四种常用的基础估计器——逻辑回归（LR）、随机森林（RF）、梯度提升决策树（GBDT）和 LightGBM（LGB）——在三个真实世界的数据集上进行实验：银行营销数据集、信用卡客户违约数据集和成人普查数据集。对于逻辑回归（LR），使用了 L2 惩罚系数为 0.8、学习率为 0.001、批处理大小为 64 的配置，确

保在联邦和非联邦运行中收敛行为保持一致。对于基于决策树的算法（RF、GBDT和 LGB），将树的数量设置为 50，最大深度为 6，学习率为 0.1。这些设置经过精心选择，以平衡模型复杂度、训练时间和泛化能力。

表 3-1 有无联邦学习的 PU 学习性能比较

Table 3-1 Performance comparison of PU Learning with and without federation

Base Estimator	Metrics	The Bank Marketing		The Default of Credit Card Clients		The Adult Census	
		Fed	No_Fed	Fed	No_Fed	Fed	No_Fed
LR	acc ↑	0.923	0.948	0.822	0.843	0.814	0.852
	recall↑	0.194	0.219	0.179	0.206	0.126	0.141
	precision↑	0.520	0.546	0.502	0.513	0.804	0.820
	AUC↑	0.658	0.685	0.621	0.643	0.854	0.858
RF	acc ↑	0.935	0.952	0.826	0.851	0.848	0.852
	recall↑	0.219	0.248	0.190	0.212	0.154	0.165
	precision↑	0.587	0.613	0.546	0.562	0.755	0.770
	AUC↑	0.882	0.909	0.625	0.639	0.805	0.854
GBDT	acc ↑	0.943	0.945	0.847	0.851	0.862	0.873
	recall↑	0.239	0.241	0.224	0.229	0.188	0.205
	precision↑	0.594	0.599	0.581	0.586	0.922	0.933
	AUC↑	0.886	0.891	0.639	0.647	0.854	0.869
LGB	acc ↑	0.896	0.914	0.815	0.843	0.828	0.852
	recall↑	0.186	0.197	0.167	0.182	0.142	0.155
	precision↑	0.518	0.542	0.496	0.524	0.354	0.373
	AUC↑	0.587	0.612	0.569	0.582	0.705	0.721

表 3-1 展示了在三个数据集上的实验结果，详细列出了每种算法-数据集组合在联邦（Fed）和非联邦（No_Fed）模式下的四种性能指标——准确率（acc）、召回率（recall）、精确率（precision）和 AUC。发现一致表明，联邦变体（Fed）在准确率、召回率、精确率和 AUC 方面略低于非联邦版本（No_Fed）。尽管这种差异在所有基础估计器和数据集上都有观察到，但重要的是要强调，在实际应用中，这种差异非常小。

仔细观察数据可以揭示这种微妙的性能差距。例如，在成人普查数据集中，使用 GBDT 作为基础估计器，联邦方法达到了 0.862 的准确率、0.188 的召回率、0.922 的精确率和 0.854 的 AUC。相比之下，非联邦方法报告了略高的 0.873、0.205、0.933 和 0.869 的准确率、召回率、精确率和 AUC。差距分别为 0.011、0.017、0.011 和 0.015。从数值上看，这些指标的平均差异为 0.0135，仍然非常小，表明性能的妥协微不足道。

此外，在所有数据集、四种指标和四种基础估计器上平均，差异为 0.0174——进一步证明虽然性能下降是持续观察到的，但这种下降的幅度是适度的。这些观察结果强调了 VFPU 算法在实际应用中的潜在可行性，在这些应用中，数据隐私和安全问题可能超过预测性能的微小下降。

实验证据因此表明，将联邦学习集成到 PU 学习中仍然保留了稳健的分类效果。尽管非联邦方法保持了轻微的优势，但这种差异不足以掩盖联邦学习提供的隐私和数据保密优势。因此，对于在敏感领域——如医疗保健或金融——操作的从业者和研究人员来说，VFPU 代表了一个引人注目的选择，因为在这些领域，数据访问受到严格监管。该方法使多个参与方能够在不共享原始数据的情况下进行协作建模，从而在模型准确性和隐私要求合规性之间取得了吸引人的平衡。

