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

摘 要 ····· I
ABSTRACT ····· II
图目录 ····· V
表目录 ····· VI
主要符号表 ····· VII
缩略词表 ····· VIII
第 1 章 绪论 ····· 1
1.1 研究背景及意义 ····· 1
1.2 国内外研究现状 ····· 1
1.3 论文研究的主要内容 ····· 1
1.4 论文组织结构 ····· 1
第 2 章 论文结构及文字格式 ····· 2
2.1 本章引言 ····· 2
2.2 论文结构 ····· 2
2.3 字数要求 ····· 2
2.3.1 硕士论文要求 ····· 2
2.3.2 博士论文要求 ····· 2
2.4 字体和段落 ····· 2
2.5 章节标题 ····· 3
2.5.1 三级标题 ····· 3
2.6 本章小结 ····· 3
第 3 章 基于多方联邦的半监督学习方法研究 ····· 4
3.1 引言 ····· 4
3.2 未标记数据缺失问题（UDD-PU）的分析与定义 ····· 6
3.3 基于多方联邦的半监督学习方法 ····· 8
3.3.1 数据预处理与加密样本对齐 ····· 9
3.3.2 基于正样本与未标记数据的纵向联邦学习 ····· 10
3.4 本章小结 ····· 15
第 4 章 总结与展望 ····· 16
4.1 主要结论 ····· 16
4.2 研究展望 ····· 16

参考文献 …… 18
附录 A 各学院中英文名称对照表 …… 21
作者简介 …… 22
1. 基本情况 …… 22
2. 教育和工作经历 …… 22
3. 攻读学位期间的研究成果 …… 22
3.1 发表的学术论文和著作 …… 22
3.2 申请（授权）专利 …… 22
3.3 参与的科研项目及获奖 …… 22
致　谢 …… 24

图目录

图 2-1 不同章节图片排版测试 …… 3

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

学位论文……

1.2 国内外研究现状

学位论文……

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　论文结构及文字格式

2.1 本章引言

本章引言……

2.2 论文结构

学位论文包括前置部分、主体部分和结尾部分共三大部分，各部分组成及顺序如所示。

学位论文各部分独立为一部分，每部分应从新的一页开始。

论文的正文（中间各章）是论文的核心部分，一般由标题、文字叙述、图、表格和公式等部分构成。由于涉及的学科、选题、研究方法等有很大的差异，可以有不同的写作表达方式，但应遵循本学科通行的学术规范，必须实事求是，客观真切，准确完备，合乎逻辑，层次分明，简练可读。引用他人研究成果时，应注明出处，不得将其与本人的工作混淆。

2.3 字数要求

字数要求

2.3.1 硕士论文要求

各学科和学部自定。

2.3.2 博士论文要求

各学科和学部自定。

2.4 字体和段落

学位论文中的中文统一用宋体，数字和英文统一用 Times New Roman 字体。从中文摘要开始，所有文字段落和标题行间距均取固定值 20 磅；所有段落按两端对齐、首行缩进 2 个全角字符方式书写内容。

中、英文混排时，除小数点以及引用的分图序号、公式序号等外，宜使用全角标点符号（逗号、冒号、括号、引号等）；英文段落中，符号使用应遵循英文书写习惯，统一使用半角符号，并规范使用空格。

其他要求：

（1）各级标题不得置于页面的最后一行，即须与下段同页；

（2）两个标题之间无正文时，第二个标题的段前距设置为 0 磅；

（3）图、表、公式统一采用单倍行距；

（4）只有一、两行文字的，不得单独作为一页内容；除各章最后一页外，中间页面不得出现较大空白；

（5）必要时，可在规定的格式要求基础上适当微调，以利于排版，但显示效果不得与规定的格式要求存在明显差距。

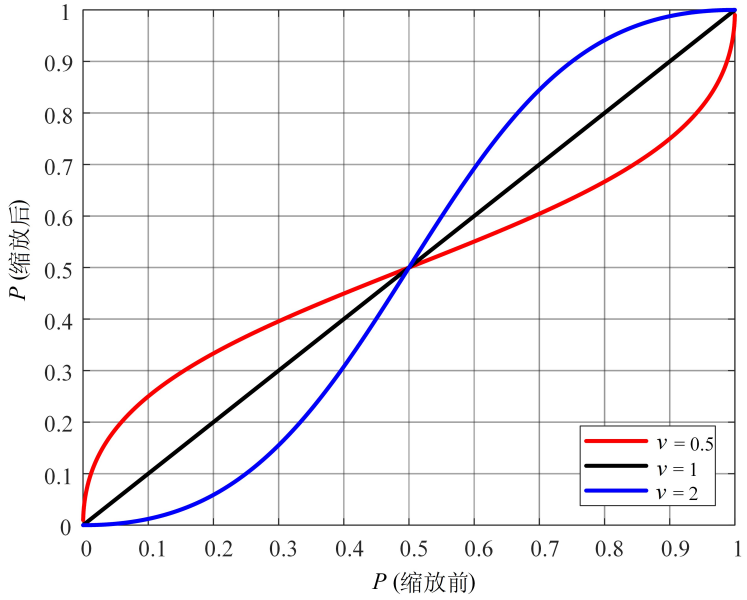

图 2-1 图片排版测试

Fig. 2-1 Scaling results with different scaling coefficients ν

2.5 章节标题

目录中章节标题只显示到 3 级标题，正文中最多显示到 4 级标题。

2.5.1 三级标题

2.5.1.1 四级标题

2.6 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 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[1,2]、医疗诊断[3,4]、智能制造[5,6]以及智能交通[7,8]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[9]，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[10]：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[11]：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[12]：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[13]：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[14]）。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）[14]，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 [15–18] 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 **??** 中得到了可视化展示，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。

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 [19]。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 [16] 提出的两步技术，以及 Mordelet

和 Vert [15] 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 **??** 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（$y_i^C = 1$）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（$y_i^C = -1$）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 **??** 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）[15] 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 [14] 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[14]。其中一些著名的例子包括逻辑回归（LR）[20,21]、随机森林（RF）[22]、梯度提升决策树（GBDT）[20]、XGBoost（XGB）[23,24] 和 LightGBM（LGB）[25]。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，我们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我们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。

3.4 本章小结

本章介绍了……

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

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！