总体而言，这些实验强调了虽然联邦和非联邦 PU 学习之间存在适度的性能差距，但 VFPU 算法仍然提供了稳健的结果。这为数据隐私法规不可妥协的领域提供了一个实用且有效的解决方案，确认了在准确率、召回率、精确率和 AUC 方面轻微的降低是增强数据安全和保密性的合理权衡。

3.4.4 实验二：分析不同基分类器器对 VFPU 性能的影响

在这一部分，对集成在 VFPU 框架中的不同基估计器进行了深入评估，这些估计器包括 VFPU_LR、VFPU_GBDT、VFPU_RF 和 VFPU_LGB，它们被用于解决 UDD-PU 问题。为了全面评估它们的性能，进行了一系列实验，重点关注四个关键评估指标：精确度（Precision）、召回率（Recall）、F-score 以及精确度-召回率曲线。本次评估的主要目标是分析这些指标随着可靠正样本数量增加的变化趋势，并根据这些观察到的趋势，确定方法的最佳基估计器。

在实验中，采用随机抽样策略从一个未标记的数据集中选择可靠的正样本子集。这种方法不仅模拟了现实情况下可能已知少量正实例的条件，还严格测试了每个基估计器在利用这些有限样本时的鲁棒性。实验结果详见图 **??** 至图 **??**。

对于基于逻辑回归（LR）的算法，配置了以下特定超参数：应用了 L2 惩罚，系数设置为 0.8，学习率为 0.001，小批量大小为 64。这些参数基于初步实验精心选择，以确保收敛速度和模型泛化能力之间的平衡。另一方面，对于所有基于树的算法——即梯度提升决策树（GBDT）、随机森林（RF）和 LightGBM（LGB）——将树的数量设置为 50，树深限制为 6 以避免过拟合，并使用 0.1 的学习率。这种在三个数据集上的统一设置，使得在一致的实验条件下公平比较它们的性能成为可能。

??展示了在银行营销数据集上的实验结果。在图 2(1) 至图 2(3) 中，x 轴表示分类器获得的可靠正样本数量，范围从大约 100 到 2293 个可靠样本。这些子图中的 y 轴分别对应于准确度、召回率和 F-score 的评估指标；在此背景下，y 轴上的

值越高表示分类性能越好。此外，图 2(4) 展示了精确度-召回率曲线，其中 x 轴为精确度，y 轴为召回率。曲线上更接近图右上角的模型表明其性能更强，特别是在数据分布不平衡的情况下。

对**??**的探索表明，VFPU_GBDT 估计器在银行营销数据集上的表现优于其他模型。例如，如图 2(1)-(3) 所示，当利用 1500 个可靠正样本时，VFPU_LR、VFPU_GBDT、VFPU_RF、VFPU_LGB 和基于随机抽样的模型分别实现了 44.43%、51.94%、49.81%、43.72% 和 4.73% 的精确度。相应的召回率记录为 14.78%、17.28%、15.24%、14.55% 和 4.61%，而这些模型的 F-score 分别为 0.22、0.26、0.23、0.22 和 0.05。尽管 VFPU_RF 估计器在某些情况下显示出相对可比较的精确度，但随着可靠样本数量的增加，其性能（特别是在召回率和 F-score 方面）下降得更快。相比之下，VFPU_LR 和 VFPU_LGB 模型虽然在不同样本大小下表现稳定，但并未达到 VFPU_GBDT 在精确度和召回率方面表现出的整体性能。因此，图 2(1)-(3) 清楚地表明，VFPU_GBDT 不仅提供了更高的精确度，而且在召回率之间保持了平衡的权衡，从而在评估的模型中实现了最高的 F-score。

此外，图 2(4) 中的精确度-召回率曲线进一步验证了 VFPU_GBDT 的鲁棒性。通过在各种阈值设置下绘制精确度对召回率的曲线，该曲线有效地捕捉了分类器的性能动态，特别是在类不平衡是关键问题的场景中。VFPU_GBDT 曲线更接近图的右上象限，表明其在多个阈值下的强大性能——这是不平衡分类任务中高性能模型的标志。

??和**??**提供了 VFPU_GBDT 估计器有效性的进一步支持证据。这些图报告了在信用卡违约和成人普查数据集上进行的类似实验。尽管数据集在内在特性和分布属性上有所不同，但这些情况下的实验结果反映了与**??**中观察到的相似趋势。VFPU_GBDT 在不同数据集上的一致优越性增强了其在 VFPU 框架中作为基估计器时的可靠性和有效性。

总结来说，三个基准数据集上的实验结果一致表明，VFPU_GBDT 优于 VFPU_LR、VFPU_RF、VFPU_LGB 以及基线随机抽样技术。VFPU_GBDT 的优越性能，表现为更高的精确度、召回率和平衡的 F-score，使其成为在解决 UDD-PU 问题时 VFPU 算法最可靠和推荐的基估计器。因此，在所有后续实验中，采用 GBDT 作为 VFPU 方法的基估计器，利用其在挑战性条件下实现准确和平衡分类的明显优势。

3.4.5 实验二：基线对比实验

在本节中，对所提出的针对 UDD-PU 问题的方法进行了全面探讨，特别是在垂直联邦学习（VFL）的设定下，为评估该方法的有效性设计了严谨的分析。为此，开展了两个不同的实验。第一个实验旨在识别最有效的解决正-未标记（PU）问题

的半监督学习方法。这些方法虽广泛应用于 UDD-PU 问题的研究，但通常是在非联邦环境下考虑。第二个实验则基于第一个实验的结果，将前四种半监督技术在实现 UDD-PU 问题的 VFL 环境中进行比较。

在第一个实验中，对比了 No_Fed_VFPU_GBDT 方法与另外九种在非联邦环境下运行的知名半监督学习方法。该对比结果在**??**中进行了全面展示。以下段落进一步介绍了每种方法的背景及其在本研究中的具体实现细节。

_GBDT: 此方法涉及直接将梯度提升决策树（GBDT）分类器应用于正-未标记（PU）数据集。在这种方法中，未标记的数据（记为 U）被当作负样本，而现有的标记数据（P）被视为正样本。这一基线方法为评估在没有特殊处理隐藏正样本的情况下，传统技术在 PU 学习中所面临的内在挑战提供了参考。

Bagging_GBDT: Bagging_GBDT 代表了一种利用装袋（bagging）技术进行 PU 学习的集成学习方法。在这种方法中，通过对未标记数据集的反复采样生成多个训练集。每个训练集都包含正样本和未标记样本的混合，从而生成一系列多样化的子集，并在其上分别训练独立的 GBDT 分类器。最终的预测结果由所有分类器的预测平均值决定。该方法旨在通过整合多个模型来降低方差并提高预测的鲁棒性。

2Step_GBDT: 2Step_GBDT 方法引入了一个两阶段的训练过程。在第一阶段，使用完整的 PU 数据集训练 GBDT 分类器，此时所有未标记样本暂时被当作负样本。随后，该分类器被用来识别一部分具有高正样本可能性的未标记样本。在第二阶段，利用这一精炼后的样本子集对分类器进行再训练，以提高其区分隐藏正样本与真实负样本的能力。此迭代细化过程对于在正样本被大量未标记样本掩盖的情况下提升分类性能至关重要。

Pseudo-labeling: Pseudo-labeling 是一种著名的半监督学习技术，它利用模型对未标记数据的预测生成伪标签。接着，将这些伪标签与原有的标记数据结合，用于重新训练模型或提升现有模型的学习过程。伪标签方法的核心思想是，通过将高置信度的预测视为真实标签，使模型不断迭代改进，从而逐步将未标记数据信息整合进训练过程。

MixMatch: MixMatch 是当前领先的半监督学习方法，通过对未标记样本进行增强后猜测标签，并利用 MixUp 技术混合标记和未标记数据而

实现。在实验中，将迭代参数 T 设置为 0.5，将增强次数 K 设置为 2。该方法旨在通过有效利用未标记数据生成更平滑的决策边界，从而减少过拟合并提升模型的泛化能力。

FixMatch: FixMatch 是伪标签方法的增强版本，它将对弱增强未标记数据的预测结果转换为硬性的一热编码伪标签，然后利用这些伪标签作为对强增强未标记数据的学习信号。在的实现中，采用一组超参数，包括置信阈值 $\tau = 0.95$，参数 $\nu = 7$，以及批次大小 $B = 64$。FixMatch 的主要优势在于其利用一致性正则化的能力，从而更为有效地利用未标记数据集。

CoMatch: CoMatch 通过同时引入训练数据的两种互补表示扩展了 FixMatch 的思想。它同时学习类别概率和低维嵌入，两者相互作用以通过平滑约束来提升伪标签的质量。同时，CoMatch 利用基于图的对比学习来正则化嵌入结构。为了保证一致性和便于比较，遵循与 FixMatch 相同的超参数设置 [73]。这一方法在底层数据分布复杂且呈多模态情况时尤为有效。

AdaMatch: AdaMatch 最初为领域适应任务提出，并被改进用于半监督学习。其核心特性包括相对阈值和分布对齐机制。相对阈值通过对标记数据置信度分数的指数移动平均（EMA）自适应估计 [3]。在实验中，置信阈值设置为 $\tau = 0.9$。AdaMatch 的自适应策略有助于更好地平衡标记数据和未标记数据的贡献，从而使其成为处理样本不平衡数据集任务的有力候选方法。

SoftMatch: SoftMatch 旨在在训练过程中同时保持伪标签的数量和质量。它通过使用截断高斯函数对样本按照其置信度分数进行加权来实现这一目标。此外，SoftMatch 采用均匀对齐方法来增强对代表性较弱类别的学习。在实验中，将参数 m 设置为 0.999，并将估计方差 $\hat{\sigma}_t$ 除以 4，对应于高斯函数的 2σ 区间。该方法旨在缓解噪声伪标签所带来的常见挑战，确保只有最为可靠的信息被纳入训练过程中。

通过对各方法工作原理和超参数设置的详细说明，旨在提供对这些方法各自优势和局限性的深入理解。这一详尽的比较不仅突显了 No_Fed_VFPU_GBDT 方法在揭示隐藏正样本方面的鲁棒性，同时也为未来在常规和联邦学习框架下解决 UDD-PU 问题的研究提供了基准。

在第一次实验中，使用分类器对所有未标记样本进行了打分，这些分数代表了某个样本被预测为正类的概率。打分过程采用了经过校准的模型，以计算每个未标记实例属于正类的可能性。获取这些概率分数后，将所有未标记样本按降序排序，并将分数最高的样本标记为顶级样本。图**??**中所示的 *num* 表示从该排名顶部选取的未标记样本数量。

对于本次实验中使用的逻辑回归（LR）算法，严格设置了 L2 惩罚项系数为 0.4，学习率为 0.0002，批量大小为 32。这些超参数设置是基于初步实验结果选择的，表明它们在收敛稳定性与计算效率之间提供了较为平衡的权衡。对于基于决策树的算法，特别是采用梯度提升的方法，将树的数量设置为 500，并将最大树深固定为 12。此外，决策树算法的学习率设置为 0.02。这些超参数配置对于确保模型能够充分表达数据中复杂模式，同时降低过拟合风险至关重要。

此外，对于 Match 系列算法——即 FixMatch、CoMatch、AdaMatch 和 SoftMatch，采用了 [74] 中的 Wide ResNet-28 模型。该模型因其在半监督环境中具有强大的特征提取能力而受到认可。然而，选择这一深度学习架构引入了一定的复杂性，尤其是在数据极度不平衡的情况下（如 PU 问题中常见），这可能导致模型过拟合。

图**??**显示，提出的 No_Fed_VFPU_GBDT 方法在所有测试数据集上以及对于各个不同的 *num* 值均持续优于其他九种半监督方法。例如，在 Census 数据集中，当 $num = 1000$ 时，的方法实现了最高的推荐准确率 99%。这一显著性能与基线方法形成了鲜明对比：_GBDT、2Step_GBDT 和 Bagging_GBDT 分别仅达到 64%、62% 和 55% 的准确率。同时，基于深度学习的方法——MixMatch、FixMatch、CoMatch、AdaMatch 和 SoftMatch——的推荐准确率明显较低，分别为 35.94%、37.06%、33.39%、37.28% 和 35.86%。

Match 系列算法尽管是半监督学习领域最新的进展，但表现不佳可以归因于两个关键因素。首先，这些方法本质上需要大量标记数据，涵盖正类和负类。然而，在 PU 问题情境下，负样本并未被明确标记，这使得这些算法处于明显的不利地位。其次，使用 Wide ResNet-28 模型这一相对复杂的深度学习网络，在面对类分布极度不平衡的数据集时，可能无意中导致过拟合。这种过拟合会削弱模型的泛化能力，从而导致整体性能较低。

相比之下，PU 学习方法，包括 VFPU_GBDT、_GBDT、Bagging_GBDT 和 2Step_GBDT，专门为解决在主要由未标记样本构成的数据集中揭示隐藏正样本这一挑战而设计。这些方法利用数据的内在结构，并采用定制的策略来迭代细化对潜在正实例的识别。因此，如**??**所示，表现最好的四种方法分别为 _GBDT、Bagging_GBDT、2Step_GBDT 和 VFPU_GBDT。它们的出色性能突显了 PU 学习策略在负样本不可直接获得且数据存在显著不平衡的情景中的有效性。

通过对打分机制、超参数设置以及底层模型架构的详细讨论，丰富了分析内容，从而全面理解了实验设置及各方法的相对优劣。这一详细论述凸显了提出的No_Fed_VFPU_GBDT 方法在有效揭示隐藏正样本方面的优势，为未来在常规 PU 学习和联邦 PU 学习框架下的研究提供了宝贵的启示。

在第二个实验中，将第一个实验中表现最佳的四种方法适配到垂直联邦学习（VFL）环境中，分别标记为 VF_GBDT、VF_Bagging_GBDT、VF_2Step_GBDT 和 VFPU_GBDT。此外，为了评估这些方法在 VFL 环境中部署时的时间复杂度和计算开销，在 *num* = 2400 的各种数据集上仔细记录了每种方法的训练时间。训练时间表示为 runtime(s)，其中’s’ 代表秒，详细内容见**??**。

??与**??**类似，展示了 VFL 设置下不同方法所达到的准确率百分比的比较分析。结果清晰地表明，VFPU_GBDT 方法在推荐准确率方面始终优于其他三种方法。例如，在 Census 数据集上，当 *num* = 1000 时，VFPU_GBDT 达到了显著高的 98.70% 的推荐准确率，这明显高于 VF_GBDT、VF_Bagging_GBDT 和 VF_2Step_GBDT 的准确率，它们分别为 24.20%、21.10% 和 24.30%。在 Credit 数据集上，尽管 VFPU_GBDT 展示了更高的准确率，但它带来了更高的计算成本：本章的方法记录的运行时间为 107075.9s，而 VF_GBDT、VF_Bagging_GBDT 和 VF_2Step_GBDT 的运行时间分别为 12025.47s、15791.59s 和 46954.19s。

值得注意的是，按 runtime(s) 指标衡量，VFPU_GBDT 的耗时约为其他方法的 10 倍。这种增加的时间消耗主要是由于 VFPU_GBDT 在选择可靠正样本时采用了更为谨慎和迭代的策略。该算法采用多次迭代，每次迭代只选择一小部分样本，确保只有最可靠和准确预测的正样本被选中。这种谨慎的选择过程引入了额外的计算开销。然而，这种训练时间与推荐准确率之间的权衡被认为是可接受的，因为准确率的显著提高证明了额外训练时间的合理性。

??提供了**??**中讨论的银行营销数据集实验结果的可视化表示。在该图中，x 轴表示被选为可靠正样本的评分最高的未标记样本数量，而 y 轴表示这些被选择样本中真正为正样本的百分比。图中清晰地说明，随着 *num* 从 100 增加到 1000，VFPU_GBDT 识别的隐藏正样本数量迅速减少。这种观察到的减少可归因于几个因素。最初，与正类表现出强烈相似性或与负类有明显差异的未标记数据样本被迅速识别和选择。然而，随着选择过程的继续，剩余的未标记数据往往在正负类之间表现出特征重叠，这给分类器带来干扰。这种干扰导致错误分类，从而降低了准确推荐的百分比。虽然其他方法在不同 *num* 值下表现相对稳定，但这种表面上的稳定性并不一定是有利的。这些方法始终较低的准确推荐百分比表明它们在可靠识别隐藏正样本方面能力有限。此外，考虑到未标记数据集中只有约 2500 个真正的正样本，当选择的正样本数量超过这个阈值时，VFPU_GBDT 的准确率自然会下

降。

总之，实验结果提供了强有力的证据，表明 VFPU_GBDT 算法在 VFL 环境中推荐可靠正样本方面非常有效。尽管需要额外的计算时间，但推荐准确率的显著提升使 VFPU_GBDT 相比其他半监督方法成为更优方案。这项实验突显了在联邦学习环境中应用 VFPU_GBDT 算法解决未标记数据缺乏的 PU（UDD-PU）学习推荐问题的显著优势，强调了它在高准确率至关重要的实际应用中的潜力。

3.5 本章小结

本章主要介绍了未标记数据缺乏的 PU（UDD-PU）学习问题，其中需要模型训练的一方仅持有正样本数据，而其他参与方拥有大量未标记数据。为解决 UDD-PU 问题，本章提出了一种基于半监督学习的多方联邦推荐方法。具体而言，设计了 VFPU 算法作为该方法的核心，该算法有效整合了两种 PU 学习技术，并将其适配到垂直联邦学习框架中。通过这种方式，VFPU 能够利用有限的标记数据（正样本）和丰富的未标记数据来提高推荐模型的性能。在三个数据集上评估了所提出的方法，并将其与其他半监督学习方法进行了比较。实验结果明确表明，VFPU 算法在确保数据隐私的同时，与非联邦方法相比只有很小的性能下降，就能达到令人满意的性能。此外，证明了 VFPU 算法在垂直联邦学习环境中发现隐藏正样本方面始终优于其他半监督学习方法。而且，对不同基础估计器的分析表明，梯度提升决策树（VFPU_GBDT）在精确率、召回率和 F 值方面持续展现出卓越的性能。这一发现强调了为 VFPU 算法选择适当基础估计器的重要性，以便在各种实际应用中优化其性能。

第 4 章　总结与展望

4.1 主要结论

本文主要……

4.2 研究展望

更深入的研究……

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。

以下文字用于测试。
以下文字用于测试。
以下文字用于测试。

参考文献

[1] Yilun Jin, Yang Liu, Kai Chen, and Qiang Yang. Federated learning without full labels: A survey. *arXiv preprint arXiv:2303.14453*, 2023.

[2] Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li. Rscfed: random sampling consensus federated semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10154–10163, 2022.

[3] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[4] Chenyou Fan, Junjie Hu, and Jianwei Huang. Private semi-supervised federated learning. In *IJCAI*, pages 2009–2015, 2022.

[5] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*, 2020.

[6] Xinyang Lin, Hanting Chen, Yixing Xu, Chao Xu, Xiaolin Gui, Yiping Deng, and Yunhe Wang. Federated learning with positive and unlabeled data. In *International Conference on Machine Learning*, pages 13344–13355. PMLR, 2022.

[7] Lun Wang, Yang Xu, Hongli Xu, Jianchun Liu, Zhiyuan Wang, and Liusheng Huang. Enhancing federated learning with in-cloud unlabeled data. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 136–149. IEEE, 2022.

[8] Sohei Itahara, Takayuki Nishio, Yusuke Koda, Masahiro Morikura, and Koji Yamamoto. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1):191–205, 2021.

[9] Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: Semi-supervised federated learning for unlabeled clients with alternate training. *Advances in Neural Information Processing Systems*, 35:17871–17884, 2022.

[10] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[11] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[12] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] Jonas Adler and Sebastian Lunz. Banach wasserstein gan. *Advances in neural information processing systems*, 31, 2018.

[16] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

[17] Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*, 2018.

[18] Jaehoon Lee, Jihyeon Hyeong, Jinsung Jeon, Noseong Park, and Jihoon Cho. Invertible tabular gans: Killing two birds with one stone for tabular data synthesis. *Advances in Neural Information Processing Systems*, 34:4263–4273, 2021.

[19] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. *Advances in neural information processing systems*, 30, 2017.

[20] Shreyansh Singh, Kanishka Kayathwal, Hardik Wadhwa, and Gaurav Dhama. Metgan: Memory efficient tabular gan for high cardinality categorical datasets. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*, pages 519–527. Springer, 2021.

[21] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.

[22] Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[25] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[27] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[28] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[29] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[30] Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

[31] Yang Liu, Yan Kang, Xinwei Zhang, Liping Li, Yong Cheng, Tianjian Chen, Mingyi Hong, and Qiang Yang. A communication efficient collaborative learning framework for distributed features. *arXiv preprint arXiv:1912.11187*, 2019.

[32] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4):70–82, 2020.

[33] X. Chen, S. Yin, W. Li, and R. Zhao. Vafl: A method of vertical asynchronous federated learning for heterogeneous data distribution. In *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–7. IEEE, 2020.

[34] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2009.

[35] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Sciences Technical Report*, 1530(9):1–57, 2005.

[36] Jesse E. Van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[37] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[38] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[39] Fabian Mordelet and Jean-Philippe Vert. Bagging for positive and unlabeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2402–2412, 2013.

[40] Ming Li, Wei Zhang, and Hong Chen. A survey on tabular data generation techniques. *IEEE Transactions on Knowledge and Data Engineering*, XX(XX):XX–XX, 2021.

[41] Qing Zhang, Fang Wu, and Hao Du. Tab: A hybrid framework for multi-dimensional table synthesis. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 1234–1241. AAAI, 2020.

[42] John Brown, Lisa Williams, and Mark Davis. Differentially private synthetic tabular data generation via deep generative models. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 567–576. IEEE, 2019.

[43] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

[44] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[45] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[46] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2018.

[47] Jay Lee, Behrad Bagheri, and Hung-An Kao. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3:18–23, 2015.

[48] Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, 94(9):3563–3576, 2018.

[49] Jinghui Wang, Yining Ma, Lin Zhang, Robert X. Gao, and Dazhong Wu. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48:144–156, 2018.

[50] Weinan Zhang, Xuesi Gu, Tianyi Zhou, Zongqing Sun, Jia Pan, Jian Li, Jun Wang, Pingzhong Xu, Cheng Zhang, Yang Gao, Han Zhang, Dong Wang, Zhenguo Li, and Jinhua Zhang. Short-term traffic forecasting: A survey. *arXiv preprint arXiv:1901.00502*, 2019.

[51] Fantine Mordelet and J-P Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

[52] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE international conference on data mining*, pages 179–186. IEEE, 2003.

[53] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

[54] Yixing Xu, Chang Xu, Chao Xu, and Dacheng Tao. Multi-positive and unlabeled learning. In *IJCAI*, pages 3182–3188, 2017.

[55] Emiliano De Cristofaro and Gene Tsudik. Practical private set intersection protocols with linear complexity. In *Financial Cryptography and Data Security: 14th International Conference, FC 2010, Tenerife, Canary Islands, January 25-28, 2010, Revised Selected Papers 14*, pages 143–159. Springer, 2010.

[56] Daojing He, Runmeng Du, Shanshan Zhu, Min Zhang, Kaitai Liang, and Sammy Chan. Secure logistic regression for vertical federated learning. *IEEE Internet Computing*, 26(2):61–68, 2021.

[57] Shengwen Yang, Bing Ren, Xuhui Zhou, and Liping Liu. Parallel distributed logistic regression for vertical federated learning without third-party coordinator. *arXiv preprint arXiv:1911.09824*, 2019.

[58] Houpu Yao, Jiazhou Wang, Peng Dai, Liefeng Bo, and Yanqing Chen. An efficient and robust system for vertically federated random forest. *arXiv preprint arXiv:2201.10761*, 2022.

[59] Wuxing Xu, Hao Fan, Kaixin Li, and Kai Yang. Efficient batch homomorphic encryption for vertically federated xgboost. *arXiv preprint arXiv:2112.04261*, 2021.

[60] Rui Wang, Oğuzhan Ersoy, Hangyu Zhu, Yaochu Jin, and Kaitai Liang. Feverless: Fast and secure vertical federated learning based on xgboost for decentralized labels. *IEEE Transactions on Big Data*, 2022.

[61] Zhi Feng, Haoyi Xiong, Chuanyuan Song, Sijia Yang, Baoxin Zhao, Licheng Wang, Zeyu Chen, Shengwen Yang, Liping Liu, and Jun Huan. Securegbm: Secure multi-party gradient boosting. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1312–1321. IEEE, 2019.

[62] Maulida Ayu Fitriani and Dany Candra Febrianto. Data mining for potential customer segmentation in the marketing bank dataset. *JUITA: Jurnal Informatika*, 9(1):25–32, 2021.

[63] Abdulhamit Subasi and Selcuk Cankurt. Prediction of default payment of credit card clients using data mining techniques. In *2019 International engineering conference (IEC)*, pages 115–120. IEEE, 2019.

[64] Navoneel Chakrabarty and Sanket Biswas. A statistical approach to adult census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 207–212. IEEE, 2018.

[65] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 142–144, 2016.

[66] Q Li, Y Cai, Y Han, CM Yung, T Fu, and B He. Fedtree: A fast, effective, and secure tree-based federated learning system, 2022.

[67] Yang Liu, Tao Fan, Tianjian Chen, Qian Xu, and Qiang Yang. Fate: An industrial grade platform for collaborative learning with data protection. *The Journal of Machine Learning Research*, 22(1):10320–10325, 2021.

[68] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[69] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[70] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[71] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. pages 223–238, 1999.

[72] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Dimitrios Papadopoulos, and Qiang Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 36(6):87–98, 2021.

[73] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[74] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

附录 A 各学院中英文名称对照表

序号	中文名称	英文名称
01	通信工程学院	School of Communications and Information Engineering

作者简介

1. 基本情况

张某某，男，重庆人，1993 年 8 月出生，重庆邮电大学 XX 学院 XX 专业 2018 级博士研究生。

2. 教育和工作经历

3. 攻读学位期间的研究成果

3.1 发表的学术论文和著作

3.2 申请（授权）专利

3.3 参与的科研项目及获奖

以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。

以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。
以下文字用于测试。

致　谢

感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！

感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！
感谢老师、同学们的关心、支持和帮助！